Statistical Learning Theory

Olivier Bousquet & Bernhard Schölkopf

Max-Planck-Institut für biologische Kybernetik
72076 Tübingen, Germany

olivier.bousquet@tuebingen.mpg.de

Roadmap

1. Introduction: what is learning ?

2. Statistical Learning Theory: Basics

3. Statistical Learning Theory: Advanced

4. SVM Insights

Learning and Inference

The inductive inference process:

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

$\Rightarrow$ This can be taken as a definition for natural sciences !

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

An Inference Problem

A simple example: sequences of numbers

Question:

$$3, 5, 7, \ldots$$

which numbers should follow ?

$\Rightarrow$ there is no satisfactory single answer.

Possible Solutions (I)

1. Prime numbers

$$3, 5, 7, 11, 13, 17, 19, \ldots$$

2. Odd numbers

$$3, 5, 7, 9, 11, 13, 15, \ldots$$

$\Rightarrow$ more numbers reduce uncertainty ?

Possible Solutions (II)

1. Numbers which end with $3, 5, 7$

$$3, 5, 7, 13, 15, 17, 23 \ldots$$

2. Prime numbers which do not end with 1

$$3, 5, 7, 13, 17, 19, 23 \ldots$$

$\Rightarrow$ What if we change the representation ?

Possible Solutions (II)

Binary representation

$$11, 101, 111, 1101, \ldots$$

$\rightarrow$ what does it mean to finish with $3, 5$ or 7 in this representation ? $(15 = 1111, 17 = 10001, 23 = 10110)$

A simple continuation

$$11, 101, 111, 1101, 1111, 11101, 11111, \ldots$$

which corresponds to

$$3, 5, 7, 13, 15, 29, 31, \ldots$$

$\Rightarrow$ Simplicity is relative !

Philosophy

Inductive inference: philosophical issues

- Can we discover the laws of Nature by observing it ?
- What is a scientific theory ?
- What is inference ?

Philosophy

- Aristotle: the best demonstration is the one using the least number of hypotheses (because Nature is simple and what is simple is beautiful)

- Epicurius: if several explanations are compatible with the observations, one should keep them all

- Indifference principle (probability): without information, one consider all hypotheses are equiprobable

Philosophy

- Occam's Razor: Entities should not be multiplied beyond necessity (because this is an efficient method to get to the truth)

- Mach: economy principle (simple is more economical in terms of number of experiments needed to confirm)

- Jeffreys: prior ordering of hypotheses using number of parameters

- Popper: falsifiability, more empirical content means easier to falsify (require less experiments), but number of parameters also

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
- ...

Theoretical Computer Science

A candidate universal notion of complexity
Kolmogorov Complexity

Definition: Given a binary string $x = 011010011.....$, $K(x)$ is the length of the shortest program that generates x.

- Need to choose a programming language (Universal Turing Machine)
- Non-computable

$\Rightarrow$ still relative !! (some things are easier in a language than in another)

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
  $\rightarrow$ *Stationarity* of the phenomenon

- Constraints on the phenomenon
  $\rightarrow$ 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 underlyin phenomenon (here a probability distribution)

Probabilistic Model

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

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

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

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

Probabilistic Model

Criterion to choose our function:

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

$$R(f) = P(f(X) \neq Y) = \int \mathbf{1}_{[f(X) \neq Y]} dP(X, Y)$$

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

$$R_{emp}(f) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{[f(X_i) \neq Y_i]}$$

Assumptions about P

Need assumptions about P.
Indeed, if P is $P_X \times P(Y|X)$ with P_X uniform and $P(Y|X)$ totally chaotic, there is no possible generalization from finite data.
Assumptions can be

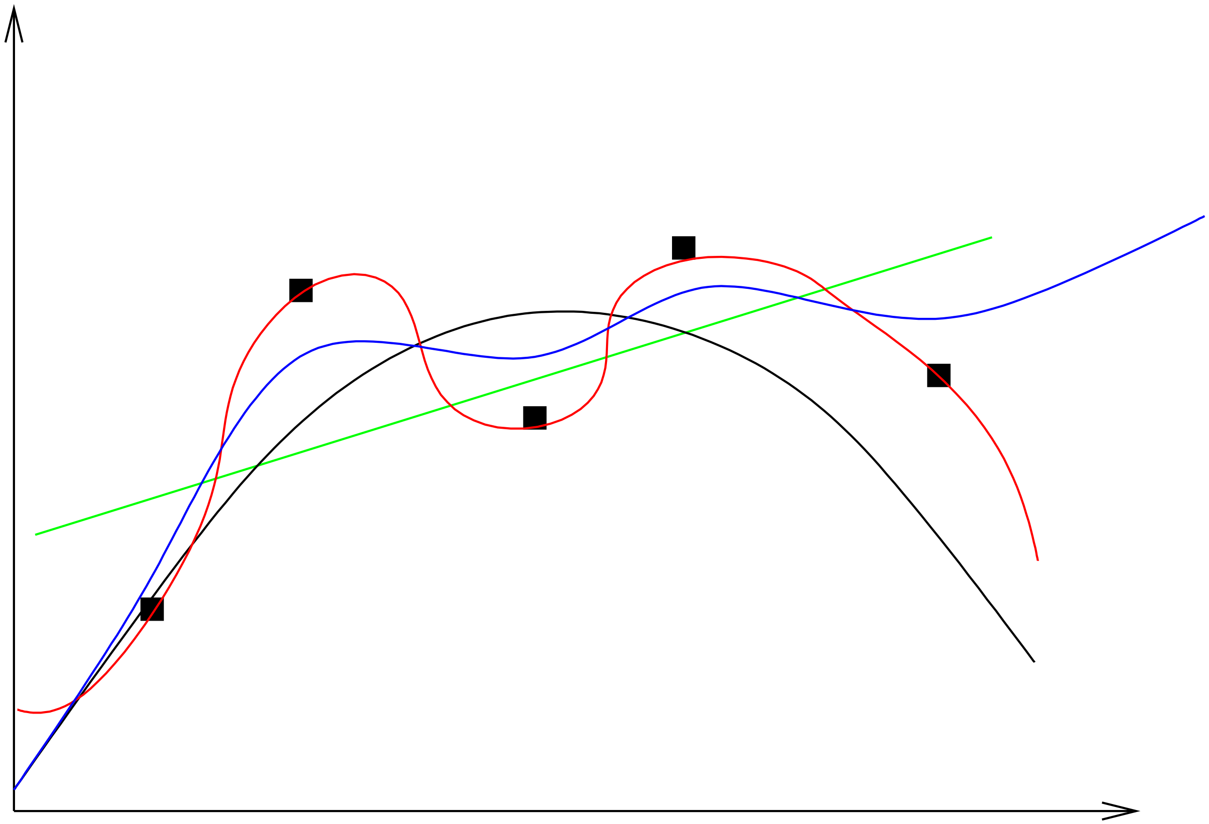
- Preference (e.g. a priori probability distribution on possible functions)
- Restriction (set of possible functions)

Treating lack of knowledge

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

Approximation/Interpolation

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 $\mathcal{F}$ (set of possible functions)

- Minimize the empirical risk in the model

$$\min_{f \in \mathcal{F}} R_{emp}(f)$$

What if the Bayes classifier is not in the model ?

Structural Risk Minimization

- Choose a collection of models $\{\mathcal{F}_d : d = 1, 2, \ldots\}$

- Minimize the empirical risk in each model

- Minimize the penalized empirical risk

$$\min_d \min_{f \in \mathcal{F}_d} R_{emp}(f) + \mathrm{pen}(d)$$

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

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

Regularization

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

$$\min_{f\in\mathcal{F}} R_{emp}(f) + \lambda \|f\|^2$$

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

Most methods can be thought of as regularization methods.

Bounds

A learning algorithm

- Takes as input the data $(X_1, Y_1), \ldots, (X_m, Y_m)$
- Produces a function f_m

Can we estimate the risk of f_m ?

- Error bounds

$$R(f_m) \leq R_{emp}(f_m) + B$$

- Relative error bounds

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

$\Rightarrow$ they are probabilistic in nature

The Law of Large Numbers

- Notice that

$$R_{emp}(f) - R(f) = \frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}[Z]$$

with $Z = \mathbf{1}_{[f(X)\neq Y]}$, is the difference between the expectation and the empirical average of a random variable.

- The law of large numbers says

$$\mathbb{P}\left[\lim_{n\to\infty}\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}[Z_1] = 0\right] = 1\,.$$

$\Rightarrow$ can we quantify it ?

Hoeffding's Inequality

Quantitative version of law of large numbers.

Assumes bounded random variables

Theorem 1 *Let* $Z_1, \ldots, Z_m$ *be* m *i.i.d. random variables with values in* $[a, b]$*. Then for all* $\varepsilon > 0$*, we have*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}\left[Z_1\right]\right| > \varepsilon\right] \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) .$$

Let's rewrite it to better understand

Hoeffding's Inequality

Write

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

Then

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}\left[Z_1\right]\right| > (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \leq \delta$$

or with probability at least $1-\delta$,

$$\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}\left[Z_1\right]\right| \leq (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Hoeffding's inequality

Let's apply to $Z = \mathbf{1}_{[f(X) \neq Y]}$, $Z \in [0, 1]$.

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

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} . \tag{1}$$

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

What we need to bound is

$$R(f_m) - R_{emp}(f_m)$$

where f_m is the function choosen by the algorithm based on the data.

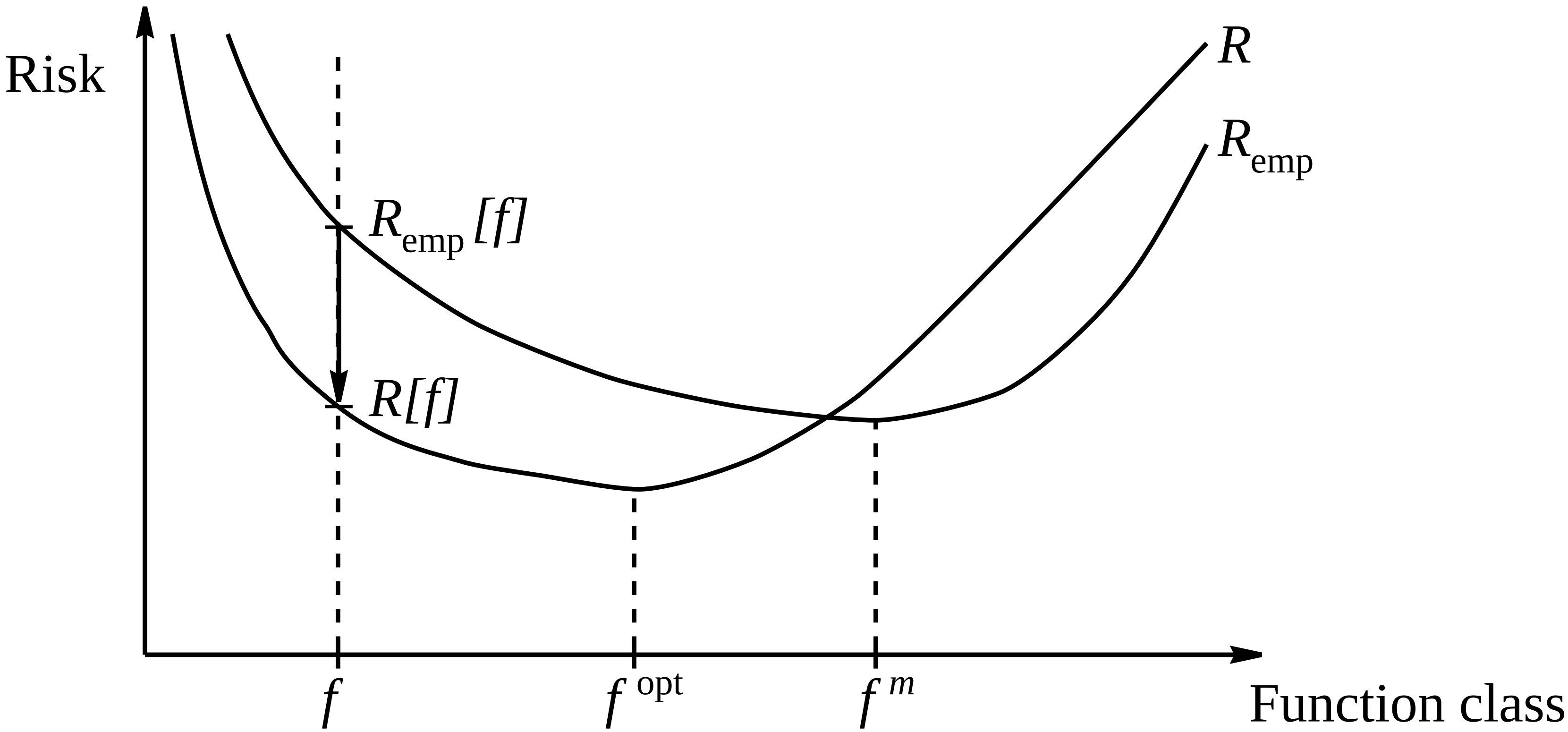
For a fixed sample, there exists a function f such that

$$R(f) - R_{emp}(f) = 1$$

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

$$R(f_m) - R_{emp}(f_m) \leq \sup_{f \in \mathcal{F}} (R(f) - R_{emp}(f))$$

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

Union Bound

Consider two functions f_1 and f_2.

For $i = 1, 2$ define the 'bad' set as

$$C_i = \left\{(x_1, y_1), \ldots, (x_m, y_m) : R(f_i) - R_{emp}(f_i) > \varepsilon\right\}$$

Hoeffding gives for each i

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

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

$$\begin{aligned} \mathbb{P}\left[C_1 \cup C_2\right] &= \mathbb{P}\left[C_1\right] + \mathbb{P}\left[C_2\right] - \mathbb{P}\left[C_1 \cap C_2\right] \\ &\leq \mathbb{P}\left[C_1\right] + \mathbb{P}\left[C_2\right] \end{aligned}$$

Finite Case

More generally

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

We have

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

Finite Case

We obtain, for $\mathcal{F} = \{f_1, \ldots, f_N\}$, for all $\delta > 0$

with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F},\ \ R(f) \leq R_{emp}(f) + \sqrt{\frac{\log N + \log \frac{2}{\delta}}{2m}}$$

This is a generalization bound !

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

Approximation/Estimation

Let

$$f^* = \arg\min_{f \in \mathcal{F}} R(f)$$

If f_m minimizes the empirical risk in $\mathcal{F}$,

$$R_{emp}(f^*) - R_{emp}(f_m) \geq 0$$

Thus

$$\begin{aligned} R(f_m) &= R(f_m) - R(f^*) + R(f^*) \\ &\leq R_{emp}(f^*) - R_{emp}(f_m) + R(f_m) - R(f^*) + R(f^*) \\ &\leq 2 \sup_{f \in \mathcal{F}} |R(f) - R_{emp}(f)| + R(f^*) \end{aligned}$$

Approximation/Estimation

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

$$R(f_m) \leq R(f^*) + 2\sqrt{\frac{\log N + \log \frac{2}{\delta}}{2m}}$$

The first term decreases if N increases

The second term increases

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

Infinite Case

Measure of the size of an infinite class ?

Consider

$$F(x_1, \ldots, x_m) = \{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}\}$$

The size of F is the number of possible ways in which the data $(x_1, \ldots, x_m)$ can be classified.

Growth function

$$S_{\mathcal{F}}(m) = \sup_{(x_1,\ldots,x_m)} |F(x_1, \ldots, x_m)|$$

Infinite Case

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

$$\forall f \in \mathcal{F}, \; R(f) \leq R_{emp}(f) + \sqrt{\frac{\log S_{\mathcal{F}}(m) + \log \frac{4}{\delta}}{8m}}$$

How to compute $S_{\mathcal{F}}(m)$?

$\Rightarrow$ use VC dimension

VC Dimension

Notice that since $f \in \{-1, 1\}$, $S_{\mathcal{F}}(m) \leq 2^m$

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

Definition 2 *The* VC-dimension *of $\mathcal{F}$ is the largest m such that*

$$S_{\mathcal{F}}(m) = 2^m$$

VC Dimension

Examples

- Hyperplanes in $\mathbb{R}^d$
  VC dimension $h = d + 1$
- $\sin(tx)$, $t \in \mathbb{R}$
  Infinite VC dimension
- Hyperplanes in $\mathbb{R}^d$ with margin ρ
  VC dimension
  $$h \leq \frac{R^2}{\rho^2}$$
  if $||x|| \leq R$.

How are $S_{\mathcal{F}}(m)$ and h related ?

Sauer Lemma

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

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

and for all $m \geq h$,

$$S_{\mathcal{F}}(m) \leq \left(\frac{em}{h}\right)^h.$$

Notice that for $m \leq h$, $S_{\mathcal{F}}(m) = 2^m$

$\Rightarrow$ phase transition

VC Bound

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

With probability at least $1-\delta$

$$\forall f \in \mathcal{F},\ R(f) \leq R_{emp}(f) + \sqrt{\frac{h \log \frac{em}{h} + \log \frac{4}{\delta}}{8m}}$$

So the error is of order

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

Interpretation

VC dimension: measure of effective dimension

- Depends on the goal to achieve (reduce overfitting)
- Gives a natural definition of simplicity
- Not related to the number of parameters
- Impossible to learn if the VC dimension is infinite (falsifiability)

Other Capacity Measures

Covering numbers

- Define a distance d between functions, e.g.

$$d(f, f') = |\{f(x_i) \neq f'(x_i) : i = 1, \ldots, n\}|$$

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

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} R(f) - R_{emp}(f) > \varepsilon\right] \leq \mathbb{E}\left[N(\mathcal{F}, \varepsilon)\right] \exp(-n\varepsilon^2/8)$$

Proof Strategy (Gurvits, 1997)

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

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

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) over $i = 1, \ldots, r$ yields

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

By the Cauchy-Schwarz inequality, on the other hand, we have

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

Combine both:

$$\frac{r}{\Lambda} \leq \left\| \sum_{i=1}^{r} y_i \mathbf{x}_i \right\|. \tag{3}$$

Part II

Consider independent random labels $y_i \in \{\pm 1\}$, uniformly distributed (*Rademacher variables*).

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

Part II, ctd.

Since $\|\mathbf{x}_i\| \leq R$, we get

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

- 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: $\left(\frac{r}{\Lambda}\right)^2 \leq \|\sum_{i=1}^r y_i \mathbf{x}_i\|^2$

Part II: $\|\sum_{i=1}^r y_i \mathbf{x}_i\|^2 \leq rR^2$

Hence

$$\frac{r^2}{\Lambda^2} \leq rR^2,$$

i.e.,

$$r \leq R^2 \Lambda^2,$$

completing the proof.

Concentration

Hoeffding's inequality is a concentration inequality
When m increases, the average is concentrated around the expectation
Generalization

Theorem 4 (McDiarmid's Inequality) *Let $Z_1, \ldots, Z_m$ be m i.i.d. random variables and let $T = F(Z_1, \ldots, Z_m)$ be a function such that there exists a constant c satisfying*

$$|F(z_1, \ldots, z_i, \ldots, z_m) - F(z_1, \ldots, z_i', \ldots, z_m)| \leq c\,,$$

for any $z_1, \ldots, z_m, z_i'$ and any $i = 1, \ldots, m$. Then we have for all $\varepsilon > 0$,

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

Application, I

We want to apply it to $Z = \sup_{f \in \mathcal{F}} R(f) - R_{emp}(f)$.
Notice that

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

Hence

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

Applied to Z this gives

$$|\sup_{f \in \mathcal{F}} (R(f) - R_{emp}(f)) - \sup_{f \in \mathcal{F}} (R(f) - R'_{emp}(f))| \leq \sup_{f \in \mathcal{F}} (R'_{emp}(f) - R'_{emp}(f))$$

R'_{emp} empirical risk with one point changed,

Application, II

For a given $f : \mathbf{X} \to \{-1, 1\}$,

$$R'_{emp}(f) - R_{emp}(f) = \frac{1}{m}(\mathbf{1}_{[f(x'_i) \neq y'_i]} - \mathbf{1}_{[f(x_i) \neq y_i]}) \leq \frac{1}{m}.$$

thus

$$|\sup_{f \in \mathcal{F}}(R(f) - R_{emp}(f)) - \sup_{f \in \mathcal{F}}(R(f) - R'_{emp}(f))| \leq \frac{1}{m}$$

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

Application

Proposition 5 *For any confidence level $\delta > 0$, with probability at least $1 - \delta$ over the random choice of the data, we have*

$$\sup_{f \in \mathcal{F}} R(f) - R_{emp}(f) \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(R[f] - R_{emp}[f]\right)\right] + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{4}$$

Bound holds *uniformly* over the class of functions $\mathcal{F}$

However, the expectation appearing on the right-hand side still has to be computed

Symmetrization

Rademacher variables

$\sigma_1, \ldots, \sigma_m$ independent random variables with

$$\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2}$$

Symmetrization lemma

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} R[f] - R_{emp}[f]\right] \leq 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \mathbf{1}_{[f(X_i) \neq Y_i]}\right].$$

Expectation is taken with respect to X_i, Y_i and σ_i

Rademacher Averages

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\mathbf{1}_{[f(X_i)\neq Y_i]}\right]$$

$$=\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\frac{1}{2}(1-Y_if(X_i))\right]$$

$$=\mathbb{E}\left[\frac{1}{2m}\sum_{i=1}^{m}\sigma_i\right]+\frac{1}{2}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_iY_if(X_i)\right]$$

Rademacher Averages

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{2m}\sum_{i=1}^{m}\sigma_i\right] + \frac{1}{2}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i Y_i f(X_i)\right] \\
&= \mathbb{E}\left[\frac{1}{2m}\sum_{i=1}^{m}\sigma_i\right] - \frac{1}{2}\mathbb{E}\left[\inf_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i Y_i f(X_i)\right] \\
&= -\frac{1}{2}\mathbb{E}\left[\inf_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(X_i)\right] \\
&= -\frac{1}{2}\mathbb{E}\left[\inf_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i f(X_i)\right]
\end{aligned}
$$

Rademacher Averages

$$
\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\inf_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i f(X_i)\right] \\
&= \frac{1}{2}-\mathbb{E}\left[\inf_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(1-\sigma_i f(X_i))\right] \\
&= \frac{1}{2}-\mathbb{E}\left[\inf_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{[f(X_i)\neq\sigma_i]}\right]
\end{aligned}
$$

Intuition: capacity of $\mathcal{F}$ to fit random noise

Concentration

Let

$$Z = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \mathbf{1}_{[f(X_i) \neq Y_i]} \right]$$

Expectation with respect to σ_i only, with (X_i, Y_i) fixed.

Z satisfies McDiarmid's assumptions
$\Rightarrow \mathbb{E}[Z]$ can be estimated by Z on the data

Data-dependent Bound

Proposition 6 *Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $[-1,1]$. For any confidence level $\delta > 0$, with probability at least $1-\delta$ over the random choice of the data, we have*

$$\sup_{f\in\mathcal{F}} \left(R[f] - R_{emp}[f]\right) \leq 2\mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \mathbf{1}_{[f(X_i)\neq Y_i]}\right] + \sqrt{\frac{2\log\frac{2}{\delta}}{m}},$$

where the expectation is taken with respect to the σ_i only.

Relationship with VC dimension

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

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(X_i) \right] \leq 2\sqrt{\log N}$$

Consequence for VC classes

Lemma 7 *Let $\mathcal{F}$ be a class of functions with finite VC-dimension h. Then for all $m \in \mathbb{N}$,*

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(X_i) \right] \leq 2\sqrt{\frac{h \log \frac{em}{h}}{m}} \, .$$

SVM Insights

Why do SVM work ?

- Computational: Convex optimization

- Capacity Control: Regularization

- Universality: Kernel

Formulation

- Soft margin

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

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

$$\xi_i \geq 0$$

Convex objective function and convex constraints

Linearization

Free vector space: define the set $V(\mathbf{X})$ of (formal) linear combinations of elements from $\mathbf{X}$

$$V(\mathbf{X}) = \{\sum_{i \in I} \alpha_i \delta_{x_i} : \alpha_i \in \mathbb{R},\ x_i \in \mathbf{X},\ |I| < \infty\}\,.$$

Any function f from $\mathbf{X}$ to $\mathbf{Y}$ can be represented as a linear function on $V(\mathbf{X})$:

$L_f(\sum \alpha_i \delta_{x_i}) = \sum \alpha_i f(x_i)$

Everything is linear

Seems like rewriting but it is at the heart of the kernel approach.

To get a kernel (reproducing kernel Hilbert space), simply define an inner product on $V(X)$ with a kernel function.

Convexity

Consider now real valued functions

Linearity eases computations

Convexity gives even simpler computations $\rightarrow$ choose a convex loss function

VC dimension

The VC dimension of the set of hyperplanes is $d + 1$.

The feature space has dimension m for RBF kernel

The VC bound does not give any information

Need scale-sensitive approach

Regularization

Capacity control by restricting the class

$$\min_{\|f\| \leq R} L_m(f)$$

Capacity control by regularization

$$\min_f L_m(f) + \lambda \|f\|^2$$

Loss Functions

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

- Convex, non-increasing
- Upper bounds $\mathbf{1}_{[Yf(X) \leq 0]}$
- Is minimized by Bayes classifier

Rademacher Averages (I)

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\|w\|\le M} \frac{1}{m}\sum_{i=1}^m \sigma_i \left\langle w, \Phi(x_i)\right\rangle\right] \\
&= \mathbb{E}\left[\sup_{\|w\|\le M} \left\langle w, \frac{1}{m}\sum_{i=1}^m \sigma_i \Phi(x_i)\right\rangle\right] \\
&\le \mathbb{E}\left[\sup_{\|w\|\le M} \|w\| \left\|\frac{1}{m}\sum_{i=1}^m \sigma_i \Phi(x_i)\right\|\right] \\
&= \frac{M}{m}\mathbb{E}\left[\sqrt{\left\langle \sum_{i=1}^m \sigma_i \Phi(x_i), \sum_{i=1}^m \sigma_i \Phi(x_i)\right\rangle}\right]
\end{aligned}
$$

Rademacher Averages (II)

$$
\begin{aligned}
&\frac{M}{m}\mathbb{E}\left[\sqrt{\left\langle \sum_{i=1}^{m} \sigma_i \Phi(x_i), \sum_{i=1}^{m} \sigma_i \Phi(x_i) \right\rangle}\right] \\
&\leq \frac{M}{m}\sqrt{\mathbb{E}\left[\left\langle \sum_{i=1}^{m} \sigma_i \Phi(x_i), \sum_{i=1}^{m} \sigma_i \Phi(x_i) \right\rangle\right]} \\
&= \frac{M}{m}\sqrt{\mathbb{E}\left[\sum_{i,j} \sigma_i \sigma_j \left\langle \Phi(x_i), \Phi(x_j) \right\rangle\right]} \\
&= \frac{M}{m}\sqrt{\sum_{i=1}^{m} \|\Phi(x_i)\|^2}
\end{aligned}
$$

Ellipsoid

Proposition 8

RBF

Geometry

- Norms

$$\|\Phi(x)\|^2 = \langle \Phi(x), \Phi(x) \rangle = e^0 = 1$$

$\rightarrow$ sphere of radius 1

- Angles

$$\cos(\widehat{\Phi(x), \Phi(y)}) = \left\| \frac{\Phi(x)}{\|\Phi(x)\|}, \frac{\Phi(y)}{\|\Phi(y)\|} \right\| = e^{-\|x-y\|^2/2\sigma^2} \geq 0$$

$\rightarrow$ positive quadrant

RBF

- Flat Riemannian metric
  $\rightarrow$ 'distance' along the sphere is equal to distance in input space
- Distances are contracted
  $\rightarrow$ 'shortcuts' by getting outside the sphere

RBF

Universality

Let k be the RBF kernel with a fixed width.
Let $\mathcal{H}$ be the corresponding reproducing kernel Hilbert space

Proposition 9 $\mathcal{H}$ *is dense in* $C(\mathbf{X})$

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

- $\sigma \to 0$ 1-nearest neighbor in input space
  Each point in a separate dimension, everything orthogonal
- $\sigma \to \infty$ linear classifier in input space
  All points very close on the sphere, initial geometry
- Tuning

RBF

Ideas

Works well if the Euclidean distance is good

Choosing the Kernel

- Major issue
- Prior knowledge
- Cross-validation
- Bound (better with convex class)

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 (e.g. invariances)