Université J.F. CHAMPOLLION Albi

Inférence Bayésienne

Licence d'informatique S6

Thierry Montaut

2022

Modèle bayésien

- L'apprentissage s'appuie sur nos observations mais rend mal compte de l'incertitude de ces observations.
- Les modèles probabilistes vont nous permettre de prendre en compte explicitement cette incertitude, en formulant l'apprentissage d'un modèle comme un problème d'inférence sur une distribution conjointe des observations et des étiquettes.
- Ce modèle se prête bien aux situations suffisamment complexes pour être considérées comme aléatoires.
- Dans cette introduction nous nous restreindrons au seul cas de la classification bayésienne naïve.

Modèle probabiliste

- On suppose ici que les observations de nos m variables $x_1, \ldots x_n$ et celles de leurs étiquettes $y_1, \ldots y_n$ sont les réalisations de deux variables aléatoires X et Y.
- On s'intéresse alors à leur loi conjointe

$$P\{(X,Y) = (x,y)\} = P(\{X = x, Y = y\}) = P(\{X = x\} \cap \{Y = y\})$$

- La probabilité qu'une observation appartienne à la classe c est alors la loi conditionnelle

$$P\{Y = c | X = x\}$$

- Deux problèmes se posent alors
  1. **L'inférence :** qui consiste à déterminer ces lois (conjointes et conditionnelles) à partir de nos observations.
  2. **La décision :** qui consiste à utiliser ces lois pour prédire la classe d'une observation. Ce sera tout simplement la classe la plus probable, et sa probabilité sera le score retenu.

Loi de Bayes

Le raisonnement probabiliste permettant l'inférence s'appuie sur la définition de la loi conditionnelle

Définition

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

et sur la loi de Bayes [1]

Proposition

$$P\{Y = c|X = x\} = \frac{P\{Y = c\}P\{X = x|Y = c\}}{P\{X = x\}}$$

1. Thomas Bayes, mathématicien et pasteur anglais du XVIIIème

Loi de Bayes

Les différents éléments mis en oeuvre dans ces équations sont :

- $P\{Y = c\}$ (la loi marginale de Y), appelé distribution a priori des étiquettes (avant les observations des données).
- $P\{Y = c|X = x\}$, appelé distribution a posteriori des étiquettes (après avoir observé les données).
- $P\{X = x|Y = c\}$, la vraisemblance de x dans la classe c.

Exercice

Un test possède une sensibilité (taux de positifs chez les malades) de 70% et une spécificité (taux de négatifs chez les non malades) de 98%. Ce test concerne une maladie qui atteint en moyenne 1 personne sur 10000. Quelle est la probabilité qu'une personne dont le test est positif soit effectivement malade ?
Il s'agit d'une problème d'inférence bayésienne...

Exercice

Ici l'observation X (1 ou 0) est le résultat du test, l'étiquette Y (1 ou 0) le fait d'être ou non malade.
On demande donc de calculer :

$$\begin{aligned} P\{Y=1|X=1\} &= \frac{P\{Y=1\}P\{X=1|Y=1\}}{P\{X=1\}} \\ &= \frac{10^{-4}.0.7}{P\{X=1\}} \end{aligned}$$

$$\begin{aligned} P\{X=1\} &= P\{X=1,Y=1\}+P\{X=1,Y=0\} \\ &= P\{X=1|Y=1\}P\{Y=1\}+P\{X=1|Y=0\}P\{Y=0\} \\ &= 10^{-4}.0.7+(1-0.98)(1-10^{-4}) \end{aligned}$$

$$P\{Y=1|X=1\} = 0.35\%$$

Quand le test est positif, il est donc beaucoup plus probable d'être en bonne santé que d'être malade. Il faut donc jeter ces tests !

Estimation de loi

Dans l'exemple précédent, $P(X|Y)$ était supposé connu. En général, nous aurons à apprendre cette loi à partir de nos données. Dans ce cas, nous ferons l'hypothèse que la loi suit une distribution particulière dépendant de certains paramètres p et nous chercherons les valeurs du paramètres qui maximise la vraisemblance des observations.

Proposition

Si on suppose que la variable X suit une loi de Bernouilli de paramètre p.

$$P(X = x) = p^x.(1-p)^{1-x}$$

Alors la valeur du paramètre p qui maximise la vraisemblance de l'observation x est

$$p = \bar{x}.$$

Classification naïve bayésienne

- On suppose désormais m observations de n variables X_i caractérisant notre problème, et de leurs étiquettes $y_1, \ldots y_m$
- L'hypothèse simplificatrice (dite naïve) que nous faisons ici est de supposer que les n variables sont indépendantes.
- Avec cette hypothèse, la probabilité qu'une observation appartienne à la classe c est :

$$P(Y = c|X = (x1, \ldots, x_p)) = \frac{P\{Y = c\}.\prod_{i=1}^{m} P\{X_i = x_i|Y = c\}}{P\{X = (x1, \ldots, x_p)\}}$$

- Puisque le dénominateur est constant, la classe décidée sera donc celle qui maximise

$$P\{Y = c\}.\prod_{i=1}^{m} P\{X_i = x_i|Y = c\}$$

- Il faut donc être en mesure d'apprendre des données les valeurs de $P\{Y = c\}$ et des $P\{X_i = x_i|Y = c\}$

Classification naïve bayésienne

Définition

On appelle classificateur naïf bayésien, un classificateur construit en maximisant la vraisemblance des observations, en supposant que les variables sont indépendantes et suivent une loi donnée de paramètre p.

Exemple : Filtrage bayésien des spams

Le filtrage des spams est un exemple classique d'application de cette méthode bayésienne. Ici $y = 1$ lorsqu'un mail est un spam et $y = 0$ sinon.

- On choisit n mots-clefs dont la présence informe sur la probabilité d'être un spam. Un mail est représenté par le vecteur binaire $X = (x_1, \ldots, x_n)$ où $x_i = 1$ si le ième mot-clef apparaît dans le spam.
- Chaque variable aléatoire X_i est modélisée par une loi de Bernouilli de paramètre p_i.
- Conformément à notre hypothèse, on suppose que l'apparition des différents mots est conditionnellement indépendante, ce qui semble effectivement très naïf dans ce contexte.

Exemple : Filtrage bayésien des spams

- Comme vu précédemment, la classe décidée sera donc celle qui maximise

$$P\{Y = c\}.\prod_{i=1}^{m} P\{X_i = x_i | Y = c\}$$

- $P\{Y = c\}$ est estimé par la fréquence des spams dans notre jeu de données
- Les variables aléatoires X_i étant modélisées par une loi de Bernouilli de paramètre p_i, $P\{X_i = x_i | Y = 1\} = p_i^{x_i}.(1 - p_i)^{1-x_i}$ où $p_i = \overline{(x_i)}$ est le taux de spams contenant le ième mot.

Avec Sklearn

- La bibliothèque *sklearn.naive_bayes* contient toutes les fonctions nécessaires pour mener à bien une modélisation bayésienne naïve.
- Les fonctions *GaussianNB*() , *BernoulliNB*() etc. permettent de définir les modèles qui seront appris comme d'habitude par $model.fit(X, y)$
- Vous pourrez trouve un data set de spams/ham $spam_h am_d ataset.csv$ sur le site de Kaggle : www.kaggle.com
- C'est une bonne occasion pour découvrir ce site et commencez à vous lancer des défis de machine learning et à participer à des concours...