Statistical Learning Theory

Olivier Bousquet
Department of Empirical Inference
Max Planck Institute of Biological Cybernetics
olivier.bousquet@tuebingen.mpg.de

Machine Learning Summer School, August 2003

Roadmap (1)

- Lecture 1: Introduction
- Part I: Binary Classification
  - ⋆ Lecture 2: Basic bounds
  - ⋆ Lecture 3: VC theory
  - ⋆ Lecture 4: Capacity measures
  - ⋆ Lecture 5: Advanced topics

Roadmap (2)

- Part II: Real-Valued Classification
  - ⋆ Lecture 6: Margin and loss functions
  - ⋆ Lecture 7: Regularization
  - ⋆ Lecture 8: SVM

Lecture 1

The Learning Problem

- Context
- Formalization
- Approximation/Estimation trade-off
- Algorithms and Bounds

Learning and Inference

The inductive inference process:

1. Observe a phenomenon
2. Construct a model of the phenomenon
3. Make predictions

$\Rightarrow$ This is more or less the definition of natural sciences !

$\Rightarrow$ The goal of Machine Learning is to automate this process

$\Rightarrow$ The goal of Learning Theory is to formalize it.

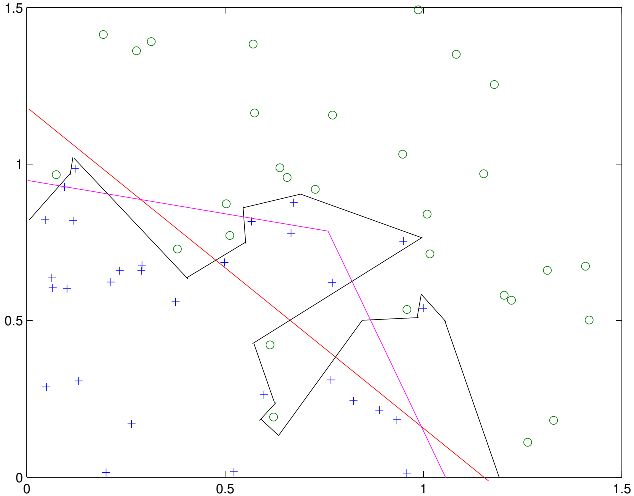
Pattern recognition

We consider here the supervised learning framework for pattern recognition:

- Data consists of pairs (instance, label)
- Label is $+1$ or -1
- Algorithm constructs a function (instance $\longrightarrow$ label)
- Goal: make few mistakes on future unseen instances

Approximation/Interpolation

It is always possible to build a function that fits exactly the data.



But is it reasonable ?

Occam's Razor

Idea: look for regularities in the observed phenomenon
These can ge generalized from the observed past to the future

$\Rightarrow$ choose the simplest consistent model

How to measure simplicity ?

- Physics: number of constants
- Description length
- Number of parameters
- ...

No Free Lunch

- No Free Lunch
  - ⋆ if there is no assumption on how the past is related to the future, prediction is impossible
  - ⋆ if there is no restriction on the possible phenomena, generalization is impossible
- We need to make assumptions
- Simplicity is not absolute
- Data will never replace knowledge
- Generalization = data + knowledge

Assumptions

Two types of assumptions

- Future observations related to past ones
  $\longrightarrow$ *Stationarity* of the phenomenon

- Constraints on the phenomenon
  $\longrightarrow$ Notion of *simplicity*

Goals

$\Rightarrow$ How can we make predictions from the past ? what are the assumptions ?

- Give a formal definition of learning, generalization, overfitting
- Characterize the performance of learning algorithms
- Design better algorithms

Probabilistic Model

Relationship between past and future observations

$\Rightarrow$ Sampled independently from the same distribution

- Independence: each new observation yields maximum information

- Identical distribution: the observations give information about the underlying phenomenon (here a probability distribution)

Probabilistic Model

We consider an input space $\mathcal{X}$ and output space $\mathcal{Y}$.
Here: classification case $\mathcal{Y} = \{-1, 1\}$.

Assumption: The pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are distributed according to P (unknown).

Data: We observe a sequence of n i.i.d. pairs (X_i, Y_i) sampled according to P.

Goal: construct a function $g : \mathcal{X} \to \mathcal{Y}$ which predicts Y from X.

Probabilistic Model

Criterion to choose our function:

Low probability of error $P(g(X) \neq Y)$.

Risk

$$R(g) = P(g(X) \neq Y) = \mathbb{E}\left[1_{[g(X)\neq Y]}\right]$$

- P is unknown so that we cannot directly measure the risk
- Can only measure the agreement on the data
- Empirical Risk

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} 1_{[g(X_i)\neq Y_i]}$$

Target function

- P can be decomposed as $P_X \times P(Y|X)$
- $\eta(x) = \mathbb{E}[Y|X = x] = 2\mathbb{P}[Y = 1|X = x] - 1$ is the regression function
- $t(x) = \operatorname{sgn} \eta(x)$ is the target function
- in the deterministic case $Y = t(X)$ ($\mathbb{P}[Y = 1|X] \in \{0, 1\}$)
- in general, $n(x) = \min(\mathbb{P}[Y = 1|X = x], 1 - \mathbb{P}[Y = 1|X = x]) = (1 - \eta(x))/2$ is the noise level

Assumptions about P

Need assumptions about P.
Indeed, if $t(x)$ is totally chaotic, there is no possible generalization from finite data.
Assumptions can be

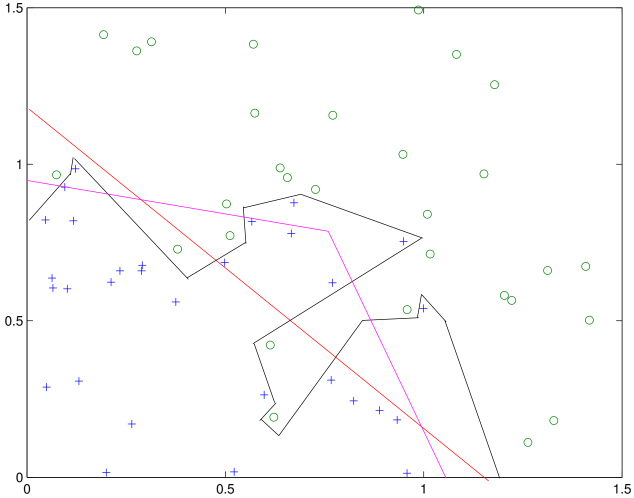
- Preference (e.g. a priori probability distribution on possible functions)
- Restriction (set of possible functions)

Treating lack of knowledge

- Bayesian approach: uniform distribution
- Learning Theory approach: worst case analysis

Approximation/Interpolation (again)

How to trade-off knowledge and data ?



Overfitting/Underfitting

The data can mislead you.

- Underfitting
  model too small to fit the data

- Overfitting
  artificially good agreement with the data

No way to detect them from the data ! Need extra validation data.

Empirical Risk Minimization

- Choose a model $\mathcal{G}$ (set of possible functions)

- Minimize the empirical risk in the model

$$\min_{g \in \mathcal{G}} R_n(g)$$

What if the Bayes classifier is not in the model ?

Approximation/Estimation

- Bayes risk

$$R^* = \inf_g R(g) \, .$$

Best risk a deterministic function can have (risk of the target function, or Bayes classifier).

- Decomposition: $R(g^*) = \inf_{g \in \mathcal{G}} R(g)$

$$R(g_n) - R^* = \underbrace{R(g) - R^*}_{\text{Approximation}} + \underbrace{R(g_n) - R(g^*)}_{\text{Estimation}}$$

- Only the estimation error is random (i.e. depends on the data).

Structural Risk Minimization

- Choose a collection of models $\{\mathcal{G}_d : d = 1, 2, \ldots\}$
- Minimize the empirical risk in each model
- Minimize the penalized empirical risk

$$\min_d \min_{g \in \mathcal{G}_d} R_n(g) + \mathrm{pen}(d, n)$$

$\mathrm{pen}(d, n)$ gives preference to models where estimation error is small

$\mathrm{pen}(d, n)$ measures the size or capacity of the model

Regularization

- Choose a large model $\mathcal{G}$ (possibly dense)
- Choose a regularizer $\|g\|$
- Minimize the regularized empirical risk

$$\min_{g \in \mathcal{G}} R_n(g) + \lambda \|g\|^2$$

- Choose an optimal trade-off λ (regularization parameter).

Most methods can be thought of as regularization methods.

Bounds (1)

A learning algorithm

- Takes as input the data $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Produces a function g_n

Can we estimate the risk of g_n ?

$\Rightarrow$ random quantity (depends on the data).

$\Rightarrow$ need probabilistic bounds

Bounds (2)

- Error bounds

$$R(g_n) \leq R_n(g_n) + B$$

$\Rightarrow$ Estimation from an empirical quantity

- Relative error bounds

  $\star$ Best in a class

$$R(g_n) \leq R(g^*) + B$$

  $\star$ Bayes risk

$$R(g_n) \leq R^* + B$$

$\Rightarrow$ Theoretical guarantees

Lecture 2

Basic Bounds

- Probability tools
- Relationship with empirical processes
- Law of large numbers
- Union bound
- Relative error bounds

Probability Tools (1)

Basic facts

- Union: $\mathbb{P}[A \text{ or } B] \leq \mathbb{P}[A] + \mathbb{P}[B]$

- Inclusion: If $A \Rightarrow B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$.

- Inversion: If $\mathbb{P}[X \geq t] \leq F(t)$ then with probability at least $1-\delta$, $X \leq F^{-1}(\delta)$.

- Expectation: If $X \geq 0$, $\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X \geq t]\, dt$.

Probability Tools (2)

Basic inequalities

- Jensen: for f convex, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

- Markov: If $X \geq 0$ then for all $t > 0$, $\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$

- Chebyshev: for $t > 0$, $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathsf{Var}[X]}{t^2}$

- Chernoff: for all $t \in \mathbb{R}$, $\mathbb{P}[X \geq t] \leq \inf_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda(X-t)}\right]$

Error bounds

Recall that we want to bound $R(g_n) = \mathbb{E}\left[1_{[g_n(X)\neq Y]}\right]$ where g_n has been constructed from $(X_1, Y_1), \ldots, (X_n, Y_n)$.

- Cannot be observed (P is unknown)
- Random (depends on the data)

$\Rightarrow$ we want to bound

$$\mathbb{P}\left[R(g_n) - R_n(g_n) > \varepsilon\right]$$

Loss class

For convenience, let $Z_i = (X_i, Y_i)$ and $Z = (X, Y)$. Given $\mathcal{G}$ define the loss class

$$\mathcal{F} = \{f : (x, y) \mapsto 1_{[g(x) \neq y]} : g \in \mathcal{G}\}$$

Denote $Pf = \mathbb{E}\left[f(X, Y)\right]$ and $P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i, Y_i)$

Quantity of interest:

$$Pf - P_n f$$

We will go back and forth between $\mathcal{F}$ and $\mathcal{G}$ (bijection)

Empirical process

Empirical process:

$$\{Pf - P_n f\}_{f \in \mathcal{F}}$$

- Process $=$ collection of random variables (here indexed by functions in $\mathcal{F}$)
- Empirical $=$ distribution of each random variable

$\Rightarrow$ Many techniques exist to control the supremum

$$\sup_{f \in \mathcal{F}} Pf - P_n f$$

The Law of Large Numbers

$$R(g) - R_n(g) = \mathbb{E}\left[f(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)$$

$\longrightarrow$ difference between the expectation and the empirical average of the r.v. $f(Z)$

Law of large numbers

$$\mathbb{P}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z)\right] = 0\right] = 1\,.$$

$\Rightarrow$ can we quantify it ?

Hoeffding's Inequality

Quantitative version of law of large numbers.

Assumes bounded random variables

Theorem 1. *Let $Z_1, \ldots, Z_n$ be n i.i.d. random variables. If $f(Z) \in [a, b]$. Then for all $\varepsilon > 0$, we have*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z)\right]\right| > \varepsilon\right] \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) .$$

$\Rightarrow$ Let's rewrite it to better understand

Hoeffding's Inequality

Write

$$\delta = 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right)$$

Then

$$\mathbb{P}\left[|P_n f - Pf| > (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right] \leq \delta$$

or [Inversion] with probability at least $1-\delta$,

$$|P_n f - Pf| \leq (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

Hoeffding's inequality

Let's apply to $f(Z) = 1_{[g(X) \neq Y]}$.

For any g, and any $\delta > 0$, with probability at least $1 - \delta$

$$R(g) \leq R_n(g) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}\,. \tag{1}$$

Notice that one has to consider a fixed function f and the probability is with respect to the sampling of the data.

If the function depends on the data this does not apply !

Limitations

- For each fixed function $f \in \mathcal{F}$, there is a set S of samples for which
$$Pf - P_n f \leq \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \quad (\mathbb{P}[S] \geq 1 - \delta)$$
- They may be different for different functions
- The function chosen by the algorithm depends on the sample

$\Rightarrow$ For the observed sample, only some of the functions in $\mathcal{F}$ will satisfy this inequality !

Limitations

What we need to bound is

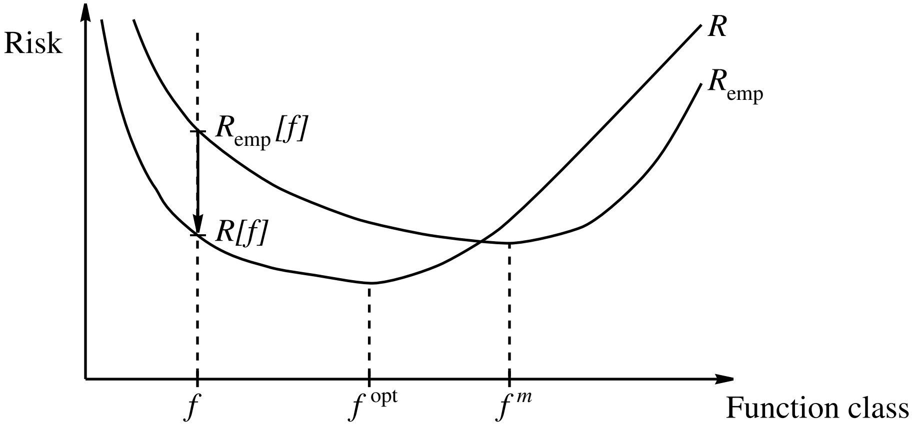
$$Pf_n - P_n f_n$$

where f_n is the function chosen by the algorithm based on the data.
For any fixed sample, there exists a function f such that

$$Pf - P_n f = 1$$

Take the function which is $f(X_i) = Y_i$ on the data and $f(X) = -Y$ everywhere else.
This does not contradict Hoeffding but shows it is not enough

Limitations



Hoeffding's inequality quantifies differences for a fixed function

Uniform Deviations

Before seeing the data, we do not know which function the algorithm will choose.

The trick is to consider uniform deviations

$$R(f_n) - R_n(f_n) \leq \sup_{f \in \mathcal{F}} (R(f) - R_n(f))$$

We need a bound which holds simultaneously for all functions in a class

Union Bound

Consider two functions f_1, f_2 and define

$$C_i = \{(x_1, y_1), \ldots, (x_n, y_n) : Pf_i - P_n f_i > \varepsilon\}$$

From Hoeffding's inequality, for each i

$$\mathbb{P}\left[C_i\right] \leq \delta$$

We want to bound the probability of being 'bad' for $i = 1$ or $i = 2$

$$\mathbb{P}\left[C_1 \cup C_2\right] \leq \mathbb{P}\left[C_1\right] + \mathbb{P}\left[C_2\right]$$

Finite Case

More generally

$$\mathbb{P}\left[C_1 \cup \ldots \cup C_N\right] \leq \sum_{i=1}^{N} \mathbb{P}\left[C_i\right]$$

We have

$$\begin{aligned}
& \mathbb{P}\left[\exists f \in \{f_1, \ldots, f_N\} : Pf - P_n f > \varepsilon\right] \\
\leq & \quad \sum_{i=1}^{N} \mathbb{P}\left[P f_i - P_n f_i > \varepsilon\right] \\
\leq & \quad N \exp\left(-2n\varepsilon^2\right)
\end{aligned}$$

Finite Case

We obtain, for $\mathcal{G} = \{g_1, \ldots, g_N\}$, for all $\delta > 0$

with probability at least $1 - \delta$,

$$\forall g \in \mathcal{G},\ \ R(g) \le R_n(g) + \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2m}}$$

This is a generalization bound !

Coding interpretation
$\log N$ is the number of bits to specify a function in $\mathcal{F}$

Approximation/Estimation

Let

$$g^* = \arg\min_{g \in \mathcal{G}} R(g)$$

If g_n minimizes the empirical risk in $\mathcal{G}$,

$$R_n(g^*) - R_n(g_n) \geq 0$$

Thus

$$\begin{aligned} R(g_n) &= R(g_n) - R(g^*) + R(g^*) \\ &\leq R_n(g^*) - R_n(g_n) + R(g_n) - R(g^*) + R(g^*) \\ &\leq 2 \sup_{g \in \mathcal{G}} |R(g) - R_n(g)| + R(g^*) \end{aligned}$$

Approximation/Estimation

We obtain with probability at least $1 - \delta$

$$R(g_n) \le R(g^*) + 2\sqrt{\frac{\log N + \log \frac{2}{\delta}}{2m}}$$

The first term decreases if N increases
The second term increases

The size of $\mathcal{G}$ controls the trade-off

Summary (1)

- Inference requires assumptions

- Data sampled i.i.d. from P

- Restrict the possible functions to $\mathcal{G}$

- Choose a sequence of models $\mathcal{G}_m$ to have more flexibility/control

Summary (2)

- Bounds are valid w.r.t. repeated sampling
- For a fixed function g, for most of the samples

$$R(g) - R_n(g) \approx 1/\sqrt{n}$$

- For most of the samples if $|\mathcal{G}| = N$

$$\sup_{g \in \mathcal{G}} R(g) - R_n(g) \approx \sqrt{\log N/n}$$

$\Rightarrow$ Extra variability because the chosen g_n changes with the data

Improvements

We obtained

$$\sup_{g\in\mathcal{G}} R(g) - R_n(g) \leq \sqrt{\frac{\log N + \log \frac{2}{\delta}}{2n}}$$

To be improved

- Hoeffding only uses boundedness, not the variance
- Union bound as bad as if independent
- Supremum is not what the algorithm chooses.

Next we improve the union bound and extend it to the infinite case

Refined union bound (1)

For each $f \in \mathcal{F}$,

$$\mathbb{P}\left[Pf - P_n f > \sqrt{\frac{\log \frac{1}{\delta(f)}}{2n}}\right] \leq \delta(f)$$

$$\mathbb{P}\left[\exists f \in \mathcal{F} : Pf - P_n f > \sqrt{\frac{\log \frac{1}{\delta(f)}}{2n}}\right] \leq \sum_{f \in \mathcal{F}} \delta(f)$$

Choose $\delta(f) = \delta p(f)$ with $\sum_{f \in \mathcal{F}} p(f) = 1$

Refined union bound (2)

With probability at least $1 - \delta$,

$$\forall f \in \mathcal{F},\ Pf \leq P_n f + \sqrt{\frac{\log \frac{1}{p(f)} + \log \frac{1}{\delta}}{2n}}$$

- Applies to countably infinite $\mathcal{F}$
- Can put knowledge about the algorithm into $p(f)$
- But p chosen before seeing the data

Refined union bound (3)

- Good p means good bound. The bound can be improved if you know ahead of time the chosen function (knowledge improves the bound)

- In the infinite case, how to choose the p (since it implies an ordering)

- The trick is to look at $\mathcal{F}$ through the data

Lecture 3

Infinite Case: Vapnik-Chervonenkis Theory

- Growth function
- Vapnik-Chervonenkis dimension
- Proof of the VC bound
- VC entropy
- SRM

Infinite Case

Measure of the size of an infinite class ?

- Consider

$$\mathcal{F}_{z_1,\ldots,z_n} = \{(f(z_1),\ldots,f(z_n)) : f \in \mathcal{F}\}$$

  The size of this set is the number of possible ways in which the data $(z_1,\ldots,z_n)$ can be classified.

- Growth function

$$S_{\mathcal{F}}(n) = \sup_{(z_1,\ldots,z_n)} |\mathcal{F}_{z_1,\ldots,z_n}|$$

- Note that $S_{\mathcal{F}}(n) = S_{\mathcal{G}}(n)$

Infinite Case

- Result (Vapnik-Chervonenkis)
  With probability at least $1 - \delta$

$$\forall g \in \mathcal{G}, \;\; R(g) \leq R_n(g) + \sqrt{\frac{\log S_{\mathcal{G}}(2n) + \log \frac{4}{\delta}}{8n}}$$

- Always better than N in the finite case
- How to compute $S_{\mathcal{G}}(n)$ in general ?

$\Rightarrow$ use VC dimension

VC Dimension

Notice that since $g \in \{-1, 1\}$, $S_{\mathcal{G}}(n) \leq 2^n$

If $S_{\mathcal{G}}(n) = 2^n$, the class of functions can generate any classification on n points (shattering)

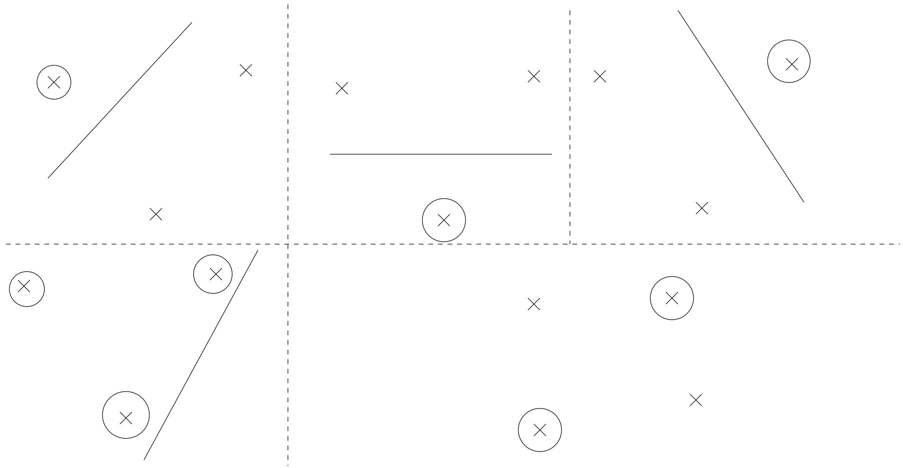
Definition 2. *The* VC-dimension *of $\mathcal{G}$ is the largest n such that*

$$S_{\mathcal{G}}(n) = 2^n$$

VC Dimension

Hyperplanes

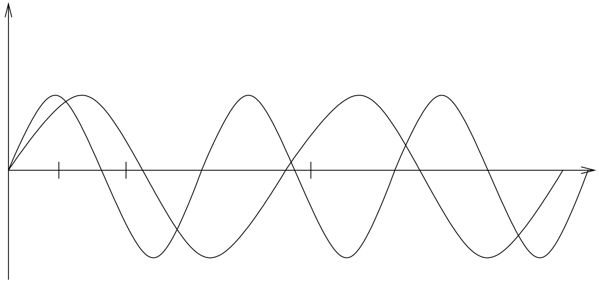
In $\mathbb{R}^d$, $VC(\text{hyperplanes}) = d + 1$



VC Dimension

Number of Parameters

Is VC dimension equal to number of parameters ?



- One parameter

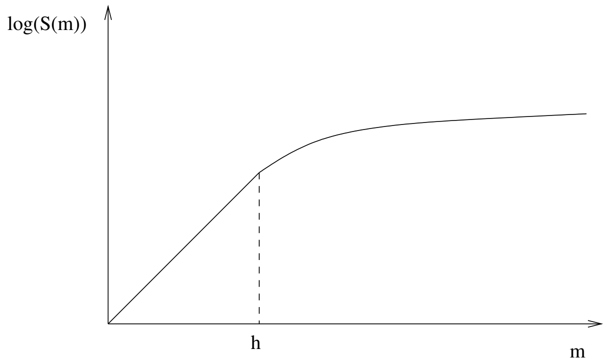
$$\{\mathrm{sgn}(\sin(tx)) : t \in \mathbb{R}\}$$

- Infinite VC dimension !

VC Dimension

- We want to know $S_{\mathcal{G}}(n)$ but we only know $S_{\mathcal{G}}(n) = 2^n$ for $n \leq h$

What happens for $n \geq h$?



Vapnik-Chervonenkis-Sauer-Shelah Lemma

Lemma 3. *Let $\mathcal{G}$ be a class of functions with finite VC-dimension h. Then for all $n \in \mathbb{N}$,*

$$S_{\mathcal{G}}(n) \leq \sum_{i=0}^{h} \binom{n}{i}$$

and for all $n \geq h$,

$$S_{\mathcal{G}}(n) \leq \left(\frac{en}{h}\right)^{h}$$

$\Rightarrow$ phase transition

VC Bound

Let $\mathcal{G}$ be a class with VC dimension h.

With probability at least $1 - \delta$

$$\forall g \in \mathcal{G}, \;\; R(g) \leq R_n(g) + \sqrt{\frac{h \log \frac{2en}{h} + \log \frac{4}{\delta}}{8n}}$$

So the error is of order

$$\sqrt{\frac{h \log n}{n}}$$

Interpretation

VC dimension: measure of effective dimension

- Depends on geometry of the class
- Gives a natural definition of simplicity (by quantifying the potential overfitting)
- Not related to the number of parameters
- Finiteness guarantees learnability under any distribution

Symmetrization (lemma)

Key ingredient in VC bounds: Symmetrization

Let $Z'_1, \ldots, Z'_n$ an independent (ghost) sample and P'_n the corresponding empirical measure.

Lemma 4. *For any $t > 0$, such that $nt^2 \geq 2$,*

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P - P_n)f \geq t\right] \leq 2\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P'_n - P_n)f \geq t/2\right]$$

Symmetrization (proof – 1)

f_n the function achieving the supremum (depends on $Z_1, \ldots, Z_n$)

$$\begin{aligned} 1_{[(P-P_n)f_n>t]} 1_{[(P-P'_n)f_n<t/2]} &= 1_{[(P-P_n)f_n>t \wedge (P-P'_n)f_n<t/2]} \\ &\leq 1_{[(P'_n-P_n)f_n>t/2]} \end{aligned}$$

Taking expectations with respect to the second sample gives

$$1_{[(P-P_n)f_n>t]} \mathbb{P}' \left[(P - P'_n) f_n < t/2 \right] \leq \mathbb{P}' \left[(P'_n - P_n) f_n > t/2 \right]$$

Symmetrization (proof – 2)

- By Chebyshev inequality,

$$\mathbb{P}'\left[(P - P'_n)f_n \geq t/2\right] \leq \frac{4\mathsf{Var}\left[f_n\right]}{nt^2} \leq \frac{1}{nt^2}$$

- Hence

$$1_{[(P-P_n)f_n>t]}(1 - \frac{1}{nt^2}) \leq \mathbb{P}'\left[(P'_n - P_n)f_n > t/2\right]$$

Take expectation with respect to first sample.

Proof of VC bound (1)

- Symmetrization allows to replace expectation by average on ghost sample
- Function class projected on the double sample

$$\mathcal{F}_{Z_1,\ldots,Z_n,Z'_1,\ldots,Z'_n}$$

- Union bound on $\mathcal{F}_{Z_1,\ldots,Z_n,Z'_1,\ldots,Z'_n}$
- Variant of Hoeffding's inequality

$$\mathbb{P}\left[P_n f - P'_n f > t\right] \leq 2e^{-nt^2/2}$$

Proof of VC bound (2)

$$
\begin{aligned}
\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P-P_n)f \geq t\right] & \\
\leq & \quad 2\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P_n'-P_n)f \geq t/2\right] \\
= & \quad 2\mathbb{P}\left[\sup_{f\in\mathcal{F}_{Z_1,\ldots,Z_n,Z_1',\ldots,Z_n'}}(P_n'-P_n)f \geq t/2\right] \\
\leq & \quad 2S_F(2n)\mathbb{P}\left[(P_n'-P_n)f \geq t/2\right] \\
\leq & \quad 4S_F(2n)e^{-nt^2/8}
\end{aligned}
$$

VC Entropy (1)

- VC dimension is distribution independent

$\Rightarrow$ The same bound holds for any distribution

$\Rightarrow$ It is loose for most distributions

- A similar proof can give a distribution-dependent result

VC Entropy (2)

- Denote the size of the projection $N(\mathcal{F}, z_1 \ldots, z_n) := \#\mathcal{F}_{z_1,\ldots,z_n}$
- The *VC entropy* is defined as

$$H_{\mathcal{F}}(n) = \log \mathbb{E}\left[N(\mathcal{F}, Z_1, \ldots, Z_n)\right],$$

- VC entropy bound: with probability at least $1-\delta$

$$\forall g \in \mathcal{G},\ R(g) \leq R_n(g) + \sqrt{\frac{H_{\mathcal{G}}(2n) + \log \frac{2}{\delta}}{8n}}$$

VC Entropy (proof)

Introduce $\sigma_i \in \{-1, 1\}$ (probability $1/2$), Rademacher variables

$$
\begin{aligned}
& 2\mathbb{P}\left[\sup_{f \in \mathcal{F}_{Z,Z'}} (P'_n - P_n)f \geq t/2\right] \\
\leq \quad & 2\mathbb{E}\left[\mathbb{P}_\sigma\left[\sup_{f \in \mathcal{F}_{Z,Z'}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i (f(Z'_i) - f(Z_i)) \geq t/2\right]\right] \\
\leq \quad & 2\mathbb{E}\left[N\left(\mathcal{F}, Z, Z'\right)\right] \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \sigma_i \geq t/2\right] \\
\leq \quad & 2\mathbb{E}\left[N\left(\mathcal{F}, Z, Z'\right)\right] e^{-nt^2/8}
\end{aligned}
$$

From Bounds to Algorithms

- For any distribution, $H_{\mathcal{G}}(n)/n \to 0$ ensures consistency of empirical risk minimizer (i.e. convergence to best in the class)
- Does it means we can learn anything ?
- No because of the approximation of the class
- Need to trade-off approximation and estimation error (assessed by the bound)

$\Rightarrow$ Use the bound to control the trade-off

Structural risk minimization

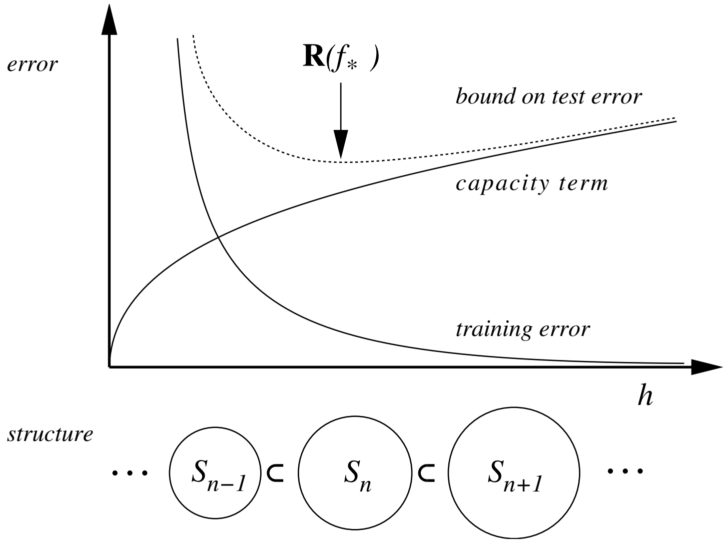
- *Structural risk minimization (SRM) (Vapnik, 1979):* minimize the right hand side of

$$R(g) \leq R_n(g) + B(h, n).$$

- To this end, introduce a structure on $\mathcal{G}$.

- Learning machine $\equiv$ a set of functions and an induction principle

SRM: The Picture



Lecture 4

Capacity Measures

- Covering numbers
- Rademacher averages
- Relationships

Covering numbers

- Define a (random) distance d between functions, e.g.

$$d(f, f') = \frac{1}{n}\#\{f(Z_i) \neq f'(Z_i) : i = 1, \ldots, n\}$$

  Normalized Hamming distance of the 'projections' on the sample

- A set $f_1, \ldots, f_N$ covers $\mathcal{F}$ at radius ε if

$$\mathcal{F} \subset \cup_{i=1}^{N} B(f_i, \varepsilon)$$

- Covering number $N(\mathcal{F}, \varepsilon, n)$ is the minimum size of a cover of radius ε

Note that $N(\mathcal{F}, \varepsilon, n) = N(\mathcal{G}, \varepsilon, n)$.

Bound with covering numbers

- When the covering numbers are finite, one can approximate the class $\mathcal{G}$ by a finite set of functions

- Result

$$\mathbb{P}\left[\exists g \in \mathcal{G} : R(g) - R_n(g) \geq t\right] \leq 8\mathbb{E}\left[N(\mathcal{G}, t, n)\right] e^{-nt^2/128}$$

Covering numbers and VC dimension

- Notice that for all t, $N(\mathcal{G}, t, n) \leq \#\mathcal{G}_Z = N(\mathcal{G}, Z)$

- Hence $N(\mathcal{G}, t, n) \leq h \log \frac{en}{h}$

- Haussler
$$N(\mathcal{G}, t, n) \leq Ch(4e)^h \frac{1}{t^h}$$

- Independent of n

Refinement

- VC entropy corresponds to log covering numbers at minimal scale
- Covering number bound is a generalization where the scale is adapted to the error
- Is this the right scale ?
- It turns out that results can be improved by considering all scales ($\rightarrow$ chaining)

Rademacher averages

- Rademacher variables: $\sigma_1, \ldots, \sigma_n$ independent random variables with
$$\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2}$$
- Notation (randomized empirical measure) $R_n f = \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)$
- Rademacher average: $\mathcal{R}(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} R_n f\right]$
- Conditional Rademacher average $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} R_n f\right]$

Result

- Distribution dependent

$$\forall f \in \mathcal{F}, Pf \leq P_n f + 2\mathcal{R}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}},$$

- Data dependent

$$\forall f \in \mathcal{F}, Pf \leq P_n f + 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{2 \log \frac{2}{\delta}}{n}},$$

Concentration

- Hoeffding's inequality is a concentration inequality

- When n increases, the average is concentrated around the expectation

- Generalization to functions that depend on i.i.d. random variables exist

McDiarmid's Inequality

Assume for all $i = 1, \ldots, n$,

$$\sup_{z_1,\ldots,z_n,z_i'} |F(z_1, \ldots, z_i, \ldots, z_n) - F(z_1, \ldots, z_i', \ldots, z_n)| \leq c$$

then for all $\varepsilon > 0$,

$$\mathbb{P}\left[|F - \mathbb{E}\left[F\right]| > \varepsilon\right] \leq 2 \exp\left(-\frac{2\varepsilon^2}{nc^2}\right)$$

Proof of Rademacher average bounds

- Use concentration to relate $\sup_{f\in\mathcal{F}} Pf - P_n f$ to its expectation

- Use symmetrization to relate expectation to Rademacher average

- Use concentration again to relate Rademacher average to conditional one

Application (1)

$$\sup_{f \in \mathcal{F}} A(f) + B(f) \leq \sup_{f \in \mathcal{F}} A(f) + \sup_{f \in \mathcal{F}} B(f)$$

Hence

$$|\sup_{f \in \mathcal{F}} C(f) - \sup_{f \in \mathcal{F}} A(f)| \leq \sup_{f \in \mathcal{F}} (C(f) - A(f))$$

this gives

$$|\sup_{f \in \mathcal{F}} (Pf - P_n f) - \sup_{f \in \mathcal{F}} (Pf - P'_n f)| \leq \sup_{f \in \mathcal{F}} (P'_n f - P_n f)$$

Application (2)

$f \in \{0, 1\}$ hence,

$$P'_n f - P_n f = \frac{1}{n}(f(Z'_i) - f(Z_i)) \leq \frac{1}{n}$$

thus

$$|\sup_{f \in \mathcal{F}}(Pf - P_n f) - \sup_{f \in \mathcal{F}}(Pf - P'_n f)| \leq \frac{1}{n}$$

McDiarmid's inequality can be applied with $c = 1/n$

Symmetrization (1)

- Upper bound

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right] \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} R_n f\right]$$

- Lower bound

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |Pf - P_n f|\right] \geq \frac{1}{2}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathcal{R}_n f\right] - \frac{1}{2\sqrt{n}}$$

Symmetrization (2)

$$
\begin{aligned}
& \mathbb{E}[\sup_{f\in\mathcal{F}} Pf - P_n f] \\
= \quad & \mathbb{E}[\sup_{f\in\mathcal{F}} \mathbb{E}\left[P'_n f\right] - P_n f] \\
\leq \quad & \mathbb{E}_{Z,Z'}[\sup_{f\in\mathcal{F}} P'_n f - P_n f] \\
= \quad & \mathbb{E}_{\sigma,Z,Z'}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i(f(Z'_i) - f(Z_i))\right] \leq 2\mathbb{E}[\sup_{f\in\mathcal{F}} R_n f]
\end{aligned}
$$

Loss class and initial class

$$\begin{aligned}
\mathcal{R}(\mathcal{F}) &= \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbf{1}_{[g(X_i)\neq Y_i]}\right] \\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1}{2}(1 - Y_i g(X_i))\right] \\
&= \frac{1}{2}\mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i Y_i g(X_i)\right] = \frac{1}{2}\mathcal{R}(\mathcal{G})
\end{aligned}$$

Computing Rademacher averages (1)

$$
\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i)\right] \\
&= \quad \frac{1}{2}+\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}-\frac{1-\sigma_i g(X_i)}{2}\right] \\
&= \quad \frac{1}{2}-\mathbb{E}\left[\inf_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\frac{1-\sigma_i g(X_i)}{2}\right] \\
&= \quad \frac{1}{2}-\mathbb{E}\left[\inf_{g\in\mathcal{G}} R_n(g,\sigma)\right]
\end{aligned}
$$

Computing Rademacher averages (2)

- Not harder than computing empirical risk minimizer

- Pick σ_i randomly and minimize error with respect to labels σ_i

- Intuition: measure how much the class can fit random noise

- Large class $\Rightarrow \mathcal{R}(\mathcal{G}) = \frac{1}{2}$

Concentration again

- Let
$$F = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} R_n f \right]$$
Expectation with respect to σ_i only, with (X_i, Y_i) fixed.

- F satisfies McDiarmid's assumptions with $c = \frac{1}{n}$

$\Rightarrow$ $\mathbb{E}[F] = \mathcal{R}(\mathcal{F})$ can be estimated by $F = \mathcal{R}_n(\mathcal{F})$

Relationship with VC dimension

- For a finite set $\mathcal{F} = \{f_1, \ldots, f_N\}$

$$\mathcal{R}(\mathcal{F}) \leq 2\sqrt{\log N/n}$$

- Consequence for VC class $\mathcal{F}$ with dimension h

$$\mathcal{R}(\mathcal{F}) \leq 2\sqrt{\frac{h \log \frac{en}{h}}{n}}\,.$$

$\Rightarrow$ Recovers VC bound with a concentration proof

Chaining

- Using covering numbers at all scales, the geometry of the class is better captured
- Dudley

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\mathcal{F}, t, n)}\, dt$$

- Consequence

$$\mathcal{R}(\mathcal{F}) \leq C\sqrt{\frac{h}{n}}$$

- Removes the unnecessary $\log n$ factor !

Lecture 5

Advanced Topics

- Relative error bounds
- Noise conditions
- Localized Rademacher averages
- PAC-Bayesian bounds

Binomial tails

- $P_n f \sim B(p, n)$ binomial distribution $p = Pf$
- $\mathbb{P}\left[Pf - P_n f \geq t\right] = \sum_{k=0}^{\lfloor n(p-t) \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$
- Can be upper bounded
  - ⋆ Exponential $\left(\frac{1-p}{1-p-t}\right)^{n(1-p-t)} \left(\frac{p}{p+t}\right)^{n(p+t)}$
  - ⋆ Bennett $e^{-\frac{np}{1-p}((1-t/p)\log(1-t/p)+t/p)}$
  - ⋆ Bernstein $e^{-\frac{nt^2}{2p(1-p)+2t/3}}$
  - ⋆ Hoeffding e^{-2nt^2}

Tail behavior

- For small deviations, Gaussian behavior $\approx \exp(-nt^2/2p(1-p))$
  $\Rightarrow$ Gaussian with variance $p(1-p)$

- For large deviations, Poisson behavior $\approx \exp(-3nt/2)$
  $\Rightarrow$ Tails heavier than Gaussian

- Can upper bound with a Gaussian with large (maximum) variance $\exp(-2nt^2)$

Illustration (1)

Maximum variance ($p = 0.5$)

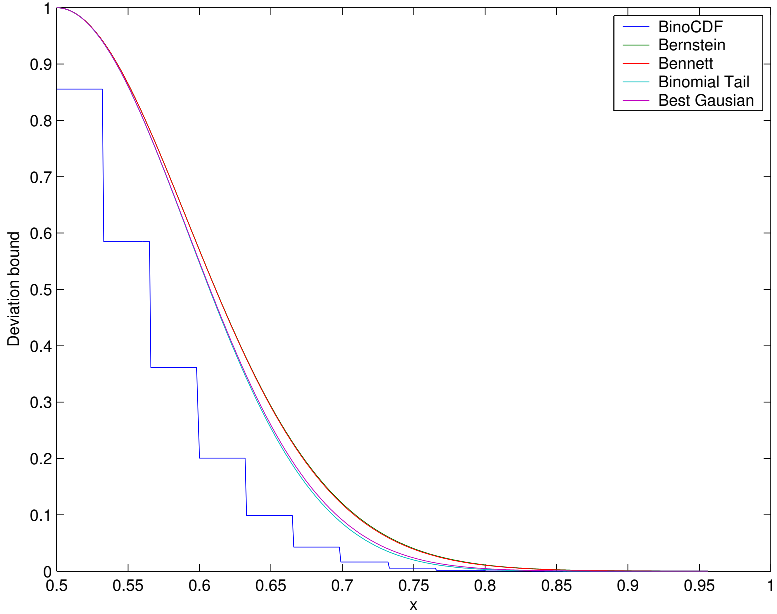
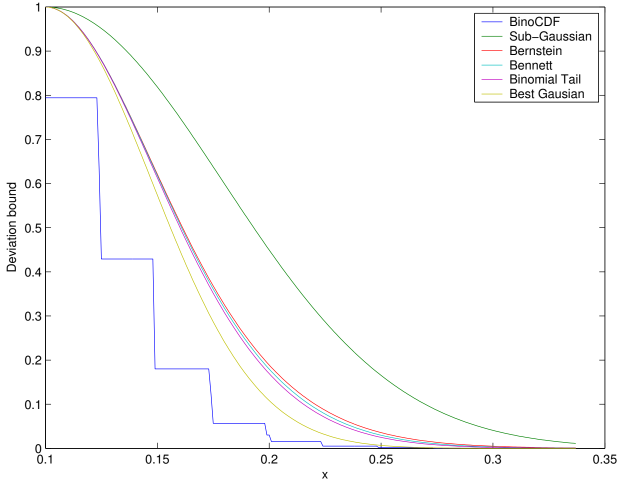
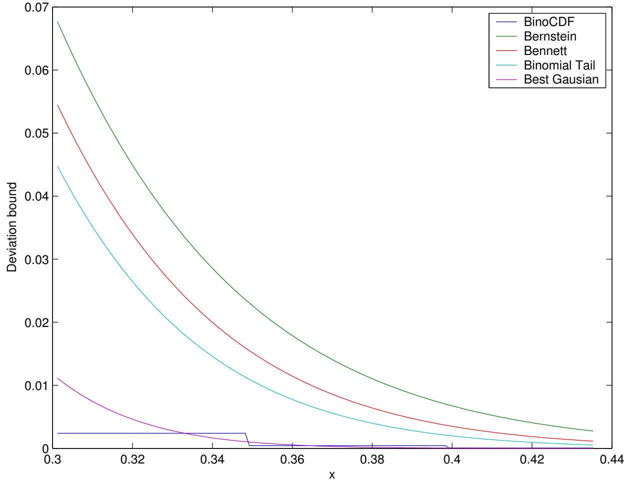


Illustration (2)

Small variance ($p = 0.1$)





Taking the variance into account (1)

- Each function $f \in \mathcal{F}$ has a different variance $Pf(1 - Pf) \leq Pf$.
- For each $f \in \mathcal{F}$, by Bernstein's inequality

$$Pf \leq P_n f + \sqrt{\frac{2Pf \log \frac{1}{\delta}}{n}} + \frac{2 \log \frac{1}{\delta}}{3n}$$

- The Gaussian part dominates (for Pf not too small, or n large enough), it depends on Pf

Taking the variance into account (2)

- Central Limit Theorem

$$\sqrt{n}\frac{Pf - P_n f}{\sqrt{Pf(1 - Pf)}} \to N(0, 1)$$

$\Rightarrow$ Idea is to consider the ratio

$$\frac{Pf - P_n f}{\sqrt{Pf}}$$

Normalization

- Here ($f \in \{0,1\}$), Var $[f] \leq Pf^2 = Pf$
- Large variance $\Rightarrow$ large risk.
- After normalization, fluctuations are more "uniform"

$$\sup_{f \in \mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}}$$

not necessarily attained at functions with large variance.

- Focus of learning: functions with small error Pf (hence small variance).

$\Rightarrow$ The normalized supremum takes this into account.

Relative deviations

Vapnik-Chervonenkis 1974
For $\delta > 0$ with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}, \; \frac{Pf - P_n f}{\sqrt{Pf}} \leq 2\sqrt{\frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}}$$

and

$$\forall f \in \mathcal{F}, \; \frac{P_n f - Pf}{\sqrt{P_n f}} \leq 2\sqrt{\frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}}$$

Proof sketch

1. Symmetrization

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}} \geq t\right] \leq 2\mathbb{P}\left[\sup_{f\in\mathcal{F}} \frac{P'_n f - P_n f}{\sqrt{(P_n f + P'_n f)/2}} \geq t\right]$$

2. Randomization

$$\cdots = 2\mathbb{E}\left[\mathbb{P}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i(f(Z'_i) - f(Z_i))}{\sqrt{(P_n f + P'_n f)/2}} \geq t\right]\right]$$

3. Tail bound

Consequences

From the fact

$$A \leq B + C\sqrt{A} \Rightarrow A \leq B + C^2 + \sqrt{B}C$$

we get

$$\begin{aligned} \forall f \in \mathcal{F},\ Pf \quad &\leq \quad P_n f + 2\sqrt{P_n f \frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}} \\ &\quad + 4\frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n} \end{aligned}$$

Zero noise

Ideal situation

- g_n empirical risk minimizer
- $t \in \mathcal{G}$
- $R^* = 0$ (no noise, $n(X) = 0$ a.s.)

In that case

- $R_n(g_n) = 0$

$\Rightarrow$ $R(g_n) = O(\frac{d \log n}{n})$.

Interpolating between rates ?

- Rates are not correctly estimated by this inequality
- Consequence of relative error bounds

$$
\begin{aligned}
Pf_n \quad &\leq \quad Pf^* + 2\sqrt{Pf^* \frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}} \\
&+4\frac{\log S_{\mathcal{F}}(2n) + \log \frac{4}{\delta}}{n}
\end{aligned}
$$

- The quantity which is small is not Pf^* but $Pf_n - Pf^*$
- But relative error bounds do not apply to differences

Definitions

- $P = P_X \times P(Y|X)$
- regression function $\eta(x) = \mathbb{E}[Y|X = x]$
- target function $t(x) = \operatorname{sgn} \eta(x)$
- noise level $n(X) = (1 - |\eta(x)|)/2$
- Bayes risk $R^* = \mathbb{E}[n(X)]$
- $R(g) = \mathbb{E}[(1 - \eta(X))/2] + \mathbb{E}\left[\eta(X) 1_{[g \leq 0]}\right]$
- $R(g) - R^* = \mathbb{E}\left[|\eta(X)| 1_{[g\eta \leq 0]}\right]$

Intermediate noise

Instead of assuming that $|\eta(x)| = 1$ (i.e. $n(x) = 0$), the deterministic case, one can assume that n is well-behaved.

Two kinds of assumptions

- n not too close to $1/2$

- n not often too close to $1/2$

Massart Condition

- For some $c > 0$, assume

$$|\eta(X)| > \frac{1}{c} \text{ almost surely}$$

- There is no region where the decision is completely random

- Noise bounded away from $1/2$

Tsybakov Condition

Let $\alpha \in [0, 1]$, equivalent conditions

$$
\begin{aligned}
(1) \quad & \exists c > 0,\ \forall g \in \{-1, 1\}^{\mathcal{X}}, \\
& \mathbb{P}\left[g(X)\eta(X) \leq 0\right] \leq c(R(g) - R^*)^\alpha \\
(2) \quad & \exists c > 0,\ \forall A \subset \mathcal{X},\ \int_A dP(x) \leq c(\int_A |\eta(x)| dP(x))^\alpha \\
(3) \quad & \exists B > 0,\ \forall t \geq 0,\ \mathbb{P}\left[|\eta(X)| \leq t\right] \leq B t^{\frac{\alpha}{1-\alpha}}
\end{aligned}
$$

Equivalence

- $(1) \Leftrightarrow (2)$ Recall $R(g) - R^* = \mathbb{E}\left[|\eta(X)|1_{[g\eta \leq 0]}\right]$. For each function g, there exists a set A such that $1_{[A]} = 1_{[g\eta \leq 0]}$
- $(2) \Rightarrow (3)$ Let $A = \{x : |\eta(x)| \leq t\}$

$$\begin{aligned}
\mathbb{P}\left[|\eta| \leq t\right] = \int_A dP(x) &\leq c(\int_A |\eta(x)| dP(x))^\alpha \\
&\leq ct^\alpha (\int_A dP(x))^\alpha
\end{aligned}$$

$$\Rightarrow \mathbb{P}\left[|\eta| \leq t\right] \leq c^{\frac{1}{1-\alpha}} t^{\frac{\alpha}{1-\alpha}}$$

- $(3) \Rightarrow (1)$

$$
\begin{aligned}
R(g) - R^* = \mathbb{E}\left[|\eta(X)|\, 1_{[g\eta \le 0]}\right] & \\
\geq \quad & t\mathbb{E}\left[1_{[g\eta\le 0]} 1_{[|\eta|>t]}\right] \\
= \quad & t\mathbb{P}\left[|\eta| > t\right] - t\mathbb{E}\left[1_{[g\eta>0]} 1_{[|\eta|>t]}\right] \\
\geq \quad & t(1 - Bt^{\frac{\alpha}{1-\alpha}}) - t\mathbb{P}\left[g\eta > 0\right] = t(\mathbb{P}\left[g\eta \le 0\right] - Bt^{\frac{\alpha}{1-\alpha}})
\end{aligned}
$$

Take $t = \left(\frac{(1-\alpha)\mathbb{P}[g\eta\le 0]}{B}\right)^{(1-\alpha)/\alpha}$

$$
\Rightarrow \mathbb{P}\left[g\eta \le 0\right] \le \frac{B^{1-\alpha}}{(1-\alpha)^(1-\alpha)\alpha^{\alpha}}(R(g) - R^*)^{\alpha}
$$

Remarks

- α is in $[0,1]$ because

$$R(g) - R^* = \mathbb{E}\left[|\eta(X)|\mathbf{1}_{[g\eta \leq 0]}\right] \leq \mathbb{E}\left[\mathbf{1}_{[g\eta \leq 0]}\right]$$

- $\alpha = 0$ no condition

- $\alpha = 1$ gives Massart's condition

Consequences

- Under Massart's condition

$$\mathbb{E}\left[\left(1_{[g(X)\neq Y]} - 1_{[t(X)\neq Y]}\right)^2\right] \leq c(R(g) - R^*)$$

- Under Tsybakov's condition

$$\mathbb{E}\left[\left(1_{[g(X)\neq Y]} - 1_{[t(X)\neq Y]}\right)^2\right] \leq c(R(g) - R^*)^\alpha$$

Relative loss class

- $\mathcal{F}$ is the loss class associated to $\mathcal{G}$

- The relative loss class is defined as

$$\tilde{\mathcal{F}} = \{f - f^* : f \in \mathcal{F}\}$$

- It satisfies

$$Pf^2 \leq c(Pf)^\alpha$$

Finite case

- Union bound on $\tilde{\mathcal{F}}$ with Bernstein's inequality would give

$$Pf_n - Pf^* \le P_n f_n - P_n f^* + \sqrt{\frac{8c(Pf_n - Pf^*)^\alpha \log \frac{N}{\delta}}{n}} + \frac{4 \log \frac{N}{\delta}}{3n}$$

- Consequence when $f^* \in \mathcal{F}$ (but $R^* > 0$)

$$Pf_n - Pf^* \le C \left(\frac{\log \frac{N}{\delta}}{n} \right)^{\frac{1}{2-\alpha}}$$

always better than $n^{-1/2}$ for $\alpha > 0$

Local Rademacher average

- Definition

$$\mathcal{R}(\mathcal{F}, r) = \mathbb{E}\left[\sup_{f \in \mathcal{F}: Pf^2 \leq r} R_n f\right]$$

- Allows to generalize the previous result

- Computes the capacity of a small ball in $\mathcal{F}$ (functions with small variance)

- Under noise conditions, small variance implies small error

Sub-root functions

Definition

A function $\psi : \mathbb{R} \longrightarrow \mathbb{R}$ is sub-root if

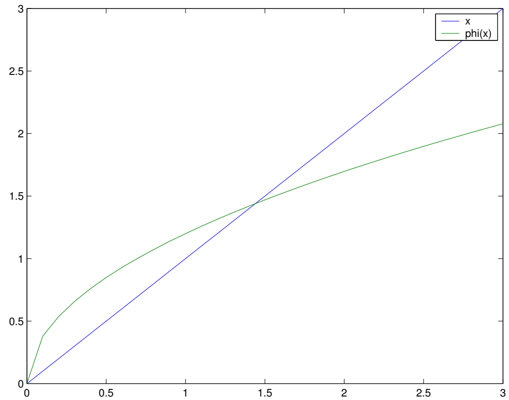
- ψ is non-decreasing
- ψ is non negative
- $\psi(r)/\sqrt{r}$ is non-increasing

Sub-root functions

Properties

A sub-root function

- is continuous
- has a unique fixed point $\psi(r^*) = r^*$



Star hull

- Definition

$$\star\mathcal{F} = \{\alpha f : f \in \mathcal{F},\ \alpha \in [0,1]\}$$

- Properties

$$\mathcal{R}_n(\star\mathcal{F}, r) \text{ is sub-root}$$

- Entropy of $\star\mathcal{F}$ is not much bigger than entropy of $\mathcal{F}$

Result

- r^* fixed point of $\mathcal{R}(\star\mathcal{F}, r)$
- Bounded functions

$$Pf - P_n f \leq C \left(\sqrt{r^* \mathrm{Var}\left[f\right]} + \frac{\log \frac{1}{\delta} + \log \log n}{n} \right)$$

- Consequence for variance related to expectation ($\mathrm{Var}\left[f\right] \leq c(Pf)^\beta$)

$$Pf \leq C \left(P_n f + (r^*)^{\frac{1}{2-\beta}} + \frac{\log \frac{1}{\delta} + \log \log n}{n} \right)$$

Consequences

- For VC classes $\mathcal{R}(\mathcal{F}, r) \leq C\sqrt{\frac{rh}{n}}$ hence $r^* \leq C\frac{h}{n}$

- Rate of convergence of $P_n f$ to Pf in $O(1/\sqrt{n})$

- But rate of convergence of Pf_n to Pf^* is $O(1/n^{1/(2-\alpha)})$

Only condition is $t \in \mathcal{G}$ but can be removed by SRM/Model selection

Proof sketch (1)

- Talagrand's inequality

$$\sup_{f\in\mathcal{F}} Pf - P_n f \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right] + c\sqrt{\sup_{f\in\mathcal{F}} \mathrm{Var}\left[f\right]/n} + c'/n$$

- Peeling of the class

$$\mathcal{F}_k = \{f : \mathrm{Var}\left[f\right] \in [x^k, x^{k+1})\}$$

Proof sketch (2)

- Application

$$\sup_{f \in \mathcal{F}_k} Pf - P_n f \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}_k} Pf - P_n f\right] + c\sqrt{x \mathsf{Var}\left[f\right]/n} + c'/n$$

- Symmetrization

$$\forall f \in \mathcal{F},\ Pf - P_n f \leq 2\mathcal{R}(\mathcal{F}, x\mathsf{Var}\left[f\right]) + c\sqrt{x\mathsf{Var}\left[f\right]/n} + c'/n$$

Proof sketch (3)

- We need to 'solve' this inequality. Things are simple if $\mathcal{R}$ behave like a square root, hence the sub-root property

$$Pf - P_n f \leq 2\sqrt{r^* \mathsf{Var}\left[f\right]} + c\sqrt{x \mathsf{Var}\left[f\right]/n} + c'/n$$

- Variance-expectation

$$\mathsf{Var}\left[f\right] \leq c(Pf)^{\alpha}$$

Solve in Pf

Data-dependent version

- As in the global case, one can use data-dependent local Rademcher averages

$$\mathcal{R}_n(\mathcal{F}, r) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}: Pf^2 \leq r} R_n f \right]$$

- Using concentration one can also get

$$Pf \leq C \left(P_n f + (r_n^*)^{\frac{1}{2-\alpha}} + \frac{\log \frac{1}{\delta} + \log \log n}{n} \right)$$

where r_n^* is the fixed point of a sub-root upper bound of $\mathcal{R}_n(\mathcal{F}, r)$

Discussion

- Improved rates under low noise conditions
- Interpolation in the rates
- Capacity measure seems 'local',
- but depends on all the functions,
- after appropriate rescaling: each $f \in \mathcal{F}$ is considered at scale r/Pf^2

Randomized Classifiers

Given $\mathcal{G}$ a class of functions

- Deterministic: picks a function g_n and always use it to predict
- Randomized
  - ⋆ construct a distribution ρ_n over $\mathcal{G}$
  - ⋆ for each instance to classify, pick $g \sim \rho_n$
- Error is averaged over ρ_n

$$R(\rho_n) = \rho_n P f$$

$$R_n(\rho_n) = \rho_n P_n f$$

Union Bound (1)

Let π be a (fixed) distribution over $\mathcal{F}$.

- Recall the refined union bound

$$\forall f \in \mathcal{F},\ Pf - P_n f \leq \sqrt{\frac{\log \frac{1}{\pi(f)} + \log \frac{1}{\delta}}{2n}}$$

- Take expectation with respect to ρ_n

$$\rho_n Pf - \rho_n P_n f \leq \rho_n \sqrt{\frac{\log \frac{1}{\pi(f)} + \log \frac{1}{\delta}}{2n}}$$

Union Bound (2)

$$
\begin{aligned}
\rho_n P f - \rho_n P_n f &\leq \rho_n \sqrt{\left(-\log \pi(f) + \log \tfrac{1}{\delta}\right) / (2n)} \\
&\leq \sqrt{\left(-\rho_n \log \pi(f) + \log \tfrac{1}{\delta}\right) / (2n)} \\
&\leq \sqrt{\left(K(\rho_n, \pi) + H(\rho_n) + \log \tfrac{1}{\delta}\right) / (2n)}
\end{aligned}
$$

- $K(\rho_n, \pi) = \int \rho_n(f) \log \frac{\rho_n(f)}{\pi(f)} \, df$ Kullback-Leibler divergence
- $H(\rho_n) = \int \rho_n(f) \log \rho_n(f) \, df$ Entropy

PAC-Bayesian Refinement

- It is possible to improve the previous bound.
- With probability at least $1 - \delta$,

$$\rho_n P f - \rho_n P_n f \leq \sqrt{\frac{K(\rho_n, \pi) + \log 4n + \log \frac{1}{\delta}}{2n - 1}}$$

- Good if ρ_n is spread (i.e. large entropy)
- Not interesting if $\rho_n = \delta_{f_n}$

Proof (1)

- Variational formulation of entropy: for any T

$$\rho T(f) \leq \log \pi e^{T(f)} + K(\rho, \pi)$$

- Apply it to $\lambda(Pf - P_n f)^2$

$$\lambda \rho_n (Pf - P_n f)^2 \leq \log \pi e^{\lambda(Pf - P_n f)^2} + K(\rho_n, \pi)$$

- Markov's inequality: with probability $1 - \delta$,

$$\lambda \rho_n (Pf - P_n f)^2 \leq \log \mathbb{E}\left[\pi e^{\lambda(Pf - P_n f)^2}\right] + K(\rho_n, \pi) + \log \frac{1}{\delta}$$

Proof (2)

- Fubini

$$\mathbb{E}\left[\pi e^{\lambda(Pf-P_nf)^2}\right] = \pi\mathbb{E}\left[e^{\lambda(Pf-P_nf)^2}\right]$$

- Modified Chernoff bound

$$\mathbb{E}\left[e^{(2n-1)(Pf-P_nf)^2}\right] \leq 4n$$

- Putting together ($\lambda = 2n - 1$)

$$(2n-1)\rho_n(Pf - P_nf)^2 \leq K(\rho_n, \pi) + \log 4n + \log \frac{1}{\delta}$$

- Jensen $(2n-1)(\rho_n(Pf - P_nf))^2 \leq (2n-1)\rho_n(Pf - P_nf)^2$

Lecture 6

Loss Functions

- Properties
- Consistency
- Examples
- Losses and noise

Motivation (1)

- ERM: minimize $\sum_{i=1}^n 1_{[g(X_i) \neq Y_i]}$ in a set $\mathcal{G}$

$\Rightarrow$ Computationally hard

$\Rightarrow$ Smoothing

  - $\star$ Replace binary by real-valued functions
  - $\star$ Introduce smooth loss function

$$\sum_{i=1}^{n} \ell(g(X_i), Y_i)$$

Motivation (2)

- Hyperplanes in infinite dimension have
  - $\star$ infinite VC-dimension
  - $\star$ but finite scale-sensitive dimension (to be defined later)

$\Rightarrow$ It is good to have a scale

$\Rightarrow$ This scale can be used to give a confidence (i.e. estimate the density)

- However, losses do not need to be related to densities
- Can get bounds in terms of margin error instead of empirical error (smoother $\longrightarrow$ easier to optimize for model selection)

Margin

- It is convenient to work with (symmetry of $+1$ and -1)

$$\ell(g(x), y) = \phi(yg(x))$$

- $yg(x)$ is the margin of g at (x, y)
- Loss

$$L(g) = \mathbb{E}\left[\phi(Yg(X))\right], \;\; L_n(g) = \frac{1}{n}\sum_{i=1}^{n}\phi(Y_i g(X_i))$$

- Loss class $\mathcal{F} = \{f : (x, y) \mapsto \phi(yg(x)) : g \in \mathcal{G}\}$

Minimizing the loss

- Decomposition of $L(g)$

$$\frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[(1+\eta(X))\phi(g(X)) + (1-\eta(X))\phi(-g(X))|X\right]\right]$$

- Minimization for each x

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} \left((1+\eta)\phi(\alpha)/2 + (1-\eta)\phi(-\alpha)/2\right)$$

- $L^* := \inf_g L(g) = \mathbb{E}\left[H(\eta(X))\right]$

Classification-calibrated

- A minimal requirement is that the minimizer in $H(\eta)$ has the correct sign (that of the target t or that of η).

- Definition
  ϕ is classification-calibrated if, for any $\eta \neq 0$

$$\inf_{\alpha:\alpha\eta\leq 0}(1+\eta)\phi(\alpha)+(1-\eta)\phi(-\alpha) > \inf_{\alpha\in\mathbb{R}}(1+\eta)\phi(\alpha)+(1-\eta)\phi(-\alpha)$$

- This means the infimum is achieved for an α of the correct sign (and not for an α of the wrong sign, except possibly for $\eta = 0$).

Consequences (1)

Results due to (Jordan, Bartlett and McAuliffe 2003)

- ϕ is classification-calibrated iff for all sequences g_i and every probability distribution P,

$$L(g_i) \to L^* \;\Rightarrow\; R(g_i) \to R^*$$

- When ϕ is convex (convenient for optimization) ϕ is classification-calibrated iff it is differentiable at 0 and $\phi'(0) < 0$

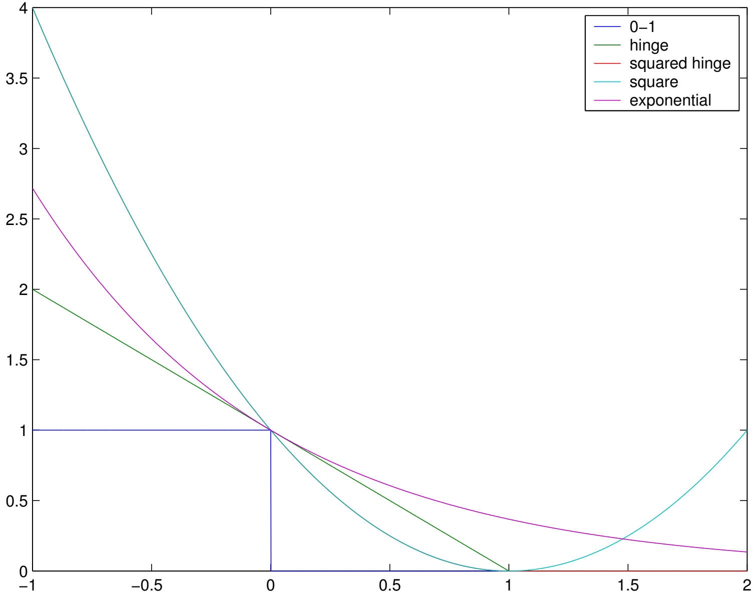
Consequences (2)

- Let $H^-(\eta) = \inf_{\alpha:\alpha\eta\leq 0}((1+\eta)\phi(\alpha)/2 + (1-\eta)\phi(-\alpha)/2)$

- Let $\psi(\eta)$ be the largest convex function below $H^-(\eta) - H(\eta)$

- One has
$$\psi(R(g) - R^*) \leq L(g) - L^*$$

Examples (1)


0−1
hinge
squared hinge
square
exponential
4
3.5
3
2.5
2
1.5
1
0.5
0
−1
−0.5
0
0.5
1
1.5
2

Examples (2)

- Hinge loss

$$\phi(x) = \max(0, 1 - x), \;\; \psi(x) = x$$

- Squared hinge loss

$$\phi(x) = \max(0, 1 - x)^2, \;\; \psi(x) = x^2$$

- Square loss

$$\phi(x) = (1 - x)^2, \;\; \psi(x) = x^2$$

- Exponential

$$\phi(x) = \exp(-x), \;\; \psi(x) = 1 - \sqrt{1 - x^2}$$

Low noise conditions

- Relationship can be improved under low noise conditions
- Under Tsybakov's condition with exponent α and constant c,

$$c(R(g) - R^*)^\alpha \psi((R(g) - R^*)^{1-\alpha}/2c) \leq L(g) - L^*$$

- Hinge loss (no improvement)

$$R(g) - R^* \leq L(g) - L^*$$

- Square loss or squared hinge loss

$$R(g) - R^* \leq (4c(L(g) - L^*))^{\frac{1}{2-\alpha}}$$

Estimation error

- Recall that Tsybakov condition implies $Pf^2 \leq c(Pf)^\alpha$ for the relative loss class (with $0-1$ loss)

- What happens for the relative loss class associated to ϕ ?

- Two possibilities
  - ⋆ Strictly convex loss (can modify the metric on $\mathbb{R}$)

  - ⋆ Piecewise linear

Strictly convex losses

- Noise behavior controlled by modulus of convexity

- Result
$$\delta\left(\frac{\sqrt{Pf^2}}{K}\right) \leq Pf/2$$
with K Lipschitz constant of ϕ and δ modulus of convexity of $L(g)$ with respect to $\|f-g\|_{L_2(P)}$

- Not related to noise exponent

Piecewise linear losses

- Noise behavior related to noise exponent

- Result for hinge loss

$$Pf^2 \leq CPf^\alpha$$

if initial class $\mathcal{G}$ is uniformly bounded

Estimation error

- With bounded and Lipschitz loss with convexity exponent γ, for a convex class $\mathcal{G}$,

$$L(g) - L(g^*) \leq C \left((r^*)^{\frac{2}{\gamma}} + \frac{\log \frac{1}{\delta} + \log \log n}{n} \right)$$

- Under Tsybakov's condition for the hinge loss (and general $\mathcal{G}$) $Pf^2 \leq CPf^\alpha$

$$L(g) - L(g^*) \leq C \left((r^*)^{\frac{1}{2-\alpha}} + \frac{\log \frac{1}{\delta} + \log \log n}{n} \right)$$

Examples

Under Tsybakov's condition

- Hinge loss

$$R(g) - R^* \leq L(g^*) - L^* + C\left((r^*)^{\frac{1}{2-\alpha}} + \frac{\log\frac{1}{\delta} + \log\log n}{n}\right)$$

- Squared hinge loss or square loss $\delta(x) = cx^2$, $Pf^2 \leq CPf$

$$R(g) - R^* \leq C\left(L(g^*) - L^* + C'(r^* + \frac{\log\frac{1}{\delta} + \log\log n}{n})\right)^{\frac{1}{2-\alpha}}$$

Classification vs Regression losses

- Consider a classification-calibrated function ϕ

- It is a classification loss if $L(t) = L^*$

- otherwise it is a regression loss

Classification vs Regression losses

- Square, squared hinge, exponential losses
  - ⋆ Noise enters relationship between risk and loss
  - ⋆ Modulus of convexity enters in estimation error
- Hinge loss
  - ⋆ Direct relationship between risk and loss
  - ⋆ Noise enters in estimation error

$\Rightarrow$ Approximation term not affected by noise in second case

$\Rightarrow$ Real value does not bring probability information in second case

Lecture 7

Regularization

- Formulation
- Capacity measures
- Computing Rademacher averages
- Applications

Equivalent problems

Up to the choice of the regularization parameters, the following problems are equivalent

$$\min_{f \in \mathcal{F}} L_n(f) + \lambda \Omega(f)$$

$$\min_{f \in \mathcal{F}: L_n(f) \leq e} \Omega(f)$$

$$\min_{f \in \mathcal{F}: \Omega(f) \leq R} L_n(f)$$

The solution sets are the same

Comments

- Computationally, variant of SRM
- variant of model selection by penalization

$\Rightarrow$ one has to choose a regularizer which makes sense

- Need a class that is large enough (for universal consistency)
- but has small balls

Rates

- To obtain bounds, consider ERM on balls
- Relevant capacity is that of balls
- Real-valued functions, need a generalization of VC dimension, entropy or covering numbers
- Involve scale sensitive capacity (takes into account the value and not only the sign)

Scale-sensitive capacity

- Generalization of VC entropy and VC dimension to real-valued functions

- Definition: a set $x_1, \ldots, x_n$ is shattered by $\mathcal{F}$ (at scale ε) if there exists a function s such that for all choices of $\alpha_i \in \{-1, 1\}$, there exists $f \in \mathcal{F}$
$$\alpha_i(f(x_i) - s(x_i)) \geq \varepsilon$$

- The fat-shattering dimension of $\mathcal{F}$ at scale ε (denoted $vc(\mathcal{F}, \varepsilon)$) is the maximum cardinality of a shattered set

Link with covering numbers

- Like VC dimension, fat-shattering dimension can be used to upper bound covering numbers

- Result

$$N(\mathcal{F}, t, n) \leq \left(\frac{C_1}{t}\right)^{C_2 vc(\mathcal{F}, C_3 t)}$$

- Note that one can also define data-dependent versions (restriction on the sample)

Link with Rademacher averages (1)

- Consequence of covering number estimates

$$\mathbb{R}_n(\mathcal{F}) \leq \frac{C_1}{\sqrt{n}} \int_0^\infty \sqrt{vc(\mathcal{F}, t) \log \frac{C_2}{t}}\, dt$$

- Another link via Gaussian averages (replace Rademacher by Gaussian N(0,1) variables)

$$\mathcal{G}_n(\mathcal{F}) = \mathbb{E}_g \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n g_i f(Z_i) \right]$$

Link with Rademacher averages (2)

- Worst case average

$$\ell_n(\mathcal{F}) = \sup_{x_1,\ldots,x_n} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} G_n f \right]$$

- Associated "dimension" $t(\mathcal{F}, \epsilon) = \sup\{n \in \mathbb{N} : \ell_n(\mathcal{F}) \geq \epsilon\}$
- Result (Mendelson & Vershynin 2003)

$$vc(\mathcal{F}, c'\epsilon) \leq t(\mathcal{F}, \epsilon) \leq \frac{K}{\epsilon^2} vc(\mathcal{F}, c\epsilon)$$

Rademacher averages and Lipschitz losses

- What matters is the capacity of $\mathcal{F}$ (loss class)
- If ϕ is Lipschitz with constant M
- then
$$\mathcal{R}_n(\mathcal{F}) \leq M\mathcal{R}_n(\mathcal{G})$$
- Relates to Rademacher average of the initial class (easier to compute)

Dualization

- Consider the problem $\min_{\|g\|\leq R} L_n(g)$
- Rademacher of ball

$$\mathbb{E}_\sigma\left[\sup_{\|g\|\leq R} R_n g\right]$$

- Duality

$$\mathbb{E}_\sigma\left[\sup_{\|f\|\leq R} R_n f\right] = \frac{R}{n}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n \sigma_i \delta_{X_i}\right\|^*\right]$$

$\|\,\|^*$ dual norm, δ_{X_i} evaluation at X_i (element of the dual under appropriate conditions)

RHKS

Given a positive definite kernel k

- Space of functions: reproducing kernel Hilbert space associated to k
- Regularizer: rkhs norm $\|\cdot\|_k$
- Properties: Representer theorem

$$g_n = \sum_{i=1}^{n} \alpha_i k(X_i, \cdot)$$

Shattering dimension of hyperplanes

- Set of functions

$$\mathcal{G} = \{g(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} : \|\mathbf{w}\| = 1\}$$

- Assume $\|\mathbf{x}\| \leq R$

- Result

$$vc(\mathcal{G}, \rho) \leq R^2/\rho^2$$

Proof Strategy (Gurvits, 1997)

Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_r$ are ρ-shattered by hyperplanes with $\|\mathbf{w}\| = 1$, i.e., for all $y_1, \ldots, y_r \in \{\pm 1\}$, there exists a $\mathbf{w}$ such that

$$y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq \rho \quad \text{for all } i = 1, \ldots, r. \tag{2}$$

Two steps:

- prove that the more points we want to shatter (2), the larger $\|\sum_{i=1}^r y_i \mathbf{x}_i\|$ must be
- upper bound the size of $\|\sum_{i=1}^r y_i \mathbf{x}_i\|$ in terms of R

Combining the two tells us how many points we can at most shatter

Part I

- Summing (2) yields $\langle \mathbf{w}, (\sum_{i=1}^r y_i \mathbf{x}_i) \rangle \geq r\rho$
- By Cauchy-Schwarz inequality

$$\left\langle \mathbf{w}, \left(\sum_{i=1}^r y_i \mathbf{x}_i \right) \right\rangle \leq \|\mathbf{w}\| \left\| \sum_{i=1}^r y_i \mathbf{x}_i \right\| = \left\| \sum_{i=1}^r y_i \mathbf{x}_i \right\|$$

- Combine both:

$$r\rho \leq \left\| \sum_{i=1}^r y_i \mathbf{x}_i \right\| . \qquad (3)$$

Part II

Consider labels $y_i \in \{\pm 1\}$, as (*Rademacher variables*).

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=1}^{r} y_i \mathbf{x}_i\right\|^2\right] &= \mathbb{E}\left[\sum_{i,j=1}^{r} y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle\right] \\
&= \sum_{i=1}^{r} \mathbb{E}\left[\langle \mathbf{x}_i, \mathbf{x}_i \rangle\right] + \mathbb{E}\left[\sum_{i \neq j}^{r} \langle \mathbf{x}_i, \mathbf{x}_j \rangle\right] \\
&= \sum_{i=1}^{r} \|\mathbf{x}_i\|^2
\end{aligned}$$

Part II, ctd.

- Since $\|\mathbf{x}_i\| \leq R$, we get $\mathbb{E}\left[\|\sum_{i=1}^r y_i \mathbf{x}_i\|^2\right] \leq rR^2$.

- This holds for the *expectation* over the random choices of the labels, hence there must be at least one set of labels for which it also holds true. Use this set.

- Hence

$$\left\|\sum_{i=1}^r y_i \mathbf{x}_i\right\|^2 \leq rR^2.$$

Part I and II Combined

- Part I: $(r\rho)^2 \leq \left\|\sum_{i=1}^r y_i \mathbf{x}_i\right\|^2$
- Part II: $\left\|\sum_{i=1}^r y_i \mathbf{x}_i\right\|^2 \leq rR^2$
- Hence

$$r^2\rho^2 \leq rR^2,$$

i.e.,

$$r \leq \frac{R^2}{\rho^2}$$

Boosting

Given a class $\mathcal{H}$ of functions

- Space of functions: linear span of $\mathcal{H}$
- Regularizer: 1-norm of the weights $\|g\| = \inf\{\sum |\alpha_i| : g = \sum \alpha_i h_i\}$
- Properties: weight concentrated on the (weighted) margin maximizers

$$g_n = \sum w_h h$$

$$\sum d_i Y_i h(X_i) = \min_{h' \in \mathcal{H}} \sum d_i Y_i h'(X_i)$$

Rademacher averages for boosting

- Function class of interest

$$\mathcal{G}_R = \{g \in \operatorname{span} \mathcal{H} : \|g\|_1 \leq R\}$$

- Result

$$\mathcal{R}_n(\mathcal{G}_R) = R\mathcal{R}_n(\mathcal{H})$$

$\Rightarrow$ Capacity (as measured by global Rademacher averages) not affected by taking linear combinations !

Lecture 8

SVM

- Computational aspects
- Capacity Control
- Universality
- Special case of RBF kernel

Formulation (1)

- Soft margin

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

- Convex objective function and convex constraints
- Unique solution
- Efficient procedures to find it

$\longrightarrow$ Is it the right criterion ?

Formulation (2)

- Soft margin

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

- Optimal value of ξ_i

$$\xi_i^* = \max(0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b))$$

- Substitute above to get

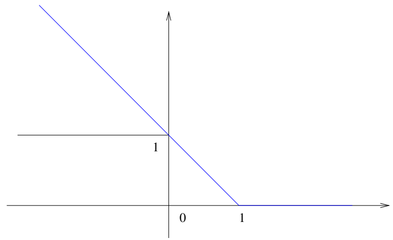
$$\min_{\mathbf{w},b} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\max(0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b))$$

Regularization

General form of regularization problem

$$\min_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{n} c(y_i f(x_i)) + \lambda \left\|f\right\|^2$$

$\longrightarrow$ Capacity control by regularization with convex cost



Loss Function

$$\phi(Yf(X)) = \max(0, 1 - Yf(X))$$

- Convex, non-increasing, upper bounds $1_{[Yf(X)\leq 0]}$
- Classification-calibrated
- Classification type ($L^* = L(t)$)

$$R(g) - R^* \leq L(g) - L^*$$

Regularization

Choosing a kernel corresponds to

- Choose a sequence (a_k)
- Set

$$\|f\|^2 := \sum_{k \geq 0} a_k \int |f^{(k)}|^2 dx$$

$\Rightarrow$ penalization of high order derivatives (high frequencies)

$\Rightarrow$ enforce smoothness of the solution

Capacity: VC dimension

- The VC dimension of the set of hyperplanes is $d+1$ in $\mathbb{R}^d$.
  Dimension of feature space ?
  ∞ for RBF kernel
- w choosen in the span of the data ($w = \sum \alpha_i y_i \mathbf{x}_i$)
  The span of the data has dimension m for RBF kernel ($k(.,x_i)$ linearly independent)
- The VC bound does not give any information

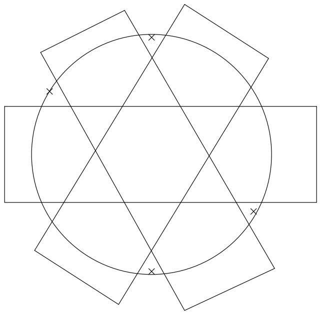
$$\sqrt{\frac{h}{m}} = 1$$

$\Rightarrow$ Need to take the margin into account

Capacity: Shattering dimension

Hyperplanes with Margin

If $\|x\| \leq R$,
$vc(\text{hyperplanes with margin } \rho, 1) \leq R^2/\rho^2$



Margin

- The shattering dimension is related to the margin

- Maximizing the margin means minimizing the shattering dimension

- Small shattering dimension $\Rightarrow$ good control of the risk

$\Rightarrow$ this control is automatic (no need to choose the margin beforehand)

$\Rightarrow$ but requires tuning of regularization parameter

Capacity: Rademacher Averages (1)

- Consider hyperplanes with $\|w\| \leq M$
- Rademacher average

$$\frac{M}{n\sqrt{2}}\sqrt{\sum_{i=1}^{n} k(x_i, x_i)} \leq \mathcal{R}_n \leq \frac{M}{n}\sqrt{\sum_{i=1}^{n} k(x_i, x_i)}$$

- Trace of the Gram matrix
- Notice that $\mathcal{R}_n \leq \sqrt{R^2/(n^2\rho^2)}$

Rademacher Averages (2)

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\|w\|\leq M} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \left\langle w, \delta_{x_i} \right\rangle\right] \\
&= \quad \mathbb{E}\left[\sup_{\|w\|\leq M} \left\langle w, \frac{1}{n}\sum_{i=1}^{n} \sigma_i \delta_{x_i} \right\rangle\right] \\
&\leq \quad \mathbb{E}\left[\sup_{\|w\|\leq M} \|w\| \left\| \frac{1}{n}\sum_{i=1}^{n} \sigma_i \delta_{x_i} \right\|\right] \\
&= \quad \frac{M}{n}\mathbb{E}\left[\sqrt{\left\langle \sum_{i=1}^{n} \sigma_i \delta_{x_i}, \sum_{i=1}^{n} \sigma_i \delta_{x_i} \right\rangle}\right]
\end{aligned}
$$

Rademacher Averages (3)

$$
\begin{aligned}
&\frac{M}{n}\mathbb{E}\left[\sqrt{\left\langle \sum_{i=1}^{n} \sigma_i \delta_{x_i}, \sum_{i=1}^{n} \sigma_i \delta_{x_i} \right\rangle}\right] \\
\leq\quad & \frac{M}{n}\sqrt{\mathbb{E}\left[\left\langle \sum_{i=1}^{n} \sigma_i \delta_{x_i}, \sum_{i=1}^{n} \sigma_i \delta_{x_i} \right\rangle\right]} \\
=\quad & \frac{M}{n}\sqrt{\mathbb{E}\left[\sum_{i,j} \sigma_i \sigma_j \left\langle \delta_{x_i}, \delta_{x_j} \right\rangle\right]} \\
=\quad & \frac{M}{n}\sqrt{\sum_{i=1}^{n} k(x_i, x_i)}
\end{aligned}
$$

Improved rates – Noise condition

- Under Massart's condition ($|\eta| > \eta_0$), with $\|g\|_\infty \leq M$

$$\mathbb{E}\left[(\phi(Yg(X)) - \phi(Yt(X)))^2\right] \leq (M - 1 + 2/\eta_0)(L(g) - L^*) \, .$$

$\longrightarrow$ If noise is nice, variance linearly related to expectation

$\longrightarrow$ Estimation error of order r^* (of the class $\mathcal{G}$)

Improved rates – Capacity (1)

- r_n^* related to decay of eigenvalues of the Gram matrix

$$r_n^* \leq \frac{c}{n} \min_{d \in \mathbb{N}} \left(d + \sqrt{\sum_{j>d} \lambda_j} \right)$$

- Note that $d = 0$ gives the trace bound

- r_n^* always better than the trace bound (equality when λ_i constant)

Improved rates – Capacity (2)

Example: exponential decay

- $\lambda_i = e^{-\alpha i}$
- Global Rademacher of order $\frac{1}{\sqrt{n}}$
- r_n^* of order

$$\frac{\log n}{n}$$

Exponent of the margin

- Estimation error analysis shows that in $\mathcal{G}_M = \{g : \|g\| \leq M\}$

$$R(g_n) - R(g^*) \leq M \ldots$$

- Wrong power (M^2 penalty) is used in the algorithm
- Computationally easier
- But does not give λ a dimension-free status
- Using M could improve the cutoff detection

Kernel

Why is it good to use kernels ?

- Gaussian kernel (RBF)

$$k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

- σ is the width of the kernel

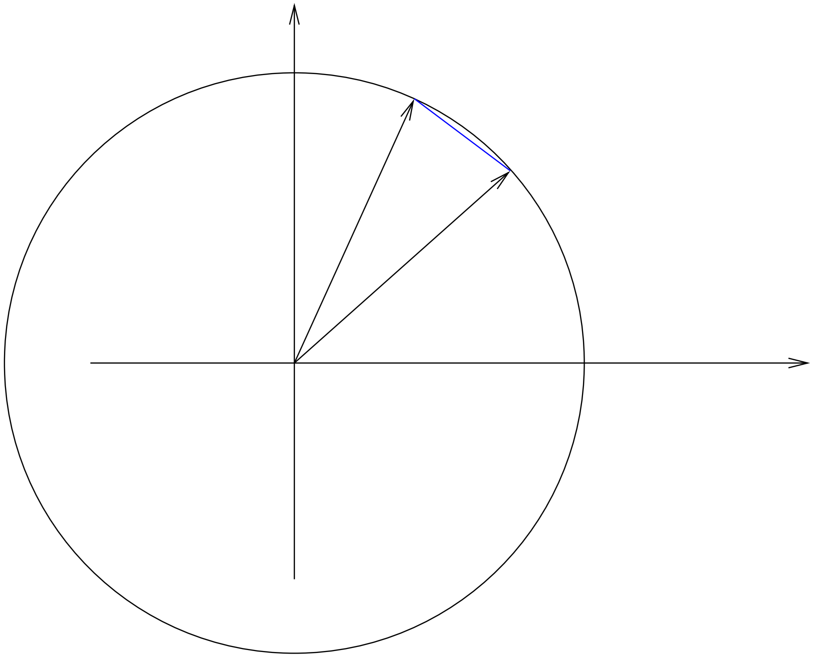
$\longrightarrow$ What is the geometry of the feature space ?

RBF

Geometry

- Norms
$$\|\Phi(x)\|^2 = \langle \Phi(x), \Phi(x) \rangle = e^0 = 1$$
$\longrightarrow$ sphere of radius 1
- Angles
$$\cos(\widehat{\Phi(x), \Phi}(y)) = \left\langle \frac{\Phi(x)}{\|\Phi(x)\|}, \frac{\Phi(y)}{\|\Phi(y)\|} \right\rangle = e^{-\|x-y\|^2/2\sigma^2} \geq 0$$
$\longrightarrow$ Angles less than 90 degrees
- $\Phi(x) = k(x, .) \geq 0$
$\longrightarrow$ positive quadrant

RBF



RBF

Differential Geometry

- Flat Riemannian metric

  $\longrightarrow$ 'distance' along the sphere is equal to distance in input space

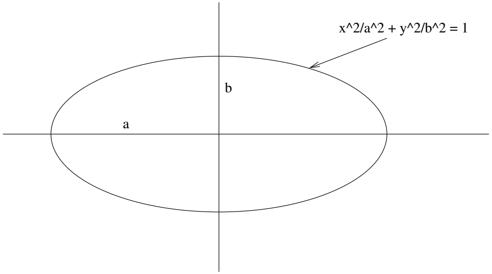
- Distances are contracted

  $\longrightarrow$ 'shortcuts' by getting outside the sphere

RBF

Geometry of the span

Ellipsoid



- $K = (k(x_i, x_j))$ Gram matrix
- Eigenvalues $\lambda_1, \ldots, \lambda_m$
- Data points mapped to ellispoid with lengths $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_m}$

RBF

Universality

- Consider the set of functions

$$\mathcal{H} = \text{span}\{k(x,\cdot) : x \in \mathcal{X}\}$$

- $\mathcal{H}$ is dense in $C(\mathcal{X})$

$\longrightarrow$ Any continuous function can be approximated (in the $\|\|_\infty$ norm) by functions in $\mathcal{H}$

$\Rightarrow$ with enough data one can construct any function

RBF

Eigenvalues

- Exponentially decreasing

- Fourier domain: exponential penalization of derivatives

- Enforces smoothness with respect to the Lebesgue measure in input space

RBF

Induced Distance and Flexibility

- $\sigma \longrightarrow 0$
  1-nearest neighbor in input space
  Each point in a separate dimension, everything orthogonal

- $\sigma \longrightarrow \infty$
  linear classifier in input space
  All points very close on the sphere, initial geometry

- Tuning σ allows to try all possible intermediate combinations

RBF

Ideas

- Works well if the Euclidean distance is good
- Works well if decision boundary is smooth
- Adapt smoothness via σ
- Universal

Choosing the Kernel

- Major issue of current research
- Prior knowledge (e.g. invariances, distance)
- Cross-validation (limited to 1-2 parameters)
- Bound (better with convex class)

$\Rightarrow$ Lots of open questions...

Learning Theory: some informal thoughts

- Need assumptions/restrictions to learn
- Data cannot replace knowledge
- No universal learning (simplicity measure)
- SVM work because of capacity control
- Choice of kernel = choice of prior/ regularizer
- RBF works well if Euclidean distance meaningful
- Knowledge improves performance (e.g. invariances)