Machine Learning

Méthode Naïve Bayes :

- Formule de Bayes
- Hypothèse d'indépendance
- Cas discret
- Cas continu

Formalisme mathématique

Comme son nom l'indique, la méthode repose sur la formule de Bayes. Il est donc nécessaire de définir un modèle probabiliste.

Notons $X = (X_1, \dots, X_p)$ le vecteur aléatoire constitué des variable aléatoires explicatives $X_1, \dots, X_p$ et Y la variable aléatoire cible.

Soit un individu ω, alors $X(\omega)$ et $Y(\omega) = y_k, k \in \{1, \dots, K\}$ sont les réalisations du vecteur X et de la variable Y pour cet individu.

Connaissant $X(\omega)$ la réalisation du vecteur X pour un individu ω, l'objectif de la méthode est de prédire la réalisation $Y(\omega)$ la plus probable . Ce qui s'écrit mathématiquement sous la forme :

$$\hat{Y}(\omega) = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} P(Y = y_k | X(\omega))$$

« On cherche k qui maximise …»

« la probabilité que ω soit dans la classe y_k sachant ses valeurs pour X …»

Considérons l'exemple de la classification du sexe en fonction de la taille, du poids et de la pointure. Alors X_1 est la variable aléatoire représentant la taille, etc.. et Y est la variable cible à 2 modalités {garçon,fille}.

Par exemple pour un individu ω=« Jean », alors on peut avoir X(ω)=(185,95,45) et Y(ω)=garçon.

Entre les classes fille et garçon, on choisit celle qui maximise

P(Y=fille|taille,poids,pointure)

P(Y=garçon|taille,poids,pointure)

Formule de Bayes

En appliquant la formule de Bayes [1], on obtient

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

P(Y=fille|taille,poids,pointure) =
P(Y=fille)× P(taille,poids,pointure|Y=fille) / P(taille,poids,pointure)

Etant donné que le dénominateur ne dépend pas de k, le problème revient donc à

$$\hat{Y}(\omega) = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \quad P(Y = y_k) \times P(X|Y = y_k)$$

La probabilité $P(Y = y_k)$ est facile à estimer, il suffit d'utiliser les fréquences observées sur l'échantillon [2],

$$\hat{P}(Y = y_k) = \frac{n_k}{n}$$

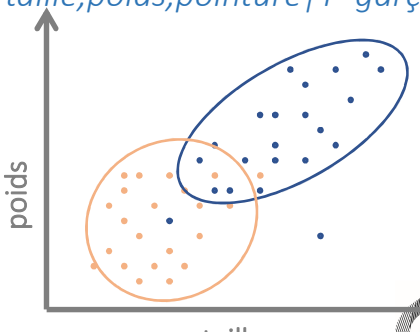
où n_k est l'effectif de la classe k et n l'effectif total.

S'il y a 45 filles et 55 garçons dans l'échantillon alors, P(Y=fille)≈0.45 et P(Y=garcon)≈0.55

P(taille,poids,pointure|Y=fille)?
P(taille,poids,pointure|Y=garçon)?

Tout le problème est donc d'estimer la probabilité conditionnelle, $P(X|Y = y_k)$

Quelle est la distribution de X à l'intérieur de chaque classe?

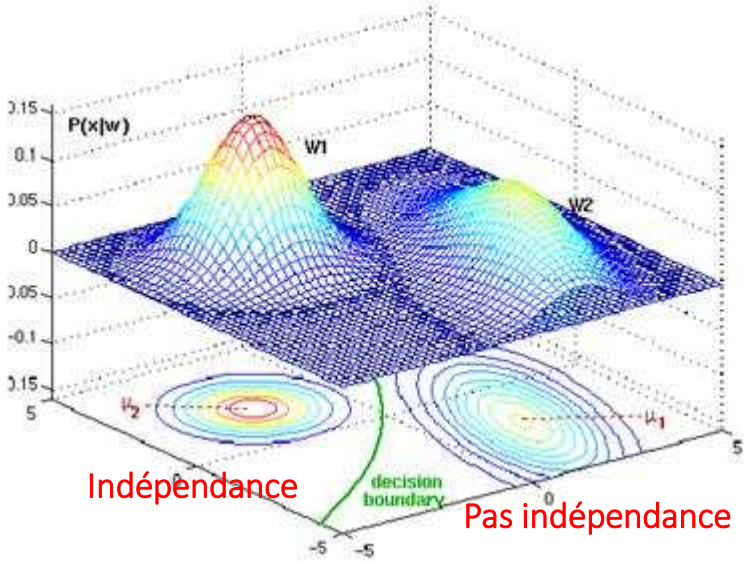


CERGY PARIS UNIVERSITÉ

[1] Formule de Bayes : P(A|B)=P(A∩B)/P(B)=P(A)×P(B|A)/P(B)

Les différentes méthodes

Les méthodes se différencient par la façon dont sont estimées les probabilités conditionnelles $P(X|Y=y_k)$.

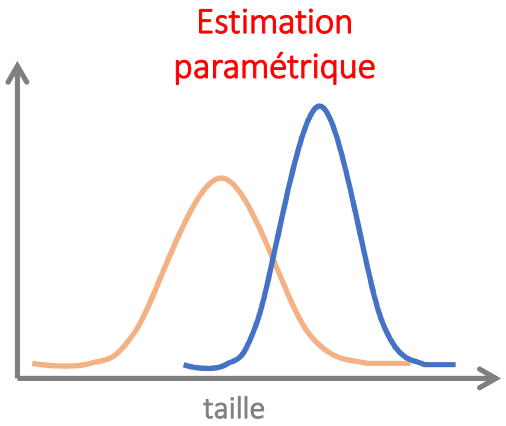


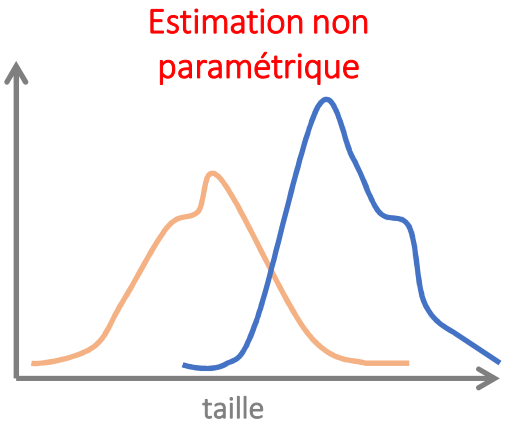
Avec deux types d'hypothèses :

Indépendance des variables X_i conditionnellement à Y
(Naive Bayes, certaines méthodes à noyaux)

Loi usuelle a priori sur X conditionnellement à Y :
loi normale(multi), binomiale,…

- Oui = méthodes paramétriques *(Naive Bayes, analyse discriminante)*
  Hypothèse forte / uniquement les paramètres de la loi à estimer
- Non = méthodes non paramétriques *(PPV, Noyaux)*
  Toute forme de distribution mais très sensible au paramétrage (bandwith)





Naïve Bayes

La méthode du classifieur bayésien naif repose sur les deux hypothèses :

1) L'indépendance des variables X_i conditionnellement à Y
2) Une distribution a priori de ces variables conditionnelles :

   Modèle uniforme en discret et modèle gaussien en continu

On suppose que les variables conditionnelles, $X_i|Y$, i=1,…,p sont indépendantes [3], d'où

$$P(X|Y = y_k) = \prod_{i=1}^{p} P(X_i|Y = y_k)$$

[3] Si A et B indépendants alors $P(A \cap B)=P(A)\times P(B)$

Le problème revient donc à estimer la distribution de chaque variable

$$X_i|Y = y_k$$

pour i=1,…,p et k=1,…,K.

Naïve Bayes : Estimation de la distribution Cas discret

Notons x_{ij} les valeurs/modalités prises par la variable X_i. Naturellement, on estime la probabilité conditionnelle par

$$\hat{P}(X_i = x_{ij} | \quad Y = y_k) = \frac{n_{ijk}}{n_k}$$

Cependant, le risque que $n_{ijk}=0$ est important. On préférera donc la version lissée (2)

$$\hat{P}(X_i = x_{ij} | \quad Y = y_k) = \frac{n_{ijk+1}}{n_k + 1 + K}$$

Exemple : Control fiscal

salaire	impôts	étudiant	contrôle
< 30	< 20 %	oui	négatif
30 - 50	< 20 %	non	positif
30 - 50	< 20 %	oui	positif
30 - 50	> 20 %	non	négatif
> 50	< 20 %	non	positif
35	6 %	oui	?

Faut-il contrôler Mr X?

Il faut classifier le nouvel individu

X = (sal=35, imp=6%,étu=oui)

et donc comparer

pp_+=P(control=pos|X) et pp_-= P(control=nég|X)

On a

pp_+ = P(control=pos)×P(sal=[30-50]|pos)×P(imp<20%|pos)×P(étu=oui|pos)
=(3/5)*(2/3)*(1)*(1/3)=0.13

pp_- = P(control=nég)×P(sal =[30-50]|nég)×P(imp<20%|nég)×P(étu=oui|nég)
=(3/5)*(1/2)*(1/2)*(1/2)=0.05

Donc on contrôle!!!!

NB les calculs ont été faits sans lissage pour simplification

Naïve Bayes : Estimation de la distribution Cas continu (1/2)

Dans le cas où les X_i sont continus, on suppose alors les variables conditionnelles suivent une loi normale

$$X_i|Y = y_k \sim N(\mu_{ik}, \sigma_{ik}^2)$$

dont on connait la fonction de densité,

$$f_{X_i|Y=y_k}(x_i) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{x_i - \mu_{ik}}{\sigma_{ik}}\right)^2$$

Les paramètres sont estimés de façon classique par la moyenne et la variance du sous-échantillon défini par la contrainte $Y = y_k$.

Sexe	Taille	Poids
M	182	81.6
M	180	86.2
M	170	77.1
M	180	74.8
F	152	45.4
F	168	68.0
F	165	59.0
F	175	68.0

	Taille		Poids	
Sexe	Moyenne	Variance	Moyenne	Variance
M	178	29.33	79.92	25.48
F	165	92.67	60.1	114.04

Pour un nouvel individu dont on connait la taille et le poids, on prédira le sexe en choisissant le max entre

pp_M = P(M) * P(taille|M) * P(poids|M)
pp_F = P(F) * P(taille|F) * P(poids|F)

On peut estimer P(Y=M)=P(Y=F)=1/2. Quant aux lois conditionnelles, on a par exemple :

Taille|F ~ N(165;92.67)

$$f_{Taille|M}(t) = \frac{1}{\sqrt{29.33 * 2\pi}} \exp -\frac{1}{2}\frac{(t-178)^2}{29.33}$$

Taille|M ~ N(178;29.33)

$$f_{Taille|F}(t) = \frac{1}{\sqrt{92.67 * 2\pi}} \exp -\frac{1}{2}\frac{(t-165)^2}{92.67}$$

Naïve Bayes : Estimation de la distribution Cas continu (2/2)

Le classifieur devient donc

$$\hat{Y}(\omega) = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \quad P(Y = y_k) \times \prod_{i=1}^{p} f_{X_i|Y=y_k}(x_i)$$

$$= \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \quad P(Y = y_k) \times \prod_{i=1}^{p} \frac{1}{\sigma_{ik}\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{x_i - \mu_{ik}}{\sigma_{ik}}\right)^2$$

On note qu'il y a un effet quadratique des X_i.

$$pp_M(t,p) = \frac{1}{2} * \frac{1}{\sqrt{29{,}33*2\pi}} \exp -\frac{1}{2}\frac{(t-178)^2}{29{,}33} * \frac{1}{\sqrt{25{,}48*2\pi}} \exp -\frac{1}{2}\frac{(p-79{,}92)^2}{25{,}48}$$

$$pp_F(t,p) = \frac{1}{2} * \frac{1}{\sqrt{92{,}67*2\pi}} \exp -\frac{1}{2}\frac{(t-165)^2}{92{,}67} * \frac{1}{\sqrt{114{,}04*2\pi}} \exp -\frac{1}{2}\frac{(p-60{,}01)^2}{114{,}04}$$

	t=188 et p=83	t=152 et p=49
pp_M	**$2{,}8*10^{-4}$**	$1{,}04*10^{-16}$
pp_F	$2{,}8*10^{-5}$	**$0{,}9*10^{-4}$**

Naïve Bayes : Bilan

Avantages

- ✓ Facile à interpréter (pas une boite noire)
- ✓ La méthode est robuste à l'hypothèse d'indépendance des variables conditionnelles
- ✓ Imbattable en temps de calcul donc parfait pour les bigdata
- ✓Gère les valeurs manquantes

Inconvénients

- ✓ Séparateur quadratique
- ✓ Modèle pouvant être trop simple pour des problèmes complexes
- ✓ Dans le cas continu, l'estimation de la densité reste problématique

La méthode naïve Bayes se décline en plusieurs méthodes suivant la façon dont est estimée cette densité :

- L'analyse discriminante qui suppose toujours une distribution gaussienne des variables mais qui introduit une corrélation entre elles
- Les méthodes des k plus proches voisins (ou k-nearest neighbors) et des noyaux qui ne pré-supposent pas de forme pour la distribution des variables (pas d'hypothèse gaussienne)

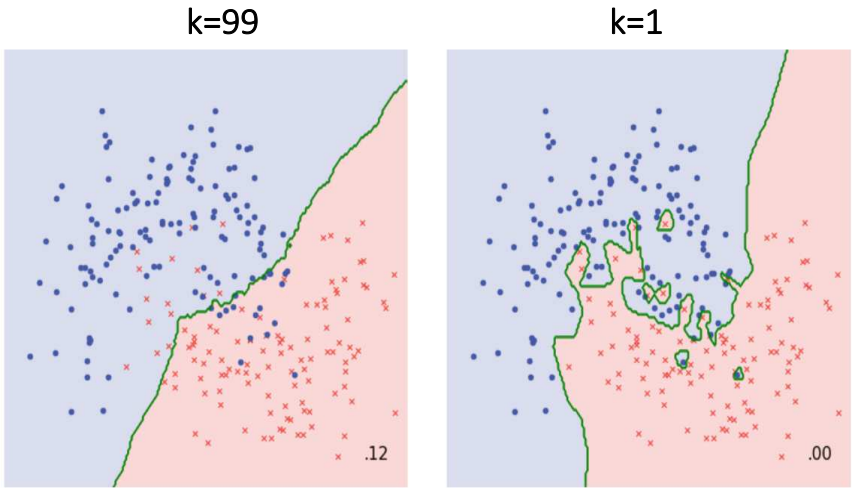
Méthode des plus proches voisins

La classification par méthode des k-plus-proches-voisins ou kNN se fait de manière très simple :

Etant donné un exemple de test x, on affecte à x la classe la plus souvent représentée parmi les k observations qui lui sont le plus proches.

La méthode repose sur le choix :

- d'une distance (cf. cours sur non supervisé)
- le nombre de voisins (attention au sur-apprentissage)



✓La méthode est sensible aux variables non pertinentes car elles ont le même poids que les autres.

✓Lorsque la dimension augmente, les points sont de moins en moins proches (cf. cours sur fléau de la dimension). On peut se poser la question de la pertinence des k points choisis par l'algorithme.

✓L'apprentissage est simple, rapide et facilement parallélisable. En revanche la méthode ne fournit par de modèle et la prédiction nécessite beaucoup de temps et de place mémoire.

Questions?

Documents ayant servis à la rédaction des slides et TD :

- *Ricco Rakotomalala : http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Naive_Bayes_Continuous_Predictors.pdf*
- *Guillaume Obozinski (aspects mathématiques) : http://www.di.ens.fr/~fbach/courses/fall2010/cours9.pdf*
- *Supports de cours de Marie Chavent (université de Bordeaux)*