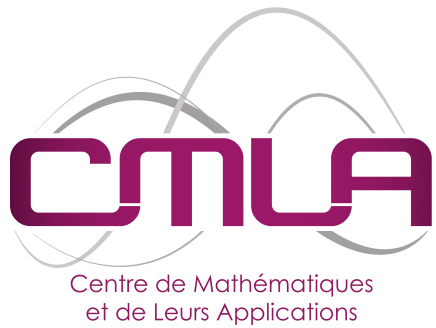
Introduction to Statistical Learning

Nicolas Vayatis (CMLA - ENS Cachan)



MVA 2016-2017

Chapter 1 - Modeling aspects

Overview of Chapter 1

- Modeling the data:
  - a *probabilistic* view
- Modeling the objective:
  - performance *metrics* and risk functionals for prediction
- The goal of learning:
  - *optimal* elements
- Assessment:
  - performance *estimation* based on empirical data

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, η)
   - 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 has density: $f_X = pf_+ + (1-p)f_-$
- Posterior probability is given by: $\eta = pf_+/f_X$

Parametric methods (1) - Discriminant Analysis

- Typical example: Mixture model with $f_+ = \mathcal{N}_d(\mu_+, \Sigma_+)$ and $f_- = \mathcal{N}_d(\mu_-, \Sigma_-)$
- Estimate class-conditional distribution with maximum likelihood estimates
- Then use estimates of posterior probabilities: $\forall x \in \mathbb{R}^d$

$$\eta(x) = \frac{p f_+(x)}{f_X(x)} \ , \quad 1 - \eta(x) = \frac{(1-p) f_-(x)}{f_X(x)}$$

- Decision rules result from the *plug-in* principle
- If d large, apply dimension reduction techniques (*e.g.* PCA)

Parametric methods (2) - Logistic Regression

- Consider a family $\{\eta_\theta \ : \ \theta \in \mathbb{R}^d\}$ such that:

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

- This is equivalent to:

$$\eta_\theta(x) = \frac{\exp(\theta^T x)}{1+\exp(\theta^T x)}$$

- Estimation $\widehat{\theta}$ by conditional likelihood maximization (Newton-Raphson algorithm)
- Estimate $\eta_{\widehat{\theta}}$ of the posterior used for classification and scoring

Which decision?

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

2. Prediction problems

a. Binary classification

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 η based on a sample D_n
- 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:

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

- But estimation of η for high dimensional data is a difficult problem!
- Q: Do we really need to do this?

Classification error = two types of error

- Decomposition of classification error

$$L(g) = \mathbb{P}\{g(X) = +1, Y = -1\} + \mathbb{P}\{g(X) = -1, Y = +1\}$$

- False positive rate

$$\alpha(g) = \mathbb{P}\{g(X) = +1 \mid Y = -1\}$$

- True positive rate

$$\beta(g) = \mathbb{P}\{g(X) = +1 \mid Y = +1\}$$

- Note that:

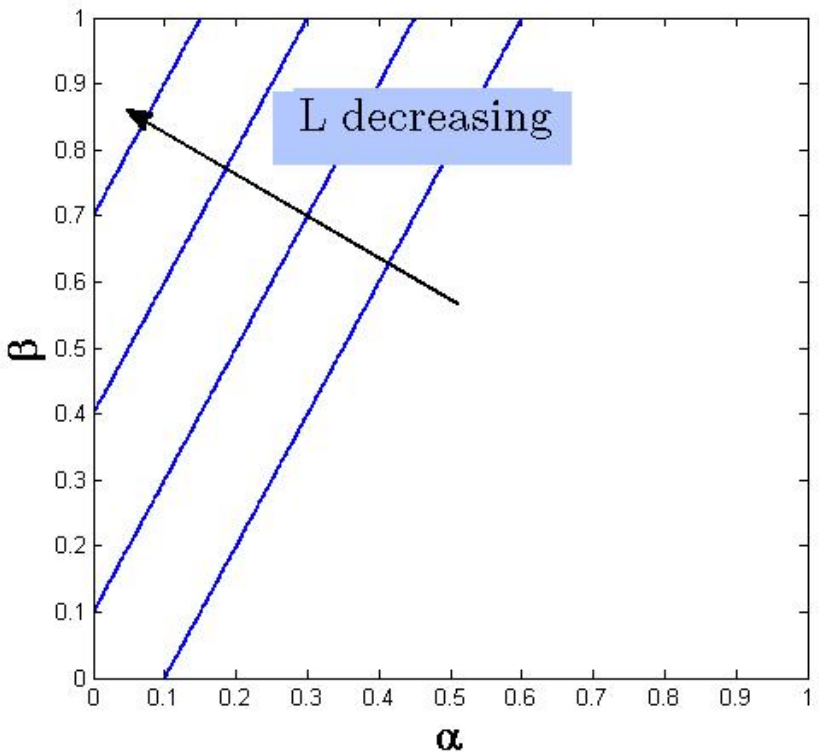
$$L(g) = (1 - p)\alpha(g) + p(1 - \beta(g))$$

where $p = \mathbb{P}(Y = +1)$

Classification - two types of error

- Fixed proportion p, fixed classification error $L(g) = L$

$$\beta = \left(\frac{1-p}{p}\right)\alpha + 1 - \frac{L}{p}$$



Hypothesis testing

- Under observation X, test

$$H_0 : Y = -1 \quad \text{against} \quad H_1 : Y = +1$$

- Optimal test statistic (Neyman-Pearson = NP)

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

- $\alpha =$ Type I error
- $\beta =$ power of the test

Hypothesis testing and optimal NP classifier

- For fixed α, rejection region given by :

$$R_\alpha^* = \{\ x\ :\ \eta(x) > Q^-(\eta, \alpha)\ \}$$

  where $Q^-(\eta, \alpha) = (1-\alpha)$-quantile of $\mathcal{L}(\eta(X) \mid Y = -1)$

- Set the classifier :

$$g_\alpha^*(x) = 2\mathbb{I}\{\ x \in R_\alpha^*\ \} - 1$$

- In general : $L(g_\alpha^*) > L^*$ except if $Q^-(\eta, \alpha) = 1/2$

Comments

- One classifier $=$ one point in the (α, β) square
- Small (or large) $p \Rightarrow$ instability in β (or α)
- Indeed, if $p \to 1$, then $g \equiv +1$ has error $L(g) \to 0$ but $\alpha(g) = 1$
- NP optimality leads to $g^*_\alpha \neq g^*$ in general
- Optimality at all levels $\Rightarrow$ ROC curve!
- Invariance property : $\psi \circ \eta$ with ψ increasing leads to the same decision

Other 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))

To be continued...