

Inférence bayésienne

Bruno Bouzy

5 février 2008

Introduction

Ce chapitre présente ce qu'est l'inférence bayésienne sur l'exemple de la classification.

Classification

Problème de départ

On cherche à classer des exemples dans des classes C_k .

Probabilité à priori $P(C_k)$.

Par l'expérience, on connaît les $P(C_k)$ où $P(C_k)$ est la probabilité d'appartenance à priori d'un exemple à la classe C_k .

Probabilités jointes

En fait, on connaît aussi des caractéristiques des exemples X^1 . Et par l'expérience, on connaît également les $P(X^1)$ où $P(X^1)$ est la probabilité à priori que l'exemple ait la caractéristique X^1 . Plus précisément, par l'expérience on connaît $P(C_k, X^1)$ et $P(X^1, C_k)$ qui sont les probabilités jointes, mesurables par l'expérience. On a l'égalité suivante :

$$P(C_k, X^1) = P(X^1, C_k) \quad (1)$$

Probabilité conditionnelle

On appelle $P(X^1|C_k)$ la probabilité conditionnelle, probabilité qu'un exemple de la classe C_k possède la caractéristique X^1 . On a :

$$P(C_k, X^1) = P(C_k).P(X^1|C_k) \quad (2)$$

Formellement $P(C_k|X^1)$ est aussi une probabilité conditionnelle : la probabilité que l'exemple soit dans la classe C_k sachant que X^1 . Mais en pratique, on l'appelle par un autre nom.

Probabilité à postériori

En effet, dans le problème de départ, on cherche à dire si l'exemple est dans la classe C_k donc $P(C_k|X^1)$ est ce que l'on cherche. Donc on appelle $P(C_k|X^1)$ la probabilité à postériori. On a :

$$P(X^1, C_k) = P(X^1).P(C_k|X^1) \quad (3)$$

Formule de Bayes

La formule de Bayes dit que :

$$P(C_k|X^1) = P(C_k).P(X^1|C_k) / P(X^1) \quad (4)$$

En pratique

En pratique, si on mesure des approximations de $P(C_k)$, $P(X^1)$, $P(C_k, X^1)$, on peut obtenir $P(X^1|C_k)$ et $P(C_k|X^1)$. Finalement, pour revenir au problème de départ (bien classer des exemples), si on veut classer un exemple connaissant ses caractéristiques, la probabilité à postériori est une meilleure mesure que la probabilité à priori. Nota Bene : en pratique la formule de Bayes (4) n'est pas obligatoirement utilisée : on utilise plutôt (2) et (3).

Maximum de vraisemblance et Maximum A Postériori (MAP)

Dans le problème de classification, étant donné X , on peut chercher la classe C maximisant $P(C|X)$, on parle plus précisément de Maximum A Posteriori (MAP). Dans ce cas, on cherche la classe la plus probable étant donné X . Ou bien on peut chercher celle maximisant $P(X|C)$, on parle encore de maximum de vraisemblance dans un sens plus restreint. (MLE = Maximum Likelihood Estimation). Dans ce cas, on cherche la classe C maximisant la probabilité de X étant donnée cette classe C .

Pour un jeu

Pour un jeu, si X^1 est la présence d'un pattern l , et si C est le fait de jouer un coup, alors la probabilité de jouer le coup sachant que le pattern l est présent est égale à la probabilité de jouer un coup, fois la probabilité que le pattern soit présent sachant que le coup est joué, divisé par la probabilité que le pattern soit présent. Donc, pourvu que l'on ait des parties de joueurs à sa disposition, on peut construire automatiquement la probabilité de jouer quand un pattern est présent.

En pratique (suite)

Si on ne considère que la caractéristique X^1 , on a :

$$\sum_k P(C_k|X^1) = 1 \quad (5)$$

(4) injectée dans (5) donne alors:

$$\sum_k P(C_k).P(X^1|C_k) = P(X^1) \quad (6)$$

En pratique, on mesure facilement $P(C_k)$ et $P(X^1|C_k)$. Donc (6) sert pour déterminer $P(X^1)$. Puis (4) sert pour calculer $P(C_k|X^1)$.

Exercice

On cherche à classer des exemples dans deux classes C_1 et C_2 en fonction de la caractéristique X qui peut prendre 10 valeurs X^0, X^1, \dots, X^9 . On dispose des données suivantes :

l	0	1	2	3	4	5	6	7	8	9
C_1	6	6	5	4	2	1	1	0	0	0
C_2	0	0	1	2	3	3	3	4	5	5

Que vaut $P(C_1)$? $P(C_2)$?

Quelle est la probabilité de se tromper en utilisant le principe du maximum de vraisemblance sans utiliser la caractéristique X ?

Que vaut $P(X^l)$ pour $l = 0, 1, \dots, 9$?

Que vaut $P(C_1|X^l)$ pour $l = 0, 1, \dots, 9$?

Que vaut $P(C_2|X^l)$ pour $l = 0, 1, \dots, 9$?

Tracer les deux courbes $P(C_1|X^l)$ et $P(C_2|X^l)$ en fonction de X^l .

Quelle est la probabilité de se tromper en utilisant le principe du maximum de vraisemblance à posteriori et la caractéristique X ?

Références

Antoine Cornuéjols, Laurent Miclet, Apprentissage artificiel, concepts et algorithmes, Eyrolles, pages 57-64.

Christopher Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995, chapitre 1, pages 17-28.