Introduction to Statistical Learning

Nicolas Vayatis

Lecture 2 - Optimality, deviation inequalities

Course overview

- **Chapter 1 : Optimality in statistical learning**

  Probabilistic view / Performance criteria / Optimal elements

- Chapter 2 : Mathematical foundations of statistical learning

  Concentration inequality / Complexity measures / Regularization

- Chapter 3 : Consistency of mainstream machine learning methods

  Boosting, SVM, Neural networks / Bagging, Random forests

Main messages of Chapter 1 (so far)

- To account for the uncertainty of evaluation, data are assumed to be sampled according to a *fixed* but *unknown* **probability distribution**.
- A prediction objective is characterized by an **error measure**, e.g. the classification *problem* is characterized by misclassification rate as an *error measure* in the case of predicting binary labels using supervised classification *data*.
- The nature of **optimal elements** does tell something about the difficulty of the prediction objective.
- Empirical Risk Minimization (ERM) can be seen as a **generic inference principle** accounting for global optimization methods (e.g. based on convex losses in the case of convex risk minimization)

Outline for today

- Follow up on Optimality (Chapter 1)
  - Variations of the plain classification problem
  - Preference learning
  - The detection problem : ROC curve, AUC & co.
- First results on ERM (Chapter 2)
  - Finite case and deviation inequalities

Chapter 1 - back to work !

Extensions of the plain classification problem

Variations on binary classification

- Asymmetric cost - set $\omega \in (0,1)$,

$$L_\omega(g) = 2\mathbb{E}\big((1-\omega)\mathbb{I}\{Y=+1\}\mathbb{I}\{g(X)=-1\} + \omega\mathbb{I}\{Y=-1\}\mathbb{I}\{g(X)=+1\}\big)$$

- Classification with mass constraint - set $u \in (0,1)$

$$\min_g \mathbb{P}(Y \neq g(X)) \quad \text{subject to} \quad \mathbb{P}(g(X)=1) = u$$

(Refer to Clémençon and Vayatis (2007))

- Classification with reject option - set $\gamma \in (0,1/2)$

$$L_d^R(g) = \mathbb{P}(Y \neq g(X)\ ,\ g(X) \neq \text{®}) + \gamma\mathbb{P}(g(X)=\text{®})$$

(Refer to Herbei and Wegkamp (2006))

Exercise

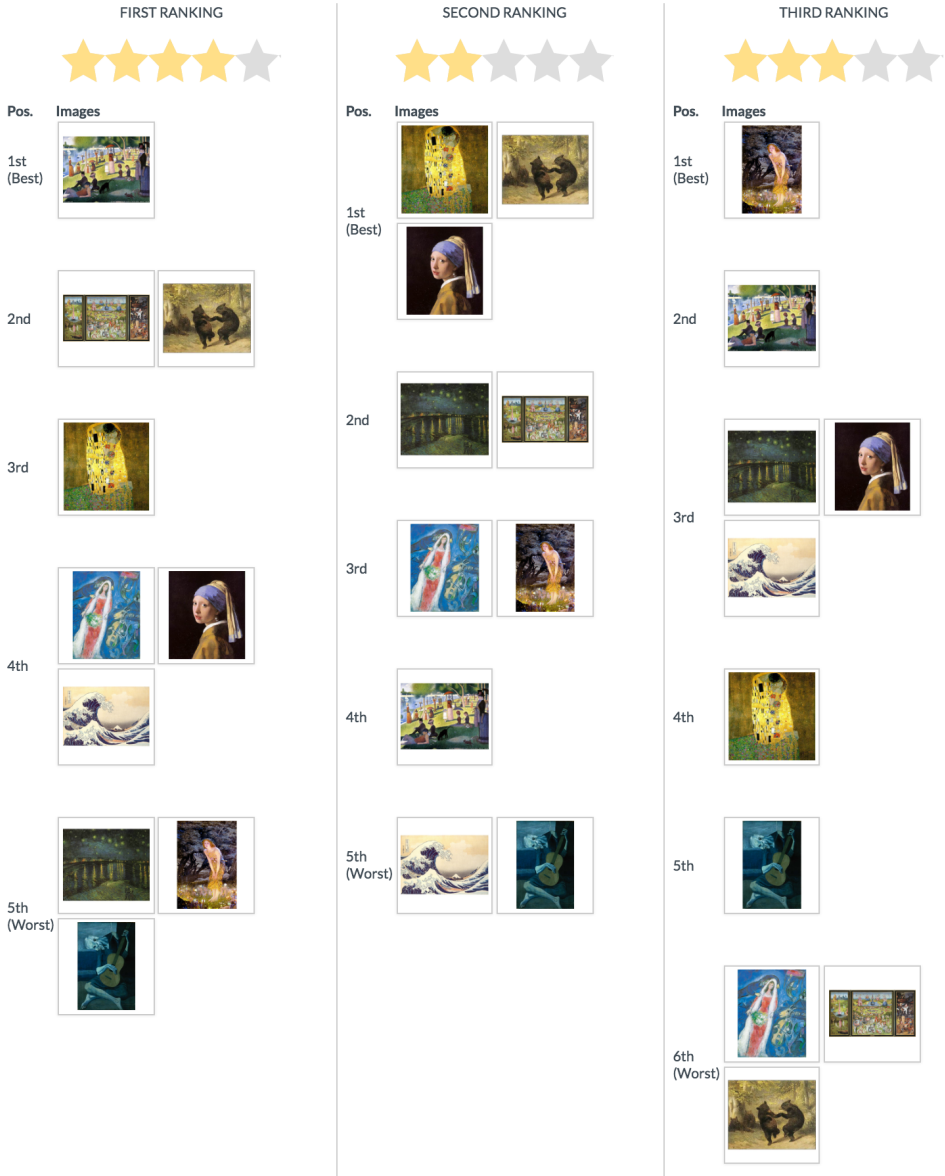
Find g^* and L^* in the three previous scenarios for binary classification :

- Asymmetric cost
- Uner mass constraint
- With reject option

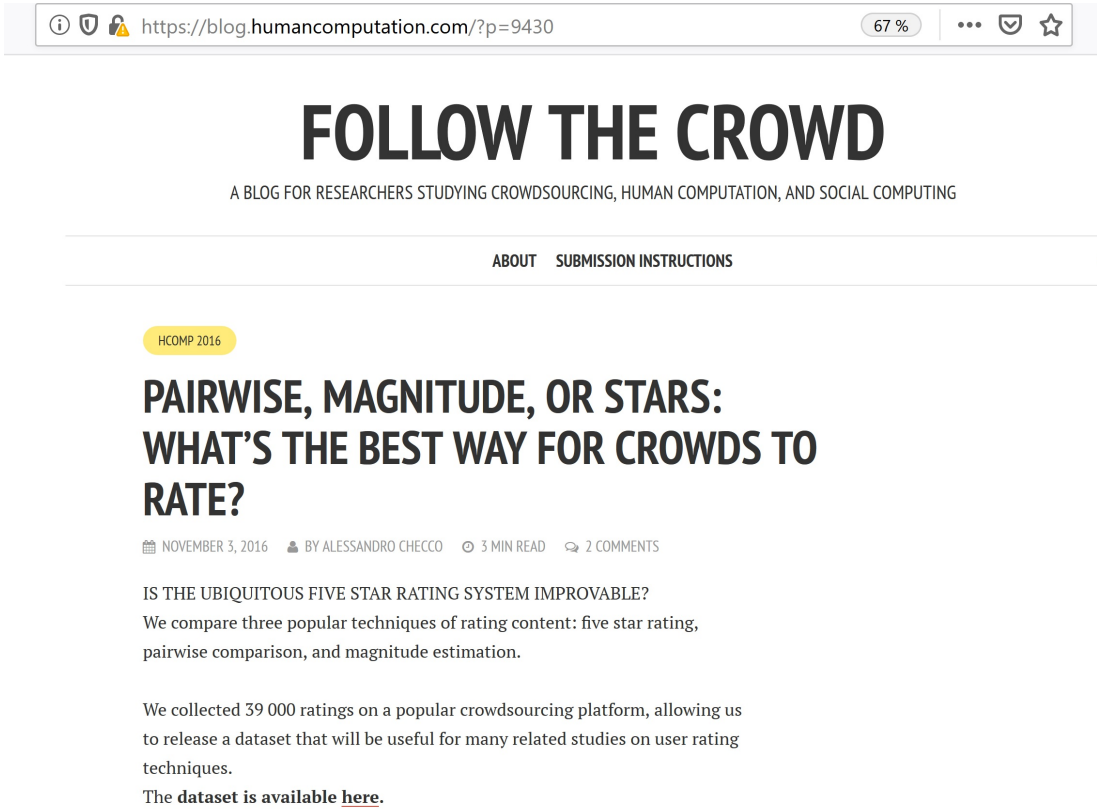
Chapter 1

Preference learning using pairwise comparisons

A study by Checco and Demartini (2016)



Motivations



https://blog.humancomputation.com/?p=9430 67 %

FOLLOW THE CROWD

A BLOG FOR RESEARCHERS STUDYING CROWDSOURCING, HUMAN COMPUTATION, AND SOCIAL COMPUTING

ABOUT SUBMISSION INSTRUCTIONS

HCOMP 2016

PAIRWISE, MAGNITUDE, OR STARS: WHAT'S THE BEST WAY FOR CROWDS TO RATE?

NOVEMBER 3, 2016 BY ALESSANDRO CHECCO 3 MIN READ 2 COMMENTS

IS THE UBIQUITOUS FIVE STAR RATING SYSTEM IMPROVABLE?
We compare three popular techniques of rating content: five star rating, pairwise comparison, and magnitude estimation.

We collected 39 000 ratings on a popular crowdsourcing platform, allowing us to release a dataset that will be useful for many related studies on user rating techniques.
The **dataset is available here.**

METHODOLOGY
We ask each worker to rate 10 popular paintings using 3 rating methods:

- **Magnitude:** Using any positive number (zero excluded).
- **Star:** Choosing between 1 to 5 stars.
- **Pairwise:** Pairwise comparisons between two images, with no ties allowed.

We run 6 different experiments (one for each combination of these three types) with 100 participants in each of them. We can thus analyze the bias given by the rating system order, and the results without order bias by using the aggregated data.

At the end of the rating activity in the task, we **dynamically build** the three painting rankings induced by the choices of the participant, and **ask them** which of the three rankings better reflects their preference (the ranking comparison is blind: There is no indication on how each ranking has been obtained, and their order is randomized).

Data available at : `https://github.com/AlessandroChecco/PairwiseMagnitudeStars`

Preference data

- X, X' , IID random variables taking values in $\mathbb{R}^d$
- $Z \in \mathbb{R}$, preference label
- $Z > 0$ means "X is better than X'"
- (X, X', Z) random triple with unknown distribution P
- Posterior distribution :

$$\forall x, x' \in \mathcal{X}, \quad \begin{aligned} \rho_+(x, x') &= \mathbb{P}\{Z > 0 \mid X = x, X' = x'\} \\ \rho_-(x, x') &= \mathbb{P}\{Z < 0 \mid X = x, X' = x'\} \end{aligned}$$

- If individual labels Y, Y' are observed, then set for instance :

$$Z = \operatorname{sgn}(Y - Y')$$

Preference error and optimal rule

- Preference rule : $r : \mathbb{R}^d \times \mathbb{R}^d \to \{-1, 0, 1\}$
- Ranking error = classification error with pairs

$$L(r) = \mathbb{P}\left\{Z \cdot r(X, X') < 0\right\}$$

- Optimal rule :

$$r^*(x, x') = 2\mathbb{I}\{\rho_+(x, x') > \rho_-(x, x')\} - 1$$

- Minimal error :

$$L^* = L(r^*) = \mathbb{E}\left\{\min\{\rho_+(X, X'), \rho_-(X, X')\}\right\}$$

Exercise

1. Classification data : $Y \in \{-1, +1\}$,
   $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$
   (i) Compute $\rho_+(x, x')$ in terms of η
   (ii) Find the optimal rule and the optimal ranking error, as well as the excess risk

2. Same questions with regression data : $Y = m(X) + \sigma(X) \cdot N$ where $N \sim \mathcal{N}(0, 1)$, $N \perp X$

Chapter 1

The detection problem : ROC curve, AUC & co.

The two types of error

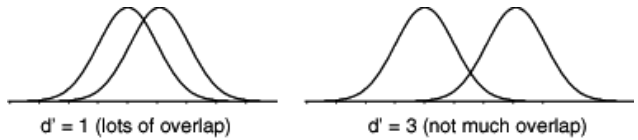
- Consider $s : \mathbb{R}^d \to \mathbb{R}$ a detector response (socring rule)
- A hit corresponds to $Y = +1$, an alarm to $\{s(X) \geq t\}$
- True positive rate and false positive rate :

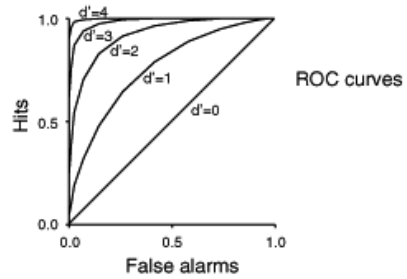
$$\begin{aligned} \beta(s,t) = &\quad \mathbb{P}\{s(X) \geq t \mid Y = +1\} \quad (\text{TPR}) \to \max \\ \alpha(s,t) = &\quad \mathbb{P}\{s(X) \geq t \mid Y = -1\} \quad (\text{FPR}) \to \min \end{aligned}$$

- Main point : trade-off required since

$$\begin{aligned} \beta(s,t) \to 1 \quad &\text{but} \quad \alpha(s,t) \to 1 \quad \text{when} t \to -\infty \\ \alpha(s,t) \to 0 \quad &\text{but} \quad \beta(s,t) \to 0 \quad \text{when} t \to +\infty \end{aligned}$$

Receiver Operating Characteristic curve





- ROC curve of a detector response s :

$$t \in \mathbb{R} \mapsto (\alpha(s,t), \beta(s,t))$$

- Property : the ROC curve is the power curve of the NP test, hence the optimal detector response is η (up to compositions with strictly increasing transformations)

Optimal elements for scoring

- $X \in \mathbb{R}^d$ - observation vector in a high dimensional space
- $Y \in \{-1, +1\}$ - binary diagnosis (i.e. classification data)
- Key theoretical quantity (posterior probability)

$$\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\} \ , \quad \forall x \in \mathbb{R}^d$$

- Optimal scoring rules :

  $\Rightarrow$ increasing transformations of η

Neyman-Pearson view on binary classification

- Hypothesis testing :

$$\mathcal{H}_0 : X \sim P_- \quad \text{against} \quad \mathcal{H}_1 : X \sim P_+$$

- Neyman-Pearson problem : for $\alpha \in (0,1)$, solve

$$\begin{aligned} \max_{T,c} \quad & \mathbb{P}\big(T(X) > c | Y = +1\big) \\ \text{subject to} \quad & \mathbb{P}\big(T(X) > c | Y = -1\big) \leq \alpha \end{aligned}$$

References :

Scott, Nowak (IEEE IT, 2005) - Clémençon, Vayatis (JMLR, 2006) - Rigollet, Tong (JMLR, 2011)

Neyman-Pearson formulation

- Likelihood ratio test

$$T^*(X) = \frac{dP_+}{dP_-}(X) = \frac{1-p}{p} \times \frac{\eta(X)}{1-\eta(X)}$$

with threshold value c^* such that

$$\mathbb{P}\big(T^*(X) > c^* | Y = -1\big) = \alpha$$

yields a **uniformly most powerful** test.

- Binary classification under constraints boils down to adjusting the threshold in a likelihood ratio test

Representation of optimal scoring rules

- Note that if $U \sim \mathcal{U}([0,1])$

$$\forall x \in \mathbb{R}^d \ , \quad \eta(x) = \mathbb{E}\left(\mathbb{I}\{\eta(x) > U\}\right)$$

- If $s^* = \psi \circ \eta$ with ψ strictly increasing, then :

$$\forall x \in \mathbb{R}^d \ , \quad s^*(x) = c + \mathbb{E}\left(w(V) \cdot \mathbb{I}\{\eta(x) > V\}\right)$$

  for some :

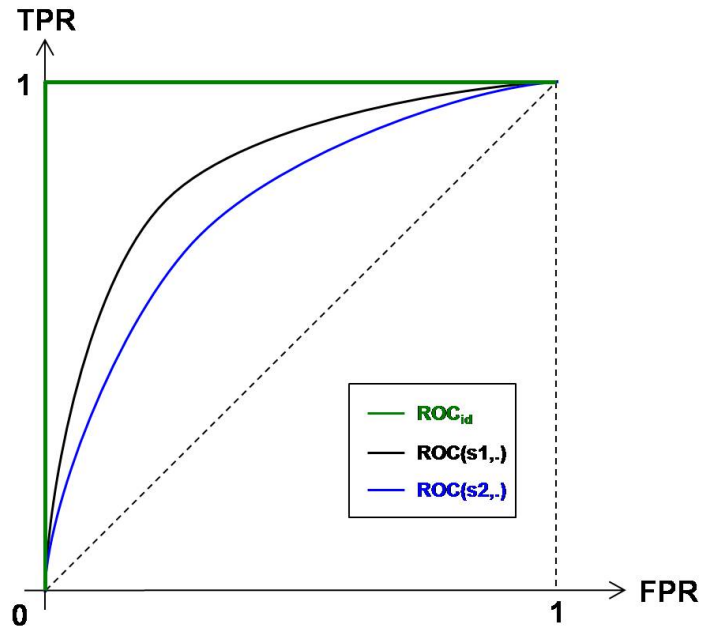
  - $c \in \mathbb{R}$,
  - V continuous random variable in $[0,1]$
  - $w : [0,1] \to \mathbb{R}_+$ integrable.

- Optimal scoring amounts to recovering the level sets of η :

$$\{x : \eta(x) > q\}_{q \in (0,1)}$$

Performance measures for scoring

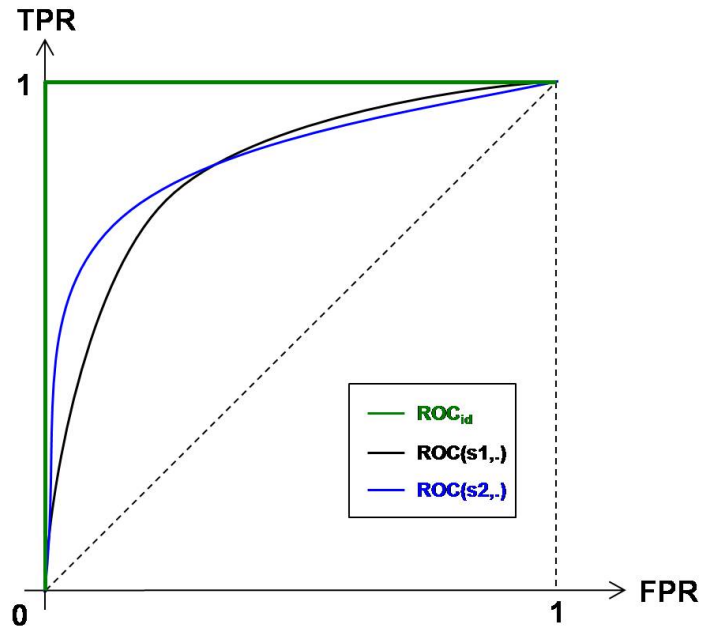
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



ROC curves.

Performance measures for scoring

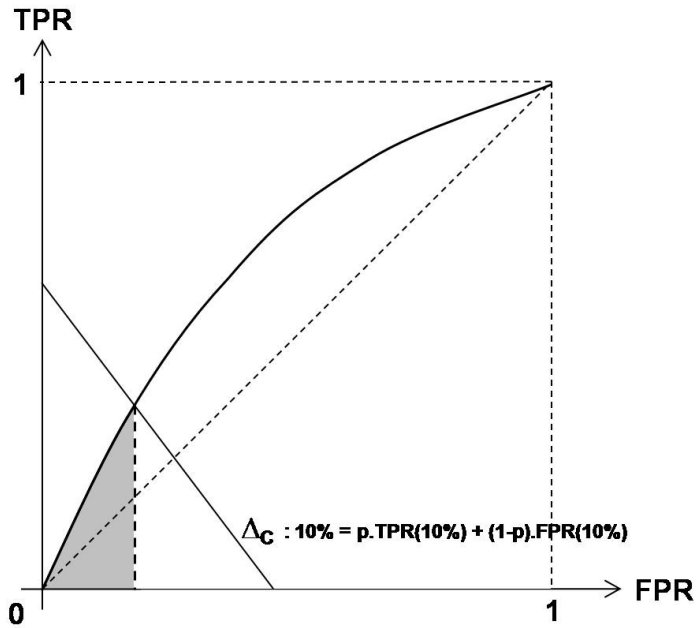
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



ROC curves.

Performance measures for scoring

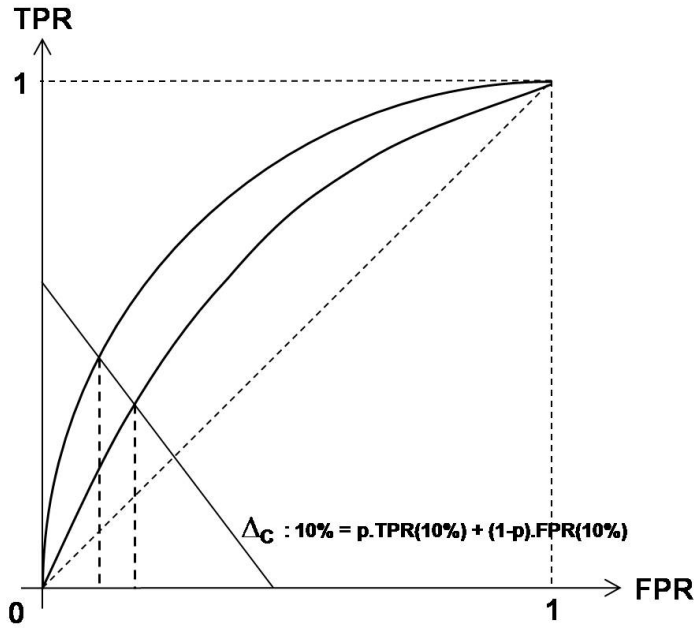
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



Partial AUC.

Performance measures for scoring

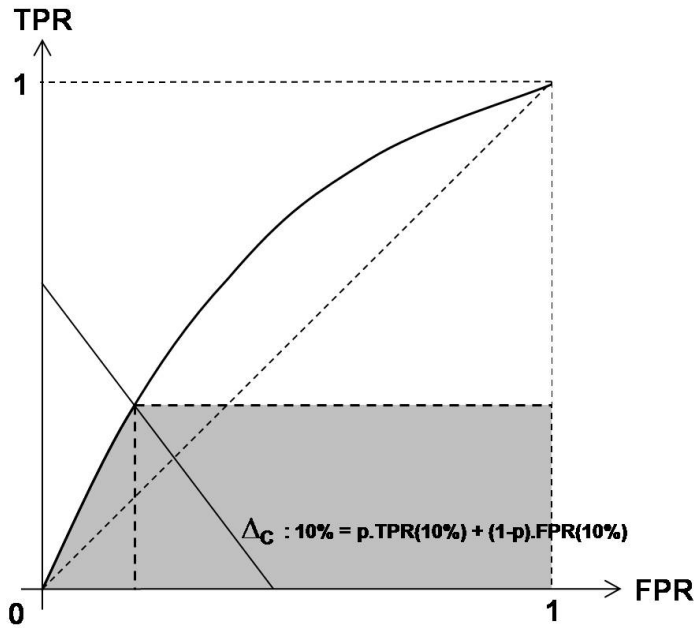
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



Inconsistency of Partial AUC.

Performance measures for scoring

- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC
    (Dodd and Pepe '03)
  - **Local AUC**
    (Clémençon and Vayatis '07)



Local AUC.

Probabilistic interpretation of AUC

- Area Under an ROC Curve (AUC)

$$\mathrm{AUC}(\mathrm{s}) = \mathbb{P}(\mathrm{s}(\mathrm{X}) \geq \mathrm{s}(\mathrm{X}') \mid (\mathrm{Y}, \mathrm{Y}') = (+1, -1))$$

(X, Y), (X', Y') i.i.d.

- The posterior probability is AUC-optimal and we have :

$$\mathrm{AUC}^* - \mathrm{AUC}(s) = \frac{1}{2p(1-p)} \mathbb{E}(\mid \eta(X) - \eta(X') \mid \cdot \mathbb{I}\{(X, X') \in \Gamma_s\})$$

where

  - $\Gamma_s = \{(x, x') \ : \ (s(x) - s(x'))(\eta(x) - \eta(x')) < 0\}$ and
  - $p = \mathbb{P}(Y = +1)$.

Chapter 2 - Mathematical tools

Probability inequalities

Complexity measures

Regularization and stability

Chapter 2

Introduction/Interlude

Learning like the twenty-question game

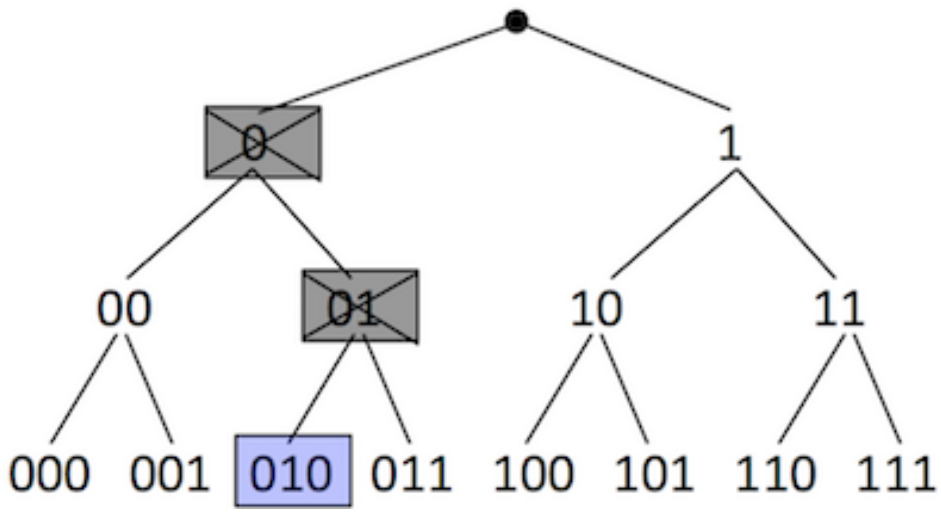
- Assume Nature has picked one function among K and we want to reveal this function
- Assume we have an oracle answering YES or NO when we ask a question about this function
- What is the optimal number n of questions to ask to find the unknown function ?

Brute force learning
Finite case

- ISSUE : How many questions with answers YES or NO one has to ask the oracle to find THE function among K functions ?
- STRATEGY : Proceed recursively by splitting the set of functions in two groups and asking whether THE function is the first group and removing the group which does not contain the function. This leads to the identification of the desired function with about $\log K$ questions.
- ANSWER : Number of questions $n = \dfrac{\log K}{\log 2}$
- NB : this quantity represents the number of bits of information characterizing the function in the set of K functions

Shannon's Information theory
The origin of the $\log K$

- Related to the entropy of a distribution P in information theory : $H(P) = -\sum_{k=1}^{K} P(k) \log P(k)$
- The entropy is the number of bits to encode a collection of K symbols (functions)



From questions to data
Zero error case

- Notations : Domain space $\mathcal{X}$ and label space $\mathcal{Y} = \{0, 1\}$
- ISSUE : How many examples $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ are required to find among a *finite* collection (size K) of indicator functions $f : \mathcal{X} \to \{0, 1\}$ the desired one ?
- SAME ANSWER : Number of examples $n = \dfrac{\log K}{\log 2}$
- STRATEGY : One has to find a vector x_i such that half of the functions take value 1 and the other half take value 0 and ask the oracle whether the desired function takes value 1 or 0 on this vector and discard those functions taking the opposite value. Apply this n times.

Probably approximately correct learning
Zero error case

- REMARK : it may be hard to find such an x_i which splits the collection of functions in two.
- NEW MODEL : Assume $X_1, \ldots, X_n$ is an IID sample
- QUESTION : How many examples (X_i, Y_i) are required to find among a finite collection of indicator functions $f : \mathcal{X} \to \{0, 1\}$ the one that with probability $1 - \delta$ is ε-close to the desired one ?
- ANSWER : Number of examples

$$n = \frac{\log K - \log \delta}{\varepsilon}$$

Probably approximately correct learning
General case

- ASSUME : among K functions, NONE of them commits zero error on the sample (X_i, Y_i).
- SAME ISSUE AS BEFORE
- ANSWER : Number of examples on average

$$n = \frac{\log K - \log \delta}{\varepsilon^2}$$

Same dependency on K, the only change is in the constant.

(Proof coming next)

Finite case (the "$\log K$")

Proposition (Uniform bound for finite classes)

Consider a finite family $\mathcal{H}$ of classifiers. We have, for any $\delta > 0$, with probability at least $1 - \delta$:

$$\forall h \in \mathcal{H} \ , \quad L(h) \leq \widehat{L}_n(h) + \sqrt{\frac{\log |\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{2n}}$$

Proof relies on : Hoeffding's inequality (see later) + union bound ($\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$)

Chapter 2

Probability inequalities

Historical perspective on probability inequalities

- Kolmogorov, Smirnov (1936) : convergence of empirical cdf to their expectations
- Dvoretsky, Kiefer, Wolfowitz (1956) : nonasymptotic version of Kolmogorov-Smirnov
- Hoeffding (1963) : deviation inequality (average of IID from its expectation
- Vapnik-Chervonenkis (1968) : equivalent of DKW for general measures (not only 1D on half lines)
- Mc Diarmid (1981) : first concentration inequality
- Massart (1990) : exact constant in DKW
- Talagrand (1996) : new concentration inequalities

Topics : uniform law of large numbers (and central limit theorem), empirical processes, large deviations, convex geometry, high dimensional probability

Reference : book by Boucheron-Lugosi-Massart (2013)

Hoeffding's lemma

Proposition

Consider Z a random variable such that :

- $\mathbb{E}(Z) = 0$
- $Z \in [a, b]$ *almost surely*

Then, for any $s > 0$, we have :

$$\mathbb{E}\left(e^{sZ}\right) \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Interpretation : the Laplace transform of bounded random variables exhibits subgaussian behavior.

Hoeffding's inequality

Proposition

Consider $Z_1, \ldots, Z_n$ *IID over* $[0,1]$ *and* $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. *We then have, for any* $t > 0$

$$\mathbb{P}\{\overline{Z}_n - \mathbb{E}(Z_1) > t\} \leq \exp(-2nt^2)$$

and

$$\mathbb{P}\{\overline{Z}_n - \mathbb{E}(Z_1) < -t\} \leq \exp(-2nt^2)$$

Consequence : This bound implies the strong law of large numbers for bounded random variables (by Borel-Cantelli lemma)

Proof technique : Chernoff's bounding method

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}(Z_1) > t\right) \leq \inf_{s>0} \exp\left(-nst + n\log\mathbb{E}(e^{s(Z_1 - \mathbb{E}(Z_1))})\right)$$

Coming next in Chapter 2

- Probability inequalities : from deviation to concentration
- Complexity measures
- Error bounds on ERM