



école —————  
normale —————  
supérieure —————  
paris-saclay —————

université  
PARIS-SACLAY

# MVA Course

## Introduction to Statistical Learning

Nicolas Vayatis

Lecture 1

## Practical information about the course

- **Course website :**

<http://nvayatis.perso.math.cnrs.fr/ISLcourse.html>

- **Schedule and location**

- Location : ENS Paris-Saclay
- Dates : Check the online agenda
- Classroom : Check out the agenda regularly
- Format : 6 lectures + 4 exercise sessions + personal research
- Office hours : on demand

- **Evaluation :**

- Two mandatory exams : Mid-term exam M + final exam F
- Final grade  $G = \max(F; (F+M)/2)$
- Mid-term on November 4 am
- Final exam on January 6 am

## Course overview

- Chapter 1 : Optimality in statistical learning  
Data / Objectives / Optimal elements / ERM
- 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

# Chapter 1 - Optimality in statistical learning

# Overview of Chapter 1

- Modeling the data :
  - a *probabilistic* view
- Modeling the prediction objective :
  - performance *metrics* and risk functionals for prediction
- The goal of learning :
  - *optimal* elements
- The mother of most Machine Learning algorithms :
  - *ERM* : Empirical Risk Minimization

## 1.1. Modeling classification data

## Generative vs. discriminative

- $(X, Y)$  random pair with distribution  $P$  over  $\mathbb{R}^d \times \{-1, +1\}$

① **Generative view** - Joint distribution  $P$  as a mixture

- Class-conditional densities :  $f_+$  and  $f_-$
- Mixture parameter :  $p = \mathbb{P}\{Y = +1\}$

② **Discriminative view** - Joint distribution  $P$  described by  $(P_X, \eta)$

- Marginal distribution :  $X \sim P_X = df_X/d\lambda_d$
- Posterior probability function :

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

- Marginal distribution of  $X$  has density :  $f_X = pf_+ + (1 - p)f_-$
- Posterior probability is given by :  $\eta = pf_+/f_X$

## Exercise

Find the expressions of  $f_+$ ,  $f_-$  and  $\eta$  in the following probabilistic models :

- Discriminant Analysis : find  $\eta$  knowing

$$f_+ = \mathcal{N}_d(\mu_+, \Sigma_+), f_- = \mathcal{N}_d(\mu_-, \Sigma_-)$$

- Logistic regression : find  $f_+$ ,  $f_-$  knowing

$$\log \left( \frac{\eta_{\theta}(x)}{1 - \eta_{\theta}(x)} \right) = h(x, \theta), \quad \text{typically } h(x, \theta) = \theta^T x$$

## 1.2. Optimality in the binary classification objective

## Classifier, Error measure, Optimal Elements

- Classifier :  $g : \mathbb{R}^d \rightarrow \{-1, +1\}$
- Classification error :  $L(g) = \mathbb{P}\{g(X) \neq Y\}$   
 $L(g) = \mathbb{E}(\eta(X) \cdot \mathbb{I}\{g(X) = -1\} + (1 - \eta(X)) \cdot \mathbb{I}\{g(X) = 1\})$
- Bayes rule :  $g^*(x) = 2\mathbb{I}\{\eta(x) > 1/2\} - 1, \quad \forall x \in \mathbb{R}^d$
- Bayes error :  $L^* = L(g^*) = \mathbb{E}\{\min(\eta(X), 1 - \eta(X))\}$
- Excess risk :

$$L(g) - L^* = 2\mathbb{E}\left\{\left|\eta(X) - \frac{1}{2}\right| \cdot \mathbb{I}\{g(X) \neq g^*(X)\}\right\}$$

## Link with parametrics : Plug-in methods do the job but...

- Let  $\hat{\eta}$  an estimate of the posterior  $\eta$  based on a sample  $D_n$  (e.g. LDA/QDA, logistic regression)
- Consider  $\hat{g}$  a plug-in estimator based on  $\hat{\eta}$

$$\hat{g}(x) = 2\mathbb{I}\{\hat{\eta}(x) > 1/2\} - 1, \quad \forall x \in \mathbb{R}^d$$

- We have, conditionally on the sample  $D_n$  :

$$L(\hat{g}) - L^* \leq 2\mathbb{E}_X(|\hat{\eta}(X) - \eta(X)|)$$

- But estimation of  $\eta$  for high dimensional data suffers of the curse of dimensionality !
- Q : Do we really need to estimate  $\eta$  ?

## 1.3. Convex risk minimization

## Convex Risk Minimization (CRM)

- Binary classification data with  $Y \in \{+1, -1\}$
- Real-valued decision rule (*soft classifier*)  $f : \mathbb{R}^d \rightarrow \mathbb{R}$
- Cost function  $\varphi : \mathbb{R} \rightarrow \mathbb{R}_+$  convex, increasing,  $\varphi(0) = 1$
- Expected  $\varphi$ -risk :

$$A(f) = \mathbb{E}(\varphi(-Y \cdot f(X)))$$

- Main examples :

$$\varphi(x) = e^x, \log_2(1 + e^x), (1 + x)_+$$

- Note that :  $L(\text{sgn}(f)) \leq A(f)$

## Exercise

Find the optimal elements for convex risk minimization :

$$f^* = \arg \min_f A(f) , \quad A^* = A(f^*)$$

in the following examples :

- (i)  $\varphi(u) = \exp(u)$
- (ii)  $\varphi(u) = \log_2(1 + \exp(u))$
- (iii)  $\varphi(u) = (1 + u)_+$

## Risk communication (1) : Zhang's lemma

- Assumption (A1) :  $\varphi$  positive, convex, increasing, such that  $\varphi(0) = 1$
- Set the 'entropy' function :

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} (\eta\varphi(-\alpha) + (1 - \eta)\varphi(\alpha))$$

- Assumption (A2) :  $\exists s \geq 1$  and  $c > 0$  such that  $\forall u \in (0, 1)$  ,

$$\left| \frac{1}{2} - u \right|^s \leq c^s (1 - H(u))$$

- Under (A1-A2), we have, for some  $s > 1$ , that any real-valued measurable  $f$  satisfies :

$$L(g_f) - L^* \leq 2c(A(f) - A^*)^{1/s}$$

## CRM optimality and examples from Zhang (2004)

- Classifier obtained by CRM :

$$g_{f^*}(x) = 2\mathbb{I}\{f^*(x) > 0\} - 1$$

- Result : if  $\varphi \in \{\text{exp, logit, hinge, \dots}\}$ , then

$$g_{f^*} = g^* \quad (\text{Bayes rule})$$

- Zhang's lemma : if  $\varphi \in \{\text{exp, logit}\}$ , then

$$L(g_f) - L^* \leq \sqrt{2}(A(f) - A^*)^{1/2}$$

- Zhang's lemma : if  $\varphi = \text{hinge}$ , then

$$L(g_f) - L^* \leq A(f) - A^*$$

## "Optimal" CRM optimality

- Assume convexity of cost function  $\varphi : \mathbb{R} \rightarrow \mathbb{R}_+$ , then :

$\forall x, f^*(x) \cdot (\eta(x) - 1/2) > 0$  if and only if  $\exists \varphi'(0)$  and  $\varphi'(0) > 0$

- One-sided version of binary entropy function :

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

## Risk communication (2) : Bartlett, Jordan, and McAuliffe (2006)

- Define the communication function :

$$\psi(x) = H^- \left( \frac{1+x}{2} \right) - H^- \left( \frac{1-x}{2} \right)$$

- Under the assumption that  $\varphi$  is convex,  $\exists \varphi'(0)$  and  $\varphi'(0) > 0$ , we have the control of excess risk through  $\psi^{-1}$  :

$$\psi(L(g_f) - L^*) \leq A(f) - A^*$$

(note that  $\psi$  convex and  $\psi(0) = 0$ )

## 1.4. Empirical Risk Minimization

## Supervised learning setup

- Goal of learning : an optimal decision function  $h^* : \mathcal{X} \rightarrow \mathcal{Y}$   
 $\mathcal{X}$  : domain set,  $\mathcal{Y}$  : label set
- Input of learning :
  - **Training data** : a set of labeled data

$$D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

of size  $n$ , where the  $(X, Y)$ 's are in  $\mathcal{X} \times \mathcal{Y}$

- **Hypothesis space** : a collection  $\mathcal{H}$  of candidate decision functions  $h : \mathcal{X} \rightarrow \mathcal{Y}$
- Output of learning : an empirical decision function  $\hat{h}$  in the hypothesis space  $\mathcal{H}$  estimated from training data  $D_n$
- Reference in  $\mathcal{H}$  : the best decision function  $\bar{h}$  in the class (the more data, the closer  $\hat{h}$  to  $\bar{h}$ )

# The ERM principle

## Definition

- Loss function :  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, +\infty]$
- Empirical risk of a decision rule  $h$  : this is a data-dependent functional

$$\widehat{L}_n(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(X_i), Y_i)$$

- ERM = Empirical Risk Minimization

Learning from training data amounts to solving the following optimization problem

$$\widehat{h}_n = \arg \min_{h \in \mathcal{H}} \widehat{L}_n(h)$$

where the minimization is restricted to the hypothesis space.

## The notion of *true error*

- Assumption :  
( $X, Y$ ) is a pair of random variables with joint distribution  $P$
- True error of a decision rule  $h$  : this is a distribution-dependent functional

$$L(h) = \mathbb{E}(\ell(h(X), Y)) = \int \ell(h(x), y) dP(x, y)$$

## Optimal elements, consistency and bounds

- Bayes rule  $h^*$  and Bayes error  $L^*$

$$h^* = \arg \min_h L(h) \quad \text{and} \quad L^* = L(h^*)$$

- (Strong) Consistency of an inference principle  $\hat{h}_n$

$$L(\hat{h}_n) \rightarrow L^* , \quad \text{almost surely}$$

- The nonasymptotic bounds Eldorado :

$$L(\hat{h}_n) - L^* \leq U(n, \mathcal{H}) \quad \text{whp}$$

# Estimation vs. approximation error

## Extension of bias-variance decomposition

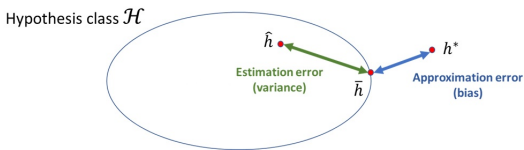
- Proof idea : Add and retrieve  $\widehat{L}_n(\widehat{h}_n)$  ,  $\widehat{L}_n(\bar{h})$ ,  $L(\bar{h})$ , then use the definition of ERM to upper bound the sum. Difference between  $L$  and  $\widehat{L}_n$  appear twice.
- We have :

$$L(\widehat{h}_n) - L^* \leq \underbrace{2 \sup_{h \in \mathcal{H}} |L(h) - \widehat{L}_n(h)|}_{\text{estimation (stochastic)}} + \underbrace{L(\bar{h}) - L^*}_{\text{approximation (deterministic)}}$$

## The key trade-off in Machine Learning

- Denote by  $L(h)$  the error measure for any decision function  $h$
- We have :  $L(\bar{h}) = \inf_{\mathcal{H}} L$  , and  $L(h^*) = \inf L$
- Bias-Variance type decomposition of error for any output  $\hat{h}$  :

$$L(\hat{h}) - L(h^*) = \underbrace{L(\hat{h}) - L(\bar{h})}_{\text{estimation (stochastic)}} + \underbrace{L(\bar{h}) - L(h^*)}_{\text{approximation (deterministic)}}$$



(Note : the argument misses numerical error - Bottou, Bousquet (2007))

## 1.5. Some other supervised learning problems

## From plain classification to...

- Classification in real life : multiclass classification, asymmetric cost, classification with mass constraint, classification with reject option, Neyman-Pearson classification
- Preference learning
- Scoring
- Regression

## Variations on binary classification

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

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

- 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 Cléménçon and Vayatis (2007))

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

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

(Refer to Herbei and Wegkamp (2006))

## Which decision ?

Build a **decision rule** to be evaluated on a new sample

### ① Predictive Classification

Given a new  $X'$ , predict the label  $Y'$

Decision rule :  $g : \mathbb{R}^d \rightarrow \{-1, +1\}$

Happy if classification error rate is low on average

### ② Predictive Ranking/Scoring

Given new data  $\{X'_1, \dots, X'_m\}$ , predict a ranking  $(X'_{i_1}, \dots, X'_{i_m})$

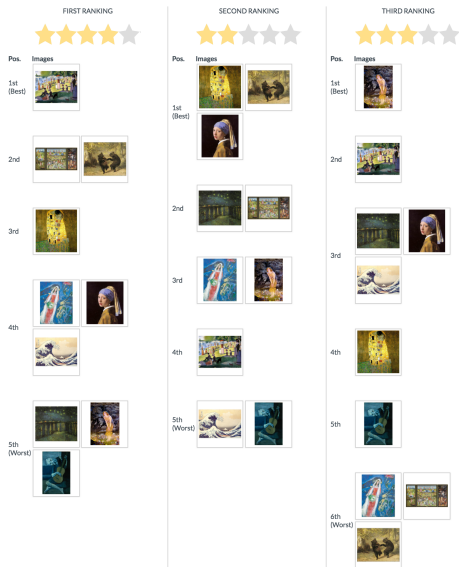
Decision rule :  $s : \mathbb{R}^d \rightarrow \mathbb{R}$  that defines the permutation  $(i_1, \dots, i_m)$

Happy if  $(Y'_{i_1}, \dots, Y'_{i_m})$  is "close" to a decreasing sequence

**Goal** : Define an order on  $\mathbb{R}^d$  from binary label information

## 1.6. Preference learning using pairwise comparisons

# A study by Checco and Demartini (2016)



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 :

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

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

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

## Preference error and optimal rule

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

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

- Optimal rule :

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

- Minimal error :

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

## 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  $\eta$
  - (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  $N \sim \mathcal{N}(0, 1)$ ,  $N \perp X$

## 1.7. The detection problem : ROC curve, AUC & co.

## The two types of error

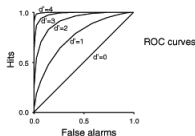
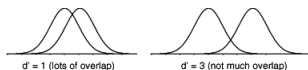
- Consider  $s : \mathbb{R}^d \rightarrow \mathbb{R}$  a detector response (scoring rule)
- A hit corresponds to  $Y = +1$ , an alarm to  $\{s(X) \geq t\}$
- True positive rate and false positive rate :

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

- Main point : trade-off required since

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

# 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  $\eta$  (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 :  
⇒ increasing transformations of  $\eta$

# Neyman-Pearson view on binary classification

- Hypothesis testing :

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

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

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

References :

Scott, Nowak (IEEE IT, 2005) - Cléménçon, 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

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

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  $U \sim \mathcal{U}([0, 1])$

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

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

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

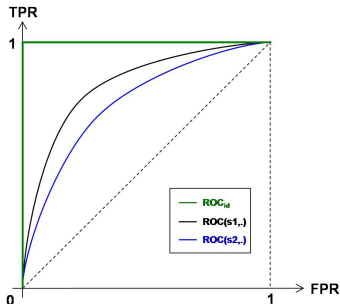
for some :

- $c \in \mathbb{R}$ ,
  - $V$  continuous random variable in  $[0, 1]$
  - $w : [0, 1] \rightarrow \mathbb{R}_+$  integrable.
- Optimal scoring amounts to recovering the level sets of  $\eta$  :

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

# Performance measures for scoring

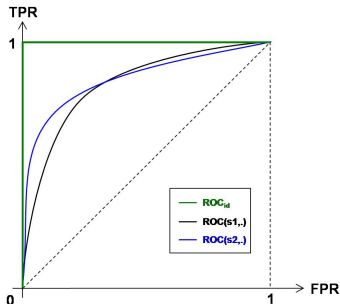
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC  
(Dodd and Pepe '03)
  - **Local AUC**  
(Cléménçon and Vayatis '07)



ROC curves.

# Performance measures for scoring

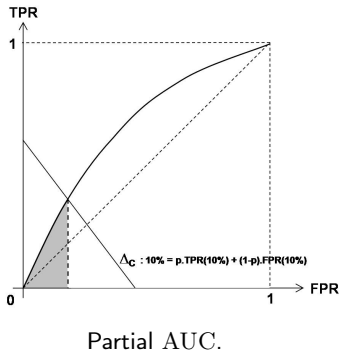
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC  
(Dodd and Pepe '03)
  - **Local AUC**  
(Cléménçon and Vayatis '07)



ROC curves.

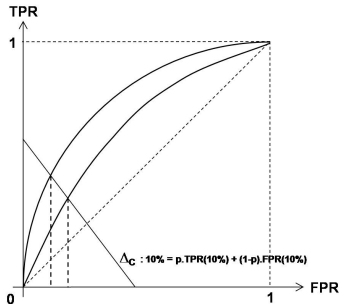
# Performance measures for scoring

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



# Performance measures for scoring

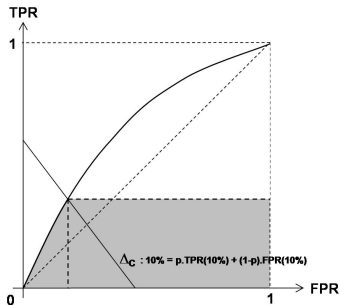
- Curves :
  - **ROC curve**
  - (Precision-Recall curve)
  - (Lift curve)
- Summaries :
  - **AUC** (global measure)
  - Partial AUC  
(Dodd and Pepe '03)
  - **Local AUC**  
(Cléménçon and Vayatis '07)



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**  
(Cléménçon and Vayatis '07)



Local AUC.

## Probabilistic interpretation of AUC

- Area Under an ROC Curve (AUC)

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

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

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

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

where

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

## Coming next

Next lecture :

- What is the complexity of learning
- Mathematical tools

⇒ Homework → Prepare Exercise Set #1