MVA Course

# Introduction to Statistical Learning

Nicolas Vayatis

Session 1

# Course objectives

- Introduction to the mathematical foundations of machine learning
- Setup a learning problem to better address users' expectations and constraints
- Get insights to understand the key principles of shallow machine learning methods

# Practical information about the course

- **Course website :**
  http://nvayatis.perso.math.cnrs.fr/ISLcourse-2022.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 8, 9-11am
  - Final exam on January 9 (TBC)

# 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

# Chapter 1

## A. 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

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

# Chapter 1

## B. Optimality in the binary classification objective

# Classifier, Error measure, Optimal Elements

- Classifier : $g : \mathbb{R}^d \to \{-1, +1\}$

- Classification error : $L(g) = \mathbb{P}\{g(X) \neq Y\}$

$$L(g) = \mathbb{E}\big(\eta(X) \cdot \mathbb{I}\{g(X) = -1\} + (1 - \eta(X)) \cdot \mathbb{I}\{g(X) = 1\}\big)$$

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

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

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

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

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

# Chapter 1

## C. Convex risk minimization

# Convex Risk Minimization (CRM)

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

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

- Main examples :

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

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

# 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) $\varphi(u) = (1 + u)_+$

# 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 :

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

# Risk communication (1) : Zhang (2004)

- Classifier obtained by CRM :

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

- Result : if $\varphi \in \{\exp, \mathrm{logit}, \mathrm{hinge}, \ldots\}$, then

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

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

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

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

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

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

- Set the 'entropy' function :

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

- ... and the one-sided version :

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

- Define the communication funtion :

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

- Control of excess risk with $\psi^{-1}$ :

$$L(g_f) - L^* \leq \psi^{-1}(A(f) - A^*)$$

and $\varphi$ convex implies $\lim_{0^+} \psi^{-1} = 0$

# Chapter 1

## D. Empirical Risk Minimization

# Supervised learning setup

- Goal of learning : an optimal decision function $h^* : \mathcal{X} \to \mathcal{Y}$
  $\mathcal{X}$ : domain set, $\mathcal{Y}$ : label set

- Input of learning :
  - **Training data :** a set of labeled data

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

    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 $h \ : \ \mathcal{X} \to \mathcal{Y}$

- Output of learning : an empirical decision function $\widehat{h}$ 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 $\widehat{h}$ to $\bar{h}$)

# The ERM principle
## Definition

- Loss function : $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty]$
- 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 $\widehat{h}_n$

$$L(\widehat{h}_n) \to L^* \;, \quad \text{almost surely}$$

- The nonasymptotic bounds Eldorado :

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

## Estimation vs. approximation error
## Extension of bias-variance decomposition

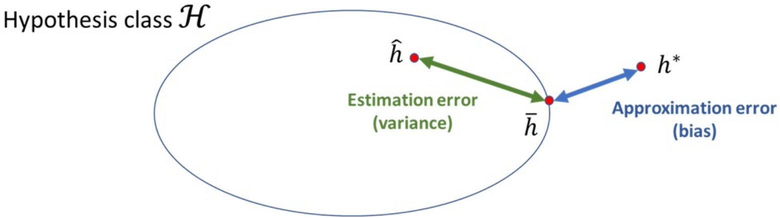
- Proof idea : Add and retrieve $\widehat{L}_n(\widehat{h}_n)$ , $\widehat{L}_n(\bar{h})$, $L(\bar{h})$, then use the definition of ERM to upper bound the sum. Difference between $L$ and $\widehat{L}_n$ appear twice.
- We have :

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

# 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 $\widehat{h}$ :

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



# Chapter 1

## E. 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 $\omega \in (0,1)$,

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

- 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 $\gamma \in (0,1/2)$

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

(Refer to Herbei and Wegkamp (2006))

- Neyman-Pearson classification

## 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 : $g \ : \ \mathbb{R}^d \to \{-1, +1\}$

   Happy if classification error rate is low on average

2. Predictive Ranking/Scoring

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

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

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

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

# Coming next

Next lecture :

- Discuss the case of different prediction objectives
- What is the complexity of learning
- Mathematical tools

$\Rightarrow$ Homework $\rightarrow$ Prepare Exercise Set #1