Modèles probabilistes de classification - LDA, QDA & régression logistique

Master parcours SSD - UE Apprentissage Statistique I

Pierre Mahé - bioMérieux & Université de Grenoble-Alpes

Plan

1. Introduction
   - théorie de la décision statistique & k-ppv
2. Modèles génératifs
   - LDA, QDA
3. Modèles discriminatifs
   - régression logistique
4. En pratique

Introduction

Rappel - Apprentissage supervisé, formalisation

Données d'entrée : échantillon $\{(x_i, y_i)\}_{i=1,\dots,n} \in \mathcal{X} \times \mathcal{Y}$.

Objectif : apprendre une fonction $f : \mathcal{X} \to \mathcal{Y}$ permettant de prédire la réponse associée à une nouvelle observation.

Rappel - Apprentissage supervisé, formalisation

Données d'entrée : échantillon $\{(x_i, y_i)\}_{i=1,\dots,n} \in \mathcal{X} \times \mathcal{Y}$.

Objectif : apprendre une fonction $f : \mathcal{X} \rightarrow \mathcal{Y}$ permettant de prédire la réponse associée à une nouvelle observation.

Critère : une fonction de perte L (pour "loss") mesurant l'erreur entre y et $f(x)$.

Typiquement :

- l'erreur quadratique pour la régression :

$$L(y, f(x)) = (y - f(x))^2$$

- le coût 0/1 pour la classification :

$$L(y, f(x)) = \mathbb{1}(y \neq f(x))$$

Rappel - Apprentissage supervisé, formalisation

Cadre probabiliste : on considère que nos observations (x_i, y_i) sont des variables aléatoires régies par une loi jointe $P(X, Y)$.

$\Rightarrow$ L'objectif de l'apprentissage supervisé est donc de trouver la fonction f minimisant l'espérance de la fonction de perte :

$$R(f) = E_{X,Y}\left[L(Y, f(X))\right],$$

à partir d'un échantillon $\{(x_i, y_i)\}$, $i = 1, ..., n$.

$R(f)$ est appelée le risque (ou la perte) de la fonction f.

Théorie de la décision statistique

La quête du Graal : quel serait notre meilleur choix si on connaissait la loi $P(X, Y)$?

Théorie de la décision statistique

La quête du Graal : quel serait notre meilleur choix si on connaissait la loi $P(X, Y)$?

Cas de la régression et de la perte quadratique :

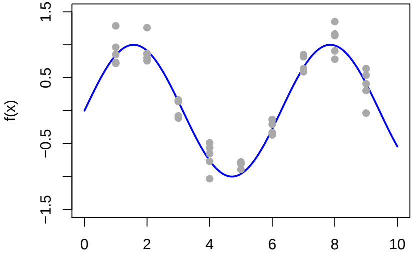
$$\begin{aligned} R(f) &= E_{X,Y}\left[L(Y, f(X))\right] \\ &= E_{X,Y}\left[(Y - f(X))^2\right] \end{aligned}$$

$\Rightarrow$ meilleure solution : $f(x) = E[Y|X = x]$

- la valeur moyenne que peut prendre Y sachant X
- la "regression function"
- NB : valable pour la perte quadratique
  - $f(x) = \text{median}(Y|X = x)$ si $L(Y, f(X)) = |Y - f(X)|$

Théorie de la décision statistique - régression

Illustration :



- $P(X, Y)$: vraie relation entre X et Y non déterministe
  - ici, $Y = f(X) + \epsilon$, $\epsilon \rightarrow \mathcal{N}(0, \sigma^2)$
  - en bleue : vraie fonction, en gris : réalisations bruitées
- On doit prendre une décision
- On minimise la perte quadratique en prenant $E[Y|X]$
  - l'espérance des points gris pour chaque valeur de x

Théorie de la décision statistique - régression

Régression et perte quadratique : démonstration

$$
\begin{aligned}
R(f) &= E_{X,Y}\left[L(Y, f(X))\right] \\
&= E_{X,Y}\left[(Y - f(X))^2\right]
\end{aligned}
$$

Théorie de la décision statistique - régression

Régression et perte quadratique : démonstration

$$
\begin{aligned}
R(f) &= E_{X,Y}\left[L(Y, f(X))\right] \\
&= E_{X,Y}\left[(Y - f(X))^2\right]
\end{aligned}
$$

$\Rightarrow$ on conditionne sur X : $E_{Y,X}(.) = E_X E_{Y|X}(.)$

$$
R(f) = E_X E_{Y|X}\left[(Y - f(X))^2 | X\right]
$$

Théorie de la décision statistique - régression

Régression et perte quadratique : démonstration

$$\begin{aligned} R(f) &= E_{X,Y}\left[L(Y, f(X))\right] \\ &= E_{X,Y}\left[(Y - f(X))^2\right] \end{aligned}$$

$\Rightarrow$ on conditionne sur X : $E_{Y,X}(.) = E_X E_{Y|X}(.)$

$$R(f) = E_X E_{Y|X}\left[(Y - f(X))^2 | X\right]$$

$\Rightarrow$ on minimise pour chaque valeur x prise par X :

$$\begin{aligned} f(x) &= \arg\min_c E_{Y|X}\left[(Y - c)^2 | X = x\right] \\ &= E[Y|X = x] \end{aligned}$$

Théorie de la décision statistique - classification

La quête du Graal : quel serait notre meilleur choix si on connaissait la loi $P(X, Y)$?

Cas de la classification et de la perte 0/1 :

$$
\begin{aligned}
R(f) &= E_{X,Y}\left[L(Y, f(X))\right] \\
&= E_{X,Y}\left[\mathbb{1}(Y \neq f(X))\right]
\end{aligned}
$$

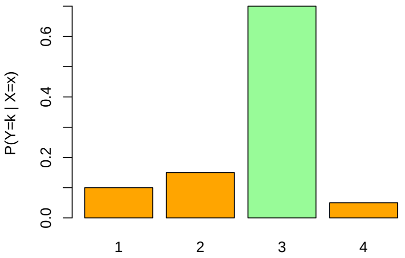
$\Rightarrow$ meilleure solution : $f(x) = \arg\max_{k=1,\dots,K} P(Y = C_k | X = x)$

- la classe la plus vraisemblable sachant X
- le classifieur de Bayes
- NB : valable pour la perte / le coût 0/1
  - se généralise pour des coûts arbitraires $L(C_i, C_j)$

Théorie de la décision statistique - classification

Illustration : $P(Y = k|X = x)$

- pour une valeur x donnée et $K = 4$ catégories



- $P(X, Y)$: vraie relation entre X et Y non déterministe
- On doit prendre une décision
- Erreur : somme des probabilités des choix qu'on rejette
- Minimisée si on choisit la probabilité la plus élevée

Théorie de la décision statistique - classification

Classification et perte 0/1 : démonstration

$$R(f) = E_{X,Y}\left[L(Y, f(X))\right] = E_{X,Y}\left[\mathbb{1}(Y \neq f(X)\right]$$

Théorie de la décision statistique - classification

Classification et perte 0/1 : démonstration

$$R(f) = E_{X,Y}\left[L(Y, f(X))\right] = E_{X,Y}\left[\mathbb{1}(Y \neq f(X)\right]$$

$\Rightarrow$ on conditionne sur X : $E_{Y,X}(.) = E_X E_{Y|X}(.)$

$$\begin{aligned} R(f) &= E_X E_{Y|X}\left[\mathbb{1}(Y \neq f(X)|X\right] \\ &= E_X \sum_{k=1}^{K} \mathbb{1}(C_k \neq f(X)) P(Y = C_k|X) \end{aligned}$$

Théorie de la décision statistique - classification

Classification et perte 0/1 : démonstration

$$R(f) = E_{X,Y}\left[L(Y, f(X))\right] = E_{X,Y}\left[\mathbb{1}(Y \neq f(X))\right]$$

$\Rightarrow$ on conditionne sur X : $E_{Y,X}(.) = E_X E_{Y|X}(.)$

$$\begin{aligned} R(f) &= E_X E_{Y|X}\left[\mathbb{1}(Y \neq f(X)|X\right] \\ &= E_X \sum_{k=1}^{K} \mathbb{1}(C_k \neq f(X)) P(Y = C_k|X) \end{aligned}$$

$\Rightarrow$ on minimise pour chaque valeur x prise par X :

$$\begin{aligned} f(x) &= \arg\min_c \sum_{k=1}^{K} \mathbb{1}(C_k \neq c) P(Y = C_k|X = x) \\ &= \arg\max_{k=1,\dots,K} P(Y = C_k|X = x) \end{aligned}$$

Théorie de la décision statistique - classification

Dans le cas de la classification binaire on doit choisir entre :

1. $Y = 1$ selon $P(Y = 1|X = x)$
2. $Y = 0$ selon $P(Y = 0|X = x) = 1 - P(Y = 1|X = x)$

Le classifieur de Bayes

$$f(x) = \arg \max_{k=1,\ldots,K} P(Y = C_k|X = x)$$

peut s'écrire comme :

$$f(x) = \begin{cases} 1 & \text{si } \; P(Y = 1|X = x) \geq 0.5 \\ 0 & \text{sinon} \end{cases}$$

Théorie de la décision statistique

En pratique...

Théorie de la décision statistique

En pratique... on ne connaît pas $P(X, Y)$! !

Théorie de la décision statistique

En pratique... on ne connaît pas $P(X, Y)$!!

Intérêt de cette démarche (théorique) :

- concevoir des algorithmes
  - définition d'estimateurs + stratégies d'estimation
- analyser des algorithmes
  - e.g., se comparer au classifieur de Bayes par simulations
  - étudier performance en fonction de n

Exemple des k-PPV

k-PPV pour la régression :

$$\hat{f}(x) = \text{Average}(y_i | i \in N_k(x))$$

$\Rightarrow$ approxime directement la regression function $E[Y|X]$

Exemple des k-PPV

k-PPV pour la régression :

$$\hat{f}(x) = \text{Average}(y_i | i \in N_k(x))$$

$\Rightarrow$ approxime directement la regression function $E[Y|X]$

Nature de l'approximation :

1. espérance $\rightarrow$ moyenne empirique
2. point $\rightarrow$ voisinage dans le conditionnement

$\Rightarrow$ convergence asymptotique

- $n, k \rightarrow +\infty$; $k/n \rightarrow 0$

Exemple des k-PPV

k-PPV pour la classification :

$$\begin{aligned}\hat{f}(x) &= \text{Majority}(y_i | i \in N_k(x)) \\ &= \arg\max_{l=1,\ldots,K} \tilde{P}_l = \frac{1}{k} \sum_{i \in N_k(x)} \mathbb{1}(y_i = l)\end{aligned}$$

$\Rightarrow$ approxime directement le classifieur de Bayes :

$$\arg\max_{l=1,\ldots,K} P(Y = C_l | X = x)$$

Exemple des k-PPV

k-PPV pour la classification :

$$\begin{aligned}\hat{f}(x) &= \text{Majority}(y_i | i \in N_k(x)) \\ &= \arg\max_{l=1,\dots,K} \tilde{P}_l = \frac{1}{k} \sum_{i \in N_k(x)} \mathbb{1}(y_i = l)\end{aligned}$$

$\Rightarrow$ approxime directement le classifieur de Bayes :

$$\arg\max_{l=1,\dots,K} P(Y = C_l | X = x)$$

Nature de l'approximation :

1. probabilité $\rightarrow$ proportion empirique
2. point $\rightarrow$ voisinage dans le conditionnement

$\Rightarrow$ convergence asymptotique

- $n, k \rightarrow +\infty$; $k/n \rightarrow 0$

Exemple des k-PPV

En dépit de sa simplicité, l'algorithme des k-PPV :

1. approxime les bonnes fonctions
   - regression function & classifieur de Bayes
2. possède des propriétés de convergence

$\Rightarrow$ pourquoi chercher plus loin ?

Exemple des k-PPV

En dépit de sa simplicité, l'algorithme des k-PPV :

1. approxime les bonnes fonctions
   - regression function & classifieur de Bayes
2. possède des propriétés de convergence

$\Rightarrow$ pourquoi chercher plus loin ?

Car la convergence est asymptotique :

- $n, k \to +\infty$; $k/n \to 0$

$\Rightarrow$ en pratique on dispose d'un nombre d'observations limité

- k-PPV rarement optimal à n fixé

Exemple des k-PPV

En dépit de sa simplicité, l'algorithme des k-PPV :

1. approxime les bonnes fonctions
    - regression function & classifieur de Bayes
2. possède des propriétés de convergence

$\Rightarrow$ pourquoi chercher plus loin ?

Car la convergence est asymptotique :

- $n, k \to +\infty$; $k/n \to 0$

$\Rightarrow$ en pratique on dispose d'un nombre d'observations limité

- k-PPV rarement optimal à n fixé

(de plus, ça se complique en haute dimension)

- "fléau de la dimension"

Modèles probabilistes de classification

Ce cours : cadre de la classification

Graal = classifieur de Bayes :

$$f(x) = \arg \max_{k=1,\dots,K} P(Y = C_k | X = x)$$

Modèles probabilistes de classification : estimer $P(Y|X)$

$\Rightarrow$ Deux approches :

- générative : modélise $P(X|Y)$
- discriminative : estime directement $P(Y|X)$

Modèles génératifs

Modèles génératifs

Objectif : estimer $P(Y|X)$

Approche générative :

1. définir un modèle pour $P(X|Y)$
   - densité conditionnelle des données au sein des classes
2. appliquer la loi de Bayes pour en déduire $P(Y|X)$:

$$\begin{aligned} P(Y = C_k|X = x) &= \frac{P(X = x, Y = C_k)}{P(X = x)} \\ &= \frac{P(X = x|Y = C_k)P(Y = C_k)}{P(X = x)} \end{aligned}$$

$\Rightarrow$ on "inverse" $P(X|Y)$ en $P(Y|X)$

- X permet de "mettre à jour" $P(Y)$ en $P(Y|X)$

Intermède

Terminologie :

- $P(Y)$ = loi a priori
- $P(X, Y)$ = loi jointe
- $P(X)$ = loi marginale
- $P(X|Y)$ = loi conditionnelle
- $P(Y|X)$ = loi a posteriori

Illustration : X=poids ; Y = sexe

	X = 50	X= 60	X = 70	X = 80	X = 90	Total
Y = ♂	1	3	15	17	12	48
Y = ♀	5	13	5	3	2	28
Total	6	16	20	20	14	76

Intermède

Illustration : X=poids ; Y = sexe

	X = 50	X= 60	X = 70	X = 80	X = 90	Total
Y = ♂	1	3	15	17	12	48
Y = ♀	5	13	5	3	2	28
Total	6	16	20	20	14	76

Intermède

Illustration : X=poids ; Y = sexe

	X = 50	X= 60	X = 70	X = 80	X = 90	Total
Y = ♂	1	3	15	17	12	48
Y = ♀	5	13	5	3	2	28
Total	6	16	20	20	14	76

- loi a priori : $P(Y = ♀) = 28/76$

Intermède

Illustration : X=poids ; Y = sexe

	X = 50	X= 60	X = 70	X = 80	X = 90	Total
Y = ♂	1	3	15	17	12	48
Y = ♀	5	13	5	3	2	28
Total	6	16	20	20	14	76

- loi a priori : $P(Y = ♀) = 28/76$
- loi jointe : $P(X = 80, Y = ♂) = 17/76$

Intermède

Illustration : X=poids ; Y = sexe

	X = 50	X= 60	X = 70	X = 80	X = 90	Total
Y = ♂	1	3	15	17	12	48
Y = ♀	5	13	5	3	2	28
Total	6	16	20	20	14	76

- loi a priori : $P(Y = ♀) = 28/76$
- loi jointe : $P(X = 80, Y = ♂) = 17/76$
- loi marginale : $P(X = 80) = 20/76$
  - $P(X = 80) = P(X = 80, Y = ♂) + P(X = 80, Y = ♀)$

Intermède

Illustration : X=poids ; Y = sexe

	X = 50	X= 60	X = 70	X = 80	X = 90	Total
Y = ♂	1	3	15	17	12	48
Y = ♀	5	13	5	3	2	28
Total	6	16	20	20	14	76

- loi a priori : $P(Y = ♀) = 28/76$
- loi jointe : $P(X = 80, Y = ♂) = 17/76$
- loi marginale : $P(X = 80) = 20/76$
  - $P(X = 80) = P(X = 80, Y = ♂) + P(X = 80, Y = ♀)$
- loi conditionnelle : $P(X = 80|Y = ♂) = 17/48$

Intermède

Illustration : X=poids ; Y = sexe

	X = 50	X= 60	X = 70	X = 80	X = 90	Total
Y = ♂	1	3	15	17	12	48
Y = ♀	5	13	5	3	2	28
Total	6	16	20	20	14	76

- loi a priori : $P(Y = ♀) = 28/76$
- loi jointe : $P(X = 80, Y = ♂) = 17/76$
- loi marginale : $P(X = 80) = 20/76$
  - $P(X = 80) = P(X = 80, Y = ♂) + P(X = 80, Y = ♀)$
- loi conditionnelle : $P(X = 80|Y = ♂) = 17/48$
- loi a posteriori : $P(Y = ♂|X = 80) = 17/20$

Modèles génératifs

Grâce à la loi de Bayes on a donc :

$$
\begin{aligned}
P(Y = C_k | X = x) &= \frac{P(X = x, Y = C_k)}{P(X = x)} \\
&= \frac{P(X = x | Y = C_k) P(Y = C_k)}{P(X = x)} \\
&= \frac{P(X = x | Y = C_k) P(Y = C_k)}{\sum_{i=1}^{K} P(X = x | Y = C_i) P(Y = C_i)}
\end{aligned}
$$

Modèles génératifs

Grâce à la loi de Bayes on a donc :

$$
\begin{aligned}
P(Y = C_k | X = x) &= \frac{P(X = x, Y = C_k)}{P(X = x)} \\
&= \frac{P(X = x | Y = C_k) P(Y = C_k)}{P(X = x)} \\
&= \frac{P(X = x | Y = C_k) P(Y = C_k)}{\sum_{i=1}^{K} P(X = x | Y = C_i) P(Y = C_i)}
\end{aligned}
$$

$\Rightarrow$ on notera :

$$
P(Y = C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i}
$$

avec $f_k(x) = P(X = x | Y = C_k)$ et $\pi_k = P(Y = C_k)$.

Modèles génératifs

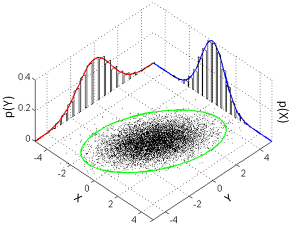
LA question : comment choisir $f_k(x)$?

Modèles génératifs

LA question : comment choisir $f_k(x)$?

Une première possibilité : lois normales (multivariées) :

$$
\begin{aligned}
f_k(x) &= \mathcal{MN}(x; \mu_k, \Sigma_k) \\
&= \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\Big(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\Big)
\end{aligned}
$$



$\Rightarrow$ approches LDA et QDA.

Modèles génératifs - LDA

LDA - définition

LDA - Linear Discriminant Analysis :

LDA - définition

LDA - Linear Discriminant Analysis :

1. modèle génératif :

$$P(Y = C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i}$$

LDA - définition

LDA - Linear Discriminant Analysis :

1. modèle génératif :

$$P(Y = C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i}$$

2. basé sur des lois normales multivariées :

$$f_k(x) = \frac{1}{\sqrt{(2\pi)^p|\Sigma_k|}} \exp\Big(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\Big)$$

LDA - définition

LDA - Linear Discriminant Analysis :

1. modèle génératif :

$$P(Y = C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i}$$

2. basé sur des lois normales multivariées :

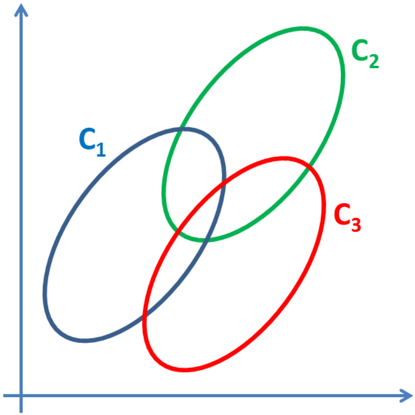
$$f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\big(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\big)$$

3. avec la même matrice de covariance par classe

$$\Sigma_k = \Sigma, \ \forall\, k \in \{1, ..., K\}.$$

LDA - Illustration

Modèle considéré :



$\Rightarrow$ toutes les classes ont la même matrice de covariance.

LDA - règle de décision

Motivation = classifieur de Bayes :

$$f(x) = \arg \max_{k=1,\dots,K} P(Y = C_k | X = x)$$

LDA - règle de décision

Motivation = classifieur de Bayes :

$$f(x) = \arg \max_{k=1,\dots,K} P(Y = C_k | X = x)$$

Approche générative :

$$P(Y = C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i} \tag{1}$$

LDA - règle de décision

Motivation = classifieur de Bayes :

$$f(x) = \arg \max_{k=1,\ldots,K} P(Y = C_k | X = x)$$

Approche générative :

$$P(Y = C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^K f_i(x)\pi_i} \qquad (1)$$

$\Rightarrow$ règle de décision :

$$\hat{f}(x) = \arg \max_{k=1,\ldots,K} \frac{f_k(x)\pi_k}{\sum_{i=1}^K f_i(x)\pi_i}$$
$$= \arg \max_{k=1,\ldots,K} f_k(x)\pi_k$$

(dénominateur de (1) = $P(X) \sim$ normalisation)

LDA - règle de décision

Règle de décision :

$$\hat{f}(x) = \arg\max_{k=1,\ldots,K} f_k(x)\pi_k = \arg\max_{k=1,\ldots,K} \ln f_k(x) + \ln \pi_k$$

LDA - règle de décision

Règle de décision :

$$\hat{f}(x) = \arg\max_{k=1,\dots,K} f_k(x)\pi_k = \arg\max_{k=1,\dots,K} \ln f_k(x) + \ln \pi_k$$

Or :

$$f_k(x) = \frac{1}{\sqrt{(2\pi)^p|\Sigma_k|}} \exp\Big(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\Big)$$

$$\begin{aligned}\ln f_k(x) &= -\ln\sqrt{(2\pi)^p} - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) \\ &= C - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}x^T\Sigma_k^{-1}x - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + x^T\Sigma_k^{-1}\mu_k\end{aligned}$$

LDA - règle de décision

Règle de décision :

$$\hat{f}(x) = \arg\max_{k=1,\dots,K} f_k(x)\pi_k = \arg\max_{k=1,\dots,K} \ln f_k(x) + \ln \pi_k$$

Or :

$$f_k(x) = \frac{1}{\sqrt{(2\pi)^p|\Sigma_k|}} \exp\Big(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\Big)$$

$$\begin{aligned}\ln f_k(x) &= -\ln\sqrt{(2\pi)^p} - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\\ &= C - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}x^T\Sigma_k^{-1}x - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + x^T\Sigma_k^{-1}\mu_k\end{aligned}$$

On a donc :

$$\Rightarrow \hat{f}(x) = \arg\max_{k=1,\dots,K}\Big\{\ln\pi_k - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}x^T\Sigma_k^{-1}x - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + x^T\Sigma_k^{-1}\mu_k\Big\}$$

LDA - règle de décision

Forme générale de la règle de décision :

$$\hat{f}(x) = \arg \max_{k=1,\dots,K} \Big\{ \ln \pi_k - \frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} x^T \Sigma_k^{-1} x - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k + x^T \Sigma_k^{-1} \mu_k \Big\}$$

1. $\ln |\Sigma_k|$ et $\frac{1}{2} x^T \Sigma_k^{-1} x$ ne dépendent plus de k

LDA - règle de décision

Forme générale de la règle de décision :

$$\hat{f}(x) = \arg \max_{k=1,\ldots,K} \Big\{ \ln \pi_k - \frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} x^T \Sigma_k^{-1} x - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k + x^T \Sigma_k^{-1} \mu_k \Big\}$$

Si $\Sigma_k = \Sigma$ (modèle LDA) alors [1] :

$$\begin{aligned} \hat{f}(x) &= \arg \max_{k=1,\ldots,K} \left\{ \ln \pi_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + x^T \Sigma^{-1} \mu_k \right\} \\ &= \arg \max_{k=1,\ldots,K} \delta_k(x) \end{aligned}$$

$\Rightarrow$ où $\delta_k(x)$ est une fonction linéaire de x.

1. $\ln |\Sigma_k|$ et $\frac{1}{2} x^T \Sigma_k^{-1} x$ ne dépendent plus de k

LDA - règle de décision

Règle de décision linéaire ?

$$\hat{f}(x) = \arg\max_{k=1,\dots,K} \delta_k(x)$$

où :

$$\begin{aligned}\delta_k(x) &= \ln \pi_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + x^T \Sigma^{-1} \mu_k \\ &= a_k + x^T b_k,\end{aligned}$$

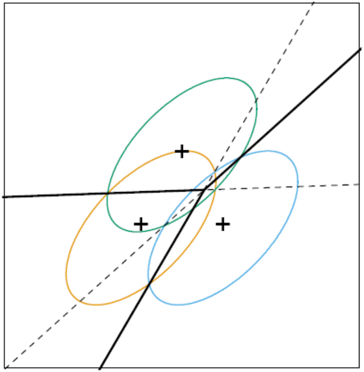
avec :

- $a_k = \ln \pi_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k \;\Rightarrow\; a_k \in \mathbb{R}$
- $b_k = \Sigma^{-1} \mu_k \;\Rightarrow\; b_k \in \mathbb{R}^p$

$\Rightarrow$ fonction définie à partir des paramètres de la loi normale

LDA - règle de décision

Conséquence :



⇒ les frontières entre les classes sont linéaires.

- NB : frontière entre les classes k et l : $\{x : \delta_k(x) = \delta_l(x)\}$
- on vérifie facilement que c'est une droite

LDA - règle de décision

Frontières linéaires ?

LDA - règle de décision

Frontières linéaires ?

Si :

$$\hat{f}(x) = \arg\max_{k=1,\dots,K} \delta_k(x)$$

alors la frontière entre les classes k et l est :

$$\begin{aligned}
\left\{ x \right. &: \delta_k(x) = \delta_l(x) \left. \right\} \\
&: \delta_k(x) - \delta_l(x) = 0 \left. \right\} \\
&: (a_k + x^T b_k) - (a_l + x^T b_l) = 0 \left. \right\} \\
&: (a_k - a_l) + x^T (b_k - b_l) = 0 \left. \right\}
\end{aligned}$$

$\Rightarrow$ une fonction linéaire (une droite, un plan ou un hyperplan)

LDA - implémentation

En pratique, il faut estimer $\{\pi_k, \mu_k\}_{k=1,\dots,K}$ et Σ.

LDA - implémentation

En pratique, il faut estimer $\{\pi_k, \mu_k\}_{k=1,\ldots,K}$ et Σ.

On considère les estimateurs standards :

- $\hat{\pi}_k = \frac{n_k}{n}$
  - les proportions relatives des classes
- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$
  - la moyenne empirique au sein de chaque classe
- $\Sigma = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$
  - la covariance empirique "poolée"
  - $\sim$ moyenne des covariances par classe

(NB : ici pas de $P(Z|X)$ dans l'estimation, on connaît les catégories)

Modèles génératifs - QDA

QDA - définition

QDA - Quadratic Discriminant Analysis :

QDA - définition

QDA - Quadratic Discriminant Analysis :

1. modèle génératif :

$$P(Y = C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i}$$

QDA - définition

QDA - Quadratic Discriminant Analysis :

1. modèle génératif :

$$P(Y = C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i}$$

2. basé sur des lois normales multivariées :

$$f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) \right)$$

QDA - définition

QDA - Quadratic Discriminant Analysis :

1. modèle génératif :

$$P(Y = C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i}$$

2. basé sur des lois normales multivariées :

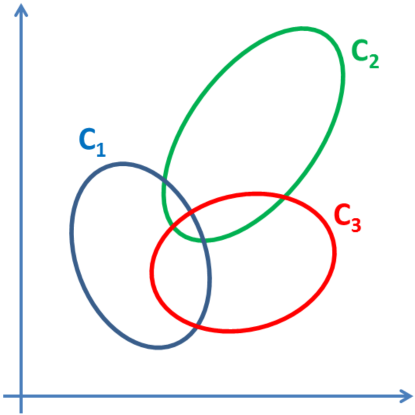
$$f_k(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_k|}} \exp\Big(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\Big)$$

3. **où chaque classe à sa propre matrice de covariance**

$$\Sigma_k \neq \Sigma, \;\; \forall\, k \in \{1, ..., K\}.$$

QDA - Illustration

Modèle considéré :



$\Rightarrow$ chaque classe à sa propre matrice de covariance.

QDA - règle de décision

(rappel) Règle de décision - forme générale :

$$\begin{aligned}
\hat{f}(x) &= \arg\max_{k=1,\ldots,K} f_k(x)\pi_k \\
&= \arg\max_{k=1,\ldots,K} \ln f_k(x) + \ln \pi_k \\
&= \arg\max_{k=1,\ldots,K} \Big\{ \ln \pi_k - \frac{1}{2}\ln|\Sigma_k| \\
&\qquad\qquad - \frac{1}{2}x^T\Sigma_k^{-1}x - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + x^T\Sigma_k^{-1}\mu_k \Big\}
\end{aligned}$$

QDA - règle de décision

(rappel) Règle de décision - forme générale :

$$\begin{aligned}\hat{f}(x) &= \arg\max_{k=1,\dots,K} f_k(x)\pi_k \\ &= \arg\max_{k=1,\dots,K} \ln f_k(x) + \ln \pi_k \\ &= \arg\max_{k=1,\dots,K} \Big\{ \ln \pi_k - \frac{1}{2}\ln|\Sigma_k| \\ &\qquad - \frac{1}{2}x^T\Sigma_k^{-1}x - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + x^T\Sigma_k^{-1}\mu_k \Big\}\end{aligned}$$

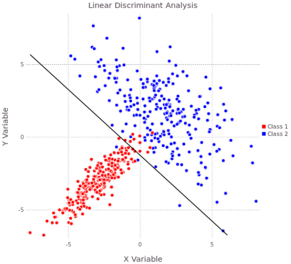
$\Rightarrow$ si $\Sigma_k \neq \Sigma$ (modèle QDA) ne se simplifie pas :

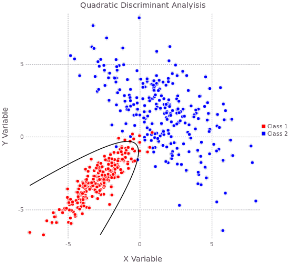
$$\hat{f}(x) = \arg\max_{k=1,\dots,K} \delta_k(x),$$

où $\delta_k(x)$ est une fonction quadratique de x.

QDA - règle de décision

Conséquence[2] :





⇒ les frontières entres classes sont non-linéaires

2. Images de http://discriminantanalysis.readthedocs.io/

QDA - implémentation

En pratique, il faut (toujours) estimer $\{\pi_k, \mu_k\}_{k=1,\ldots,K}$ et Σ.

On conserve :

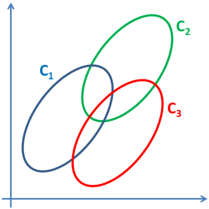
- $\hat{\pi}_k = \frac{n_k}{n}$
  - les proportions relatives des classes
- $\hat{\mu}_k = \dfrac{1}{n_k} \displaystyle\sum_{i:y_i=k} x_i$
  - la moyenne empirique au sein de chaque classe

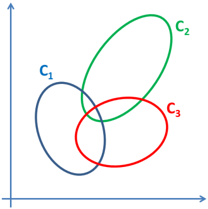
On estime une matrice de covariance par classe :

- $\Sigma_K = \dfrac{1}{n_k - 1} \displaystyle\sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$

LDA ou QDA ?

Pourquoi hésiter entre LDA et QDA ?





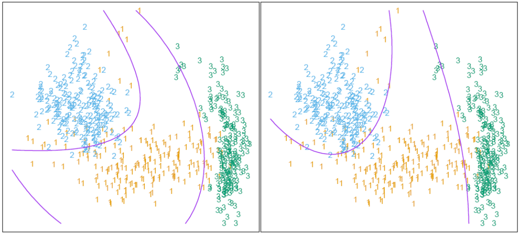
- LDA : hypothèse très contraignante (et peu réaliste)
- QDA : beaucoup plus de paramètres à estimer
  - $\frac{p(p+1)}{2}$ paramètres par Σ_k

$\Rightarrow$ dépend du jeu de données
$\Rightarrow$ compromis complexité (modèle) / stabilité (estimation)

LDA ou QDA ?

Solution intermédiaire : LDA sur données "transformées" [3]

- Gauche : LDA sur $(X_1, X_2, X_1X_2, X_1^2, X_2^2)$
- Droite : QDA sur (X_1, X_2)



- séparations non-linéaires
- contrôle du nombre de paramètres à estimer

3. example tiré de Hastie et al. (2001).

Modèles discriminatifs

Modèles discriminatifs

Objectif : estimer $P(Y|X)$

Modèles discriminatifs

Objectif : estimer $P(Y|X)$

Approche générative :

1. définir un modèle pour $P(X|Y)$
2. l'inverser en $P(Y|X)$ grâce à la loi de Bayes :

$$P(Y = C_k|X = x) = \frac{P(X = x|Y = C_k)P(Y = C_k)}{P(X = x)}$$

Modèles discriminatifs

Objectif : estimer $P(Y|X)$

Approche générative :

1. définir un modèle pour $P(X|Y)$
2. l'inverser en $P(Y|X)$ grâce à la loi de Bayes :

$$P(Y = C_k|X = x) = \frac{P(X = x|Y = C_k)P(Y = C_k)}{P(X = x)}$$

Avantages / inconvénients :

- + : LDA / QDA simple à mettre en oeuvre
- + : $P(X|Y)$ pour la détection de points abérrants
- - : sensible à l'adéquation modèle / données
- - : beaucoup de paramètres à estimer

Modèles discriminatifs

Approche générative : estimer directement $P(Y|X)$

Modèles discriminatifs

Approche générative : estimer directement $P(Y|X)$

Motivation : s'attacher à la fonction de classification

- (en soi connaître $P(X|Y)$ est accessoire)

Modèles discriminatifs

Approche générative : estimer directement $P(Y|X)$

Motivation : s'attacher à la fonction de classification

- (en soi connaître $P(X|Y)$ est accessoire)

Objectif : obtenir de meilleures performances prédictives

1. modèle optimisé en ce sens
2. en pratique, permet de limiter le nombre de paramètres

Modèles discriminatifs

Approche générative : estimer directement $P(Y|X)$

Motivation : s'attacher à la fonction de classification

- (en soi connaître $P(X|Y)$ est accessoire)

Objectif : obtenir de meilleures performances prédictives

1. modèle optimisé en ce sens
2. en pratique, permet de limiter le nombre de paramètres

$\Rightarrow$ modèle incontournable : la régression logistique

Régression logistique - formalisation

On considèrera la classification binaire : $Y \in \{C_1, C_2\}$

Régression logistique - formalisation

On considèrera la classification binaire : $Y \in \{C_1, C_2\}$

Le classifieur de Bayes :

$$f(x) = \arg\max_{k=1,\dots,K} P(Y = C_k | X = x)$$

peut s'écrire

$$f(x) = \begin{cases} C_1 & \text{si} \quad \frac{P(Y=C_1|X=x)}{P(Y=C_2|X=x)} > 1 \\ C_2 & \text{sinon} \end{cases}$$

Régression logistique - formalisation

On considèrera la classification binaire : $Y \in \{C_1, C_2\}$

Le classifieur de Bayes :

$$f(x) = \arg\max_{k=1,\ldots,K} P(Y = C_k | X = x)$$

peut s'écrire

$$f(x) = \begin{cases} C_1 & \text{si} \quad \frac{P(Y=C_1|X=x)}{P(Y=C_2|X=x)} > 1 \\ C_2 & \text{sinon} \end{cases}$$

$\Rightarrow$ la quantité :

$$\frac{P(Y = C_1|X = x)}{P(Y = C_2|X = x)} = \frac{P(Y = C_1|X = x)}{1 - P(Y = C_1|X = x)}$$

est appelée "rapport des cotes", ou odds-ratio.

Régression logistique - formalisation

Le "rapport des cotes", ou odds-ratio :

$$\frac{P(Y = C_1|X = x)}{P(Y = C_2|X = x)} = \frac{P(Y = C_1|X = x)}{1 - P(Y = C_1|X = x)}$$

Régression logistique - formalisation

Le "rapport des cotes", ou odds-ratio :

$$\frac{P(Y = C_1|X = x)}{P(Y = C_2|X = x)} = \frac{P(Y = C_1|X = x)}{1 - P(Y = C_1|X = x)}$$

1. est compris entre 0 et $+\infty$
   - 0 si $P(C_1|x) = 0$; $+\infty$ si $P(C_1|x) = 1$

Régression logistique - formalisation

Le "rapport des cotes", ou odds-ratio :

$$\frac{P(Y = C_1|X = x)}{P(Y = C_2|X = x)} = \frac{P(Y = C_1|X = x)}{1 - P(Y = C_1|X = x)}$$

1. est compris entre 0 et $+\infty$
   - 0 si $P(C_1|x) = 0$; $+\infty$ si $P(C_1|x) = 1$
2. définit la fonction de décision
   - $f(x) = C_1$ si > 1

Régression logistique - formalisation

Le "rapport des cotes", ou odds-ratio :

$$\frac{P(Y = C_1|X = x)}{P(Y = C_2|X = x)} = \frac{P(Y = C_1|X = x)}{1 - P(Y = C_1|X = x)}$$

1. est compris entre 0 et $+\infty$
   - 0 si $P(C_1|x) = 0$; $+\infty$ si $P(C_1|x) = 1$
2. définit la fonction de décision
   - $f(x) = C_1$ si > 1
3. quantifie la confiance dans la prédiction
   - si ~ 1, alors $P(C_1|x) \sim P(C_2|x)$

Régression logistique - formalisation

Inconvénient de l'odd-ratio : pas symétrique

- si $P(C_2|x) \geq P(C_1|x)$: entre 0 et 1
- si $P(C_1|x) \geq P(C_2|x)$: entre 1 et $+\infty$

Régression logistique - formalisation

Inconvénient de l'odd-ratio : pas symétrique

- si $P(C_2|x) \geq P(C_1|x)$: entre 0 et 1
- si $P(C_1|x) \geq P(C_2|x)$: entre 1 et $+\infty$

$\Rightarrow$ on le considère souvent en échelle log :

$$score(x) = \ln \left(\frac{P(Y = C_1|X = x)}{1 - P(Y = C_1|X = x)} \right)$$

Régression logistique - formalisation

Inconvénient de l'odd-ratio : pas symétrique

- si $P(C_2|x) \geq P(C_1|x)$: entre 0 et 1
- si $P(C_1|x) \geq P(C_2|x)$: entre 1 et $+\infty$

$\Rightarrow$ on le considère souvent en échelle log :

$$score(x) = \ln\left(\frac{P(Y = C_1|X = x)}{1 - P(Y = C_1|X = x)}\right)$$

Ainsi :

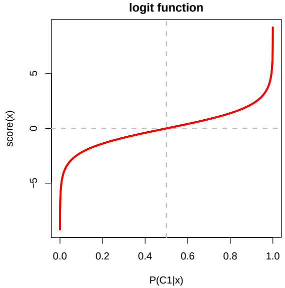
1. il est compris entre $-\infty$ et $+\infty$ (et est symétrique)
2. la règle de décision devient :
   - $f(x) = C_1$ si $score(x) > 0$
   - ou encore $f(x) = \textbf{signe}(score(x))$
3. la confiance est directement liée à $|score(x)|$

Régression logistique - formalisation

Lien entre $p = P(C_1|x)$ et $score(x) = \ln(p/(1-p))$:

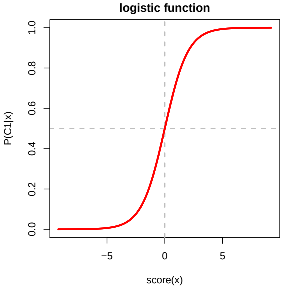
- On passe de p à $score(x)$ par la fonction logit :

$$s = logit(p) = \ln \frac{p}{1-p}$$



- On passe de $score(x)$ à p par la fonction logistique :

$$p = \sigma(s) = \frac{\exp^s}{1 + \exp^s}$$



Régression logistique - formalisation

En bref :

Régression logistique - formalisation

En bref :

1. classifieur de Bayes & odd-ratios - fonction de score :

$$score(x) = \ln\left(\frac{P(Y = C_1|X = x)}{1 - (Y = C_1|X = x)}\right) \in \left]-\infty, +\infty\right[$$

Régression logistique - formalisation

En bref :

1. classifieur de Bayes & odd-ratios - fonction de score :

$$score(x) = \ln\left(\frac{P(Y = C_1|X = x)}{1 - (Y = C_1|X = x)}\right) \in \,]-\infty, +\infty[$$

2. règle de décision :

$$f(x) = \begin{cases} C_1 & \text{si} \quad score(x) > 0 \\ C_2 & \text{sinon} \end{cases}$$

Régression logistique - formalisation

En bref :

1. classifieur de Bayes & odd-ratios - fonction de score :

$$score(x) = \ln\left(\frac{P(Y = C_1|X = x)}{1 - (Y = C_1|X = x)}\right) \in \,]-\infty, +\infty[$$

2. règle de décision :

$$f(x) = \begin{cases} C_1 & \text{si} \quad score(x) > 0 \\ C_2 & \text{sinon} \end{cases}$$

$\Rightarrow$ Objectif de la régression logistique : modéliser $score(x)$

- (et ainsi éviter de définir $P(X|Y = C_k)$ et $P(Y = C_k)$)

Régression logistique - formalisation

La régression logistique considère un modèle linéaire :

$$score(x) = \ln\left(\frac{P(Y = C_1|X = x)}{1 - (Y = C_1|X = x)}\right) = \langle w, x\rangle = \sum_{j=1}^{p} w_j x_j$$

$\Rightarrow$ le modèle met donc en jeu p paramètres pour $x \in \mathbb{R}^p$.

Régression logistique - formalisation

La régression logistique considère un modèle linéaire :

$$score(x) = \ln\left(\frac{P(Y = C_1|X = x)}{1 - (Y = C_1|X = x)}\right) = \langle w, x\rangle = \sum_{j=1}^{p} w_j x_j$$

$\Rightarrow$ le modèle met donc en jeu p paramètres pour $x \in \mathbb{R}^p$.

Remarques :

- en général, on introduit un biais/intercept
  - $score(x) = \langle \tilde{w}, \tilde{x}\rangle$ avec $\tilde{x} = [1 \;\; x] \Rightarrow p + 1$ paramètres
- moins de paramètres que les approches génératives
  - LDA : 2 pour π_1/π_2 + $2p$ pour μ_1/μ_2 + $p(p+1)/2$ pour Σ
  - QDA : $p(p+1)/2$ en plus pour Σ_1/Σ_2

Régression logistique - formalisation

Frontière de décision :

$$\left\{ x \;:\; P(Y = C_1|X = x) = P(Y = C_2|X = x) \right\}$$

$$: \;\ln\left(\frac{P(Y = C_1|X = x)}{1 - P(Y = C_1|X = x)}\right) = 0 \right\}$$

$$: \;\langle w, x\rangle = 0 \right\}$$

Régression logistique - formalisation

Frontière de décision :

$$\left\{ x \;:\; P(Y = C_1|X = x) = P(Y = C_2|X = x) \right\}$$

$$: \; \ln\left(\frac{P(Y = C_1|X = x)}{1 - P(Y = C_1|X = x)}\right) = 0 \right\}$$

$$: \; \langle w, x\rangle = 0 \right\}$$

$\Rightarrow$ une fonction linéaire (une droite, un plan ou un hyperplan)

Régression logistique - formalisation

Frontière de décision :

$$\left\{ x \;:\; P(Y = C_1|X = x) = P(Y = C_2|X = x) \right\}$$

$$: \;\ln\left(\frac{P(Y = C_1|X = x)}{1 - P(Y = C_1|X = x)}\right) = 0 \;\}$$

$$: \;\langle w, x\rangle = 0 \;\}$$

$\Rightarrow$ une fonction linéaire (une droite, un plan ou un hyperplan)

Régression logistique vs LDA ?

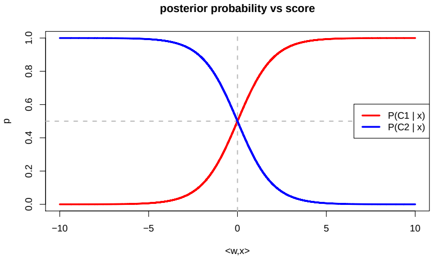
- ici : on estime directement les p paramètres
- LDA : on estime les paramètres des gaussiennes et on en déduit les p paramètres

Régression logistique - formalisation

Expression des probabilités a posteriori :

$$P(Y = C_1|X = x) = \sigma(\langle w, x\rangle) = \frac{\exp^{\langle w,x\rangle}}{1 + \exp^{\langle w,x\rangle}}$$

$$P(Y = C_2|X = x) = 1 - P(Y = C_1|X = x) = \frac{1}{1 + \exp^{\langle w,x\rangle}}$$



Rappel : $\langle w, x\rangle = \ln \frac{P(C_1|x)}{1-P(C_1|x)}$

Régression logistique - implémentation

En pratique il faut donc estimer le vecteur w.

Régression logistique - implémentation

En pratique il faut donc estimer le vecteur w.

On procède par maximum de (log-) vraisemblance :

$$\begin{aligned}\hat{w} &= \arg\max_w \prod_{i=1}^{n} P(Y = y_i | X = x_i) \\ &= \arg\max_w \sum_{i=1}^{n} \ln P(Y = y_i | X = x_i)\end{aligned}$$

$\Rightarrow$ problème d'optimisation, pas de solution analytique

Régression logistique - implémentation

En pratique il faut donc estimer le vecteur w.

On procède par maximum de (log-) vraisemblance :

$$\hat{w} = \arg\max_w \prod_{i=1}^{n} P(Y = y_i | X = x_i)$$
$$= \arg\max_w \sum_{i=1}^{n} \ln P(Y = y_i | X = x_i)$$

$\Rightarrow$ problème d'optimisation, pas de solution analytique

Remarques :

- problème convexe, solution unique (minimum global)
- algorithme itératif de moindres carrés pondérés
  - Iterative Reweighted Least-Squares (IRLS)

Régression logistique en pratique

1. Apprentissage : estimer w / calculer $\hat{w}$

Régression logistique en pratique

1. Apprentissage : estimer w / calculer $\hat{w}$

2. Prédiction d'une instance x' :

Régression logistique en pratique

1. Apprentissage : estimer w / calculer $\hat{w}$

2. Prédiction d'une instance x' :

   2.1 calcul du score :

   - $score(x') = \langle \hat{w}, x' \rangle$

Régression logistique en pratique

1. Apprentissage : estimer w / calculer $\hat{w}$

2. Prédiction d'une instance x' :

   2.1 calcul du score :
      - $score(x') = \langle \hat{w}, x' \rangle$

   2.2 calcul des probabilités a posteriori :
      - $P(C_1|x') = \exp^{score(x')} / (1 + \exp^{score(x')})$
      - $P(C_2|x') = 1/(1 + \exp^{score(x')})$

Régression logistique en pratique

1. Apprentissage : estimer w / calculer $\hat{w}$
2. Prédiction d'une instance x' :
   - 2.1 calcul du score :
     - $score(x') = \langle \hat{w}, x' \rangle$
   - 2.2 calcul des probabilités a posteriori :
     - $P(C_1|x') = \exp^{score(x')} / (1 + \exp^{score(x')})$
     - $P(C_2|x') = 1/(1 + \exp^{score(x')})$
   - 2.3 prise de décision – on prédit C_1 si : $score(x') > 0$
     - ce qui est équivalent à $P(C_1|x') > 0.5$
     - ou encore $P(C_1|x') > P(C_2|x')$

En pratique

Classification multi-classe (vs binaire)

Approches génératives : "nativement" multi-classes

- même expression si $K = 2$ ou $K > 2$

Régression logistique : un peu plus compliqué

- on modèlise $K - 1$ odds-ratio
- on estime $K - 1$ vecteurs w_k de p (ou $p + 1$) paramètres
- (reste bien moins de paramètres que LDA/QDA)

$\Rightarrow$ voir Hastie et al. (2001)[sec. 4.4] pour plus de détails.

Approches probabilistes et critères de rejet

Modèles probabilistes de classification : fournissent $P(C_k|x)$

$\Rightarrow$ définit un critère de confiance dans la prédiction

Approches probabilistes et critères de rejet

Modèles probabilistes de classification : fournissent $P(C_k|x)$

$\Rightarrow$ définit un critère de confiance dans la prédiction

Illustration (cadre binaire) :

- $P(C_1|x) \sim 0.5 \rightarrow$ confiance faible
- $P(C_1|x) \sim 1$ ou $\sim 0 \rightarrow$ confiance forte (dans C_1 ou C_2)

Approches probabilistes et critères de rejet

Modèles probabilistes de classification : fournissent $P(C_k|x)$

$\Rightarrow$ définit un critère de confiance dans la prédiction

Illustration (cadre binaire) :

- $P(C_1|x) \sim 0.5 \rightarrow$ confiance faible
- $P(C_1|x) \sim 1$ ou $\sim 0 \rightarrow$ confiance forte (dans C_1 ou C_2)

Critère de rejet : on ne prédit que si $\max_k P(C_k|x) > \delta$

- seuil de décision $\delta \in [1/K, 1]$
- $\delta = 1/K$ / $\delta = 1$: on prédit toujours / jamais

Approches probabilistes et critères de rejet

Modèles probabilistes de classification : fournissent $P(C_k|x)$

$\Rightarrow$ définit un critère de confiance dans la prédiction

Illustration (cadre binaire) :

- $P(C_1|x) \sim 0.5 \rightarrow$ confiance faible
- $P(C_1|x) \sim 1$ ou $\sim 0 \rightarrow$ confiance forte (dans C_1 ou C_2)

Critère de rejet : on ne prédit que si $\max_k P(C_k|x) > \delta$

- seuil de décision $\delta \in [1/K, 1]$
- $\delta = 1/K$ / $\delta = 1$: on prédit toujours / jamais

$\Rightarrow$ compromis - en augmentant δ :

1. on effectue moins de prédictions
2. mais on peut <u>espérer</u> faire moins d'erreurs

Impact du nombre de paramètres

Inconvénient du cadre génératif : nombre de paramètres

- LDA/QDA : augmente quadratiquement avec p

Conséquence :

- estimation instable
- mauvaise généralisation

Solution : sélection de variable ou réduction de dimension

- e.g., commencer par faire une ACP puis choisir entre LDA et QDA

Quel modèle choisir ?

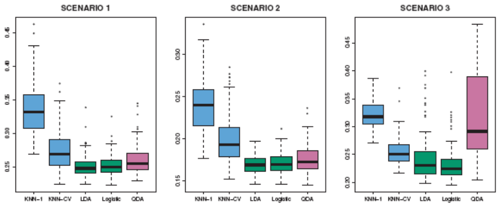


FIGURE 4.10. *Boxplots of the test error rates for each of the linear scenarios described in the main text.*

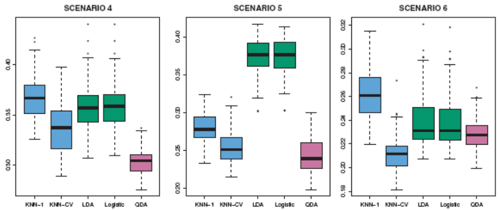


FIGURE 4.11. *Boxplots of the test error rates for each of the non-linear scenarios described in the main text.*

$\Rightarrow$ voir James et al. (2013)[4.5] : dépend du jeu de données !

- (non) linéaire, gaussiennes, bruit, nombre d'observations, ...

Mise en oeuvre en R

LDA et QDA : implémentés dans le package `MASS`.

- apprentissage : fonction `lda` et `qda`
  - arguments principaux : `X` et `y`
- prédiction : fonctions `predict.lda` et `predict.qda`
  - NB : possibilité de spécifier les a priori à utiliser
  - utile pour dé-conditionner des effectifs d'apprentissage

Mise en oeuvre en R

LDA et QDA : implémentés dans le package `MASS`.

- apprentissage : fonction `lda` et `qda`
  - arguments principaux : `X` et `y`
- prédiction : fonctions `predict.lda` et `predict.qda`
  - NB : possibilité de spécifier les a priori à utiliser
  - utile pour dé-conditionner des effectifs d'apprentissage

Régression logistique : un modèle linéaire généralisé (`glm`)

- apprentissage : fonction `glm`
  - avec option **family = "binomial"**
  - même syntaxe que fonction `lm`
- prédiction : fonction `predict.glm`
  - avec option **type = "link"** : renvoie $score(x)$
  - avec option **type = "response"** : renvoie $P(C_1|x)$
  - même syntaxe que fonction `lm`

Conclusion

Conclusion

Théorie statistique de la décision :

- meilleur choix si on connaissait $P(X, Y)$
- régression : $E(Y|X)$ - regression function
- classification : $\arg\max_k P(Y_k|X)$ - classifieur de Bayes

Conclusion

Théorie statistique de la décision :

- meilleur choix si on connaissait $P(X, Y)$
- régression : $E(Y|X)$ - regression function
- classification : $\arg\max_k P(Y_k|X)$ - classifieur de Bayes

Modèles probabilistes de classification : estimer $P(Y_k|X)$

- des modèles incontournables - populaires et performants

Conclusion

Théorie statistique de la décision :

- meilleur choix si on connaissait $P(X, Y)$
- régression : $E(Y|X)$ - regression function
- classification : $\arg\max_k P(Y_k|X)$ - classifieur de Bayes

Modèles probabilistes de classification : estimer $P(Y_k|X)$

- des modèles incontournables - populaires et performants

Modèles génératifs :

- définir $P(X|Y_k)$ et l'inverser en $P(Y_k|X)$ - loi de Bayes
- modèle classique : $P(X|Y_k) = \mathcal{MN}(x; \mu_k, \Sigma_k)$
- LDA : $\Sigma_k = \Sigma \Rightarrow$ frontières linéaires
- QDA : $\{\Sigma_k\}_{k=1,\dots,K} \Rightarrow$ frontières non-linéaires

Conclusion

LDA vs QDA :

- compromis complexité (modèle) / stabilité (estimation)
- hypothèse QDA moins contraignante, mais plus de paramètres à estimer

Conclusion

LDA vs QDA :

- compromis complexité (modèle) / stabilité (estimation)
- hypothèse QDA moins contraignante, mais plus de paramètres à estimer

Modèles discriminatifs :

- modéliser $P(Y_k|X)$ - la fonction de prédiction
  - pas d'hypothèse sur la distribution des données (X)
- modèle de la régression logistique
  - modélise log(odds-ratio) comme une fonction linéaire
  - p (ou $p+1$) paramètres à estimer
  - estimation = problème d'optimisation (convexe)

Conclusion

LDA vs QDA :

- compromis complexité (modèle) / stabilité (estimation)
- hypothèse QDA moins contraignante, mais plus de paramètres à estimer

Modèles discriminatifs :

- modéliser $P(Y_k|X)$ - la fonction de prédiction
  - pas d'hypothèse sur la distribution des données (X)
- modèle de la régression logistique
  - modélise log(odds-ratio) comme une fonction linéaire
  - p (ou $p+1$) paramètres à estimer
  - estimation = problème d'optimisation (convexe)

TP : classifieur de Bayes & loi normales, comparaison LDA/QDA/LogReg.

Références

T. Hastie, R. Tibshirani, and J.. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.