Notes TC4

Adrien Pavao

September 2017

Contents

1 Inférence Bayesienne

Différents niveaux d'inférence...

1.1 Niveau 1 : Classification Bayesienne

- Y : La classe à prédire (catégorielle)
- $\vec{X}$: Vecteur aléatoire, $\vec{X} \begin{pmatrix} x_1 \\ ... \\ x_2 \end{pmatrix}$

On cherche à choisir y de façon à maximiser :

$$P(Y = y|\vec{X} = \vec{x}) = \frac{P(\vec{X} = \vec{x}|Y = y)P(Y = y)}{P(\vec{X} = \vec{x})}$$

Dans cette formule, on remarque des termes particuliers :

- La **vraisemblance** : $P(\vec{X} = \vec{x}|Y = y)$.
- L'**a priori** : $P(Y = y)$.
- L'**évidence** : $P(\vec{X} = \vec{x})$.

La vraisemblance et l'a priori sont à estimer. On estime une ditribution sur X pour chaque classe y. On peut donc faire l'hypothèse naïve suivante :

$$P(\vec{X} = \vec{x}|Y = y) = \Pi_{i=1}^{d} P(\vec{X}_i = \vec{x}_i|Y = y)$$

Estimer les paramètres

Cas Bernouilli : $\Theta_{iy} = \frac{n(1,i,y)}{N(i,y)}$

$n(1, i, y) =$ nombre de fois où $\vec{X}_i = 1$ dans la classe y.

Si $n(1, i, y) = 0$ alors $\Theta_{iy} = 0$ Donc $P(\vec{X} = \vec{x}|Y = y) = 0$, ce qui est mauvais. On estime Θ sur les données et on vient à la conclusion qu'un evenement est impossible sous pretexte qu'on ne l'a jamais observé. Il faut éviter ce problème.

Ce type d'estimation est appelée une estimation MLE : Maximum Likelihood Estimate. Il s'agit de l'interprétation **fréquentiste** des données.

Autrement dit, on cherche les paramètres Θ_{iy} qui maximisent $P(D|\Theta_{iy})$. (D la réalisation des données ..)

1.2 Niveau 2 : Inférence Bayesienne des paramètres

On cherche $P(X_i|Y)$ -¿ $P(X_i|Y_i\Theta_{iy})$. L'apprentissage revient à l'estimation d'une distribution sur les paramètres.

Estimer $P(\Theta_{iy}|D)$.

$$P(\Theta_{iy}|D) = \frac{P(D|\Theta_{iy})P(\Theta_{iy})}{P(D)}$$

1.2.1 A priori sur les paramètres

Cas Bernouilli : $\Theta_{iy} \in [0, 1]$, continu. Donc $P(\Theta_{iy})$ _ une loi continue de support $[0, 1]$. Le choix : Loi Beta.

$$P(\Theta_{iy}; \alpha_0, \alpha_1) = \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\Theta_{iy}^{\alpha_1-1}(1 - \Theta_{iy})^{\alpha_0-1}$$

(Dénominateur et game -¿ Normalisation)

α_0 et α_1 sont les paramètres de la loi Beta. On a $\alpha_0, \alpha_1 > 0, \in R$ (R reel, D majuscule ...)

- **Fonction de densité symétrique :** $\alpha_0 = \alpha_1$ et $\alpha_0, \alpha_1 > 1$.

  Graphe 1

- **A priori non-informatif :**

  $\alpha_0 = \alpha_1 = 1$.

  Graphe 2

- **A priori parcimonieux (sparse) :**

  $\alpha_0, \alpha_1 < 1$

  Graphe 3

1.2.2 A posteriori sur les paramètres

$P(\Theta_{iy}|D)$ gamealpha $P(D|\Theta_{iy})P(\Theta_{iy};\alpha_1,\alpha_0)$ (vraimsemblance et a priori). gamealpha -¿ proportionnel à $P(\Theta_{iy}|D)$ propor $\Theta_{iy}^{N_1+\alpha_1-1}(1-\Theta_{iy})^{N_0+\alpha_0-1}$

- N_0 : Nombre de x_i à 0 dans D.
- N_1 : Nombre de x_i à 1 dans D.

(defition importante) La loi a posteriori est comme la loi a priori, une loi Beta. La loi Beta est l'a priori **conjugué** de Bernouilli (conjugated prior).

1.2.3 Retour à la classification

1. **Maximum a Posteriori des Paramètres (MAP)**

   $\Theta_{iy} = argmaxP(\Theta_{iy}|D)$ (chapeau sur le theta !) $\Theta_{iy} = \frac{N_1+\alpha_1-1}{N_1+N_0+\alpha_1+\alpha_0-2}$ α_1 et α_0 agissent comme des "pseudo-comptes". Lissage (smoothing) de distibution. $\Theta_{iy}! = 0$ Si $N_1, N_0 >> \alpha_1, \alpha_0$ alors l'a priori est négligeable.

   -¿ Régularisation, eviter le sur-apprentissage.

2. **Loi prédictive (inférence Bayesienne 3)**

   $P(X_i = x_i|Y = y;\Theta_{iy})$ avec Θ_{iy} estimés à partir des données (MAP).

   Le paramètre n'existe pas et ne doit donc pas apparaitre dans la prédiction. La vraie prédiction :

   $P(X_i = x_i|D) = integrale01P(X_i = x_i;\Theta_{iy}|D)d\Theta_{iy}$, en marginalisant les paramètres.

   $P(X_i;\Theta_{iy}|D) = P(X_i|\Theta_{iy};D)P(\Theta_{iy}|D)$ (vraisemblance et a priori).

   $P(X_i = x_i|D) = \frac{N_1+\alpha_1}{N_1+N_0+\alpha_1+\alpha_0}$, $\forall\alpha_1$ et $\alpha_0 > 0$.