Cours 4 – LA METHODE NAIVE BAYES

Présentation

Méthode de classification supervisée

Callcul de probabilités conditionnelles pour la prédiction

Utilisation du théorème de Bayes

Hypothèse de l'indépendance des variables explicatives

=> Méthode qualifiée de naïve

Description de la méthode

Echantillon d'apprentissage

variables explicatives Xj

variable à expliquer Y

Y(ωi) est égal à yi pour chaque individu ωi

	X1	...	Xj	...	XP	Y
ω1	X1(ω1)	...	Xj(ω1)	...	XP(ω1)	y1
ω2	X1(ω2)	...	Xj(ω2)	...	XP(ω2)	y2
...	...	...	...	...	...	...
ωi	X1(ωi)	...	Xj(ωi)	...	XP(ωi)	yi
...	...	...	...	...	...	...
ωN	X1(ωN)	...	Xj(ωN)	...	XP(ωN)	yN

Problème

Pour un nouvel individu ω

prédire (trouver) sa classe y , c'est à dire Y(ω).

ω	X1(ω)	...	Xj(ω)	...	XP(ω)	?

On note

X =(X1, … , Xj , …, XP)

X(ω) =(X1(ω), … , Xj(ω) , …, XP(ω))

On veut trouver y, c'est-à-dire Y(ω) étant donné X(ω)

Solution

Ensemble C des classes distinctes

C = { y1 , y2 , … , yL }, avec L ≤ N

La classe y à trouver est la classe yk ∈ C qui a la plus grande probabilité d'être la classe de l'individu ω

y = argmax $_{yk \in C}$ P(yk | X(ω))

Théorème de Bayes

$$P(yk/X(\omega)) = \frac{P(X(\omega)/yk)\,P(yk)}{P(X(\omega))}$$

même dénominateur P(X(ω)) quelle que soit la classe yk

=> Comparer les numérateurs P(X(ω)| yk) P(yk)

Calcul de P(yk)

P(yk) est la probabilité a priori de la classe yk

= proportion d'individus de la classe yk

= Nk / N

avec

Nk le nombre d'individus dans la classe yk

N le nombre d'individus

Calcul de P(X(ω) | yk)

Hypothèse : indépendance des variables Xj

=> P(X(ω)|yk)=P(X1(ω)|yk)×P(X2(ω)| yk) ×…× P(XP(ω)| yk)

Pour une variable qualitative Xj

P(Xj(ω) | yk)

= probabilité d'avoir la valeur Xj(ω) étant donnée la classe yk

= proportion d'individus de la classe yk pour lesquelles Xj a pour valeur Xj(ω)

PourXj variable quantitative

Xj : variable aléatoire avec une distribution gaussienne

P(Xj(ω) | yk) est la fonction de densité de probabilité

$$P(X_j(\omega)|y_k)=\frac{1}{\sqrt{2\Pi}\sigma(X_j|y_k)}\exp\left(\frac{-(X_j(\omega)-\mu(X_j|y_k))^2}{2\sigma(X_j|y_k)^2}\right.$$

où :

μ(Xj|yk) est la moyenne de Xj dans la classe yk

$$\mu(X_j|y_k)=\frac{1}{n_k}\sum_{i=1}^{n_k}X_j(\omega_i)$$

σ (Xj |yk) est la déviation standard (écart-type) de Xj dans la classe yk

$$\sigma(X_j|y_k)=\sqrt{\frac{1}{n_k-1}\sum_{i=1}^{n_k}(X_j(\omega_i)-\mu(X_j|y_k))^2}$$

Exemple d'application

Soit le tableau suivant regroupant des données sur la possibilité de jouer au tennis en fonction de la météo.

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Prédire la valeur de la variable Play pour un nouvel exemple ω, avec X(ω) = (sunny ,74, 66 , true) regroupant les valeurs des variables Outlook, Temperature, Humidity et Windy respectivement.

Solution

Comparer P(yes | X(ω)) et P(no | X(ω))

Ceci revient à comparer P(X(ω) | yes)×P(yes) et P(X(ω) | no)×P(no)

P(yes) = 9/14

P(no) = 5/14

P(X(ω)|yes) =P(Outlook=sunny,Temperature=74,Humidity=66,Windy=true| yes)

=P(Outlook=sunny | yes)×P(Temperature=74 | yes)

×P(Humidity=66 | yes)×P(Windy=true | yes)

P(Outlook=sunny | yes) = 2/9

$$\mu(Temperature|yes)=73$$

$$\sigma(Temperature|yes)=6.16$$

$$P(Temperature=74|yes)=\frac{1}{\sqrt{2\Pi}6.16}\exp\left(\frac{-(74-73)^2}{2\times 6.16^2}\right)=0.064$$

$$\mu(Humidity|yes)=79.1$$

$$\sigma(Humidity|yes)=10.2$$

$$P(Humidity=66|yes)=\frac{1}{\sqrt{2\Pi}10.2}\exp\left(\frac{-(66-79.1)^2}{2\times 10.2^2}\right)=0.045$$

P(Windy=true | yes) = 3/9

Donc P(X(ω) | yes)×P(yes) = (2/9) × 0.064 × 0.045 × (3/9) × (9/14) = 0.64

P(X(ω) | yes)×P(yes) = 0.64

Les mêmes calculs donnent pour la classe no :

P(X(ω) | no) ×P(no) = 0.00027

P(X(ω)|yes)×P(yes) est supérieure à P(X(ω)|no)×P(no)

donc P(yes | X(ω)) est supérieure à P(no | X(ω))

donc Play = yes pour X(ω) = (sunny ,74, 66 , true)