

# Classificateur des spam par « Naive bayes »

Le but est :

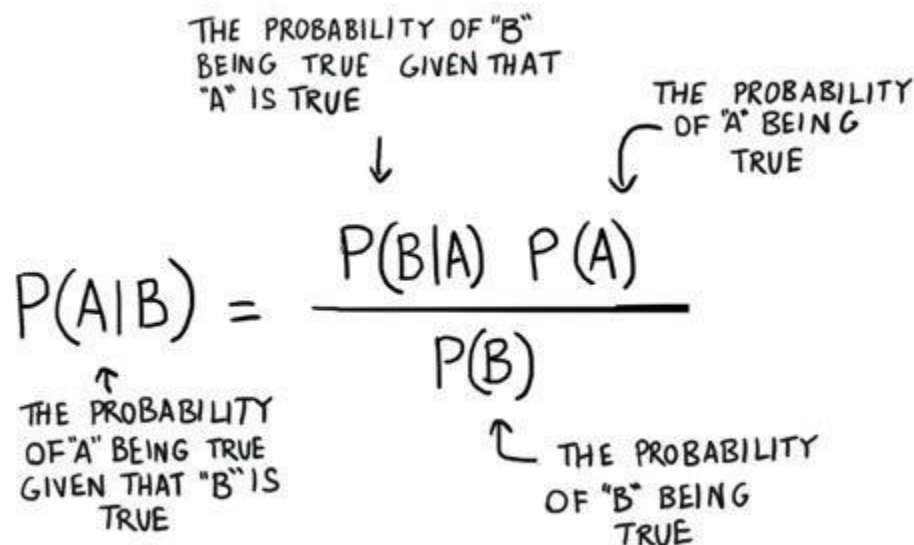
- Écrivez un modèle qui prédit si le texte « courriel » est du spam ou non.
- Enveloppez le modèle dans un service afin que d'autres services puissent communiquer avec le modèle - envoyez-lui des messages et recevez « spam/pas spam » en réponse.

## Comment fonctionnent les filtres anti-spam ?

La plupart des filtres anti-spam sont basés sur le théorème bayésien. Il vous permet de trouver la probabilité d'un événement, à condition qu'un autre événement se soit produit. Ceci est symboliquement exprimé par  $P(A|B)$ , c'est-à-dire la probabilité que l'événement A se produise étant donné que l'événement B s'est déjà produit. Dans notre cas, un événement est la probabilité de rencontrer un texte. c.-à-d. la condition « le texte soit un spam ».

Les raisons d'utiliser le théorème de Naive Bayes :

- Le théorème de Bayes fournit une perspective utile pour comprendre et évaluer de nombreux algorithmes d'apprentissage.
- Il calcule des probabilités explicites pour l'hypothèse et il est robuste au bruit dans les données d'entrée.
- Dans la classification statistique, cela minimise la probabilité d'erreur de classification.


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

THE PROBABILITY OF "B" BEING TRUE GIVEN THAT "A" IS TRUE

THE PROBABILITY OF "A" BEING TRUE

THE PROBABILITY OF "A" BEING TRUE GIVEN THAT "B" IS TRUE

THE PROBABILITY OF "B" BEING TRUE

- $P(A)$  - probabilité de spam
- $P(B)$  - probabilité du texte
- $P(A|B)$  - probabilité que le texte soit du spam
- $P(B|A)$  - probabilité de rencontrer ce texte parmi les spams

## Comment compter ces probabilités ?

Le classificateur bayésien est basé sur l'hypothèse que les attributs d'une classe sont indépendants les uns des autres. En réalité, presque tous les attributs de quoi que ce soit dépendent les uns des autres, et il est très naïf de l'admettre. Mais avec la classification des spams, l'algorithme bayésien est très efficace (et ce malgré le fait que les mots dans le texte dépendent toujours les uns des autres). En effet, pour déterminer le caractère « spam » d'un texte, nous fonctionnons sur des probabilités. Et si nous croyons que ces probabilités sont indépendantes, nous avons une arme très puissante - les probabilités indépendantes peuvent être multipliées !

Pour calculer la probabilité qu'une phrase donnée soit du spam, nous calculons la probabilité de spam pour chaque mot de la phrase, puis multiplions simplement ces probabilités. Ainsi, la probabilité qu'un texte donné pour être spam est le produit de la probabilité de spam et de la probabilité de rencontrer un texte donné parmi l'ensemble de toutes les variantes de spam possibles.

Ainsi, on calcule

$$P(B|A) = \prod_i P(B_i|A)$$

- $\prod_i$  - product of several elements,
- $B$  - one sentence,
- $B_i$  - one word in sentence.

Et puis calculons et comparons les valeurs des expressions. Cette inégalité nous montre qu'une phrase donnée est spam :

$$P(A|B) > P(\bar{A}|B)$$

- $P(\bar{A}|B)$  - probabilité que la phrase donnée ne soit pas du spam.

Ou nous pourrions réécrire la formule :

$$P(A) \cdot P(B|A) > P(\bar{A}) \cdot P(B|\bar{A})$$