Machine Learning IX : Naive Bayes

Nicolas Bourgeois

1. Modèle Y=f(X)
2. Modèle génératif
3. Application en Lexicométrie

Rappels de Notations

Variables aléatoires (X, Y) à valeurs dans $E \times G$
Dissimilarité $d : E^2 \to \mathbb{R}+$
Fonction de perte $LF : G^2 \to \mathbb{R}+$
Expériences $\tau = (X_i, Y_i)_{i \leq n} \in (E \times G)^n$
Risque du modèle $D(g) = \mathbb{E}\left(LF(g(X), Y)\right)$
Risque optimal / de Bayes $ROPT = \min_g \mathbb{E}\left(LF(g(X), Y)\right)$
Risque empirique $\tilde{D}(g, \tau) = \frac{1}{n} \sum_{j \leq n} LF(g(X_j), Y_j)$

Objectif (théorique)

$ROPT = \min_g \mathbb{E}\left(LF(g(X), Y)\right)$

Trouver g^* tel que $D(g^*) = ROPT$

Objectif (théorique)

$ROPT = \min_g \mathbb{E}\left(LF(g(X), Y)\right)$

Trouver g^* tel que $D(g^*) = ROPT$

Facile (dans le cas G discret) :

$g^* : x \mapsto \arg\min_{y \in G} \sum_{y' \neq y} LF(y, y')\mathbb{P}(Y = y' \mid X = x)$

Objectif (théorique)

$ROPT = \min_g \mathbb{E}\left(LF(g(X), Y)\right)$

Trouver g^* tel que $D(g^*) = ROPT$

Facile (dans le cas G discret) :

$g^* : x \mapsto \arg\min_{y \in G} \sum_{y' \neq y} LF(y, y')\mathbb{P}(Y = y' \mid X = x)$

Où est le piège ?

Exercice

Exercice

On pioche un dé dans un sac contenant certains à 6 faces et d'autres à 10 faces, puis on jette celui-ci et on lit le résultat obtenu. Quel sont les estimateurs optimaux et quel est le risque optimal pour une fonction de perte uniforme ?

Exercice

Qu'en est-il si pour chaque dé la probabilité est croissante avec la valeur ?

Exercice

Exercice

A partir du tableau d'observations ci-dessous quel estimateur $\tilde{f}$ avez-vous a priori envie de construire ?

	1	2	3	4	5	6	7	8	9	10
1	12	11	12	13	12	11	0	0	0	0
2	7	7	8	5	13	7	8	6	8	7

Question subsidiaire : pourquoi tout ceci est-il très décevant ?

Hypothèse alternative

E et G sont des espaces finis.

(X, Y) sont des variables aléatoires non indépendantes, et on veut caractériser cette dépendance.

Hypothèse alternative

E et G sont des espaces finis.

(X, Y) sont des variables aléatoires non indépendantes, et on veut caractériser cette dépendance.

Cette dépendance ne peut pas être caractérisée par une simple fonction $Y = f(X)$ on cherche plutôt $P(Y \mid X)$

L'espace des probabilités est trop vaste pour une simple optimisation.

exercice

Considérez un espace où X comporte p champs, chacun possédant m modalités. Y possède m' modalités. On suppose qu'on discrétise l'ensemble des scalaires pour se ramener à k possibilités.

Exercice

Si on se restreignait aux lois $Y = f(X)$, quelle serait la taille de l'espace des fonctions ?

Exercice

Si on cherche un modèle joint (X, Y), quelle est la taille de l'espace des distributions possibles ?

Réduction de complexité

$$\mathbb{P}(X, Y) = \mathbb{P}(X_1 \mid X_{2\ldots n}, Y)\mathbb{P}(X_2 \mid X_{3\ldots n}, Y)\ldots\mathbb{P}(X_n \mid Y)\mathbb{P}(Y)$$

Réduction de complexité

$$\mathbb{P}(X, Y) = \mathbb{P}(X_1 \mid X_{2\ldots n}, Y)\mathbb{P}(X_2 \mid X_{3\ldots n}, Y)\ldots\mathbb{P}(X_n \mid Y)\mathbb{P}(Y)$$

Hypothèse (forte) d'indépendance :

$$\mathbb{P}(X, Y = k) = \prod_{j \leq p} \mathbb{P}(X_j \mid Y = k)\mathbb{P}(Y = k)$$

D'où

$$\mathbb{P}(Y = k \mid X) = \frac{1}{P_\Omega} \prod_{j \leq p} \mathbb{P}(X_j \mid Y = k)\mathbb{P}(Y = k)$$

Notation complète

Distribution catégorique :

$$\mathbb{P}_{X\sim C(\gamma)} : \mathbb{P}(X = x) = \gamma_x$$

Distribution bayésienne naïve :

$$\mathbb{P}_{(X,Y)\sim NB(\gamma)}(X = x, Y = k) =$$

$$\mathbb{P}_{Y\sim C(\gamma)}(Y = k) \prod_{j\leq p} \mathbb{P}_{X_j|(Y=k)\sim C(\gamma)}(X_j = x_j \mid Y = k)$$

Minimiseur du reste empirique

Pour :

$$\tilde{D}(\tilde{f},\tau) = \frac{1}{n}\sum_{j\leq n} LF(\tilde{f}(X_j), Y_j)$$

On peut prendre :

$$\tilde{F} = \arg\min_{\tilde{f}} \tilde{D}(\tilde{f},\tau)$$

Minimiseur du reste empirique

Cas bayésien naïf :

$$\forall x, \tilde{F}(x) = \arg\max_{k \in G} \mathbb{P}(Y = k) \prod_{j \leq p} \mathbb{P}(X_j \mid Y = k)$$

Minimiseur du reste empirique

Cas bayésien naïf :

$$\forall x, \tilde{F}(x) = \arg\max_{k \in G} \mathbb{P}(Y = k) \prod_{j \leq p} \mathbb{P}(X_j \mid Y = k)$$

$$= \arg\max_{k \in G} \log \mathbb{P}(Y = k) + \sum_{j \leq p} \log \mathbb{P}(X_j \mid Y = k)$$

exercice

On considère les tableaux d'observations ci-dessous :

Y=True	X=True	X=False
X1	17	43
X2	31	29
X3	11	49
X4	40	20

Y=False	X=True	X=False
X1	1	24
X2	20	5
X3	11	14
X4	3	22

Quelle serait la meilleure estimation bayésienne associée à l'observation *True,True,False,False* ?

Présentation

Soit un vocabulaire $\mathcal{W}$ et un ensemble de documents $\mathcal{X}$, une observation est une distribution avec $x_w = \mid \{w \in x\} \mid$. On suppose que les documents sont répartis en K classes (inconnues).

Présentation

Soit un vocabulaire $\mathcal{W}$ et un ensemble de documents $\mathcal{X}$, une observation est une distribution avec $x_w = \mid \{w \in x\} \mid$. On suppose que les documents sont répartis en K classes (inconnues).

Modèle multinomial :

$$\forall w \in \mathcal{W}, \forall k \in 1 \ldots K, \forall x \in C_k, \mathbb{P}(X_w = n) = p_{k,w}^n$$

Présentation

Soit un vocabulaire $\mathcal{W}$ et un ensemble de documents $\mathcal{X}$, une observation est une distribution avec $x_w = \mid \{w \in x\} \mid$. On suppose que les documents sont répartis en K classes (inconnues).

Modèle multinomial :

$$\forall w \in \mathcal{W}, \forall k \in 1 \dots K, \forall x \in C_k, \mathbb{P}(X_w = n) = p_{k,w}^n$$

Pourquoi l'estimateur bayésien semble-t-il adapté ?

Estimateur

$$\mathbb{P}(X = x \mid Y = k) = \frac{(\sum_w x_w)!}{\prod_w x_w!} \prod_w p_{k,w}^{x_w}$$

Estimateur

$$\mathbb{P}(X = x \mid Y = k) = \frac{(\sum_w x_w)!}{\prod_w x_w!} \prod_w p_{k,w}^{x_w}$$

$$\mathbb{P}(Y = k \mid X = x) = \frac{\mathbb{P}(Y = k)}{P_\Omega} \frac{(\sum_w x_w)!}{\prod_w x_w!} \prod_w p_{k,w}^{x_w}$$

Estimateur

$$\mathbb{P}(X = x \mid Y = k) = \frac{(\sum_w x_w)!}{\prod_w x_w!} \prod_w p_{k,w}^{x_w}$$

$$\mathbb{P}(Y = k \mid X = x) = \frac{\mathbb{P}(Y = k)}{P_\Omega} \frac{(\sum_w x_w)!}{\prod_w x_w!} \prod_w p_{k,w}^{x_w}$$

$$\log \mathbb{P}(Y = k \mid X = x) = c + \log \mathbb{P}(Y = k) + \sum_w \left(\log p_{k,w} \times x_w\right)$$

exercice

Entrainez (à la main) un naive Bayes sur les données du titanic en vous restreignant aux champs *pClass* et *sex* pour expliquer *survived*.

solution

```
import pandas as pd
df = pd.read_csv("data2.csv")[['sex','pclass','survived']]
pX_Y,sx,cl = [],['male','female'],[1,2,3]
for i in range(2):
    pX_Y.append([[df[(df.survived==i) & (df.pclass==p)].shape[0]
        * df[(df.survived==i) & (df.sex==s)].shape[0]
        / (df[df.survived==i].shape[0]*df.shape[0]) for p in cl]
         for s in sx])
for s in range(2):
    for p in range(3):
        print (sx[s],cl[p],"Survit" if pX_Y[1][s][p]>pX_Y[0][s][p]
            else "Disparait", pX_Y[1][s][p],"vs",pX_Y[0][s][p])
```

Exercice

Proposez des exemples pour lesquels l’hypothèse d’indépendance est trop forte, entraînant une faiblesse de l’estimateur bayésien.