Bayes

November 3, 2020

1 Proba conditionnelles et Bayes

On note -

- $P(A)$ la probabilité d'un événement A
- $P(A \cap B)$ ou $P(A, B)$ la probabilité d'avoir à la fois les événements A et B.
- $P(A \mid B)$ la probabilité d'avoir l'évenement A sachant B.

La définition des probabilités conditionnelles est :

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

Donc on a aussi

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

Du coup

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

et le théorème de Bayes

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

2 Un peu de notations

A et B sont des événements. Je fais maintenant un abus de notation : je vais écrire $P(\vec{x})$ pour dire que j'ai un variable aléatoire X qui prend comme valeur $\vec{x}$ donc l'événement $X = \vec{x}$ et je note donc $P(X = \vec{x})$ simplement par $P(\vec{x})$.

Imaginons que X soit la variable qui sur deux attributs x_1 et x_2 peuvent prendre uniquement des valeurs binaires. Donc en gros $P(X)$ c'est soit $P(X = (0,0))$, $P(X = (0,1))$, etc.. que je note simplement $P(0,0)$, $P(0,1)$, etc...

Je vais noter aussi $P(\vec{x})$ ou $P(x_1, x_2)$ quand je veux parler d'un de ces 4 cas.

3 Revenant à l'apprentissage

Maintenant on sait que les données sont générées par une proba jointe d'avoir une description des données $\vec{x}$ et une classe y, écrite $P(\vec{x}, y)$. On veut aussi trouver le bon y quand on observe un $\vec{x}$...

Donc on va chercher à résoudre : trouver le meilleur y connaissant $\vec{x}$, soit

$$\text{argmax}_y P(y \mid \vec{x})$$

C'est ce qu'on appelle La règle de Bayes. C'est la meilleure règle qu'on puisse imaginer et l'erreur de cette règle, est l'erreur de Bayes, est la plus petite erreur qu'on puisse faire pour cet apprentissage si les exemples sont décrits par $\vec{x}$.

4 Difficile à calculer

Mais cette règle est difficile à appliquer car on ne sait pas estimer $P(y \mid \vec{x})$ car P est inconnue. On pourrait l'estimer par le principe ERM. En appliquant la règle de Bayes on a

$$P(y \mid \vec{x}) = \frac{P(y)P(\vec{x} \mid y)}{P(\vec{x})}$$

Quand on cherche la valeur de y qui maximise cette quantité, on se moque du calcul de $P(\vec{x})$ qui ne dépend pas de y. Donc on va chercher

$$\text{argmax}_y P(y)P(\vec{x} \mid y)$$

Et là c'est simple car on peut estimer $P(y)$, $P(\vec{x} \mid y)$ et $P(\vec{x})$ sur notre échantillon... En effet si y est binaire, il suffit de compter combien de fois on a $y = 1$ sur le nombre total d'exemple pour estimer ce $P(y = 1)$ et $P(y = 0)$. Si $\vec{x}$ c'est 5 attributs binaires et on va compter combien on a de fois (0,0,0,0,0) parmi tous les exemples qui sont tels que $y = 1$ pour estimer $P(\vec{x} = (0, 0, 0, 0, 0) \mid y = 1)$ etc...

Mais sauf cas particulier, c'est "intractable"... à cause de la dimension du vecteur $\vec{x}$ et des valeurs qu'il peut prendre. Car par exemple dans le cas binaire avec 5 attributs, on a 2^5 possibilités et donc 2 fois 2^5 quantités à estimer pour tous les cas de $P(\vec{x} \mid y)$.

5 La chain rule

Par la chaine rule en probabilités on peut calculer $P(\vec{x})$ en appliquant plusieurs fois $P(A, B) = P(A|B)P(B)$ en se distant que B est un événement qui peut être une conjonction d'événements :

$$P(x_1, x_2, \dots x_n) = P(x_1 \mid x_2, \dots x_n)P(x_2, \dots x_n) = P(x_1 \mid x_2, \dots x_n)P(x_2, \mid x_3, \dots x_n)P(x_3, \dots x_n)...$$

De ce fait pour calculer $P(\vec{x} \mid y)$ on peut appliquer cette chain rule. On va obtenir

$$P(\vec{x} \mid y) = \prod_{k=1}^{n} P(x_k \mid x_{k-1}, ..., x_1, y)$$

6 Indépendance et Naive Bayes

Si A et B sont indépendants alors $P(A \mid B) = P(A)$.

Si on fait l'hypothèse que tous les attributs sont indépendants, c'est-à-dire si x_i et x_j sont indépendants pour tout $i \neq j$, alors le produit s'écrit :

$$P(\vec{x} \mid y) = \prod_{k=1}^{n} P(x_k \mid y)$$

7 Maintenant c'est facile à calculer...

Pour calculer chacun de ces $P(x_k \mid y)$ il suffit de compter!

[ ]: