

TP R : classifieur bayésien naïf

Fabrice Rossi

L'objectif de ce TP est d'acquérir la maîtrise du classifieur bayésien naïf à travers son implémentation et sa mise œuvre sur des données artificielles et réelles. Les données suggérées sont décrites en préliminaire.

1 Données NIPS

1.1 Présentation

On s'intéresse aux données NIPS qui correspondent à une représentation de type *sac de mots* de l'ensemble des 5812 articles parus dans NIPS entre 1987 et 2015 sur un vocabulaire de 11463 mots.

Dans le fichier d'origine, chaque ligne commence par un mot puis contient un vecteur d'entiers. La i -ème coordonnée du vecteur vaut k si le mot associé est apparu exactement k fois dans l'article numéro i . Les noms de colonnes sont de la forme année_numéro.

1.2 Préparation

Question 1 Écrire une fonction d'extraction de l'année de publication de chaque article à partir de noms au format utilisé dans le fichier.

Question 2 Écrire une fonction d'extraction qui prend en paramètre la base nips et un ensemble d'années, et renvoie une matrice au format classique avec en ligne les textes des années concernées et en colonne les mots. On réduira le vocabulaire aux mots présents au moins une fois dans chaque année concernée et on ajoutera une colonne $\mathbf{Y}$ contenant l'année de l'article.

2 Rappels

2.1 Notations et définitions

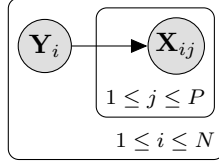
On suppose que les données étudiées sont à valeurs dans $\mathcal{X} \times \mathcal{Y}$ avec :

- $\mathcal{Y}$ est un ensemble fini ;
- $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \mathcal{X}_P$;
- chaque $\mathcal{X}_j$ est soit $\mathbb{R}$, soit $\mathbb{N}$, soit un ensemble fini.

Si $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$ est un ensemble fini et γ un vecteur de $\mathbb{R}^{|\mathcal{A}|}$ tel que $\gamma = (\gamma_1, \dots, \gamma_{|\mathcal{A}|})^T$ avec $\forall j, \gamma_j \in [0, 1]$ and $\sum_{j=1}^{|\mathcal{A}|} \gamma_j = 1$, on note $C(\mathcal{A}, \gamma)$ la loi sur $\mathcal{A}$ tel que $\mathbf{X} \sim C(\mathcal{A}, \gamma)$ si et seulement si pour tout j , $\mathbb{P}(\mathbf{X} = a_j) = \gamma_j$. Quand cela n'est pas ambigu on note $\mathbb{P}(\mathbf{X} = a) = \gamma_a$ quand $\mathbf{X} \sim C(\mathcal{A}, \gamma)$. Notons que toute loi de probabilité sur $\mathcal{A}$ est de la forme $C(\mathcal{A}, \gamma)$ pour une valeur unique de γ .

2.2 Classifieur bayésien naïf

Une loi sur $(\mathcal{X} \times \mathcal{Y})^N$ suit le modèle bayésien naïf si elle est compatible avec le modèle graphique orienté suivant :



La loi vérifie donc

$$p((\mathbf{x}_i, \mathbf{y}_i)_{1 \leq i \leq N}) = \prod_{i=1}^N p(\mathbf{y}_i) \prod_{j=1}^P p(\mathbf{x}_{ij} | \mathbf{y}_i)$$

2.3 Maximum de vraisemblance

Soit une loi paramétrique en Θ et suivant le modèle bayésien naïf sur $(\mathcal{X} \times \mathcal{Y})^N$ de sorte que

$$p((\mathbf{x}_i, \mathbf{y}_i)_{1 \leq i \leq N} | \Theta) = \prod_{i=1}^N p(\mathbf{y}_i | \Theta_{\mathbf{Y}}) \prod_{j=1}^P p(\mathbf{x}_{ij} | \Theta_{\mathbf{x}_j | \mathbf{y}_i}).$$

Autrement dit les paramètres Θ se répartissent en des paramètres $\Theta_{\mathbf{Y}}$ ne concernant que la loi marginale de $\mathbf{Y}$ et des paramètres $\Theta_{\mathbf{x}_j | \mathbf{y}}$ ne concernant que la loi conditionnelle de $\mathbf{X}_j$ sachant $\mathbf{Y} = \mathbf{y}$.

Alors on peut montrer que l'estimation par maximum de vraisemblance de Θ peut se faire en estimant indépendamment chaque groupe de paramètres sur les données concernées :

- $\Theta_{\mathbf{Y}}$ est obtenu en maximisant la vraisemblance $\prod_{i=1}^N p(\mathbf{y}_i | \Theta_{\mathbf{Y}})$
- $\Theta_{\mathbf{x}_j | \mathbf{y}}$ est obtenu en maximisant la vraisemblance $\prod_{i=1, \mathbf{y}_i = \mathbf{y}}^N p(\mathbf{x}_{ij} | \Theta_{\mathbf{x}_j | \mathbf{y}})$.

3 Cas binaire

On s'intéresse dans un premier temps au cas totalement binaire, c'est-à-dire quand $\mathcal{Y} = \{0, 1\}$ et pour tout j $\mathcal{X}_j = \{0, 1\}$.

3.1 Estimation

Question 1 Indiquer les paramètres nécessaires pour représenter n'importe quelle loi sur $(\mathcal{X} \times \mathcal{Y})^N$ qui suit le modèle bayésien naïf.

Question 2 Rappeler l'estimateur du maximum de vraisemblance pour ces paramètres.

3.2 Mise en œuvre

Question 3 Écrire une fonction R ou Python qui prend comme paramètre N , P et des paramètres numériques appropriés et qui renvoie une réalisation de $(\mathcal{X} \times \mathcal{Y})^N$ (avec $\mathcal{X} = \{0, 1\}^P$) selon le modèle bayésien naïf binaire spécifié par les paramètres.

Question 4 Écrire une fonction R ou Python qui prend en paramètre un tableau de données binaires et une variable cible $\mathbf{Y}$, et renvoie les paramètres du modèle bayésien naïf le plus vraisemblable pour ces données. On pourra s'appuyer sur les fonctionnalités de type *group by* fournies par le langage choisi (`aggregate` en R de base, `group_by` dans `dplyr` et `groupby` dans `pandas` pour Python).

3.3 Décision

Question 5 Rappeler la formule donnant le classifieur optimal à partir de la loi $p(\mathbf{X}, \mathbf{Y})$ pour une fonction de perte quelconque l . On rappelle que le classifieur optimal est la fonction g^* de $\mathcal{X}$ dans $\mathcal{Y}$ qui minimise le risque

$$R(g) = \mathbb{E}_{p(\mathbf{X}, \mathbf{Y})} \{l(g(\mathbf{X}), \mathbf{Y})\}.$$

Question 6 Écrire une fonction R ou Python qui prend en paramètre une observation $\mathbf{x}$ et les paramètres d'un modèle bayésien naïf binaire, et qui renvoie une estimation de $\log \mathbb{P}(\mathbf{Y} = 1 | \mathbf{X} = x)$ selon ce modèle.

Question 7 Écrire une fonction R ou Python qui prend en paramètre une observation $\mathbf{x}$, les paramètres d'un modèle bayésien naïf binaire et une fonction de perte l , et qui renvoie la décision optimale obtenue en supposant que les données sont issues du modèle.

3.4 Mise en œuvre

Question 8 Étudier le comportement de base du modèle en fonction de N et P . On pourra par exemple dessiner l'évolution du risque du modèle sur des données synthétiques engendrées par un modèle bayésien naïf en fonction de N .

Question 9 Étudier la robustesse au bruit du modèle. On pourra tout d'abord injecter du bruit en basculant aléatoirement certaines observations de 0 à 1 ou vice versa. On pourra aussi ajouter des variables indépendantes de $\mathbf{Y}$.

Question 10 Étudier la robustesse du classifieur à une rupture du modèle. On pourra par exemple ajouter des variables fortement dépendantes des variables initiales.

Question 11 Étudier le comportement du modèle sur les données réelles NIPS en les binarisant. On opposera par exemple quelques années récentes aux premières années de la conférence. On réduira le vocabulaire à celui actif sur les années concernées.

3.5 Extension au cas discret fini

On suppose dans cette section que les $\mathcal{X}_j$ et $\mathcal{Y}$ sont tous des ensembles finis.

Question 12 Indiquer les paramètres nécessaires pour représenter n'importe quelle loi sur $(\mathcal{X} \times \mathcal{Y})^N$ qui suit le modèle bayésien naïf.

Question 13 Rappeler l'estimateur du maximum de vraisemblance pour ces paramètres.

Question 14 Décrire brièvement les adaptations à réaliser pour le code développé pour le cas binaire.

4 Cas entier positif

On suppose maintenant que $\mathcal{Y}$ est fini et que certains $\mathcal{X}_j$ sont égaux à $\mathbb{N}$ (les autres sont supposés finis). D'après les résultats généraux sur le classifieur bayésien naïf, l'estimation des paramètres associés à $\mathbf{Y}$ et aux $\mathbf{X}_j$ pour $\mathcal{X}_j$ fini n'est pas modifiée par cette hypothèse. Il faut en revanche spécifier $p(\mathbf{X}_j | \boldsymbol{\Theta}_{\mathbf{X}_j | \mathbf{Y}})$ pour $\mathcal{X}_j = \mathbb{N}$.

4.1 Modèle et estimation

Question 1 Proposer deux lois simples sur $\mathbb{N}$ utilisables pour $p(\mathbf{X}_j | \boldsymbol{\Theta}_{\mathbf{X}_j | \mathbf{Y}})$.

Question 2 Rappeler les estimateurs du maximum de vraisemblance pour les paramètres de ces lois.

4.2 Mise en œuvre

Question 3 Écrire une fonction qui estime par maximum de vraisemblance les paramètres d'un modèle bayésien naïf sur un ensemble de données, dans le cas particulier étudié dans cette section. On pourra s'appuyer sur les types des différentes variables pour déterminer automatiquement les paramètres à estimer quand cela est possible. Une spécification du modèle à utiliser pour chaque variable entière pourra être nécessaire.

Question 4 Écrire une fonction qui estime $\log \mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x})$ pour ses paramètres $\mathbf{x}$ et $\mathbf{y}$, et pour un modèle bayésien naïf adapté au type des données (et dont les paramètres sont passés à la fonction).

Question 5 Étudier le comportement du modèle sur les données réelles NIPS.

5 Ouverture

5.1 Cas réel

Si pour certains j on a $\mathcal{X}_j = \mathbb{R}$, la situation ne se complique pas particulièrement, ni d'un point de vue théorique, ni au niveau théorique. En effet, il suffit de choisir un modèle adapté pour $p(\mathbf{X}_j | \boldsymbol{\Theta}_{\mathbf{X}_j | \mathbf{y}})$ avec $\mathbf{X}_j \in \mathbb{R}$. On peut ainsi choisir un modèle paramétrique (comme une loi normale) ou passer par une estimation non paramétrique à noyau, par exemple.

5.2 Estimation bayésienne

il est aussi assez facile de passer à une estimation bayésienne des paramètres par maximum à posteriori, à condition de disposer d'une loi conjuguée à celle choisie pour les données. C'est le cas pour beaucoup de situation, comme par exemple :

- binaire : loi de Bernoulli et loi conjuguée bêta
- fini : loi catégorielle et loi conjuguée de Dirichlet
- entier : loi de Poisson et loi conjuguée Gamma
- réel : loi normale et loi conjuguée normale (pour une variance fixée)