Basic methods for Classification

B. Michel

Ecole Centrale de Nantes

Statistical Learning

Classification

Framework for this chapter:

- Target variable: $Y \in \{1, \ldots, K\}$.
- Features: $(X^1, \ldots X^p) \in \mathcal{X}$.
- Data: n observations (target, features).
- We want to define classification rules to predict Y from X.

MAP rule

The **MAP** (Maximum a posteriori) rule consists in assigning a point x to the group for which the posterior probability is the greatest :

$$\tilde{k}(x) = \underset{k=1\ldots K}{\operatorname{argmax}} P(Y = k|X = x).$$

Proposition

The MAP rule is a Bayes rule for the 0-1 loss.

Outline

Outline

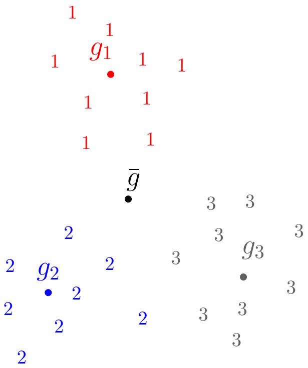
1. **Generative Models**
   - Generative approach and the Bayes's rule
   - Naive Bayes

Generative Models

- Based on the **conditional distribution** $(X|Y)$ distribution (X, Y) (Bayes Theorem)
- “Generative" because it is based on the joint distribution that generates the observations.
- Popular models : Gaussians, Naive Bayes, Linear / Quadratic Discriminant Analysis, Hidden Markov Models (HMM), Bayesian networks, Markov random fields ...

Generative Models

- Outcome variable :
  - $Y \in \mathcal{Y} = \{1, \ldots, K\}$.
  - Y follows a discrete distribution $\sum_{k=1}^{K} \pi_k \delta_{y=k}$ on $\{1, \ldots, K\}$.
  - Density $\sum_{k=1}^{K} \pi_k \mathbb{1}_{y=k}$ with respect to $\delta_{\{1,\ldots,K\}} := \sum_{k=1}^{K} \delta_{y=k}$.
  - $\pi_k = P(Y = k)$: a priori probabilities.
- Vector of features :
  - $X = (X_1, \ldots, X_j, \ldots, X_p) \in \mathcal{X}$
  - The X_j's are continuous or categorical variables.
  - The conditional distribution of $(X|Y = k)$ admits a density f_k with respect to a ref. measure μ on $\mathcal{X}$.
  - e.g. $\mathcal{X} = \mathbb{R}^p$, $\mu = \lambda^p$ (Lebesgue on $\mathbb{R}^p$).



MAP rule for generative models

- The **MAP** (Maximum a posteriori) rule consists in assigning a point x to the group for which the posterior probability is the greatest :

$$\tilde{k}(x) = \operatorname*{argmax}_{k=1\ldots K} P(Y = k | X = x).$$

- Bayes' theorem: $P(B)P(A|B) = P(B|A)P(A)$. Application to compare posterior distributions :

$$\begin{aligned}
\frac{P(Y = k|X = x)}{P(Y = k'|X = x)} &= \frac{P(Y = k)}{P(Y = k')} \frac{P(X = x|Y = k)}{P(X = x|Y = k')} \\
&= \frac{\pi_k}{\pi_{k'}} \frac{f_k(x)}{f_{k'}(x)}
\end{aligned}$$

Thus

$$\tilde{k}(x) = \operatorname*{argmax}_{k=1\ldots K} \ \pi_k f_k(x).$$

Inference

- MAP rule
$$\tilde{k}(x) = \underset{k=1\ldots K}{\operatorname{argmax}} \ \pi_k f_k(x).$$

- In practice the π_k's and the posterior probabilities $P(X = x | Y = k)$'s are unknown.

- We infer these quantities in parametric settings (Maximum Likelihood !)

- Effective MAP rule by plug-in
$$\hat{k}(x) = \underset{k=1\ldots K}{\operatorname{argmax}} \ \hat{\pi}_k \hat{f}_k(x).$$

Outline

1. **Generative Models**
   - Generative approach and the Bayes's rule
   - Naive Bayes

Naive Bayes Assumption

- Framework :
  - Y : K classes
  - p features $X_1, \dots, X_j, \dots, X_p$, continuous or categorical variables.

- Naive Bayes assumes a crude modeling for $(X|Y)$: features X_j are **independent conditionally on** Y, i.e.

$$P(X = x|Y = k) = \prod_{j=1}^{p} P(X_j = x_j|Y = k)$$

For the densities :

$$f_k(x) = \prod_{j=1}^{p} f_{j,k}(x_j)$$

where $f_{j,k}$ is the density of $(X_j|Y = k)$.

Exercice

Draw pictures (scatter plot, boxplots ...) to illustrate this assumption for a bivariate distribution (X_1, X_2) (numerical or categorical).

Additive model

- Ratio of posterior probabilities :

$$\frac{P(Y=k|X=x)}{P(Y=k'|X=x)} = \frac{P(Y=k)}{P(Y=k')} \frac{f_k(x)}{f_{k'}(x)}$$

- Log-ratio of the posterior probabilities is an additive function of the univariate log-ratios :

$$\log \frac{P(Y=k|X=x)}{P(Y=k'|X=x)} = \log \frac{\pi_k}{\pi_{k'}} + \sum_{j=1}^{p} \log \frac{f_{j,k}(x_j)}{f_{j,k'}(x_j)}$$

Parametric assumptions and Maximum Likelihood Estimation

- The observations (X_i, Y_i) are independent.
- For any (j, k), assume that the distribution of $(X_j | Y = k)$ is in a parametric model such that
  - it admits a density $x_j \mapsto f_{j,k}(\eta_{j,k}, x_j)$ (w.r.t. a reference measure μ_j)
  - the vector of parameters $\eta_{j,k}$ lies in a parameter space $E_{j,k}$.

  e.g : Univariate Gaussian distributions, Bernoulli or Multinomial distributions ...
- $\eta = (\eta_{1,1}, \dots, \eta_{K,p}) \in E_{1,1} \times \cdots \times E_{K,p}$: meta parameter of the parametric models.
- For Naive Bayes : features X_j are **independent conditionally on** Y. Consequently Maximum Likelihood Estimation corresponds to $p \times K$ independent estimation problems.

Example: Spam detection

- A given dictionary of words $\mathcal{W}$ of size p.
- p can be very large : $\sim 10^4,\ 10^5$.
- $Y(t) = 1$ if t is a spam $Y(t) = 0$ otherwise.
- For a text t (e-mail) and a word $w \in \mathcal{W}$: $X_w(t) = 1$ if $w \in t$, $X_w(t) = 0$ otherwise.
- Naive Bayes assumption with the Bernoulli variables :

$$(X_w|Y = 1) \sim \mathcal{B}(\theta_w^1) \quad \text{and} \quad (X_w|Y = 0) \sim \mathcal{B}(\theta_w^0).$$

- Learning set : $(X(1), Y(1)), \dots, (X(n), Y(n))$.

Exercice

Solve the MLE problem and give the expression of the NB :

$$\hat{\theta}_w^0 = \frac{\sum_{i=1}^n X_w(i)(1 - Y(i))}{\sum_{i=1}^n 1 - Y(i)} \text{ and } \hat{\theta}_w^1 = \frac{\sum_{i=1}^n X_w(i)Y(i)}{\sum_{i=1}^n Y(i)}$$

Outline

Outline

2 **Discriminative methods**

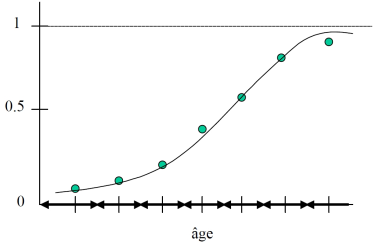
- Logistic Regression
- Nearest neighbors
- Other discriminative methods

Logistic Regression

- Discriminative approach : we define a model for the distribution of $(Y|X)$.
- Assume that $y \in \{-1, 1\}$ (not in $\{0, 1\}$).
- we consider the model

$$P(Y = 1|X = x) = \sigma(x'w + b) = \sigma(\langle x, w\rangle + b)$$

where $w \in R^d$ is a vector of model weights and $b \in \mathbb{R}$ is the intercept, and where σ is the sigmoid function :



$$z \in \mathbb{R} \mapsto \sigma(z) = \frac{1}{1 + e^{-z}} \in [0, 1]$$

Logistic Regression

- The sigmoid choice is a way to map $\mathbb{R}$ into $[0, 1]$.
- It is a modeling choice. Alternative :

$$P(Y = 1|X = x) = F(\langle x, w\rangle + b)$$

  with F the c.d.f. of a Gaussian distribution (probit model).

- The sigmoid choice has a nice interpretation on the ratio of posterior distributions :

$$\log \frac{P(Y = 1|X = x)}{P(Y = -1|X = x)} = \langle x, w\rangle + b$$

- MAP rule is linear with respect to the features x_j :

$$\begin{aligned} \hat{y}(x) = 1 \quad &\Leftrightarrow \quad P(Y = 1|X = x) > P(Y = -1|X = x) \\ &\Leftrightarrow \quad \langle x, w\rangle + b \geq 0 \end{aligned}$$

Logistic Regression : inference

- Maximum Likelihood for **conditional distributions** $(Y|X)$: maximize
$$\prod_{i=1}^{n} P(Y = y_i | X = x_i)$$
leads to find
$$(\hat{w}, \hat{b}) \in \underset{w \in \mathbb{R}^p,\, b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i(\langle x_i, w\rangle + b)}\right)$$

- It is a convex and smooth problem $\to$ convex optimization algorithms.

- Note that $(\hat{w}, \hat{b})$ is the ERM for the logistic loss
$$(u, v) \in \mathbb{R}^2 \mapsto \ell(u, v) = \log(1 + e^{-uv})$$
over the class of linear functions $\{x \mapsto \langle x, w\rangle + b, w \in \mathbb{R}^p,\ b \in \mathbb{R}\}$.

Outline

2 **Discriminative methods**
- Logistic Regression
- Nearest neighbors
- Other discriminative methods

Nearest neighbors in one slide

- Distance d on $\mathcal{X}$.
- Observations $(X_1, Y_1), \dots, (X_n, Y_n)$, $X_i \in \mathcal{X}$ and $Y_i \in \{1, \dots, L\}$.
- For some $K \in \mathbb{N}^*$, the K-NN classifier is defined by :

  $\hat{y}(x)$ = Majority vote on Y over the K-NN of x in the sample.
- It works in any metric space but ... you need to choose the metric !
- Require to choose K. Can be tuned by cross validation.

To sum up ...

- Logistic regression is very popular for classification, especially when $K = 2$. In particular, logistic regression with binary variables and Lasso penalization can be very efficient in practice.
- LDA (LQA) is useful when the classes are well separated, and Gaussian assumptions are reasonable.
- Naive Bayes is useful when p is very large.
- K-NN works in general metric spaces.

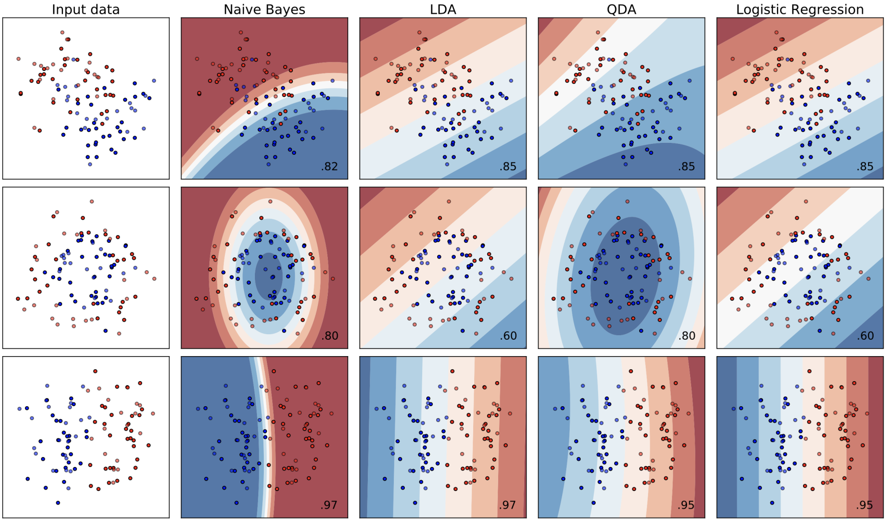
Outline

2. **Discriminative methods**
   - Logistic Regression
   - Nearest neighbors
   - Other discriminative methods

Other methods : coming soon ...

- SVMs and kernel methods
- Random Forests
- Boosting
- Neural networks and Deep Learning

A comparison of classifiers on toy datasets



Outline

Multiclass and multilabel classification

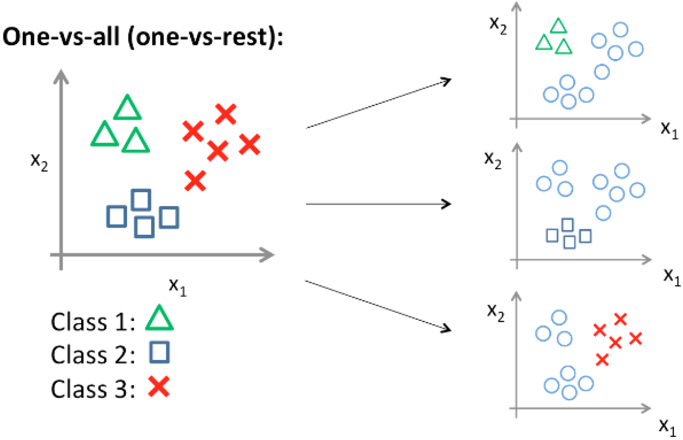
- Multiclass classification :
  - a classification task with more than two classes (e.g. several animals)
  - each sample is assigned to one and only one label: a fruit can be either an apple or a pear but not both at the same time.
- Multilabel classification :
  - each sample is assigned to a set of target labels.
  - label = properties that are not mutually exclusive, e.g. (topics of a collection of books)

(from scikit-learn doc)

One Versus All strategy

- How can we solve a multiclass or a multilabel classification using binary classifiers ?
- The **one versus all** strategy consists in fitting one (binary) classifier per class: for each classifier, the class is fitted against all the other classes.
- For multiclass classification, the final decision corresponds to the classifier with the highest score (e.g. posterior distribution i.e. MAP rule)
- Computational efficiency : only $n_{classes}$ classifiers are needed
- Interpretability : since each class is represented by one and only one classifier, it is possible to gain knowledge about the class by inspecting its corresponding classifier.
- Most commonly used strategy.

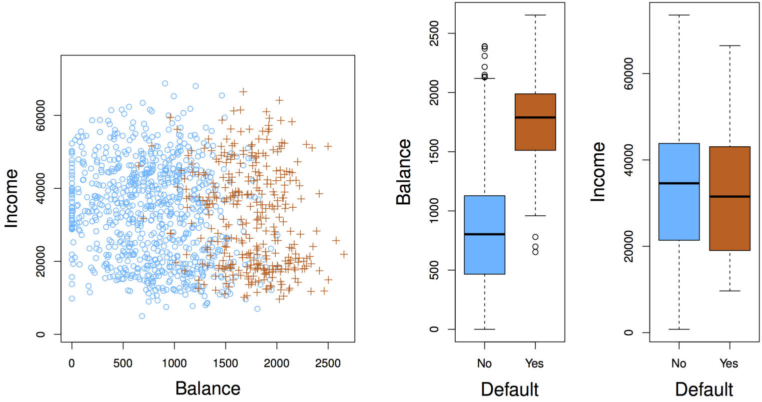
One Versus All strategy



[houxianxu.github.io]

Outline

Credit Card Default dataset (From Hastie et.al)



Gauche : revenus annuels (income) et montants mensuels crédités sur les cartes de crédit (balance de 10 000) individus.

Droite : boxplots de Balance et Income en fonction de la variable défaut de paiement (default).

Confusion Matrix

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

- In this example, *Postive* corresponds to the default status. Note that the two classes are *unbalanced*
- False Positive (FP) rate: The fraction of negative examples that are classified as positive.
- False Negative (FN) rate: The fraction of positive examples that are classified as negative.
- True Positive (TP) and True Negative (TN) : idem.

Precision, Recall, Accuracy

$$\text{Precision} = \frac{\text{TP}}{|\text{Predicted as P}|} = \frac{\text{TP}}{\text{TP + FP}}$$

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{|\text{ Real P}|}$$

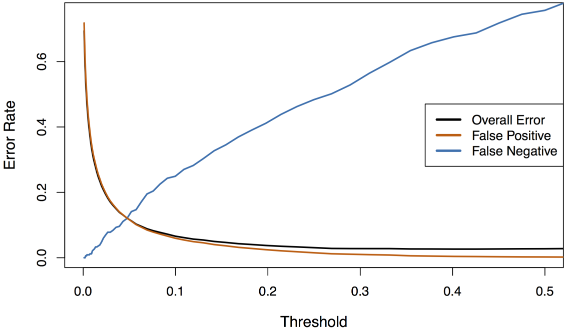
$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}}$$

Varying the threshold (two classes problem)

- For two classes, MAP rule is $\hat{y}(x) = 1$ if $\hat{P}(Y = 1|X = x) > 1/2$.
- We can change this threshold : $\hat{y}_\eta(x) = 1$ if $\hat{P}(Y = 1|X = x) > \eta$ for
  - Improving the performances (MAP rule is a bayes rule only for the true posterior distribution),
  - Giving an advantage to a class.

Varying the threshold (two classes problem)

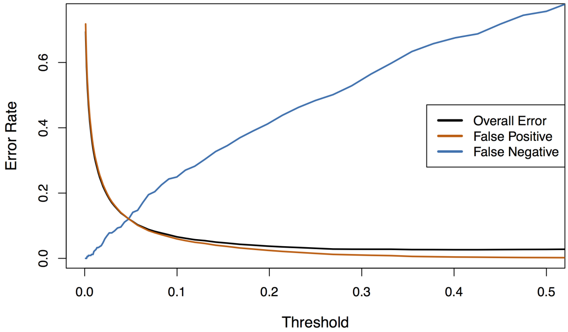
- For two classes, MAP rule is $\hat{y}(x) = 1$ if $\hat{P}(Y = 1|X = x) > 1/2$.
- We can change this threshold : $\hat{y}_\eta(x) = 1$ if $\hat{P}(Y = 1|X = x) > \eta$



[Hastie et.al]

Varying the threshold (two classes problem)

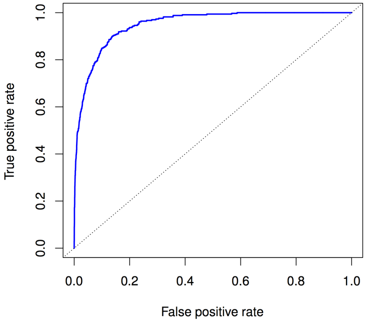
- For two classes, MAP rule is $\hat{y}(x) = 1$ if $\hat{P}(Y = 1|X = x) > 1/2$.
- We can change this threshold : $\hat{y}_\eta(x) = 1$ if $\hat{P}(Y = 1|X = x) > \eta$



[Hastie et.al]

Question: Why is the Overall Error very close to the False Positive rate in this example ?

ROC Curve (Receiver Operating Characteristic)



[Hastie et.al]

- Each point of the curve has coordinates $(\mathrm{FP}_\eta, \mathrm{TP}_\eta)$, computed from the classification rule with threshold η.
- The curve is non decreasing.
- Classification rule with zero error corresponds to the point $(0, 1)$.
- AUC score is the Area Under the ROC Curve.

Unbalanced classes in classification

- Unbalanced data refers to situation where the classes are not represented equally.
- E.g. medical dataset : 5% disease / 95% healthy.
- In these situations, classification rules tend to predict only the majority class : accuracy is not enough !
- Solution 1: rebalance the metric. Empirical risk with rebalanced loss :

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{n_1}\ell(y_i,\hat{y}_i)\mathbb{1}_{y_i=1} + \frac{1}{n_{-1}}\ell(y_i,\hat{y}_i)\mathbb{1}_{y_i=-1}$$

- Solution 2 : oversampling methods : creates copy data or create synthetic samples (SMOTE) from the minor class.