Introduction au Machine Learning

C3 - Méthodes linéaires pour la classification

Lucas Gnecco Heredia

LAMSADE - Université Paris-Dauphine
lucas.gnecco-heredia@dauphine.psl.eu - C602

* *Remerciement spécial à Florian Yger*

Cadre de travail

On observe les valeurs $x_1, \cdots, x_n$ pour les variables aléatoires (iid) $X_1, \cdots, X_n$.
En posant une hypothèse sur la distribution des variables X, on cherche à estimer θ, le paramètre de la loi à partir des observations.

Exemples

Cadre de travail

On observe les valeurs $x_1, \cdots, x_n$ pour les variables aléatoires (iid) $X_1, \cdots, X_n$.
En posant une hypothèse sur la distribution des variables X, on cherche à estimer θ, le paramètre de la loi à partir des observations.

Exemples

- observations de loi gaussienne $\mathcal{N}(\mu, 1)$ alors, $\theta = \mu$

Cadre de travail

On observe les valeurs $x_1, \cdots, x_n$ pour les variables aléatoires (iid) $X_1, \cdots, X_n$.
En posant une hypothèse sur la distribution des variables X, on cherche à estimer θ, le paramètre de la loi à partir des observations.

Exemples

- observations de loi gaussienne $\mathcal{N}(\mu, 1)$ alors, $\theta = \mu$
- observations de loi gaussienne $\mathcal{N}(\mu, \sigma)$ alors, $\theta = (\mu, \sigma)$ (vecteur de paramètres)

Cadre de travail

On observe les valeurs $x_1, \cdots, x_n$ pour les variables aléatoires (iid) $X_1, \cdots, X_n$.
En posant une hypothèse sur la distribution des variables X, on cherche à estimer θ, le paramètre de la loi à partir des observations.

Exemples

- observations de loi gaussienne $\mathcal{N}(\mu, 1)$ alors, $\theta = \mu$
- observations de loi gaussienne $\mathcal{N}(\mu, \sigma)$ alors, $\theta = (\mu, \sigma)$ (vecteur de paramètres)
- observations de loi Bernouilli $\mathcal{B}(p)$ alors, $\theta = p$ (observation d'un ensemble de jets de pièce)

Vraisemblance

Définition

La fonction de vraisemblance [1], notée $\mathcal{L}(x_1, \cdots, x_n|\theta)$ (ou par abus de notation $\mathcal{L}(\theta)$) est une fonction de probabilité conditionnelle qui décrit les valeurs d'une loi en X_i en fonction des paramètres θ de cette loi.
Elle s'exprime à partir des densités $f(x|\theta)$ ou de loi de probabilité $P(X = x|\theta)$.

$$\mathcal{L}(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

1. En anglais : Likelihood

Vraisemblance

Définition

La fonction de vraisemblance [1], notée $\mathcal{L}(x_1, \cdots, x_n|\theta)$ (ou par abus de notation $\mathcal{L}(\theta)$) est une fonction de probabilité conditionnelle qui décrit les valeurs d'une loi en X_i en fonction des paramètres θ de cette loi.
Elle s'exprime à partir des densités $f(x|\theta)$ ou de loi de probabilité $P(X = x|\theta)$.

$$\mathcal{L}(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

avec

$$f(x_i; \theta) = \begin{cases} f_\theta(x) & \text{si } X \text{ est continue} \\ P_\theta(X = x) & \text{sinon.} \end{cases}$$

1. En anglais : Likelihood

Interprétation

En pratique

$\mathcal{L}$ indique à quel point, si les observations suivent la loi $f_\theta(x)$, il serait vraisemblance/crédible/probable d'observer l'échantillon.

Interprétation

En pratique

$\mathcal{L}$ indique à quel point, si les observations suivent la loi $f_\theta(x)$, il serait vraisemblance/crédible/probable d'observer l'échantillon.

Important

$\mathcal{L}$ est la vraisemblance d'une loi pour un échantillon.

Interprétation

En pratique

$\mathcal{L}$ indique à quel point, si les observations suivent la loi $f_\theta(x)$, il serait vraisemblance/crédible/probable d'observer l'échantillon.

Important

$\mathcal{L}$ est la vraisemblance d'une loi pour un échantillon.

Utilisation pratique

Pour deux valeurs θ_1 et θ_2 du paramètre θ, si on a

$$\mathcal{L}(\theta_1) < \mathcal{L}(\theta_2)$$

il semble plus vraisemblable que l'échantillon suive une loi $f_{\theta_2}(x)$

Vraisemblance

Exemple

On dispose d'une pièce et on se demande si elle est truquée ou pas à partir d'observations.
Soit p, le paramètre de la loi de Bernouilli (ie. probabilité d'obtenir "face") et un échantillon constitué de 49 "face" et 31 "pile".
La vraisemblance de l'échantillon s'écrit $\mathcal{L}(p) = p^{49} \times (1-p)^{31}$.
On pourrait alors tester la valeur de cette vraisemblance pour

- $p_1 = \frac{1}{3}$
- $p_2 = \frac{1}{2}$
- $p_3 = \frac{2}{3}$

Vraisemblance

Exemple

On dispose d'une pièce et on se demande si elle est truquée ou pas à partir d'observations.
Soit p, le paramètre de la loi de Bernouilli (ie. probabilité d'obtenir "face") et un échantillon constitué de 49 "face" et 31 "pile".
La vraisemblance de l'échantillon s'écrit $\mathcal{L}(p) = p^{49} \times (1-p)^{31}$.
On pourrait alors tester la valeur de cette vraisemblance pour

- $p_1 = \frac{1}{3}$
- $p_2 = \frac{1}{2}$
- $p_3 = \frac{2}{3}$

On peut montrer que $\mathcal{L}(p_1) < \mathcal{L}(p_2) < \mathcal{L}(p_3)$

On peut faire mieux que cela

Maximum de Vraisemblance

Estimation de paramètre

Pour inférer le paramètre d'une distribution de probabilité à partir d'un échantillon donné, on cherche $\theta^\star$ qui maximise la vraisemblance :

$$\theta_{\mathsf{MV}} = \arg\max_{\theta} \mathcal{L}(x_1, \cdots, x_n|\theta)$$

Maximum de Vraisemblance

Estimation de paramètre

Pour inférer le paramètre d'une distribution de probabilité à partir d'un échantillon donné, on cherche $\theta^\star$ qui maximise la vraisemblance :

$$\theta_{\text{MV}} = \arg\max_{\theta} \mathcal{L}(x_1, \cdots, x_n|\theta)$$

En pratique

- suivant la forme de $f(x|\theta)$, on pourra privilégier la maximisation de la log-vraisemblance $\log(\mathcal{L}(\theta))$
- chercher l'arg max de $\mathcal{L}$ revient à trouver le point $\hat{\theta}$ tel que $\nabla_\theta \mathcal{L} = 0$

Retour à l'exemple

Exemple

On dispose d'une pièce et on se demande si elle est truquée ou pas à partir d'observations.
Soit p, le paramètre de la loi de Bernouilli (ie. probabilité d'obtenir "face") et un échantillon constitué de 49 "face" et 31 "pile".
La vraisemblance de l'échantillon s'écrit $\mathcal{L}(p) = p^{49} \times (1-p)^{31}$.
Quelle est la valeur la plus vraisemblable pour p ?

Retour à l'exemple

Exemple

On dispose d'une pièce et on se demande si elle est truquée ou pas à partir d'observations.
Soit p, le paramètre de la loi de Bernouilli (ie. probabilité d'obtenir "face") et un échantillon constitué de 49 "face" et 31 "pile".
La vraisemblance de l'échantillon s'écrit $\mathcal{L}(p) = p^{49} \times (1-p)^{31}$.
Quelle est la valeur la plus vraisemblable pour p ?

Réponse

$$\nabla_p \mathcal{L} = 49p^{48}(1-p)^{31} - 31(1-p)^{30}p^{49}$$

Retour à l'exemple

Exemple

On dispose d'une pièce et on se demande si elle est truquée ou pas à partir d'observations.
Soit p, le paramètre de la loi de Bernouilli (ie. probabilité d'obtenir "face") et un échantillon constitué de 49 "face" et 31 "pile".
La vraisemblance de l'échantillon s'écrit $\mathcal{L}(p) = p^{49} \times (1-p)^{31}$.
Quelle est la valeur la plus vraisemblable pour p ?

Réponse

$$\begin{aligned}\nabla_p \mathcal{L} &= 49p^{48}(1-p)^{31} - 31(1-p)^{30}p^{49} \\ &= p^{48}(1-p)^{30}(49(1-p) - 31p)\end{aligned}$$

Retour à l'exemple

Exemple

On dispose d'une pièce et on se demande si elle est truquée ou pas à partir d'observations.
Soit p, le paramètre de la loi de Bernouilli (ie. probabilité d'obtenir "face") et un échantillon constitué de 49 "face" et 31 "pile".
La vraisemblance de l'échantillon s'écrit $\mathcal{L}(p) = p^{49} \times (1-p)^{31}$.
Quelle est la valeur la plus vraisemblable pour p ?

Réponse

$$\begin{aligned}\nabla_p \mathcal{L} &= 49p^{48}(1-p)^{31} - 31(1-p)^{30}p^{49} \\ &= p^{48}(1-p)^{30}(49-80p)\end{aligned}$$

Retour à l'exemple

Réponse

$$\begin{aligned}\nabla_p \mathcal{L} &= 49p^{48}(1-p)^{31} - 31(1-p)^{30}p^{49} \\ &= p^{48}(1-p)^{30}(49-80p)\end{aligned}$$

On pose alors $\nabla_p \mathcal{L} = 0$ et on a :

- $p = 0$ (racine évidente du polynôme) pour lequel $\mathcal{L}(0) = 0$
- $p = 1$ (racine évidente du polynôme) pour lequel $\mathcal{L}(1) = 0$
- $p = \frac{49}{80}$ et pour lequel $\mathcal{L}(\frac{49}{80}) > 0$

Retour à l'exemple

Réponse

$$\nabla_p \mathcal{L} = 49p^{48}(1-p)^{31} - 31(1-p)^{30}p^{49}$$
$$= p^{48}(1-p)^{30}(49-80p)$$

On pose alors $\nabla_p \mathcal{L} = 0$ et on a :

- $p = 0$ (racine évidente du polynôme) pour lequel $\mathcal{L}(0) = 0$
- $p = 1$ (racine évidente du polynôme) pour lequel $\mathcal{L}(1) = 0$
- $p = \frac{49}{80}$ et pour lequel $\mathcal{L}(\frac{49}{80}) > 0$

Dans ce cas, la fréquence empirique est un estimateur du MV de la probabilité (pour une distribution de Bernouilli).

Application

Soient $X_1, \cdots, X_n$ des variables aléatoires i.i.d. suivant une loi normale $\mathcal{N}(\mu, \sigma)$. À partir de l'échantillon $x_1, \cdots, x_n$, trouver l'estimateur du Maximum de Vraisemblance pour μ.

Cadre de travail

On dispose de l'échantillon $(x_1, y_1), \cdots, (x_n, y_n)$ et on fait l'hypothèse

$$y_i = w_0 + \sum_{j=1}^{p} w_j x_i^{(j)} + \epsilon_i$$

et tels que $\epsilon_i \sim \mathcal{N}(0, \sigma)$
Dans ce contexte,

$$y_i \sim \mathcal{N}(w_0 + \sum_{j=1}^{p} w_j x_i^{(j)}, \sigma)$$

log-vraisemblance du modèle

$$\log \mathcal{L}(w) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - w_0 - \sum_{j=1}^{p} w_j x_i^{(j)})^2}{\sigma^2}}\right)$$

$$= -\frac{1}{2} \underbrace{\sum_i (y_i - w_0 - \sum_{j=1}^{p} w_j x_i^{(j)})^2}_{J(w)} + C(\sigma)$$

log-vraisemblance du modèle

$$\log \mathcal{L}(w) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - w_0 - \sum_{j=1}^{p} w_j x_i^{(j)})^2}{\sigma^2}}\right)$$

$$= -\frac{1}{2} \underbrace{\sum_i (y_i - w_0 - \sum_{j=1}^{p} w_j x_i^{(j)})^2}_{J(w)} + C(\sigma)$$

En posant $x_i^{(0)} = 1$, on peut ré-écrire :

$$J(w) = \sum_i (y_i - \sum_{j=0}^{p} w_j x_i^{(j)})^2$$

$$= \sum_i (y_i - x_i^\top w)^2$$

Estimation du MV - méthode 1 - forme algébrique

On pose
$\nabla_w J = 0$

Estimation du MV - méthode 1 - forme algébrique

On pose
$-2X^\top y + 2X^\top Xw = 0$

Estimation du MV - méthode 1 - forme algébrique

On pose
$-2X^\top y + 2X^\top Xw = 0$

Si $X^\top X$ est inversible, alors
$\hat{w} = \left(X^\top X\right)^{-1} X^\top y$

Estimation du MV - méthode 2 - descente de gradient

Minimisation d'une fonction convexe par descente de gradient

Pour résoudre $\arg\min_x f(x)$, on produit une suite d'itéré $x_{t+1} = x_t - \alpha \nabla f(x_t)$ (avec le pas $\alpha > 0$).

Estimation du MV - méthode 2 - descente de gradient

Minimisation d'une fonction convexe par descente de gradient

Pour résoudre $\arg\min_x f(x)$, on produit une suite d'itéré $x_{t+1} = x_t - \alpha \nabla f(x_t)$ (avec le pas $\alpha > 0$).

Éléments de démonstration

$f(x+s) \approx f(x) + \nabla f(x)^\top s$ (série de Taylor à l'ordre 1)
Si on pose $s = -\alpha \nabla f(x)$, alors

$$f\left(x - \alpha \nabla f(x)\right) \approx f(x) - \alpha \overbrace{\nabla f(x)^\top \nabla f(x)}^{>0} < f(x)$$

Notations

- ensemble d'apprentissage $\mathcal{D} = \{(x_1, y_1), \cdot, (x_n, y_n)\}$
- $x_i = (x_i^1, \cdot, x_i^d) \in \mathcal{X} \subset \mathbb{R}^d$
- y_i prenant ses valeurs dans un ensemble fini (et non ordonné) $\mathcal{Y}$ (par ex. $\mathcal{Y} \subset \{C_1, \cdot, C_K\}$)

But

Pour toute nouvelle observation, on souhaite prédire Y via une fonction (apprise à partir des données) $h(X)$

En pratique

Considérons le cas $\mathcal{Y} = \{0, 1\}$ et soit

$$r(x) = P(Y = 1|X = x)$$

D'après le théorème de Bayes, nous avons :

$$\begin{aligned} & r(x) \\ =& P(Y = 1|X = x) \\ =& \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x|Y = 1)P(Y = 1) + P(X = x|Y = 0)P(Y = 0)} \\ =& \frac{\pi p_1(x)}{\pi p_1(x) + (1 - \pi)p_0(x)} \end{aligned}$$

avec

$$p_0(x) = P(x|Y = 0), p_1(x) = P(x|Y = 1), \pi = P(Y = 1)$$

Classifieur de Bayes

Définition

La règle de classification de Bayes $h^\star$ s'écrit :

$$h^\star(x) = \begin{cases} 1 \text{ si } r(x) > \frac{1}{2} \\ 0 \text{ sinon} \end{cases}$$

Ré-écritures équivalentes

- $$h^\star(x) = \begin{cases} 1 \text{ si } P(Y=1|X=x) > P(Y=0|X=x) \\ 0 \text{ sinon} \end{cases}$$

Ré-écritures équivalentes

- $$h^\star(x) = \begin{cases} 1 \text{ si } P(Y=1|X=x) > P(Y=0|X=x) \\ 0 \text{ sinon} \end{cases}$$
- $$h^\star(x) = \begin{cases} 1 \text{ si } \pi p_1(x) > (1-\pi) p_0(x) \\ 0 \text{ sinon} \end{cases}$$

Ré-écritures équivalentes

- $$h^\star(x) = \begin{cases} 1 \text{ si } P(Y=1|X=x) > P(Y=0|X=x) \\ 0 \text{ sinon} \end{cases}$$
- $$h^\star(x) = \begin{cases} 1 \text{ si } \pi p_1(x) > (1-\pi) p_0(x) \\ 0 \text{ sinon} \end{cases}$$
- $$h^\star(x) = \arg\max_{y \in \mathcal{Y}} P(Y=y|X=x)$$

Classifieur de Bayes

Théorème (admis)

Le classifieur de Bayes est optimal.

Classifieur de Bayes

Théorème (admis)

Le classifieur de Bayes est optimal.

Cependant en pratique, π, p_0 et p_1 sont inconnus et on va alors chercher à s'approcher de ce classifieur de Bayes en introduisant certaines hypothèses (ex. l'hypothèse d'indépendance conditionnelle des caractéristiques dans le classifieur bayésien naïf).

Dans ce chapitre

Nous allons chercher à estimer les différentes densités (à partir des données) et produire les estimateurs suivants :

- $\hat{r}(x) = \hat{P}(Y = 1|X = x) = \frac{\hat{\pi}\hat{p}_1(x)}{\hat{\pi}\hat{p}_1(x)+(1-\hat{\pi})\hat{p}_0(x)}$

Dans ce chapitre

Nous allons chercher à estimer les différentes densités (à partir des données) et produire les estimateurs suivants :

- $\hat{r}(x) = \hat{P}(Y = 1|X = x) = \frac{\hat{\pi}\hat{p}_1(x)}{\hat{\pi}\hat{p}_1(x)+(1-\hat{\pi})\hat{p}_0(x)}$
- $\hat{h}(x) \begin{cases} 1 \text{ si } \hat{r}(x) > \frac{1}{2} \\ 0 \text{ sinon} \end{cases}$

Introduction

En anglais

Linear Discriminant Analysis (LDA [2])

2. Attention, en ML/TAL, il existe aussi un modèle génératif probabiliste appelé *Latent Dirichlet Allocation* (LDA) qui partage le même acronyme.

Introduction

En anglais

Linear Discriminant Analysis (LDA [2])

Hypothèses

$p_0(x)$ et $p_1(x)$ sont des lois gaussiennes (multivariées)

$$\forall k = 0, 1,\ p_k(x) = \frac{e^{-\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)}}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}}$$

Et on suppose en plus que $\Sigma_0 = \Sigma_1 = \Sigma$

Ainsi, $X|Y = 0 \sim \mathcal{N}(\mu_0, \Sigma)$ et $X|Y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$

2. Attention, en ML/TAL, il existe aussi un modèle génératif probabiliste appelé *Latent Dirichlet Allocation* (LDA) qui partage le même acronyme.

Illustrations

Différentes formes de gaussiennes

Illustration (2D)

Illustrations

Différentes formes de gaussiennes

Illustration (2D)

Hypothèse de distribution des données

Illustration (2D)

Impact des hypothèses sur la règle de décision

Théorème

Avec les hypothèses précédentes [3], la règle de décision devient :

$$h_{\text{LDA}}(x) = \underset{k \in \{0,1\}}{\arg\max}\, \delta_k(x)$$

avec $\delta_k(x) = x^\top \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \log(\pi_k)$

3. On note $\pi_1 = P(Y = 1)$ et $\pi_0 = P(Y = 0) = (1 - \pi_1)$

Impact des hypothèses sur la règle de décision

Théorème

Avec les hypothèses précédentes [3], la règle de décision devient :

$$h_{\text{LDA}}(x) = \underset{k \in \{0,1\}}{\arg\max}\, \delta_k(x)$$

avec $\delta_k(x) = x^\top \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \log(\pi_k)$

Interprétation en termes de distance de Mahalanobis

3. On note $\pi_1 = P(Y = 1)$ et $\pi_0 = P(Y = 0) = (1 - \pi_1)$

Mise en œuvre

On estime les paramètres de la façon suivante :

- $\hat{\pi}_0 = \frac{1}{n}\sum_i (1 - y_i) = \frac{n_0}{n}$ $\qquad$ $\hat{\pi}_1 = \frac{1}{n}\sum_i y_i = \frac{n_1}{n}$

Mise en œuvre

On estime les paramètres de la façon suivante :

- $\hat{\pi}_0 = \frac{1}{n}\sum_i (1 - y_i) = \frac{n_0}{n}$ $\qquad$ $\hat{\pi}_1 = \frac{1}{n}\sum_i y_i = \frac{n_1}{n}$
- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i \backslash y_i = k} x_i$

Mise en œuvre

On estime les paramètres de la façon suivante :

- $\hat{\pi}_0 = \frac{1}{n}\sum_i (1 - y_i) = \frac{n_0}{n}$ $\qquad$ $\hat{\pi}_1 = \frac{1}{n}\sum_i y_i = \frac{n_1}{n}$
- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i \backslash y_i = k} x_i$
- $\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i \backslash y_i = k} (x_i - \mu_k)(x_i - \mu_k)^\top$

  $\hat{\Sigma} = \frac{n_0 \hat{\Sigma}_0 + n_1 \hat{\Sigma}_1}{n}$

Frontière de décision

Définition

Pour tout classifieur probabiliste $\hat{h}$, l'ensemble

$$\mathcal{D}(\hat{h}) = \{x \in \mathcal{X} \quad \backslash \quad \hat{P}(Y = 1|X = x) = \hat{P}(Y = 0|X = x)\}$$

est appelé frontière de décision.

Frontière de décision de l'Analyse Discriminante Linéaire

$$\mathcal{D} = \{x \in \mathcal{X} \quad \backslash \quad \delta_0(x) = \delta_1(x)\}$$

Frontière de décision de l'Analyse Discriminante Linéaire

$$\mathcal{D} = \{x \in \mathcal{X} \quad \setminus \quad \delta_0(x) = \delta_1(x)\}$$

$$\Leftrightarrow x^\top \Sigma^{-1}\mu_0 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0 + \log(\pi_0) = x^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 + \log(\pi_1)$$

Frontière de décision de l'Analyse Discriminante Linéaire

$$\mathcal{D} = \{x \in \mathcal{X} \quad \setminus \quad \delta_0(x) = \delta_1(x)\}$$

$$\Leftrightarrow x^\top \Sigma^{-1}\mu_0 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0 + \log(\pi_0) = x^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 + \log(\pi_1)$$

$$\Leftrightarrow x^\top \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 + \mu_1) + \log(\frac{\pi_0}{\pi_1}) = 0$$

Frontière de décision de l'Analyse Discriminante Linéaire

$$\mathcal{D} = \{x \in \mathcal{X} \quad \setminus \quad \delta_0(x) = \delta_1(x)\}$$

$$\Leftrightarrow x^\top \Sigma^{-1}\mu_0 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0 + \log(\pi_0) = x^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 + \log(\pi_1)$$

$$\Leftrightarrow x^\top \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 + \mu_1) + \log(\frac{\pi_0}{\pi_1}) = 0$$

$$\Leftrightarrow x^\top \underbrace{\Sigma^{-1}(\mu_0 - \mu_1)}_{w} - \underbrace{\frac{1}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 + \mu_1) + \log(\frac{\pi_0}{\pi_1})}_{b} = 0$$

Frontière de décision de l'Analyse Discriminante Linéaire

$$\mathcal{D} = \{x \in \mathcal{X} \quad \setminus \quad \delta_0(x) = \delta_1(x)\}$$

$$\Leftrightarrow x^\top \Sigma^{-1}\mu_0 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0 + \log(\pi_0) = x^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 + \log(\pi_1)$$

$$\Leftrightarrow x^\top \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 + \mu_1) + \log(\frac{\pi_0}{\pi_1}) = 0$$

$$\Leftrightarrow x^\top \underbrace{\Sigma^{-1}(\mu_0 - \mu_1)}_{w} \underbrace{- \frac{1}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 + \mu_1) + \log(\frac{\pi_0}{\pi_1})}_{b} = 0$$

La frontière de décision définit un hyperplan (séparant les classes) : $x^\top w + b = 0$

Frontière de décision de l'Analyse Discriminante Linéaire

$$\mathcal{D} = \{x \in \mathcal{X} \quad \setminus \quad \delta_0(x) = \delta_1(x)\}$$

$$\Leftrightarrow x^\top \Sigma^{-1}\mu_0 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0 + \log(\pi_0) = x^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 + \log(\pi_1)$$

$$\Leftrightarrow x^\top \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 + \mu_1) + \log(\frac{\pi_0}{\pi_1}) = 0$$

$$\Leftrightarrow x^\top \underbrace{\Sigma^{-1}(\mu_0 - \mu_1)}_{w} \underbrace{- \frac{1}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 + \mu_1) + \log(\frac{\pi_0}{\pi_1})}_{b} = 0$$

La frontière de décision définit un hyperplan (séparant les classes) : $x^\top w + b = 0$

Illustration

Généralisation au cas multi-classe

Dans ce cas, $\mathcal{Y} \in \{1, \cdots, K\}$.
$\forall k$, $p_k(x) = P(X = x|Y = k) \sim \mathcal{N}(\mu_k, \Sigma)$ (loi gaussienne de paramètres (μ_k, Σ), la règle de décision devient alors :

$$h_{\text{LDA}}(x) = \underset{k \in \{1,\cdot,K\}}{\arg\max}\, \delta_k(x)$$

avec $\delta_k(x) = x^\top \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \log(\pi_k)$

Bilan

- une méthode simple basée sur une approche probabiliste

Bilan

- une méthode simple basée sur une approche probabiliste
- fonction de décision linéaire

Bilan

- une méthode simple basée sur une approche probabiliste
- fonction de décision linéaire
  - décision rapide pour des nouveaux échantillons

Bilan

- une méthode simple basée sur une approche probabiliste
- fonction de décision linéaire
  - décision rapide pour des nouveaux échantillons
  - pas forcément adaptée à tous les jeux de données

Bilan

- une méthode simple basée sur une approche probabiliste
- fonction de décision linéaire
  - décision rapide pour des nouveaux échantillons
  - pas forcément adaptée à tous les jeux de données
- hypothèse réaliste (lois gaussiennes) mais réductrice ($\Sigma_0 = \Sigma_1$) (cependant moins réductrice que le classifieur bayésien naïf - les interactions entre variables sont prises en compte)

TD 1 - Analyse Discriminante Quadratique

En anglais

Quadratic Discriminant Analysis (QDA).

En reprenant les notations pour la classification binaire, on applique les hypothèses précédentes de gaussianité ($\mathcal{N}(\mu_k, \Sigma_k)$) sans imposer l'hypothèse $\Sigma_0 = \Sigma_1$.

TD 1 - Analyse Discriminante Quadratique

En reprenant les notations pour la classification binaire, on applique les hypothèses précédentes de gaussianité ($\mathcal{N}(\mu_k, \Sigma_k)$) sans imposer l'hypothèse $\Sigma_0 = \Sigma_1$.

1. Montrer que la règle de décision peut s'écrire :

$$h_{\text{QDA}}(x) = \underset{k \in \{0,1\}}{\arg\max}\, \delta_k(x)$$

avec $\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log(\pi_k)$

2. Montrer que cette régle peut se ré-écrire :
   - 1 si $(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0) < 2\log\left(\frac{\pi_1}{\pi_0}\right) + \log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right)$
   - 0 sinon

   et donner une interprétation géométrique de cette règle.

TD 1 - Analyse Discriminante Quadratique

1. Montrer que la règle de décision peut s'écrire :

$$h_{\text{QDA}}(x) = \underset{k \in \{0,1\}}{\arg\max}\, \delta_k(x)$$

avec $\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log(\pi_k)$

2. Montrer que cette régle peut se ré-écrire :
   - 1 si $(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0) < 2\log\left(\frac{\pi_1}{\pi_0}\right) + \log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right)$
   - 0 sinon

   et donner une interprétation géométrique de cette règle.
3. quelle est la forme de la frontière de décision ?
4. Montrer que si $\Sigma_0 = \Sigma_1 = \Sigma$, on retrouve l'Analyse Discriminante Linéaire.

TD 2 - Application numérique

On a les données suivantes :

X^1	X^2	Y
0	-1	0
-1	$-\frac{1}{2}$	0
-2	-1	0
-1	$-\frac{3}{2}$	0
$\frac{1}{2}$	1	1
1	2	1
1	0	1
$\frac{3}{2}$	1	1

1. représenter ces données
2. calculer les estimateurs de $\hat{\pi}_0$, $\hat{\pi}_1$, $\hat{\mu}_0$, $\hat{\mu}_1$ et $\hat{\Sigma}$
3. calculer w et b et réprésenter l'hyperplan séparateur
4. prédire la classe pour $x = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$
5. commenter l'utilisation de ce modèle sur ces données.

TD 2 - Application numérique

À partir de $\hat{\mu_0} = \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}$, $\hat{\mu_1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, $\hat{\pi}_0 = \hat{\pi}_1 = \frac{1}{2}$

calculer w et b et prédire la classe de $x = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ pour

1. $\hat{\Sigma} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix}$
2. $\hat{\Sigma} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

Cadre de travail

On dispose de l'échantillon $(x_1, y_1), \cdots, (x_n, y_n)$, mais étant dans un cadre de classification, on a $\forall i, y_i \in \{0, 1\}$.

Cadre de travail

On dispose de l'échantillon $(x_1, y_1), \cdots, (x_n, y_n)$, mais étant dans un cadre de classification, on a $\forall i, y_i \in \{0, 1\}$.

Cette fois, il n'est pas conseillé d'appliquer directement un modèle linéaire (car alors on pourrait avoir $\hat{y}_i \notin \{0, 1\}$)

Cadre de travail

On va se placer dans un modèle probabiliste (comme pour l'analyse discriminante linéaire) et on va chercher à estimer $P(Y = y|X = x)$ comme une combinaison linéaire des variables x.

Cadre de travail

On va se placer dans un modèle probabiliste (comme pour l'analyse discriminante linéaire) et on va chercher à estimer $P(Y = y|X = x)$ comme une combinaison linéaire des variables x.

Cependant, l'estimation de $P(Y = y|X = x)$ devra être comprise entre 0 et 1 et devra donc être modélisée par une fonctionne non-linéaire.

Fonctions logit et logistiques

Définition

On appelle *fonction logit* la fonction :

$$\text{logit} : [0,1] \to \mathbf{R}$$

$$p \mapsto \log \frac{p}{1-p}$$

Fonctions logit et logistiques

Définition

On appelle *fonction logit* la fonction :

$$\text{logit} : [0,1] \to \mathbf{R}$$
$$p \mapsto \log \frac{p}{1-p}$$

illustration

Fonctions logit et logistiques

Définition

On appelle *fonction logistique* la fonction :

$$\sigma : \mathbf{R} \to [0,1]$$
$$u \mapsto \frac{1}{1+e^{-u}} = \frac{e^u}{1+e^u}$$

Fonctions logit et logistiques

Définition

On appelle *fonction logistique* la fonction :

$$\sigma : \mathbf{R} \to [0,1]$$
$$u \mapsto \frac{1}{1+e^{-u}} = \frac{e^u}{1+e^u}$$

illustration

Fonctions logit et logistiques

Définition

On appelle *fonction logit* la fonction :

$$\begin{aligned} \text{logit} : [0,1] &\to \mathbf{R} \\ p &\mapsto \log \frac{p}{1-p} \end{aligned}$$

Définition

On appelle *fonction logistique* la fonction :

$$\begin{aligned} \sigma : \mathbf{R} &\to [0,1] \\ u &\mapsto \frac{1}{1+e^{-u}} = \frac{e^u}{1+e^u} \end{aligned}$$

Fonctions logit et logistiques

Propriéts

- La fonction logistique est la fonction réciprique de la fonction logit.
  $\text{logit}(x) = \sigma^{-1}(x)$
- la fonction logistique est une version mise à l'échelle (et décallée) de la fonction tanh
- $\sigma(-x) = 1 - \sigma(x)$

Remarques

- *logit* et *tanh* ont été utilisées comme fonction d'activation dans les réseaux de neurones

Hypothèse sur la distribution des données

Distribution des données

On fait l'hypothèse que $P(Y = 1|X = x) = \sigma(\beta_0 + \sum_{j=1}^{p} \beta_j x^{(j)})$.

Hypothèse sur la distribution des données

Distribution des données

On fait l'hypothèse que $P(Y = 1|X = x) = \sigma(\beta_0 + \sum_{j=1}^{p} \beta_j x^{(j)})$.

En posant $x^{(0)} = 1$, on peut écrire $P(Y = 1|X = x) = \sigma(\beta^\top x)$

Hypothèse sur la distribution des données

Distribution des données

On fait l'hypothèse que $P(Y = 1|X = x) = \sigma(\beta_0 + \sum_{j=1}^{p} \beta_j x^{(j)})$.

En posant $x^{(0)} = 1$, on peut écrire $P(Y = 1|X = x) = \sigma(\beta^\top x)$

Règle de décision

Une fois $\hat{\beta}$ estimé, on appliquera comme règle :

$$\hat{y} = \begin{cases} 1 & \text{si } \sigma(\beta^\top x) > \frac{1}{2} \\ 0 & \text{sinon} \end{cases}$$

log-vraisemblance

$$\log \mathcal{L}(\beta)$$
$$= \log \prod_{i=1}^{n} P(X = x_i, Y = y_i|\beta)$$

log-vraisemblance

$$\log \mathcal{L}(\beta)$$
$$= \log \prod_{i=1}^{n} P(X = x_i, Y = y_i|\beta)$$
$$= \log \prod_{i=1}^{n} P(Y = y_i|X = x_i, \beta) + \log \prod_{i=1}^{n} P(X = x_i)$$

log-vraisemblance

$$
\begin{aligned}
&\log \mathcal{L}(\beta) \\
&= \log \prod_{i=1}^{n} P(X = x_i, Y = y_i | \beta) \\
&= \log \prod_{i=1}^{n} P(Y = y_i | X = x_i, \beta) + \underbrace{\log \prod_{i=1}^{n} P(X = x_i)}_{C}
\end{aligned}
$$

log-vraisemblance

$$\log \mathcal{L}(\beta)$$

$$= \log \prod_{i=1}^{n} P(Y = y_i | X = x_i, \beta) + \underbrace{\log \prod_{i=1}^{n} P(X = x_i)}_{C}$$

$$= \sum_{i=1}^{n} \log P(Y = y_i | X = x_i, \beta) + C$$

log-vraisemblance

$$
\begin{aligned}
&\log \mathcal{L}(\beta) \\
&= \sum_{i=1}^{n} \log P(Y = y_i | X = x_i, \beta) + C \\
&= \sum_{i=1}^{n} \log \left(P(Y = 1 | X = x_i, \beta)^{y_i} \left(1 - P(Y = 1 | X = x_i, \beta)\right)^{1-y_i} \right) + C
\end{aligned}
$$

log-vraisemblance

$$\log \mathcal{L}(\beta)$$
$$= \sum_{i=1}^{n} \log \left(P(Y = 1|X = x_i, \beta)^{y_i} \left(1 - P(Y = 1|X = x_i, \beta)\right)^{1-y_i} \right) + C$$
$$= \sum_{i=1}^{n} y_i \log P(Y = 1|X = x_i, \beta) + (1 - y_i) \log \left(1 - P(Y = 1|X = x_i, \beta)\right)$$

log-vraisemblance

$$
\begin{aligned}
&\log \mathcal{L}(\beta) \\
&= \sum_{i=1}^{n} \log \left(P(Y=1|X=x_i,\beta)^{y_i} \left(1 - P(Y=1|X=x_i,\beta)\right)^{1-y_i} \right) + C \\
&= \sum_{i=1}^{n} y_i \log P(Y=1|X=x_i,\beta) + (1-y_i) \log \left(1 - P(Y=1|X=x_i,\beta)\right) \\
&= \sum_{i=1}^{n} y_i \log \sigma(\beta^\top x_i) + (1-y_i) \log \left(1 - \sigma(\beta^\top x_i)\right) + C
\end{aligned}
$$

Maximum de vraisemblance

On cherche alors :

$$\arg\max_{\beta \in \mathbf{R}^{p+1}} \sum_{i=1}^{n} y_i \log \sigma(\beta^\top x_i) + (1 - y_i) \log \left(1 - \sigma(\beta^\top x_i)\right)$$

Solution

Ce problème de maximisation est concave et admet un unique maximum global.

Maximum de vraisemblance

On cherche alors :

$$\arg\max_{\beta \in \mathbf{R}^{p+1}} \sum_{i=1}^{n} y_i \log \sigma(\beta^\top x_i) + (1 - y_i) \log \left(1 - \sigma(\beta^\top x_i)\right)$$

Solution

Ce problème de maximisation est concave et admet un unique maximum global. Cependant, il n'a pas de solution analytique/explicite.

Maximum de vraisemblance

On cherche alors :

$$\arg\max_{\beta \in \mathbf{R}^{p+1}} \sum_{i=1}^{n} y_i \log \sigma(\beta^\top x_i) + (1 - y_i) \log \left(1 - \sigma(\beta^\top x_i)\right)$$

Solution

Ce problème de maximisation est concave et admet un unique maximum global. Cependant, il n'a pas de solution analytique/explicite.

Gradient

$$\nabla_\beta \log \mathcal{L} = \sum_{i=1}^{n} \left(y_i - \frac{1}{1+e^{-\beta^\top x_i}}\right) x_i$$

Maximum de vraisemblance

On cherche alors :

$$\arg\max_{\beta \in \mathbf{R}^{p+1}} \sum_{i=1}^{n} y_i \log \sigma(\beta^\top x_i) + (1 - y_i) \log \left(1 - \sigma(\beta^\top x_i)\right)$$

Gradient

$$\nabla_\beta \log \mathcal{L} = \sum_{i=1}^{n} \left(y_i - \underbrace{\frac{1}{1 + e^{-\beta^\top x_i}}}_{\sigma(\beta^\top x_i)} \right) x_i$$

Maximum de vraisemblance

On cherche alors :

$$\arg\max_{\beta \in \mathbf{R}^{p+1}} \sum_{i=1}^{n} y_i \log \sigma(\beta^\top x_i) + (1 - y_i) \log \left(1 - \sigma(\beta^\top x_i)\right)$$

Gradient

$$\nabla_\beta \log \mathcal{L} = \sum_{i=1}^{n} \left(y_i - \underbrace{\frac{1}{1 + e^{-\beta^\top x_i}}}_{\hat{y}_i} \right) x_i$$

Résolution

Résolution numérique

On appliquera une méthode de descente de gradient pour annuler numériquement le gradient et trouver une solution.

Réécrire le gradient

Comment 'écrire le gradient en termes des matrices X et Y ?

Conclusion

LDA vs. regression logistique

LDA	**Régression logistique**
modèle *génératif* hypothèse sur la distribution de $P(X, Y)$	modèle *discriminatif* hypothèse sur la distribution de $P(Y\|X)$
paramètres à estimer : $\frac{p(p-1)}{2} + 2p \quad (\mu_1, \mu_2, \Sigma)$	paramètres à estimer : $p + 1 \qquad (\beta)$
modèle linéaire	modèle linéaire

La régression logistique est souvent plus robuste aux valeurs aberrantes et aux distributions non gaussiennes.

C'est une méthode très utilisée en pratique.

Extensions

- Une régression logistique peut être vue comme un réseau de neurones (sans couche cachée) avec la fonction logistique comme fonction d'activation.

Extensions

- Une régression logistique peut être vue comme un réseau de neurones (sans couche cachée) avec la fonction logistique comme fonction d'activation.
- on peut ajouter un terme de régularisation dans la fonction de coût pour rendre le modèle plus robuste.