Classification and statistical machine learning

Sylvain Arlot

http://www.di.ens.fr/~arlot/

[1]CNRS

[2]École Normale Supérieure (Paris), DI/ENS, Équipe SIERRA

CEMRACS 2013, July 26th, 2013

Outline

Hand-written digit recognition (MNIST)

 $\Rightarrow$?

http://yann.lecun.com/exdb/mnist/

Object recognition

American flag:      ...

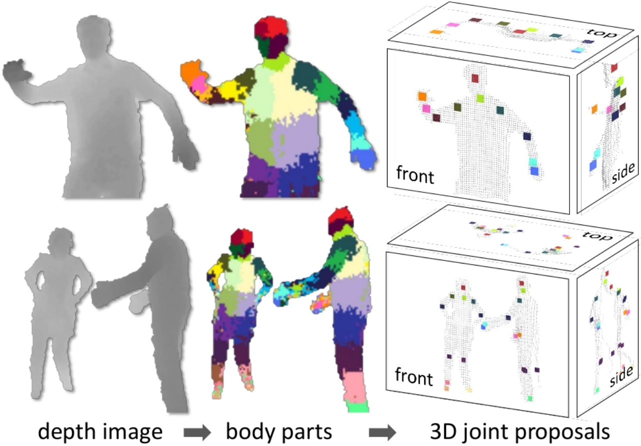
Butterfly:      ...

Teddy bear:      ...

$\Rightarrow$ American flag? Butterfly? Teddy bear? ...

http://www.vision.caltech.edu/Image_Datasets/Caltech256/

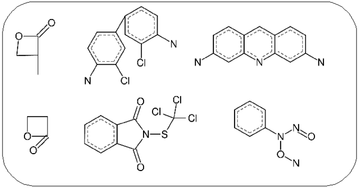
Kinect: body part recognition [Shotton et al., 2011]



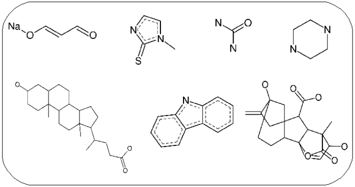
http://research.microsoft.com/en-us/projects/vrkinect/

Predict biochemical properties of molecules from structure

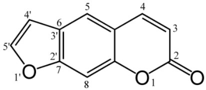
Mutagenic compounds



Non-mutagenic compounds



A compound with unknown properties:



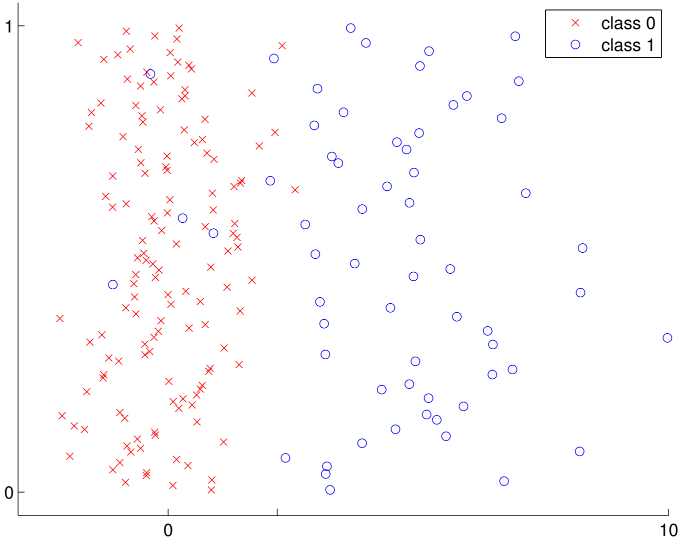
Is it likely to be mutagenic or not?

[Mahé et al., 2005, Shervashidze et al., 2011] Figure obtained from Koji Tsuda

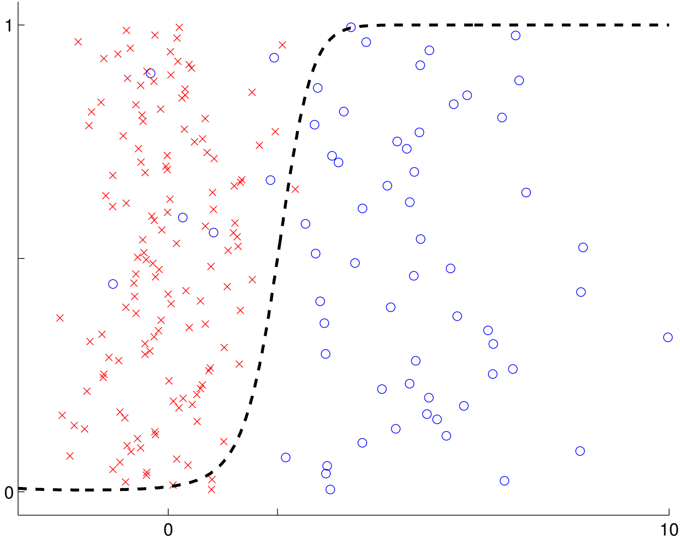
Many other applications

- Bioinformatics:
  - sequencing data for diagnosis and prognosis (cancer, ...)
  - personalized medicine
  - ...
- Text classification:
  - Spam detection
  - Google ads
  - Automatic document classification
- Action recognition in videos
- Speech recognition
- Credit scoring
- ...

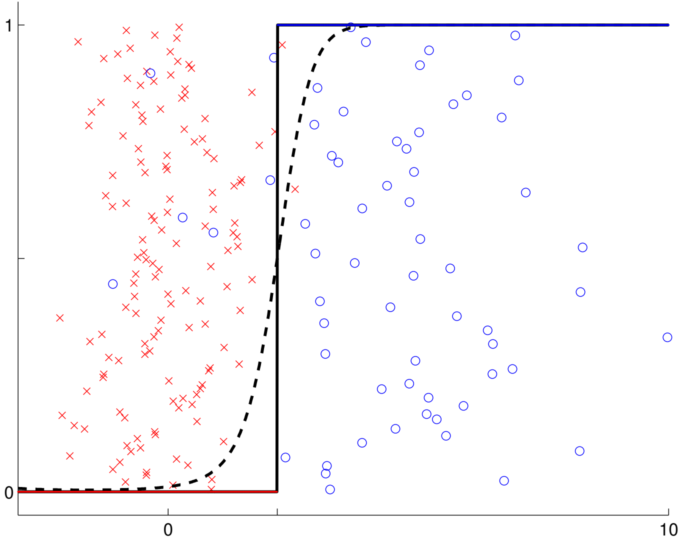
Classification in $\mathbb{R}$: data



Classification in $\mathbb{R}$: regression function



Classification in $\mathbb{R}$: Bayes classifier



Binary supervised classification

- Data D_n : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ (i.i.d. $\sim P$)

Binary supervised classification

- Data D_n : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ (i.i.d. $\sim P$)
- Classifier: $f : \mathcal{X} \to \{0, 1\}$ measurable

Binary supervised classification

- Data D_n : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ (i.i.d. $\sim P$)
- Classifier: $f : \mathcal{X} \to \{0, 1\}$ measurable
- Cost/Loss function $\ell(f(x), y)$ measures how well $f(x)$ "predicts" y
  For this talk: $\ell(y, y') = \mathbb{1}_{y \neq y'}$ (0–1 loss)

Binary supervised classification

- Data D_n : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ (i.i.d. $\sim P$)
- Classifier: $f : \mathcal{X} \to \{0, 1\}$ measurable
- Cost/Loss function $\ell(f(x), y)$ measures how well $f(x)$ "predicts" y
  For this talk: $\ell(y, y') = \mathbb{1}_{y \neq y'}$ (0–1 loss)
- Goal: learn $f \in \mathbb{S} = \{$measurable functions $\mathcal{X} \to \{0, 1\}\}$ s.t. the risk

$$\mathcal{R}(f) := \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)] = \mathbb{P}(f(X) \neq Y)$$

is minimal.

Binary supervised classification

- Data D_n : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ (i.i.d. $\sim P$)
- Classifier: $f : \mathcal{X} \to \{0, 1\}$ measurable
- Cost/Loss function $\ell(f(x), y)$ measures how well $f(x)$ "predicts" y
  For this talk: $\ell(y, y') = \mathbb{1}_{y \neq y'}$ (0–1 loss)
- Goal: learn $f \in \mathbb{S} = \{\text{measurable functions } \mathcal{X} \to \{0, 1\}\}$ s.t. the risk

$$\mathcal{R}(f) := \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)] = \mathbb{P}(f(X) \neq Y)$$

  is minimal.

- Remark: asymmetric cost $\ell_w(f(x), y) = w(y)\mathbb{1}_{f(x)\neq y}$ with $w(0) \neq w(1) > 0$ (spams, medical diagnosis).

Bayes estimator and excess risk

- Bayes classifier: $f^\star \in \operatorname{argmin}_{f \in \mathbb{S}} \{ \mathcal{R}(f) \}$

Bayes estimator and excess risk

- Bayes classifier: $f^\star \in \operatorname{argmin}_{f \in \mathbb{S}} \{ \mathcal{R}(f) \}$

Proposition

In binary classification with the 0–1 loss,

$$f^\star(X) = \mathbb{1}_{\eta(X) \geq 1/2} \qquad (\textit{except maybe on } \{ \eta(X) = 1/2 \})$$

where $\eta(X) = \mathbb{P}(Y = 1 \mid X)$ *is the regression function.*

Bayes estimator and excess risk

- Bayes classifier: $f^\star \in \operatorname{argmin}_{f \in \mathbb{S}} \{ \mathcal{R}(f) \}$

Proposition

In binary classification with the 0–1 loss,

$$f^\star(X) = \mathbb{1}_{\eta(X) \geq 1/2} \qquad (\textit{except maybe on } \{ \eta(X) = 1/2 \})$$

where $\eta(X) = \mathbb{P}(Y = 1 \mid X)$ *is the regression function.*
The Bayes risk is $\mathcal{R}(f^\star) = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}]$
and the excess risk of any $f \in \mathbb{S}$ *is*

$$\mathcal{R}(f) - \mathcal{R}(f^\star) = \mathbb{E}\left[|2\eta(X) - 1| \, \mathbb{1}_{f(X) \neq f^\star(X)} \right] .$$

Remark: for the asymmetric cost ℓ_w, a similar result holds with $1/2$ replaced by $w(0)/(w(0) + w(1))$.

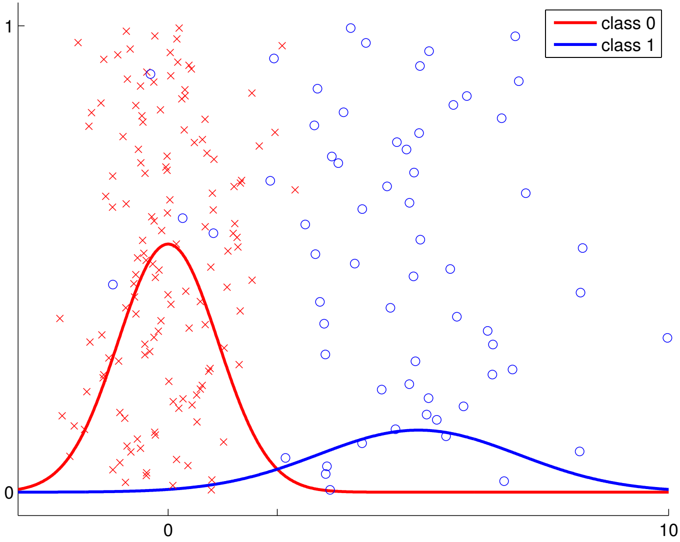
Bayes estimator and excess risk: proof

$$\begin{aligned}\mathbb{P}\left(f(X) \neq Y \mid X\right) &= \mathbb{P}\left(Y = 1\right) \mathbb{1}_{f(X)\neq 1} + \mathbb{P}\left(Y = 0\right) \mathbb{1}_{f(X)\neq 0} \\ &= \eta(X)\mathbb{1}_{f(X)=0} + (1-\eta(X))\mathbb{1}_{f(X)=1} \\ &\geq \min\left\{\eta(X), 1-\eta(X)\right\}\end{aligned}$$

with equality if and only if $\eta(X) = 1/2$ or $f(X) = \mathbb{1}_{\eta(X)\geq 1/2}$. The first two results follow by integrating over X.
Then, the excess risk is equal to

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{f(X)\neq Y} - \mathbb{1}_{f^\star(X)\neq Y}\right] \\ &= \mathbb{E}\left[\mathbb{1}_{f(X)\neq f^\star(X)}\left(\mathbb{1}_{f(X)\neq Y} - \mathbb{1}_{f^\star(X)\neq Y}\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{f(X)\neq f^\star(X)}\left(\mathbb{1}_{f(X)\neq Y} - \mathbb{1}_{f^\star(X)\neq Y}\right) \,\middle|\, X\right]\right] \\ &= \mathbb{E}\left[\mathbb{1}_{f(X)\neq f^\star(X)}\left(\max\left\{\eta(X), 1-\eta(X)\right\} - \min\left\{\eta(X), 1-\eta(X)\right\}\right)\right] \\ &= \mathbb{E}\left[\left|2\eta(X)-1\right| \mathbb{1}_{f(X)\neq f^\star(X)}\right] \quad \square\end{aligned}$$

Classification seen as a testing problem



Classification seen as a testing problem

- f_i: density of $P_i = \mathcal{L}(X \mid Y = i)$ for $i = 0, 1$
- Regression function

$$\eta(x) = \frac{\mathbb{P}(Y=1) f_1(x)}{\mathbb{P}(Y=0) f_0(x) + \mathbb{P}(Y=1) f_1(x)}$$

- Bayes predictor

$$f^\star(x) = \mathbb{1}_{\eta(x) \geq \frac{1}{2}} = \mathbb{1}_{\frac{f_1(x)}{f_0(x)} \geq \frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)}}$$

Classification seen as a testing problem

- f_i: density of $P_i = \mathcal{L}(X \mid Y = i)$ for $i = 0, 1$
- Regression function

$$\eta(x) = \frac{\mathbb{P}(Y=1) f_1(x)}{\mathbb{P}(Y=0) f_0(x) + \mathbb{P}(Y=1) f_1(x)}$$

- Bayes predictor

$$f^\star(x) = \mathbb{1}_{\eta(x) \geq \frac{1}{2}} = \mathbb{1}_{\frac{f_1(x)}{f_0(x)} \geq \frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)}}$$

$\Leftrightarrow$ likelihood-ratio test $\mathbb{1}_{\frac{f_1(x)}{f_0(x)} \geq t}$ of
H_0: "$X \sim P_0$" against H_1: "$X \sim P_1$".

Outline

Classification rule/algorithm

- Classification rule

$$\widehat{f} : \bigcup_{n \geq 1} (\mathcal{X} \times \{0,1\})^n \to \mathbb{S}$$

- Input: a data set D_n (of any size $n \geq 1$)
- Output: a classifier $\widehat{f}(D_n)$: $\mathcal{X} \to \{0,1\}$

Classification rule/algorithm

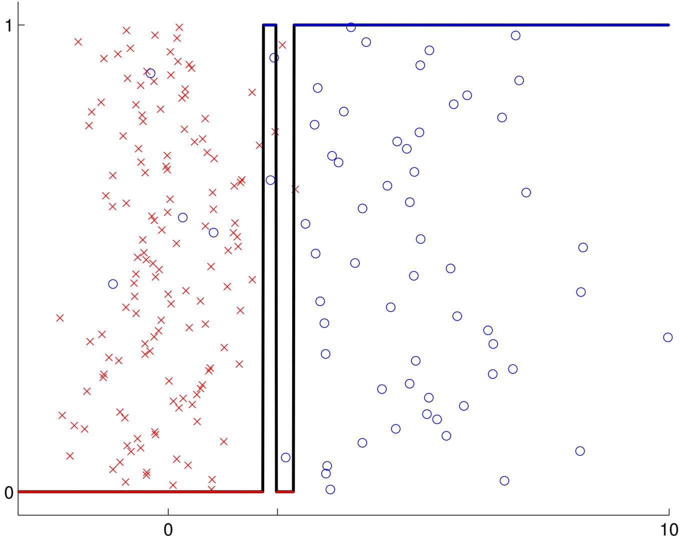
- Classification rule

$$\widehat{f} : \quad \bigcup_{n \geq 1} (\mathcal{X} \times \{0,1\})^n \to \mathbb{S}$$

- Input: a data set D_n (of any size $n \geq 1$)
- Output: a classifier $\widehat{f}(D_n)$: $\mathcal{X} \to \{0,1\}$

- Example: k-nearest neighbours (k-NN):
$x \in \mathcal{X} \to$ majority vote among the Y_i such that X_i is one of the k nearest neighbours of x in $X_1, \ldots, X_n$

Example: 3-nearest neighbours



Universal consistency

- weak consistency: $\mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n))\right] \xrightarrow[n\to\infty]{} \mathcal{R}(f^\star)$

Universal consistency

- weak consistency: $\mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n))\right] \xrightarrow[n\to\infty]{} \mathcal{R}(f^\star)$
- strong consistency: $\mathcal{R}(\widehat{f}(D_n)) \xrightarrow[n\to\infty]{a.s.} \mathcal{R}(f^\star)$

Universal consistency

- weak consistency: $\mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n))\right] \xrightarrow[n\to\infty]{} \mathcal{R}(f^\star)$
- strong consistency: $\mathcal{R}(\widehat{f}(D_n)) \xrightarrow[n\to\infty]{a.s.} \mathcal{R}(f^\star)$

- universal (weak) consistency: for all P, $\mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n))\right] \xrightarrow[n\to\infty]{} \mathcal{R}(f^\star)$
- universal strong consistency: for all P, $\mathcal{R}(\widehat{f}(D_n)) \xrightarrow[n\to\infty]{a.s.} \mathcal{R}(f^\star)$

Universal consistency

- weak consistency: $\mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n))\right] \xrightarrow[n\to\infty]{} \mathcal{R}(f^\star)$
- strong consistency: $\mathcal{R}(\widehat{f}(D_n)) \xrightarrow[n\to\infty]{a.s.} \mathcal{R}(f^\star)$

- universal (weak) consistency: for all P, $\mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n))\right] \xrightarrow[n\to\infty]{} \mathcal{R}(f^\star)$
- universal strong consistency: for all P, $\mathcal{R}(\widehat{f}(D_n)) \xrightarrow[n\to\infty]{a.s.} \mathcal{R}(f^\star)$

- Stone's theorem [Stone, 1977]: If $\mathcal{X} = \mathbb{R}^d$ with the Euclidean distance, k_n-NN is (weakly) universally consistent if $k_n \to +\infty$ and $k_n/n \to 0$ as $n \to +\infty$.

Uniform universal consistency?

- universal weak consistency:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \overline{\lim_{n \to +\infty}} \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) = 0$$

- uniform universal weak consistency:

$$\overline{\lim_{n \to +\infty}} \sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \right\} = 0$$

that is, a common learning rate for all P?

Uniform universal consistency?

- universal weak consistency:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \overline{\lim_{n \to +\infty}} \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) = 0$$

- uniform universal weak consistency:

$$\overline{\lim_{n \to +\infty}} \sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \right\} = 0$$

  that is, a common learning rate for all P?

- Yes if $\mathcal{X}$ is finite.
- No otherwise (see Chapter 7 of [Devroye et al., 1996]).

Classification on $\mathcal{X}$ finite

Theorem

If $\mathcal{X}$ is finite and $\widehat{f}^{\mathrm{maj}}$ is the majority vote rule (for each $x \in \mathcal{X}$, majority vote among $\{ Y_i \,/\, X_i = x \}$),

$$\sup_P \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}^{\mathrm{maj}}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \leq \sqrt{\frac{\mathrm{Card}(\mathcal{X}) \log(2)}{2n}} \enspace .$$

Classification on $\mathcal{X}$ finite

Theorem

If $\mathcal{X}$ is finite and $\widehat{f}^{\mathrm{maj}}$ is the majority vote rule (for each $x \in \mathcal{X}$, majority vote among $\{ Y_i / X_i = x \}$),

$$\sup_P \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}^{\mathrm{maj}}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \leq \sqrt{\frac{\mathrm{Card}(\mathcal{X}) \log(2)}{2n}} \enspace .$$

Proof: standard risk bounds (see next section) + maximal inequality

$$\mathbb{E}\left[\sup_{t \in T} \left\{ \sum_{i=1}^{n} \xi_{i,t} \right\} \right] \leq \sqrt{\frac{\log(\mathrm{Card}(T))}{2n}}$$

if for all t, $(\xi_{i,t})_i$ are independent, centered and in $[0,1]$.

See e.g. `http://www.di.ens.fr/~arlot/2013orsay.htm`

Classification on $\mathcal{X}$ finite

Theorem

If $\mathcal{X}$ is finite and $\widehat{f}^{\mathrm{maj}}$ is the majority vote rule (for each $x \in \mathcal{X}$, majority vote among $\{ Y_i \,/\, X_i = x \}$),

$$\sup_P \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}^{\mathrm{maj}}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \leq \sqrt{\frac{\mathrm{Card}(\mathcal{X}) \log(2)}{2n}} \enspace .$$

Constants matter: $\mathrm{Card}(\mathcal{X})$ can be larger than $n \Rightarrow$ beware of asymptotic results and $\mathcal{O}(\cdot)$ that can hide such constants in first or second order terms.

No Free Lunch Theorem

Theorem

If $\mathcal{X}$ is infinite, for any classification rule $\widehat{f}$ and any $n \geq 1$,

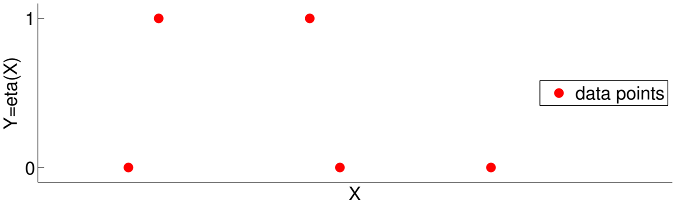
$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \geq \frac{1}{2} \ .$$

No Free Lunch Theorem

Theorem

If $\mathcal{X}$ is infinite, for any classification rule $\widehat{f}$ and any $n \geq 1$,

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \geq \frac{1}{2} \enspace .$$

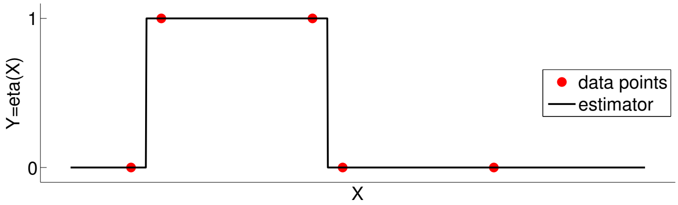


No Free Lunch Theorem

Theorem

If $\mathcal{X}$ is infinite, for any classification rule $\widehat{f}$ and any $n \geq 1$,

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \geq \frac{1}{2} \enspace .$$

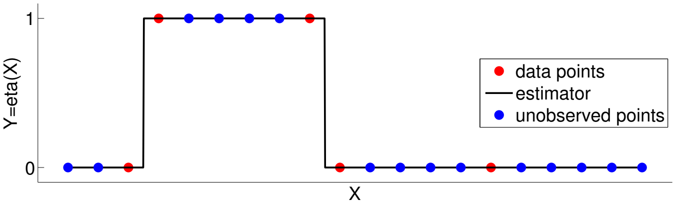


No Free Lunch Theorem

Theorem

If $\mathcal{X}$ is infinite, for any classification rule $\widehat{f}$ and any $n \geq 1$,

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \geq \frac{1}{2} .$$

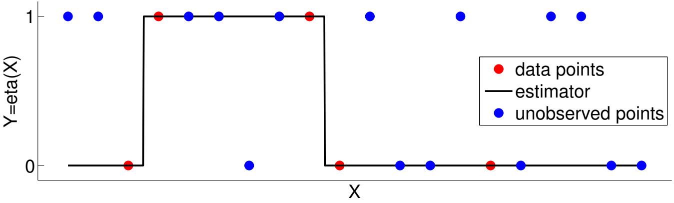


No Free Lunch Theorem

Theorem

If $\mathcal{X}$ is infinite, for any classification rule $\widehat{f}$ and any $n \geq 1$,

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \geq \frac{1}{2} \ .$$

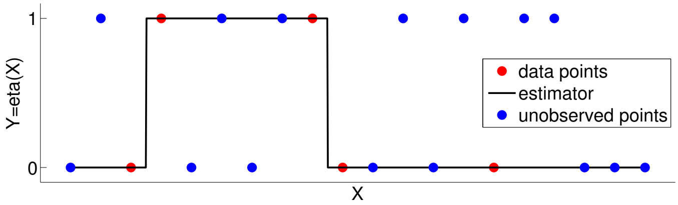


No Free Lunch Theorem

Theorem

If $\mathcal{X}$ is infinite, for any classification rule $\widehat{f}$ and any $n \geq 1$,

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \geq \frac{1}{2} \ .$$

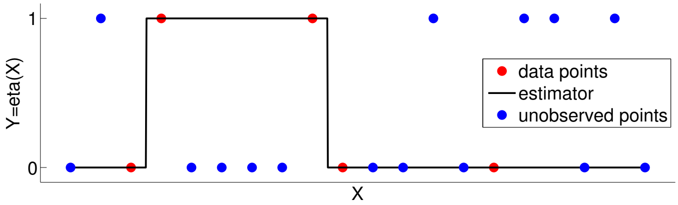


No Free Lunch Theorem

Theorem

If $\mathcal{X}$ is infinite, for any classification rule $\widehat{f}$ and any $n \geq 1$,

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \geq \frac{1}{2} \enspace .$$



No Free Lunch Theorem

Theorem

If $\mathcal{X}$ is infinite, for any classification rule $\widehat{f}$ and any $n \geq 1$,

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \right\} \geq \frac{1}{2} \enspace .$$

Remark: for any (a_n) decreasing to zero and any $\widehat{f}$, some P exists such that $\mathbb{E}\left[\mathcal{R}(\widehat{f}(D_n)) \right] - \mathcal{R}(f^\star) \geq a_n$. See Chapter 7 of [Devroye et al., 1996].

$\Rightarrow$ impossible to have $\frac{C(P)}{\log \log n}$ as a universal risk bound!

No Free Lunch Theorem: proof

Assume $\mathbb{N} \subset \mathcal{X}$ and let $K \geq 1$. For any $r \in \{0,1\}^K$, define P_r by X uniform on $\{1, \ldots, K\}$ and $\mathbb{P}(Y = r_i \mid X = i) = 1$ for all $i = 1, \ldots, K$.

Under P_r, $f^\star(x) = r_x$ and $\mathcal{R}(f^\star) = 0$. So,

$$\sup_P \left\{ \mathbb{E}_P \left[\mathcal{R}_P(\widehat{f}(D_n)) \right] - \mathcal{R}_P(f^\star) \right\} \geq \sup_{P_r} \left\{ \mathbb{P}_{P_r} \left(\widehat{f}(X; D_n) \neq r_X \right) \right\}$$

$$\geq \mathbb{E}_{r \sim R} \left\{ \mathbb{P}_{P_r} \left(\widehat{f}(X; D_n) \neq r_X \right) \right\}$$

$$\geq \mathbb{E} \left[\mathbb{1}_{X \notin \{X_1, \ldots, X_n\}} \mathbb{E} \left[\mathbb{1}_{\widehat{f}(X; (X_i, r_{X_i})_{i=1\ldots n}) \neq r_X} \,\middle|\, X, (X_i, r_{X_i})_{i=1\ldots n} \right] \right]$$

$$= \frac{1}{2} \mathbb{P}(X \notin \{X_1, \ldots, X_n\}) = \frac{1}{2} \left(1 - \frac{1}{K} \right)^n \qquad \square$$

Learning rates

- How can we get a bound such as $\mathcal{R}\left(\widehat{f}(D_n)\right) - \mathcal{R}\left(f^\star\right) \leq C(P)n^{-1/2}$?

Learning rates

- How can we get a bound such as $\mathcal{R}\left(\widehat{f}(D_n)\right) - \mathcal{R}\left(f^\star\right) \leq C(P) n^{-1/2}$?
- No Free Lunch Theorems $\Rightarrow$ must make assumptions on P

Learning rates

- How can we get a bound such as $\mathcal{R}\left(\widehat{f}(D_n)\right) - \mathcal{R}(f^\star) \leq C(P) n^{-1/2}$?
- No Free Lunch Theorems $\Rightarrow$ must make assumptions on P
- Minimax rate: given a set $\mathcal{P} \subset \mathcal{M}_1(\mathcal{X} \times \{0,1\})$,

$$\inf_{\widehat{f}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}\left[\mathcal{R}\left(\widehat{f}(D_n)\right) - \mathcal{R}(f^\star) \right] \right\}$$

Learning rates

- How can we get a bound such as $\mathcal{R}\left(\widehat{f}(D_n)\right) - \mathcal{R}(f^\star) \leq C(P) n^{-1/2}$?
- No Free Lunch Theorems $\Rightarrow$ must make assumptions on P
- Minimax rate: given a set $\mathcal{P} \subset \mathcal{M}_1(\mathcal{X} \times \{0,1\})$,

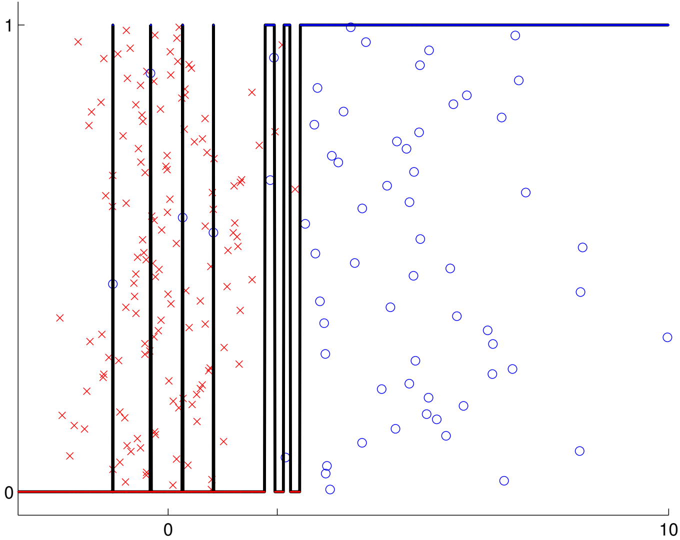
$$\inf_{\widehat{f}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}\left[\mathcal{R}\left(\widehat{f}(D_n)\right) - \mathcal{R}(f^\star) \right] \right\}$$

- Examples:
  - $\sqrt{V/n}$ when $f^\star \in S$ known and $\dim_{\mathrm{VC}}(S) = V$ [Devroye et al., 1996]
  - $V/(nh)$ when in addition $\mathbb{P}(|\eta(X) - 1/2| \leq h) = 0$ (margin assumption) [Massart and Nédélec, 2006]

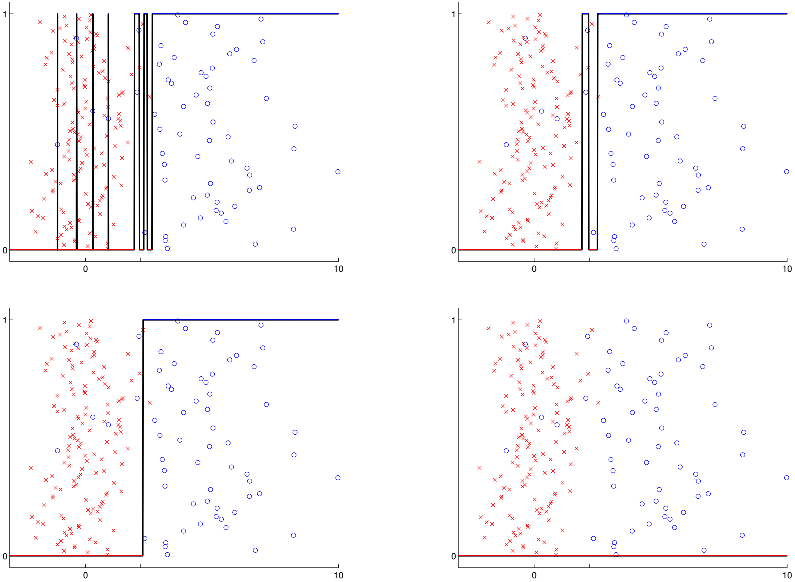
Outline

1. Introduction
2. Goals
3. Overfitting
4. Examples
5. Key issues

Overfitting with k-nearest-neighbours: $k = 1$



Choosing $k \in \{1, 3, 20, 200\}$ for k-NN ($n = 200$)



Empirical risk minimization

- Empirical risk

$$\widehat{\mathcal{R}}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

- Empirical risk minimizer over a model $S \subset \mathbb{S}$:

$$\widehat{f}_S \in \operatorname{argmin}_{f\in S}\left\{ \widehat{\mathcal{R}}_n(f) \right\}$$

Empirical risk minimization

- Empirical risk

$$\widehat{\mathcal{R}}_n (f) := \frac{1}{n} \sum_{i=1}^{n} \ell (f(X_i), Y_i)$$

- Empirical risk minimizer over a model $S \subset \mathbb{S}$:

$$\widehat{f}_S \in \operatorname{argmin}_{f \in S} \left\{ \widehat{\mathcal{R}}_n (f) \right\}$$

- Examples:
  - partitioning rule: $S = \left\{ \sum_{k \geq 1} \alpha_k \mathbb{1}_{A_k} \, / \, \alpha_k \in \{ 0, 1 \} \right\}$ for some partition $(A_k)_{k \geq 1}$ of $\mathcal{X}$
  - linear discrimination ($\mathcal{X} = \mathbb{R}^d$): $S = \left\{ x \mapsto \mathbb{1}_{\beta^\top x + \beta_0 \geq 0} \, / \, \beta \in \mathbb{R}^d, \ \beta_0 \in \mathbb{R} \right\}$
  - ...

Example: linear discrimination

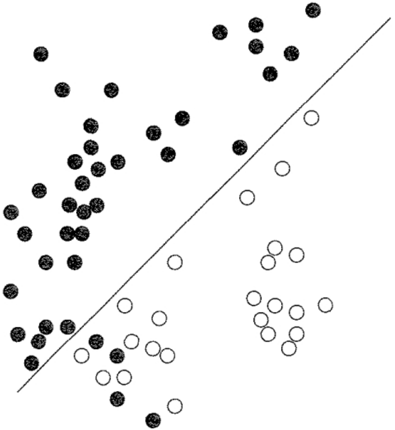


Fig. 4.3 of [Devroye et al., 1996]

Bias-variance trade-off

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{f}_S\right) - \mathcal{R}\left(f^\star\right)\right] = \text{Bias} + \text{Variance}$$

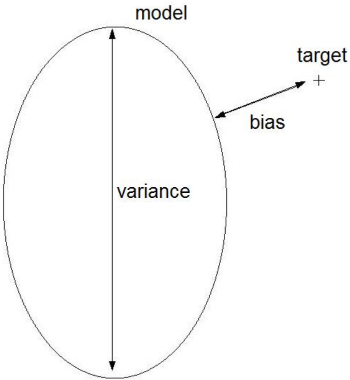
Bias or Approximation error

$$\mathcal{R}\left(f_S^\star\right) - \mathcal{R}\left(f^\star\right) = \inf_{f \in S} \mathcal{R}\left(f\right) - \mathcal{R}\left(f^\star\right)$$

Variance or Estimation error

OLS in regression: $\dfrac{\sigma^2 \dim(S)}{n}$

k-NN in regression: $\dfrac{\sigma^2}{k}$



Bias-variance trade-off

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{f}_S\right) - \mathcal{R}(f^\star)\right] = \text{Bias} + \text{Variance}$$

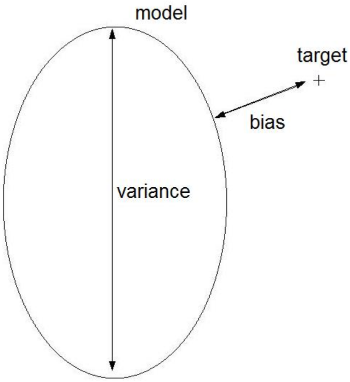
Bias or Approximation error

$$\mathcal{R}(f_S^\star) - \mathcal{R}(f^\star) = \inf_{f \in S} \mathcal{R}(f) - \mathcal{R}(f^\star)$$

Variance or Estimation error

OLS in regression: $\dfrac{\sigma^2 \dim(S)}{n}$

k-NN in regression: $\dfrac{\sigma^2}{k}$



Bias-variance trade-off $\Leftrightarrow$ avoid overfitting and underfitting

Outline

1. Introduction
2. Goals
3. Overfitting
4. Examples
   - Plug in rules
   - Empirical risk minimization and model selection
   - Convexification and support vector machines
   - Decision trees and forests
5. Key issues

Plug in classifiers

- Idea:

$$f^{\star}(x) = \mathbb{1}_{\eta(x) \geq \frac{1}{2}}$$

$\Rightarrow$ if $\widehat{\eta}(D_n)$ estimates η (regression problem),

$$\widehat{f}(x; D_n) = \mathbb{1}_{\widehat{\eta}(x; D_n) \geq \frac{1}{2}}$$

- Examples: partitioning, k-NN, local average classifiers [Devroye et al., 1996], [Audibert and Tsybakov, 2007]...

Risk bound for plug in

Proposition (Theorem 2.2 in [Devroye et al., 1996])

For a plug in classifier $\widehat{f}$,

$$\mathcal{R}\left(\widehat{f}(D_n)\right) - \mathcal{R}(f^\star) \leq 2\mathbb{E}\left[|\eta(X) - \widehat{\eta}(X; D_n)| \,\middle|\, D_n\right]$$

$$\leq 2\sqrt{\mathbb{E}\left[\left(\eta(X) - \widehat{\eta}(X; D_n)\right)^2 \,\middle|\, D_n\right]}$$

(First step for proving Stone's theorem [Stone, 1977])

Proof:

$$\mathcal{R}\left(\widehat{f}(D_n)\right) - \mathcal{R}(f^\star) = \mathbb{E}\left[|2\eta(X) - 1| \, \mathbb{1}_{\widehat{f}(X;D_n) \neq f^\star(X)} \,\middle|\, D_n\right]$$

and $\widehat{f}(X; D_n) \neq f^\star(X)$ implies $|2\eta(X) - 1| \leq 2|\eta(X) - \widehat{\eta}(X; D_n)|$. $\square$

Empirical risk minimization (ERM)

- ERM over S: $\widehat{f}_S \in \operatorname{argmin}_{f \in S} \left\{ \widehat{\mathcal{R}}_n (f) \right\}$

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{f}_S \right) - \mathcal{R}\left(f^{\star} \right) \right] = \text{Approximation error} + \text{Estimation error}$$

Empirical risk minimization (ERM)

- ERM over S: $\widehat{f}_S \in \operatorname{argmin}_{f \in S} \left\{ \widehat{\mathcal{R}}_n (f) \right\}$

$$\mathbb{E} \left[\mathcal{R} \left(\widehat{f}_S \right) - \mathcal{R} (f^\star) \right] = \text{Approximation error} + \text{Estimation error}$$

- Approximation error $\mathcal{R} (f_S^\star) - \mathcal{R} (f^\star)$: bounded thanks to approximation theory, or assumed equal to zero

Empirical risk minimization (ERM)

- ERM over S: $\widehat{f}_S \in \operatorname{argmin}_{f \in S} \left\{ \widehat{\mathcal{R}}_n(f) \right\}$

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{f}_S\right) - \mathcal{R}(f^\star) \right] = \text{Approximation error} + \text{Estimation error}$$

- Approximation error $\mathcal{R}(f_S^\star) - \mathcal{R}(f^\star)$: bounded thanks to approximation theory, or assumed equal to zero
- Estimation error

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{f}_S\right) - \mathcal{R}(f_S^\star) \right] \leq \mathbb{E}\left[\sup_{f \in S} \left\{ \mathcal{R}(f) - \widehat{\mathcal{R}}_n(f) \right\} \right]$$

Proof: $\mathcal{R}\left(\widehat{f}_S\right) - \mathcal{R}(f_S^\star)$

$$= \mathcal{R}\left(\widehat{f}_S\right) - \widehat{\mathcal{R}}_n\left(\widehat{f}_S\right) - \mathcal{R}(f_S^\star) + \widehat{\mathcal{R}}_n(f_S^\star) + \widehat{\mathcal{R}}_n\left(\widehat{f}_S\right) - \widehat{\mathcal{R}}_n(f_S^\star)$$

$$\leq \sup_{f \in S} \left\{ \mathcal{R}(f) - \widehat{\mathcal{R}}_n(f) \right\} + \widehat{\mathcal{R}}_n(f_S^\star) - \mathcal{R}(f_S^\star) \quad \square$$

Bounds on the estimation error (1): global approach

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{f}_S\right) - \mathcal{R}\left(f_S^\star\right)\right]$$

$$\leq \mathbb{E}\left[\sup_{f\in S}\left\{\mathcal{R}(f) - \widehat{\mathcal{R}}_n(f)\right\}\right] \qquad \text{(global complexity of } S\text{)}$$

$$\leq 2\mathbb{E}\left[\sup_{f\in S}\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \ell(f(X_i), Y_i)\right\}\right] \qquad \text{(symmetrization)}$$

$$\leq \frac{2\sqrt{2}}{\sqrt{n}}\mathbb{E}\left[\sqrt{H(S; X_1, \ldots, X_n)}\right] \qquad \text{(combinatorial entropy)}$$

$$\leq 2\sqrt{\frac{2V(S)\log\left(\frac{en}{V(S)}\right)}{n}} \qquad \text{(VC dimension)}$$

References: Section 3 of [Boucheron et al., 2005], Chapters 12–13 of [Devroye et al., 1996]

See also lectures 1–2 of `http://www.di.ens.fr/~arlot/2013orsay.htm`

Bounds on the estimation error (2): localization

- $\sup_{f \in S} \{ \operatorname{var}(\mathcal{R}(f) - \widehat{\mathcal{R}}_n(f)) \} \geq C n^{-1/2} \Rightarrow$ no faster rate

Bounds on the estimation error (2): localization

- $\sup_{f \in S} \{ \operatorname{var}(\mathcal{R}(f) - \widehat{\mathcal{R}}_n(f)) \} \geq C n^{-1/2} \Rightarrow$ no faster rate
- Margin condition: $\mathbb{P}(|\eta(X) - 1/2| \leq h) = 0$ with $h > 0$ [Mammen and Tsybakov, 1999]
- Localization idea: use that $\widehat{f}_S$ is not anywhere in S

Bounds on the estimation error (2): localization

- $\sup_{f \in S} \{ \mathrm{var}(\mathcal{R}(f) - \widehat{\mathcal{R}}_n(f)) \} \geq C n^{-1/2} \Rightarrow$ no faster rate
- Margin condition: $\mathbb{P}(|\eta(X) - 1/2| \leq h) = 0$ with $h > 0$ [Mammen and Tsybakov, 1999]
- Localization idea: use that $\widehat{f}_S$ is not anywhere in S

$$\widehat{f}_S \in \{ f \in S \,/\, \mathcal{R}(f) - \mathcal{R}(f^\star) \leq \varepsilon \}$$
$$\subset \{ f \in S \,/\, \mathrm{var}(\ell(f(X), Y) - \ell(f^\star(X), Y)) \leq \varepsilon / h \}$$

by the margin condition.

Bounds on the estimation error (2): localization

- $\sup_{f \in S} \{ \operatorname{var}(\mathcal{R}(f) - \widehat{\mathcal{R}}_n(f)) \} \geq C n^{-1/2} \Rightarrow$ no faster rate
- Margin condition: $\mathbb{P}(|\eta(X) - 1/2| \leq h) = 0$ with $h > 0$ [Mammen and Tsybakov, 1999]
- Localization idea: use that $\widehat{f}_S$ is not anywhere in S

$$\widehat{f}_S \in \{ f \in S \,/\, \mathcal{R}(f) - \mathcal{R}(f^\star) \leq \varepsilon \}$$
$$\subset \{ f \in S \,/\, \operatorname{var}(\ell(f(X), Y) - \ell(f^\star(X), Y)) \leq \varepsilon / h \}$$

by the margin condition. + Talagrand concentration inequality [Talagrand, 1996, Bousquet, 2002] + ...

$\Rightarrow$ fast rates (depending on the assumptions), e.g.,
$\kappa \frac{V(S)}{nh} \left(1 + \log \left(\frac{nh^2}{V(S)} \right) \right)$

[Boucheron et al., 2005, Sec. 5], [Massart and Nédélec, 2006].

Model selection

- family of models $(S_m)_{m \in \mathcal{M}}$

$\Rightarrow$ family of classifiers $(\widehat{f}_m(D_n))_{m \in \mathcal{M}_n}$

$\Rightarrow$ choose $\widehat{m} = \widehat{m}(D_n)$ such that $\mathcal{R}\left(\widehat{f}_{\widehat{m}}(D_n)\right)$ is minimal?

Model selection

- family of models $(S_m)_{m \in \mathcal{M}}$

$\Rightarrow$ family of classifiers $(\widehat{f}_m(D_n))_{m \in \mathcal{M}_n}$

$\Rightarrow$ choose $\widehat{m} = \widehat{m}(D_n)$ such that $\mathcal{R}\left(\widehat{f}_{\widehat{m}}(D_n)\right)$ is minimal?

- Goal: minimize the risk, i.e.,
  Oracle inequality (in expectation or with a large probability):

$$\mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - \mathcal{R}\left(f^{\star}\right) \leq C \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}\left(\widehat{f}_m\right) - \mathcal{R}\left(f^{\star}\right) \right\} + R_n$$

- Interpretation of $\widehat{m}$: the best model can be wrong / the true model can be worse than smaller ones.

Penalization for model selection

- Penalization:

$$\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n \left(\widehat{f}_m \right) + \mathrm{pen}(m) \right\}$$

- Ideal penalty:

$$\mathrm{pen}_{\mathrm{id}}(m) = \mathcal{R}\left(\widehat{f}_m \right) - \widehat{\mathcal{R}}_n \left(\widehat{f}_m \right) \quad \Leftrightarrow \quad \widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R} \left(\widehat{f}_m \right) \right\}$$

Penalization for model selection

- Penalization:

$$\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n \left(\widehat{f}_m \right) + \operatorname{pen}(m) \right\}$$

- Ideal penalty:

$$\operatorname{pen}_{\mathrm{id}}(m) = \mathcal{R}\left(\widehat{f}_m \right) - \widehat{\mathcal{R}}_n \left(\widehat{f}_m \right) \quad \Leftrightarrow \quad \widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}\left(\widehat{f}_m \right) \right\}$$

- General idea: choose pen such that $\operatorname{pen}(m) \approx \operatorname{pen}_{\mathrm{id}}(m)$ or at least $\operatorname{pen}(m) \geq \operatorname{pen}_{\mathrm{id}}(m)$ for all $m \in \mathcal{M}$.

  Lemma (see next slide): if $\operatorname{pen}(m) \geq \operatorname{pen}_{\mathrm{id}}(m)$ for all $m \in \mathcal{M}$,

$$\mathcal{R}\left(\widehat{f}_{\widehat{m}} \right) - \mathcal{R}\left(f^{\star} \right) \leq \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}\left(\widehat{f}_m \right) - \mathcal{R}\left(f^{\star} \right) + \operatorname{pen}(m) - \operatorname{pen}_{\mathrm{id}}(m) \right\}$$

Penalization for model selection: lemma

Lemma

If $\forall m \in \mathcal{M}$, $-B(m) \leq \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m) \leq A(m)$, *then,*

$$\mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - \mathcal{R}\left(f^{\star}\right) - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}\left(\widehat{f}_m\right) - \mathcal{R}\left(f^{\star}\right) + A(m) \right\} \enspace .$$

Proof: For all $m \in \mathcal{M}$, by definition of $\widehat{m}$,

$$\widehat{\mathcal{R}}_n\left(\widehat{f}_{\widehat{m}}\right) + \mathrm{pen}(\widehat{m}) \leq \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right) + \mathrm{pen}(m) \enspace .$$

$$\begin{aligned}
\text{So,} \quad \widehat{\mathcal{R}}_n\left(\widehat{f}_{\widehat{m}}\right) + \mathrm{pen}(\widehat{m}) &= \mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - \mathrm{pen}_{\mathrm{id}}(\widehat{m}) + \mathrm{pen}(\widehat{m}) \\
&\geq \mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - B(\widehat{m}) \\
\text{and} \quad \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right) + \mathrm{pen}(m) &= \mathcal{R}\left(\widehat{f}_m\right) - \mathrm{pen}_{\mathrm{id}}(m) + \mathrm{pen}(m) \\
&\leq \mathcal{R}\left(\widehat{f}_m\right) + A(m) \enspace . \quad \square
\end{aligned}$$

Penalization for model selection

- Structural risk minimization (Vapnik):

$$\mathrm{pen}_{\mathrm{id}}(m) \leq \sup_{f \in S_m} \left\{ \mathcal{R}(f) - \widehat{\mathcal{R}}_n(f) \right\}$$

$\Rightarrow$ can use previous bounds
[Koltchinskii, 2001, Bartlett et al., 2002, Fromont, 2007]
but remainder terms $\geq Cn^{-1/2} \Rightarrow$ no fast rates.

Penalization for model selection

- Structural risk minimization (Vapnik):

$$\mathrm{pen}_{\mathrm{id}}(m) \leq \sup_{f \in S_m} \left\{ \mathcal{R}(f) - \widehat{\mathcal{R}}_n(f) \right\}$$

$\Rightarrow$ can use previous bounds
[Koltchinskii, 2001, Bartlett et al., 2002, Fromont, 2007]
but remainder terms $\geq Cn^{-1/2} \Rightarrow$ no fast rates.

- Tighter estimates of $\mathrm{pen}_{\mathrm{id}}(m)$ for fast rates: localization [Koltchinskii, 2006], resampling [Arlot, 2009].

See also Section 8 of [Boucheron et al., 2005].

Convexification of the classification problem

Convention: $Y_i \in \{-1, 1\}$ so that $\mathbb{1}_{y \neq y'} = \mathbb{1}_{yy' < 0} = \Phi_{0-1}(yy')$

$$\min_f \frac{1}{n} \sum_{i=1}^{n} \Phi_{0-1}(Y_i f(X_i))$$ computationally heavy in general.

Convexification of the classification problem

Convention: $Y_i \in \{-1, 1\}$ so that $\mathbb{1}_{y \neq y'} = \mathbb{1}_{yy' < 0} = \Phi_{0-1}(yy')$

$$\min_f \frac{1}{n} \sum_{i=1}^{n} \Phi_{0-1}(Y_i f(X_i)) \quad \text{computationally heavy in general.}$$

- Classifier $f : \mathcal{X} \to \{-1, 1\} \Rightarrow$ prediction function $f : \mathcal{X} \to \mathbb{R}$ such that $\mathrm{sign}(f(x))$ will be used to classify x
- Risk $\mathcal{R}_{0-1}(f) = \mathbb{E}[\Phi_{0-1}(Yf(X))]$ $\Rightarrow$ Φ-risk $\mathcal{R}_\Phi(f) = \mathbb{E}[\Phi(Yf(X))]$ for some $\Phi : \mathbb{R} \to \mathbb{R}^+$

Convexification of the classification problem

Convention: $Y_i \in \{-1, 1\}$ so that $\mathbb{1}_{y \neq y'} = \mathbb{1}_{yy' < 0} = \Phi_{0-1}(yy')$

$$\min_f \frac{1}{n} \sum_{i=1}^{n} \Phi_{0-1}(Y_i f(X_i)) \quad \text{computationally heavy in general.}$$

- Classifier $f : \mathcal{X} \to \{-1, 1\} \Rightarrow$ prediction function $f : \mathcal{X} \to \mathbb{R}$ such that $\operatorname{sign}(f(x))$ will be used to classify x
- Risk $\mathcal{R}_{0-1}(f) = \mathbb{E}[\Phi_{0-1}(Yf(X))]$
  $\Rightarrow$ Φ-risk $\mathcal{R}_\Phi(f) = \mathbb{E}[\Phi(Yf(X))]$ for some $\Phi : \mathbb{R} \to \mathbb{R}^+$

$$\Rightarrow \quad \min_{f \in S} \frac{1}{n} \sum_{i=1}^{n} \Phi(Y_i f(X_i)) \quad \text{with } S \text{ and } \Phi \text{ convex.}$$

Examples of functions Φ

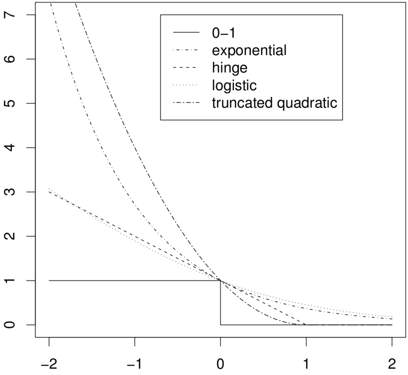


Figure from [Bartlett et al., 2006].

- exponential: $\Phi(u) = e^{-u}$
  $\Rightarrow$ AdaBoost
- hinge:
  $\Phi(u) = \max\{1 - u, 0\}$
  $\Rightarrow$ support vector machines
- logistic/logit:
  $\Phi(u) = \log(1 + \exp(-u))$
  $\Rightarrow$ logistic regression
- truncated quadratic:
  $\Phi(u) = (\max\{1 - u, 0\})^2$

References: [Bartlett et al., 2006] and Section 4 of [Boucheron et al., 2005].

Links between 0–1 and convex risks

Definition

Φ is classification-calibrated if for any x with $\eta(x) \neq 1/2$,

$$\operatorname{sign}(f_\Phi^\star(x)) = f^\star(x) \qquad \text{for any} \quad f_\Phi^\star \in \operatorname{argmin}_f \mathcal{R}_\Phi(f)$$

Links between 0–1 and convex risks

Definition

Φ is classification-calibrated if for any x with $\eta(x) \neq 1/2$,

$$\operatorname{sign}(f^\star_\Phi(x)) = f^\star(x) \qquad \text{for any} \quad f^\star_\Phi \in \operatorname{argmin}_f \mathcal{R}_\Phi\left(f\right)$$

Theorem ([Bartlett et al., 2006])

Φ convex is classification-calibrated $\Leftrightarrow$ Φ differentiable at 0 and $\Phi'(0) < 0$.
Then, a function ψ exists such that

$$\psi\left(\mathcal{R}_{0-1}\left(f\right) - \mathcal{R}_{0-1}\left(f^\star_{0-1}\right)\right) \leq \mathcal{R}_\Phi\left(f\right) - \mathcal{R}_\Phi\left(f^\star_\Phi\right) \quad .$$

Examples:

- exponential loss: $\psi(\theta) = 1 - \sqrt{1 - \theta^2}$
- hinge loss: $\psi(\theta) = |\theta|$
- truncated quadratic: $\psi(\theta) = \theta^2$

Support Vector Machines: linear classifier

$\mathcal{X} = \mathbb{R}^d$, linear classifier: $\operatorname{sign}(\beta^\top x + \beta_0)$ with $\beta \in \mathbb{R}^d$, $\beta_0 \in \mathbb{R}$

Support Vector Machines: linear classifier

$\mathcal{X} = \mathbb{R}^d$, linear classifier: $\mathrm{sign}(\beta^\top x + \beta_0)$ with $\beta \in \mathbb{R}^d$, $\beta_0 \in \mathbb{R}$

$$
\begin{aligned}
& \operatorname{argmin}_{\beta,\beta_0 \,/\, \|\beta\| \leq R} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Phi_{\mathrm{hinge}} \left(Y_i \left(\beta^\top X_i + \beta_0 \right) \right) \right\} \\
\Leftrightarrow \quad & \operatorname{argmin}_{\beta,\beta_0} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Phi_{\mathrm{hinge}} \left(Y_i \left(\beta^\top X_i + \beta_0 \right) \right) + \lambda \left\| \beta \right\|^2 \right\}
\end{aligned}
$$

up to some (random) reparametrization ($\lambda = \lambda(R; D_n)$).

$\Rightarrow$ quadratic program with $2n$ linear constraints.

Support Vector Machines: linear classifier

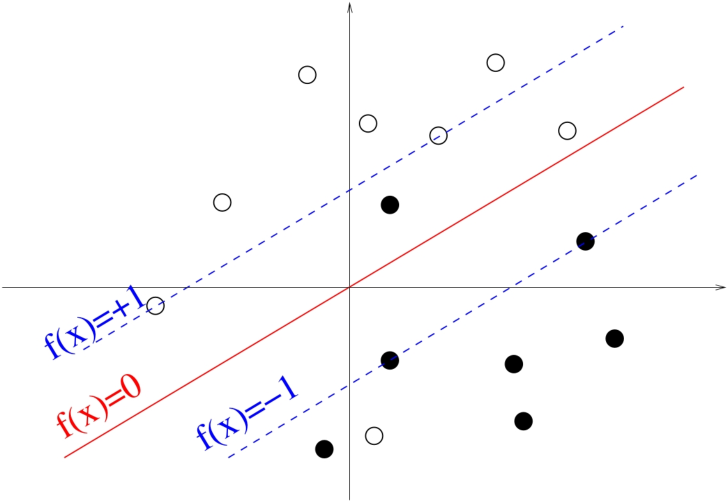


Figure from `http://cbio.ensmp.fr/~jvert/svn/kernelcourse/slides/master/master.pdf`

Support Vector Machines: kernel trick

Positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.t. $(k(X_i, X_j))_{i,j}$ symmetric positive definite

Reproducing Kernel Hilbert Space (RKHS) $\mathcal{F}$: space of functions $\mathcal{X} \to \mathbb{R}$ spanned by the $\Phi(x) = k(x, \cdot)$, $x \in \mathcal{X}$.

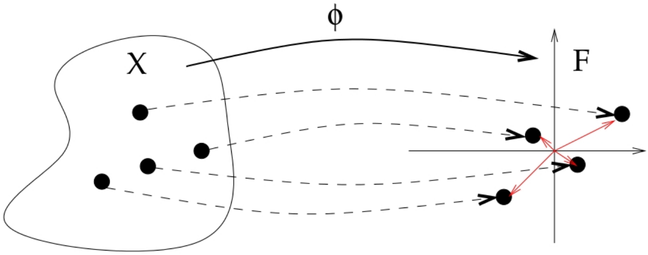


Figure from http://cbio.ensmp.fr/~jvert/svn/kernelcourse/slides/master/master.pdf

Support Vector Machines: kernel trick

Positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.t. $(k(X_i, X_j))_{i,j}$ symmetric positive definite

Reproducing Kernel Hilbert Space (RKHS) $\mathcal{F}$: space of functions $\mathcal{X} \to \mathbb{R}$ spanned by the $\Phi(x) = k(x, \cdot)$, $x \in \mathcal{X}$.

Theorem (Representer theorem)

For any cost function ℓ,

$$\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) + \lambda \left\| f \right\|_{\mathcal{F}}^2 \right\}$$

is attained at some f of the form $\sum_{i=1}^{n} \alpha_i k(X_i, \cdot)$

$\Rightarrow$ any algorithm for $\mathcal{X} = \mathbb{R}^d$ relying only on the dot products $(\langle X_i, X_j \rangle)_{i,j}$ can be kernelized.

Kernel examples

- linear kernel: $\mathcal{X} = \mathbb{R}^d$, $k(x,y) = \langle x,\ y \rangle \Rightarrow \mathcal{F} = \mathbb{R}^d$ Euclidean
- polynomial kernel: $\mathcal{X} = \mathbb{R}^d$, $k(x,y) = (\langle x,\ y \rangle + 1)^r \Rightarrow$ $\mathcal{F} = \mathbb{R}_r[X_1, \ldots, X_d]$

Kernel examples

- linear kernel: $\mathcal{X} = \mathbb{R}^d$, $k(x,y) = \langle x, y \rangle \Rightarrow \mathcal{F} = \mathbb{R}^d$ Euclidean
- polynomial kernel: $\mathcal{X} = \mathbb{R}^d$, $k(x,y) = (\langle x, y \rangle + 1)^r \Rightarrow$ $\mathcal{F} = \mathbb{R}_r[X_1, \ldots, X_d]$
- Gaussian kernel: $\mathcal{X} = \mathbb{R}^d$, $k(x,y) = e^{-\lVert x-y \rVert^2/(2\sigma^2)}$
- Laplace kernel: $\mathcal{X} = \mathbb{R}$, $k(x,y) = e^{-\lvert x-y \rvert/2}$ $\Rightarrow \mathcal{F} = H^1$ (Sobolev space), $\lVert f \rVert_{\mathcal{F}}^2 = \lVert f \rVert_{L^2}^2 + \lVert f' \rVert_{L^2}^2$.
- min kernel: $\mathcal{X} = [0,1]$, $k(x,y) = \min\{x,y\}$ $\Rightarrow \mathcal{F} = \{f \in \mathcal{C}^0([0,1]),\ f' \in L^2,\ f(0) = 0\}$, $\lVert f \rVert_{\mathcal{F}} = \lVert f' \rVert_{L^2}$.

Kernel examples

- Gaussian kernel: $\mathcal{X} = \mathbb{R}^d$, $k(x, y) = e^{-\|x-y\|^2/(2\sigma^2)}$
- Laplace kernel: $\mathcal{X} = \mathbb{R}$, $k(x, y) = e^{-|x-y|/2}$
  $\Rightarrow \mathcal{F} = H^1$ (Sobolev space), $\|f\|_{\mathcal{F}}^2 = \|f\|_{L^2}^2 + \|f'\|_{L^2}^2$.
- min kernel: $\mathcal{X} = [0, 1]$, $k(x, y) = \min\{x, y\}$
  $\Rightarrow \mathcal{F} = \{f \in \mathcal{C}^0([0, 1]),\ f' \in L^2,\ f(0) = 0\}$, $\|f\|_{\mathcal{F}} = \|f'\|_{L^2}$.

$\Rightarrow$ intersection kernel: $\mathcal{X} = \{ p \in [0, 1]^d \,/\, p_1 + \cdots + p_d = 1 \}$, $k(p, q) = \sum_{i=1}^d \min(p_i, q_i)$, useful in computer vision [Hein and Bousquet, 2004, Maji et al., 2008].

- other kernels on non-vectorial data (graphs, words / DNA sequences, ...): see for instance [Schölkopf et al., 2004, Mahé et al., 2005, Shervashidze et al., 2011] and `http://cbio.ensmp.fr/~jvert/svn/kernelcourse/slides/master/master.pdf`

Support Vector Machines: results / references

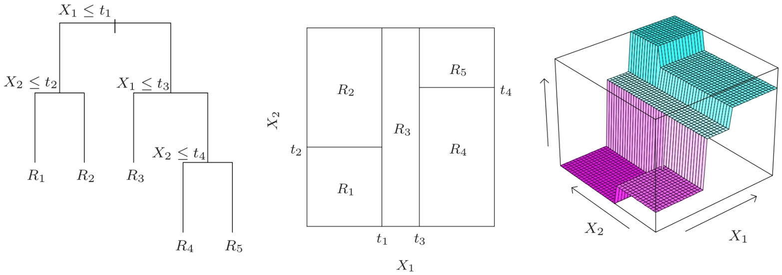
Main mathematical tools for SVM analysis: probability in Hilbert spaces (RKHS), functional analysis.

Some references:

- Risk bounds: e.g., [Blanchard et al., 2008] (SVM as a penalization procedure for selecting among balls).
  see also [Boucheron et al., 2005, Section 4]
- Tutorials and lecture notes: [Burges, 1998], `http://cbio.ensmp.fr/~jvert/svn/kernelcourse/slides/master/master.pdf`
- Books: e.g., [Steinwart and Christmann, 2008, Hastie et al., 2009, Scholkopf and Smola, 2001]

Decision / classification tree

- piecewise constant predictor
- partition obtained by recursive splitting of $\mathcal{X} \subset \mathbb{R}^p$, orthogonally to one axis ($X^j < t$ vs. $X^j \geq t$)
- empirical risk minimization



Figures from [Hastie et al., 2009]

CART (Classification And Regression Trees)

CART [Breiman et al., 1984]:

1. generate one large tree by splitting recursively the data (minimization of some impurity measure),

$\Rightarrow$ over-adapted to data

CART (Classification And Regression Trees)

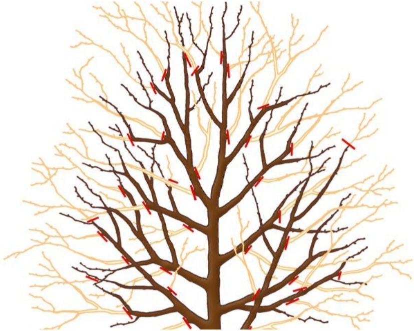
CART [Breiman et al., 1984]:

1. generate one large tree by splitting recursively the data (minimization of some impurity measure),

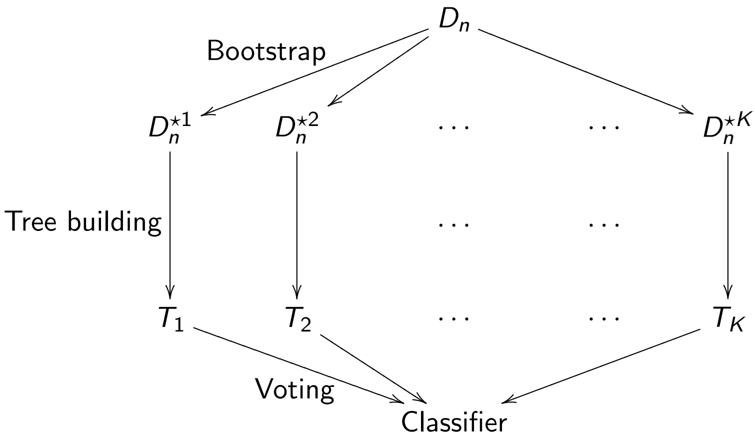
$\Rightarrow$ over-adapted to data

2. pruning ($\Leftrightarrow$ model selection)

Model selection results: e.g., [Gey and Nédélec, 2005, Sauvé and Tuleau-Malot, 2011, Gey and Mary-Huard, 2011].



Random forests [Breiman, 2001]



Various ways to build individual trees (subset of variables...)

Purely random forests: partitions independent from training data.

Results on random forests (classification and regression)

- Most theoretical results on purely random forests (partitions independent from training data: by data splitting or with simpler models)

- Consistency result in classification [Biau et al., 2008]
- Convergence rate and some combination with variable selection [Biau, 2012]

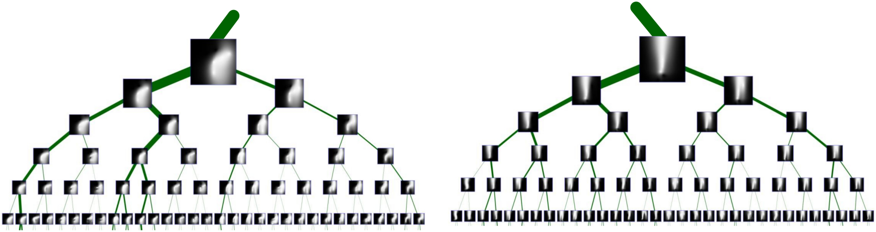
- From a single tree to a large forest:
  - estimation error reduction (at least a constant factor) [Genuer, 2012]
  - approximation error reduction (A. & Genuer, work in progress)

  $\Rightarrow$ sometimes improvement in the learning rate

See also [Breiman, 2004, Genuer et al., 2008, Genuer et al., 2010].

Kinect: depth features $\Rightarrow$ body part

Depth image  $\Rightarrow$ depth comparison features at each pixel



$\Rightarrow$ body part at each pixel  $\Rightarrow$ body part positions  $\Rightarrow \cdots$

Figure from [Shotton et al., 2011]

Outline

Hyperparameter choice

- Always one or several parameters to choose:
  k for k-NN, model selection, λ for SVM, kernel bandwidth for SVM with Gaussian kernel, tree size in random forests, ...
- No universal choice possible (No Free Lunch Theorems apply)
  $\Rightarrow$ must use some prior knowledge at some point

Hyperparameter choice

- Always one or several parameters to choose:
  k for k-NN, model selection, λ for SVM, kernel bandwidth for SVM with Gaussian kernel, tree size in random forests, ...
- No universal choice possible (No Free Lunch Theorems apply)
  $\Rightarrow$ must use some prior knowledge at some point

- Most general ideas: data splitting (cross-validation) [Arlot and Celisse, 2010]
- Sometimes specific approaches (penalization...): more efficient (for risk and computational cost) but also dependent on stronger assumptions

Hyperparameter choice

- Always one or several parameters to choose:
  k for k-NN, model selection, λ for SVM, kernel bandwidth for SVM with Gaussian kernel, tree size in random forests, ...
- No universal choice possible (No Free Lunch Theorems apply)
  $\Rightarrow$ must use some prior knowledge at some point

- Most general ideas: data splitting (cross-validation) [Arlot and Celisse, 2010]
- Sometimes specific approaches (penalization...): more efficient (for risk and computational cost) but also dependent on stronger assumptions
- Important to choose a good parametrization (e.g., for cross-validation, the optimal parameter should not vary too much from a sample to another)

Computational complexity

Most classifiers are defined as $\widehat{f} \in \operatorname{argmin}_{f \in S} C(f)$

- Optimization algorithms: usually faster (polynomial) when C and S convex. Often NP hard with 0–1 loss. Counterexample: interval classification [Kearns et al., 1997].

Computational complexity

Most classifiers are defined as $\widehat{f} \in \operatorname{argmin}_{f \in S} C(f)$

- Optimization algorithms: usually faster (polynomial) when C and S convex. Often NP hard with 0–1 loss. Counterexample: interval classification [Kearns et al., 1997].
- General convex optimization algorithms usually too slow if n or $p = \dim(\mathcal{X})$ are $> 10^3$.

$\Rightarrow$ Need for specific faster algorithms (e.g., for SVM, consider the dual problem and take advantage of the "sparsity" of the solution).
Constants matter! (e.g., dependence on p).

Computational complexity

Most classifiers are defined as $\widehat{f} \in \operatorname{argmin}_{f \in S} C(f)$

- Optimization algorithms: usually faster (polynomial) when C and S convex. Often NP hard with 0–1 loss. Counterexample: interval classification [Kearns et al., 1997].

- General convex optimization algorithms usually too slow if n or $p = \dim(\mathcal{X})$ are $> 10^3$.

$\Rightarrow$ Need for specific faster algorithms (e.g., for SVM, consider the dual problem and take advantage of the "sparsity" of the solution).
Constants matter! (e.g., dependence on p).

- Choice of a classification learning algorithm: trade-off between statistical performances and computational cost.
Also depends on the confidence in the modelling chosen.

Optimization error

Risk = Approximation error + Estimation error

Optimization error

Risk = Approximation error + Estimation error + Optimization error

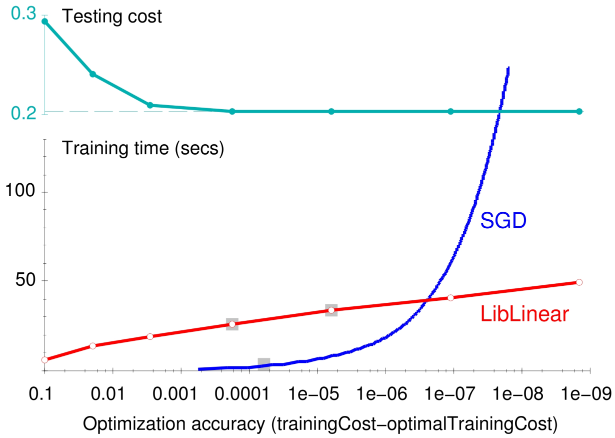


Figure from [Bottou and Bousquet, 2011]

The big data setting

- Given $\varepsilon > 0$, what do we need to get $\mathcal{R}\left(\widehat{f}\right) - \mathcal{R}\left(f^{\star}\right) \leq \varepsilon$?

The big data setting

- Given $\varepsilon > 0$, what do we need to get $\mathcal{R}\left(\widehat{f}\right) - \mathcal{R}\left(f^{\star}\right) \leq \varepsilon$?

- Traditional statistical learning: sample complexity, i.e., $n \geq n_0(\varepsilon)$, whatever the computational cost

The big data setting

- Given $\varepsilon > 0$, what do we need to get $\mathcal{R}\left(\widehat{f}\right) - \mathcal{R}(f^\star) \leq \varepsilon$?

- Traditional statistical learning: sample complexity, i.e., $n \geq n_0(\varepsilon)$, whatever the computational cost

- Big data: n so large that exploring all data is impossible (and unnecessary) $\Rightarrow$ better to throw away some data! [Bottou and Bousquet, 2008, Shalev-Shwartz and Srebro, 2008]

$\Rightarrow$ time complexity, i.e., minimal number of computations, whatever n

The big data setting

- Given $\varepsilon > 0$, what do we need to get $\mathcal{R}\left(\widehat{f}\right) - \mathcal{R}\left(f^{\star}\right) \leq \varepsilon$?
- Traditional statistical learning: sample complexity, i.e., $n \geq n_0(\varepsilon)$, whatever the computational cost
- Big data: n so large that exploring all data is impossible (and unnecessary) $\Rightarrow$ better to throw away some data! [Bottou and Bousquet, 2008, Shalev-Shwartz and Srebro, 2008]

$\Rightarrow$ time complexity, i.e., minimal number of computations, whatever n

- A very active field: Big Data Research and Development Initiative (US government), MASTODONS (CNRS), AMPLab (UC Berkeley), ...

Computational trade-offs, from statistics to big data

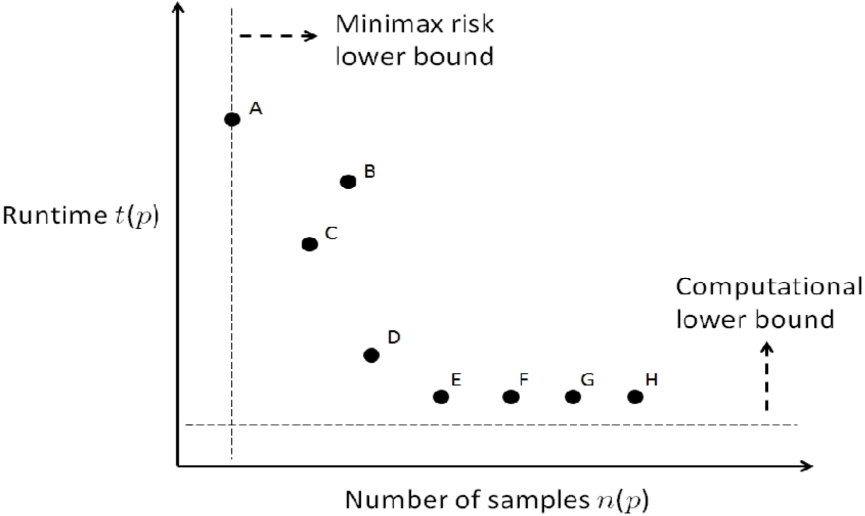


Figure from [Chandrasekaran and Jordan, 2012]

Conclusion

- Learning theory: assumptions $\Rightarrow$ learning rates (NFLT)
- Main danger: overfitting

Conclusion

- Learning theory: assumptions $\Rightarrow$ learning rates (NFLT)
- Main danger: overfitting
- Various ways to model the data:
  - k-NN: $f^\star$ locally constant w.r.t. d
  - ERM/model selection: family of possible $f^\star$
  - SVM: kernel $\Rightarrow$ smoothness of $f^\star$ / feature space
  - random forests: weak modelling (trees) + aggregation
  - many other approaches: Bayesian statistics, neural networks, deep learning, ...

Conclusion

- Learning theory: assumptions $\Rightarrow$ learning rates (NFLT)
- Main danger: overfitting
- Various ways to model the data:
  - k-NN: $f^\star$ locally constant w.r.t. d
  - ERM/model selection: family of possible $f^\star$
  - SVM: kernel $\Rightarrow$ smoothness of $f^\star$ / feature space
  - random forests: weak modelling (trees) + aggregation
  - many other approaches: Bayesian statistics, neural networks, deep learning, ...
- Key issues: tuning parameters & computational complexity
  Big data $\Rightarrow$ new challenges

Conclusion

- Learning theory: assumptions ⇒ learning rates (NFLT)
- Main danger: overfitting
- Various ways to model the data:
  - k-NN: $f^\star$ locally constant w.r.t. d
  - ERM/model selection: family of possible $f^\star$
  - SVM: kernel ⇒ smoothness of $f^\star$ / feature space
  - random forests: weak modelling (trees) + aggregation
  - many other approaches: Bayesian statistics, neural networks, deep learning, ...
- Key issues: tuning parameters & computational complexity
  Big data ⇒ new challenges
- Main mathematical domains involved (outside statistics): probability theory (concentration of measure, ...), approximation theory, functional analysis, optimization, ...

More references

These slides: `http://www.di.ens.fr/~arlot/`

- Devroye, L., Györfi, L., and Lugosi, G. (1996).
  A probabilistic theory of pattern recognition, volume 31 of *Applications of Mathematics (New York)*.
  Springer-Verlag, New York.

- Boucheron, S., Bousquet, O., and Lugosi, G. (2005).
  Theory of classification: a survey of some recent advances.
  ESAIM Probab. Stat., 9:323–375 (electronic).

- Hastie, T., Tibshirani, R., and Friedman, J. (2009).
  The elements of statistical learning.
  Springer Series in Statistics. Springer, New York, second edition.
  Data mining, inference, and prediction.

Arlot, S. (2009).
Model selection by resampling penalization.
Electron. J. Stat., 3:557–624 (electronic).

Arlot, S. and Celisse, A. (2010).
A survey of cross-validation procedures for model selection.
Statist. Surv., 4:40–79.

Audibert, J.-Y. and Tsybakov, A. (2007).
Fast learning rates for plug-in classifiers.
Annals of Statistics, 35(2):608–633.

Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002).
Model selection and error estimation.
Machine Learning, 48:85–113.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006).
Convexity, classification, and risk bounds.

Journal of the American Statistical Association, 101(473):138–156.
(Was Department of Statistics, U.C. Berkeley Technical Report number 638, 2003).

Biau, G. (2012).
Analysis of a random forests model.
J. Mach. Learn. Res., 13:1063–1095.

Biau, G., Devroye, L., and Lugosi, G. (2008).
Consistency of random forests and other averaging classifiers.
J. Mach. Learn. Res., 9:2015–2033.

Blanchard, G., Bousquet, O., and Massart, P. (2008).
Statistical performance of support vector machines.
Ann. Statist., 36(2):489–531.

Bottou, L. and Bousquet, O. (2008).
The tradeoffs of large scale learning.

In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20, pages 161–168. NIPS Foundation (http://books.nips.cc).

Bottou, L. and Bousquet, O. (2011).
The tradeoffs of large scale learning.
In Sra, S., Nowozin, S., and Wright, S. J., editors, *Optimization for Machine Learning*, pages 351–368. MIT Press.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005).
Theory of classification: a survey of some recent advances.
ESAIM Probab. Stat., 9:323–375 (electronic).

Bousquet, O. (2002).
A Bennett concentration inequality and its application to suprema of empirical processes.
C. R. Math. Acad. Sci. Paris, 334(6):495–500.

Breiman, L. (2001).
Random forests.
Machine Learning, 45:5–32.
10.1023/A:1010933404324.

Breiman, L. (2004).
Consistency for a simple model of random forests.
Technical Report Technical Report 670, U.C. Berkeley Department of Statistics.
available at
http://www.stat.berkeley.edu/tech-reports/670.pdf.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984).
Classification and Regression Trees.
Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.

Burges, C. (1998).

A tutorial on support vector machines for pattern recognition.
Data Mining and Knowledge Discovery, 2(2):121–167.
http://research.microsoft.com/en-us/um/people/cburges/papers/SVMTutorial.pdf.

Chandrasekaran, V. and Jordan, M. I. (2012).
Computational and statistical tradeoffs via convex relaxation.
arXiv:1211.1073.

Devroye, L., Györfi, L., and Lugosi, G. (1996).
A probabilistic theory of pattern recognition, volume 31 of *Applications of Mathematics (New York)*.
Springer-Verlag, New York.

Fromont, M. (2007).
Model selection by bootstrap penalization for classification.
Mach. Learn., 66(2–3):165–207.

Genuer, R. (2012).
Variance reduction in purely random forests.

Journal of Nonparametric Statistics, 24(3):543–562.

Genuer, R., Poggi, J.-M., and Tuleau, C. (2008).
Random forests: Some methodological insights.
arXiv:0811.3619.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010).
Variable selection using random forests.
Pattern Recognition Letters, 31(14):2225–2236.

Gey, S. and Mary-Huard, T. (2011).
Risk bounds for embedded variable selection in classification trees.
arXiv:1108.0757.

Gey, S. and Nédélec, É. (2005).
Model selection for CART regression trees.
IEEE Trans. Inform. Theory, 51(2):658–670.

Hastie, T., Tibshirani, R., and Friedman, J. (2009).
The elements of statistical learning.

Springer Series in Statistics. Springer, New York, second edition.
Data mining, inference, and prediction.

Hein, M. and Bousquet, O. (2004).
Hilbertian metrics and positive-definite kernels on probability measures.
In *AISTATS*.

Kearns, M., Mansour, Y., Ng, A. Y., and Ron, D. (1997).
An Experimental and Theoretical Comparison of Model Selection Methods.
Mach. Learn., 7:7–50.

Koltchinskii, V. (2001).
Rademacher penalties and structural risk minimization.
IEEE Trans. Inform. Theory, 47(5):1902–1914.

Koltchinskii, V. (2006).

Local Rademacher complexities and oracle inequalities in risk minimization.
Ann. Statist., 34(6):2593–2656.

Mahé, P., Ueda, N., Akutsu, T., Perret, J.-L., and Vert, J.-P. (2005).
Graph kernels for molecular structure-activity relationship analysis with support vector machines.
Journal of Chemical Information and Modeling, 45(4):939–951.

Maji, S., Berg, A. C., and Malik, J. (2008).
Classification using intersection kernel support vector machines is efficient.
In *CVPR*.

Mammen, E. and Tsybakov, A. B. (1999).
Smooth discrimination analysis.
Ann. Statist., 27(6):1808–1829.

Massart, P. and Nédélec, É. (2006).

Risk bounds for statistical learning.
Ann. Statist., 34(5):2326–2366.

Sauvé, M. and Tuleau-Malot, C. (2011).
Variable selection through cart.
arxiv:1101.0689.

Scholkopf, B. and Smola, A. J. (2001).
Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.
MIT Press, Cambridge, MA, USA.

Schölkopf, B., Tsuda, K., and Vert, J.-P., editors (2004).
Kernel Methods in Computational Biology.
MIT Press.

Shalev-Shwartz, S. and Srebro, N. (2008).
Svm optimization: Inverse dependence on training set size.
In *25th International Conference on Machine Learning (ICML)*.

Shervashidze, N., Schweitzer, P., van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. (2011).
Weisfeiler-lehman graph kernels.
Journal of Machine Learning Research, 12(Sep):2539–2561.

Shotton, J., Fitzgibbon, A. W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011).
Real-time human pose recognition in parts from single depth images.
In *CVPR*, pages 1297–1304.

Steinwart, I. and Christmann, A. (2008).
Support vector machines.
Information Science and Statistics. Springer, New York.

Stone, C. J. (1977).
Consistent nonparametric regression.
Ann. Statist., 5(4):595–645.
With discussion and a reply by the author.

Talagrand, M. (1996).
New concentration inequalities in product spaces.
Invent. Math., 126(3):505–563.