Apprentissage par réseau Bayésien Naif

. La ***classification naïve bayésienne*** *est un type de classification bayésienne probabiliste simple basée sur le Théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses.*

. Elle met en œuvre un classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des classifieurs linéaires.

. Un terme plus approprié pour le modèle probabiliste sous-jacent pourrait être « modèle à caractéristiques statistiquement indépendantes ».

. En termes simples, un *classifieur bayésien naïf* suppose que l'existence d'une caractéristique pour une classe, est *indépendante de l'existence d'autres caractéristiques.*

Exemple : Un fruit peut être considéré comme une *pomme* s'il est *rouge*, *arrondi*, et fait une *dizaine de centimètres*. Même si ces caractéristiques sont liées dans la réalité, un classifieur bayésien naïf déterminera que le fruit est une pomme en considérant indépendamment ces caractéristiques de *couleur*, de *forme* et de *taille*.

. Malgré leur modèle de conception « naïf » et ses hypothèses de base extrêmement simplistes, les classifieurs bayésiens naïfs ont fait preuve d'une efficacité plus que suffisante dans beaucoup de situations réelles complexes.

Le Modèle bayésien naïf

Rappel du principe d'un classifieur supervisé.

. Etant donnée une nouvelle entrée X (n'ayant pas servie durant la phase d'entrainement) et définie par des caractéristiques X_1, X_2, ...X_n.

. Le classifieur tente de classer (avec une forte probabilité) l'appartenance de l'entrée X à une classe c_i particulière de l'ensemble C de toutes les classes. La classe c_i étant le label associé à X.

. Le modèle probabiliste pour un classifieur Bayésien naif est le modèle conditionnel suivant :

$P(C/X_1, X_2, ...X_n)$ *et qui correspond au calcul de la probabilité d'appartenir à une classe C sachant qu'on a observé une entrée X définie par les caractéristiques* X_1, X_2, ...X_n.

. Lorsque le nombre de caractéristiques n est grand, baser ce modèle sur des tableaux de probabilités devient impossible. Par conséquent, nous le dérivons pour qu'il soit plus facilement soluble.

À l'aide de la règle de Bayes, on peut écrire :

$P(C/ X_1, X_2, ...X_n) = P(C)*P(X_1, X_2, ...X_n/C)) / P(X_1, X_2, ...X_n)$

. En pratique, seul le numérateur nous intéresse, puisque le dénominateur ne dépend pas de C et les valeurs des caractéristiques Xi sont données.

. Le dénominateur est donc en réalité constant. Le numérateur est soumis à la loi de probabilité à plusieurs variables et peut être factorisé de la façon suivante, en utilisant plusieurs fois la définition de la probabilité conditionnelle :

$P(C)*P(X_1, X_2, ...X_n/C)$ =

$P(C)*P(X1/C)*P(X_2, ...X_n/C,X1)$

=

$P(C)*P(X1/C)*P(X2/C,X1)*P(X_3, ...X_n/C,X1,X2)$

…..

=

$P(C)*P(X_1/C)*P(X_2/C,X_1)*P(X_3/C,X_1,X_2)*..P(X_n/C,X_1,X_2,...X_{n-1})$

. C'est là que nous faisons intervenir l'hypothèse naïve :

. Si chaque Xi est indépendan des autres caractéristiques X$j \neq i$, conditionnellement à C alors :

$P(X_i/C,X_j)=P(X_i/C)$ pour tout $j \neq i$,

. Par conséquent la probabilité conditionnelle peut s'écrire

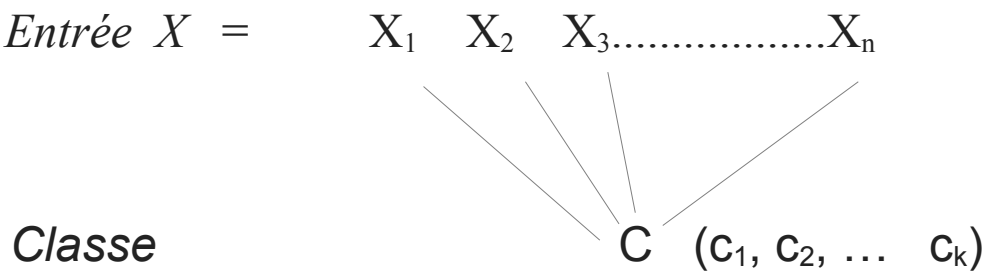
$P(X_1,X_2,...X_n/C) = P(X_1/C)*P(X_2/C)*P(X_3/C)*...P(X_n/C) = \Pi P(X_i/C)$

. Par conséquent, en tenant compte de l'hypothèse d'indépendance ci-dessus, la probabilité conditionnelle de la variable de classe C peut être exprimée par

$$P(C/ X_1, X_2, ...X_n) = 1/Z*P(C)* \Pi P(X_i/C)$$

où Z est une constante (appelé « évidence ») qui est un facteur d'échelle qui dépend uniquement de $X_1,...,Xn$.

. Le modèle du ***classifieur Bayésien naif*** peut être assimilé à un réseau Bayésien défini par le schéma suivant :



Estimation de la valeur des paramètres

Cas de paramètres discrets

. Tous les paramètres du modèle peuvent faire l'objet d'une approximation par rapport aux *fréquences relatives des classes et caractéristiques dans l'ensemble des données d'entraînement.*

. Soit D = base de données d'apprentissage = ensemble d'éléments de D de la forme (X,C)

. Formules de calcul des paramètres du modèle du réseau :

P(C=ci) *= (nombre d'éléments de D tel que C=ci))/(nombre total d'élements de D)*

P(X_i=v/C=c_j) *= (nombre d'éléments de D tel que X_i=v et C=c_j))/(nombre d'éléments de D tel que C=ci)*

Cas de paramètres continus

Lorsque l'on travaille avec des caractéristiques qui sont des variables aléatoires continues, on suppose généralement que les lois de probabilités correspondantes sont des lois normales, dont on estimera l'espérance et la variance.

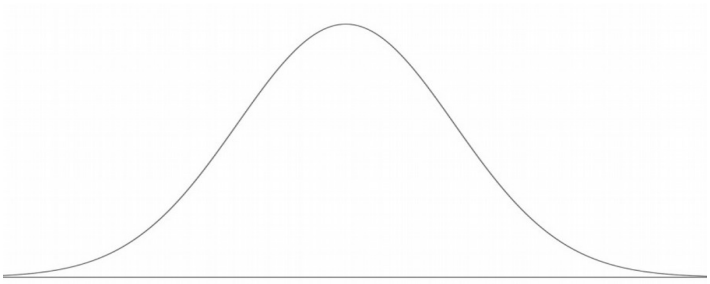
Rappel sur la loi normale

En théorie des probabilités et en statistique, les **lois normales** sont parmi les lois de probabilité les plus utilisées pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires.

Plus formellement, une loi normale est une loi de probabilité continue qui dépend de deux paramètres : **son espérance**, un nombre réel noté μ, et **son écart type**, un nombre réel positif noté σ. La densité de probabilité de la loi normale d'espérance μ, et d'écart type σ est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

La courbe de cette densité est appelée *courbe de Gauss* ou *courbe en cloche*, entre autres



Lorsqu'une variable aléatoire X suit une loi normale, elle est dite *gaussienne* ou *normale* et il est habituel d'utiliser la notation avec la variance $\sigma 2$

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

L'espérance μ se calcule avec

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i,$$

où N est le nombre d'échantillons et xi est la valeur d'un échantillon donné.

La variance $\sigma 2$ se calcule avec:

$$\sigma^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \mu)^2.$$

Construction du classifieur

. Le classifieur bayésien naïf se construit en appliquant une règle de décision couramment employée : la règle du *maximum a posteriori* définie par :

Soit une donnée ***X = (X1=x1, X2=x2, ...Xn=xn)*** à classer.

Alors

Classifieur $(x_1,x_2,...x_n)$ = *ArgMax* $P(C=c_j) \prod P(X_i = x_i/C=c_j)$

Analyse de la méthode

. Malgré les hypothèses d'indépendance relativement simplistes, le classifieur bayésien naïf a plusieurs propriétés qui le rendent très pratique dans les cas réels.

. En particulier, la dissociation des lois de probabilités conditionnelles de classe entre les différentes caractéristiques aboutit au fait que chaque loi de probabilité peut être estimée indépendamment en tant que loi de probabilité à une dimension.

Cela permet d'éviter nombre de problèmes venant du fléau de la dimension.