Big Data Analytics

Naive Bayes

Rappels - Probabilités

Rappels :

- Une variable aléatoire
  $A \in \mathbb{R}$
- Probabilité
  $0 \leq P(A \in [a_1\ a_2]) \leq 1$
- Probabilité conditionnelle
  $P(\ A > 0 \mid B < -3\ )$
- Évènements indépendants
  $P(A|B) = P(A)$ et $P(B|A) = P(B)$
- Probabilité jointe
  $P(A, B) = P(B|A) * P(A)$
  $P(A, B) = P(A|B) * P(B)$
- A et B Indépendants
  $\iff P(A, B) = P(A) * P(B)$

Théorème de Bayes

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Naive Bayes

Prenons un exemple :
Soit une base de données de fruits contenant uniquement des bananes, oranges, tomates.
Chaque élément possède des caractéristiques couleur, taille, sucré.
Appliquer Naive Bayes, c'est chercher le maximum de vraisemblance d'un élèments dont on ne connait pas la nature mais dont on connait les caractéristiques.
On cherche donc quelle est la plus grande probabilité :

- P(banane | jaune,long,sucré)
- P(orange | jaune,long,sucré)
- P(tomate | jaune,long,sucré)

Naive Bayes

Naive = toutes les variables sont considérée indépendantes, donc :

$$P(banane|jaune, long, sucre) =$$

$$\frac{P(jaune|banane)*P(long|banane)*P(sucre|banane)*P(banane)}{P(jaune)*P(long)*P(sucre)}$$

Pour estimer les différentes probabilités, on 'compte' dans notre base de donnée de fruits :

$$P(sucre|banane) = \frac{card(banane\ ET\ sucre)}{card(banane)}$$