Exercices sur le classifieur bayesien naïf

Fabrice Rossi

13 février 2015

1 Dimension deux

On étudie un problème de classement entre deux classes 1 et 2 pour des objets caractérisés par deux variables :

T la taille qui prend les valeurs S, M et L (*small*, *medium* et *large*) ;

P le poids, qui prend les valeurs 1, 2 ou 3.

1.1 Estimation complète

On observe des exemples d'objets de chaque classe selon le tableau suivant (les données d'apprentissage) :

classe	T	P
1	S	1
1	S	2
1	M	1
1	L	2
1	S	1
1	M	2
2	M	3
2	L	2
2	M	1
2	L	2
2	L	3
2	L	2

1.1.1 Questions

1. Identifier les probabilités conditionnelles nécessaires au calcul du classifieur optimal.
2. Estimer ces probabilités à partir du tableau.
3. Calculer le risque du classifieur empirique optimal (avec la fonction de perte de comptage).

4. Quel problème rencontre-t-on ?

1.2 Estimation avec indépendance conditionnelle

On fait maintenant l'hypothèse d'indépendance conditionnelle du classifieur bayesien naïf.

1.2.1 Questions

1. Calculer les lois empiriques de T et P dans les deux classes.
2. Calculer le risque du classifieur bayesien naïf obtenu à partir de lois empiriques.

2 Un exemple réaliste

2.1 Présentation des données

On considère la base de données des votes effectués par les membres de la Chambre des représentants des EUA en 1984 sur 16 propositions importantes. Chaque individu est un membre de la Chambre décrit par 17 variables nominales. La variable Parti prend les modalités Démocrate et Républicain. Les autres variables, V1 à V16 représentent les votes et prennent les valeurs OUI, NON et NSP (pour une absence de vote). Il y a 267 représentants démocrates et 168 représentants républicains.

2.2 Les données

Républicains

	NON	NSP	OUI
V1	134	3	31
V2	73	20	75
V3	142	4	22
V4	2	3	163
V5	8	3	157
V6	17	2	149
V7	123	6	39
V8	133	11	24
V9	146	3	19
V10	73	3	92
V11	138	9	21
V12	20	13	135
V13	22	10	136
V14	3	7	158
V15	142	12	14
V16	50	22	96

Démocrates

	NON	NSP	OUI
V1	102	9	156
V2	119	28	120
V3	29	7	231
V4	245	8	14
V5	200	12	55
V6	135	9	123
V7	59	8	200
V8	45	4	218
V9	60	19	188
V10	139	4	124
V11	126	12	129
V12	213	18	36
V13	179	15	73
V14	167	10	90
V15	91	16	160
V16	12	82	173

2.3 Questions

1. Combien de valeurs différentes sont possibles pour le vecteur des votes ?
2. Les tables ci-dessus représentent les votes de chaque parti aux 16 propositions. Quelles grandeurs nécessaires à la mise en œuvre d'un classifieur bayésien naïf peuvent être évaluées grâce à cette table ?
3. Soit un représentant ayant voté selon le vecteur de réponse V suivant :

V1	V2	V3	V4	V5	V6	V7	V8
OUI	NON	NSP	OUI	NON	OUI	OUI	OUI

V9	V10	V11	V12	V13	V14	V15	V16
NON	NON	OUI	NON	NON	NON	NON	OUI

Donner le rapport de probabilités $\frac{P(\text{Démocrate} \mid V)}{P(\text{Républicain} \mid V)}$ tel qu'estimé par le classifieur bayésien naïf. On rappelle qu'il y 267 représentants démocrates et 168 représentants républicains. On se contentera de donner une formule numérique sans chercher à simplifier la fraction obtenue (ni à évaluer les produits).

3 Introduction au cas continu

3.1 Point de vue théorique

On considère deux populations, les hommes H de taille moyenne 1,74m avec un écart type de 0,07m et les femmes F de taille moyenne 1,62m avec un écart type de 0,065m (chiffres INSEE 2001). La population H contient h individus et la population F, f individus. On suppose que les répartitions des tailles sont gaussiennes au sein de chaque sous-population.

On choisit aléatoirement uniformément un individu dans la population totale et on veut déterminer en fonction de sa taille uniquement de quelle sous-population il est issu : il s'agit donc de classer les individus en fonction d'une variable continue.

3.1.1 Questions

1. On note G la variable aléatoire indiquant le genre d'une personne choisie au hasard. Donner la loi de G.
2. On note T la variable aléatoire donnant la taille d'une personne choisie au hasard. Donner la densité de T.
3. Calculer $P(G = f|T = t)$.
4. Donner le classifieur bayésien optimal pour l'erreur de comptage.
5. On suppose que $h = f$. Préciser les décisions prisent par le classifieur optimal.
6. Comment interpréter cette stratégie de décision ?

3.2 Point de vue empirique

On observe maintenant des exemples de la population considérée, décrits par la table 1.

3.2.1 Questions

1. Quelles hypothèses permettent de se ramener au cas théorique ?
2. Quelles grandeurs doit-on calculer pour définir le classifieur optimal ?

4 Données multidimensionnelles

On étudie les Iris de Fisher/Anderson : il s'agit de 150 fleurs caractérisées par quatre variables numériques et appartenant à trois espèces différentes. Il y a 50 fleurs par espèce. On note C la variable de classe (l'espèce), prenant les valeurs 1, 2 ou 3, et X la variable des caractéristiques numériques, à valeur dans $\mathbb{R}^4$.

	Genre	Taille
1	Femme	1.83
2	Femme	1.72
3	Femme	1.83
4	Femme	1.83
5	Femme	1.77
6	Femme	1.63
7	Femme	1.68
8	Femme	1.72
9	Femme	1.74
10	Femme	1.91
11	Homme	1.67
12	Homme	1.57
13	Homme	1.55
14	Homme	1.60
15	Homme	1.60
16	Homme	1.59
17	Homme	1.64
18	Homme	1.56
19	Homme	1.65
20	Homme	1.54

TABLE 1 – Une population

4.1 Questions

1. Si on fait l'hypothèse que les variables numériques suivent dans chaque classe une loi normale, combien de paramètres doit-on estimer au total ?
2. On suppose que l'hypothèse du classifieur bayésien naïf est vérifiée et que la distribution de chaque variable est gaussienne au sein d'une classe. Quelle est alors la loi jointe de quatre variables dans chaque classe ? Combien de paramètres doit-on estimer dans ce cas ?
3. La figure 1 représente les 50 éléments de l'une des classes en utilisant deux des quatre variables. D'après cette représentation, l'hypothèse bayésienne naïve semble-t-elle raisonnable ?
4. On note μ_i^j la moyenne de la variable i pour la classe j, et σ_i^j l'écart type de la variable i pour la classe j. Soit un vecteur $x = (x_1, x_2, x_3, x_4)$. Donner $P(C = j|X = x)$ en supposant que l'hypothèse bayésienne naïve est vraie et avec une loi normale pour chaque variable dans chaque classe.

5 Maximum de vraisemblance

On considère un problème de classification mixte où chaque observation est décrite par une variable discrète D à valeurs dans $\{0,1\}$ et une variable continue C à valeurs dans $\mathbb{R}$. La classe de chaque observation est donnée par la variable Y à valeurs dans $\{0,1\}$.

5.1 Questions

1. Écrire la vraisemblance d'une observation (d,c,y) en notant $p(C|D,Y,\theta)$ la densité conditionnelle de C sachant D et Y, où θ désigne un vecteur de paramètres pour la densité conditionnelle.
2. Donner la forme simplifiée de la vraisemblance quand on fait l'hypothèse du classifieur bayésien naïf (que l'on maintiendra à partir de cette question).
3. On se donne N observations, $(d_i,c_i,y_i)_{1\leq i\leq N}$ supposées i.i.d. Déterminer l'estimation de $P(D=1|Y=y)$ (pour $y\in\{0,1\}$) par maximum de vraisemblance des N observations.
4. On suppose maintenant que la distribution de C sachant Y est gaussienne. Déterminer l'estimation des paramètres des gaussiennes par maximisation de la vraisemblance des N observations.

6 Approche bayésienne

6.1 Pile ou Face bayésien

On suppose que X est une variable de Bernoulli de paramètre θ (soit $P(X=1)=\theta$). On se donne N répliques i.i.d. de X, $X_1,\ldots,X_N$.

6.1.1 Questions fréquentistes

1. Donner la vraisemblance des N répliques et en déduire l'estimation de θ par maximum de vraisemblance.
2. Quelle valeur prend l'estimateur ci-dessus quand on obtient 2 fois 1 pour $N=2$?

6.1.2 Questions bayésiennes

Dans l'approche Bayésienne, on considère un modèle plus complexe où on choisit θ aléatoirement, puis où on observe N variables de Bernoulli du paramètre θ. On a donc $P(X_i=1|\Theta=\theta)=\theta$. Pour simplifier les calculs, on choisit ici pour Θ une loi Beta, c'est-à-dire

$$p(\Theta=\theta|a,b)=\frac{\theta^{a-1}(1-\theta)^{b-1}}{Beta(a,b)}.$$

On rappelle que la loi $Beta(a,b)$ est d'espérance $\frac{a}{a+b}$, et de mode $\frac{a-1}{a+b-2}$.

1. On note $\mathcal{D} = (X_1, \ldots, X_N)$ avec

$$P(\mathcal{D} = (x_1, \ldots, x_N) | \Theta = \theta) = \prod_{i=1}^{N} P(X_i = x_i | \Theta = \theta)$$

   Calculer $p(\Theta = \theta | \mathcal{D} = (x_1, \ldots, x_N))$.
2. Déduire de l'expression précédente l'estimation de θ par maximum à postériori, c'est-à-dire le mode de $p(\Theta = \theta | \mathcal{D} = (x_1, \ldots, x_N))$.
3. On tire une nouvelle valeur X_{N+1} selon la même loi (et donc selon le même θ). Donner $P(X_{N+1} = 1 | \mathcal{D} = (x_1, \ldots, x_N))$.

6.2 Modèle bayésien naïf bayésien

On considère un problème de classification binaire (variable Y à valeurs dans $\{0,1\}$) où chaque observation est décrite par p variables binaires, $X = (X_1, \ldots, X_p)$, supposées conditionnellement indépendantes sachant la classe. On a donc $2p$ paramètres θ_1^1, θ_p^1 et θ_1^0, θ_p^0, avec $P(X_i = 1 | Y = y) = \theta_i^y$.

On choisit la distribution à priori $Beta(a,b)$ pour tous les θ_i^y. On suppose $P(Y = 1) = \frac{1}{2}$ et on se donne un ensemble d'apprentissage $\mathcal{D} = ((X_1, Y_1), \ldots, (X_N, Y_N))$

1. Donner l'estimateur du maximum à posteriori pour les $2p$ paramètres.
2. Donner $P(Y = 1 | X, \mathcal{D})$.

7 Limites du classifieur bayésien

On considère le problème simple suivant : chaque observation est décrite par deux variables discrètes, X_1 et X_2, toutes deux à valeurs dans $\{0,1\}$. La classe de chaque observation est donnée par la variable Y à valeurs dans $\{0,1\}$. On observe les 4 réalisations suivantes :

X_1	X_2	Y
1	1	0
1	0	1
0	1	1
0	0	1

7.1 Questions

1. Calculer le classifieur bayésien optimal pour ces données.
2. Calculer le classifieur bayésien naïf pour ces données.

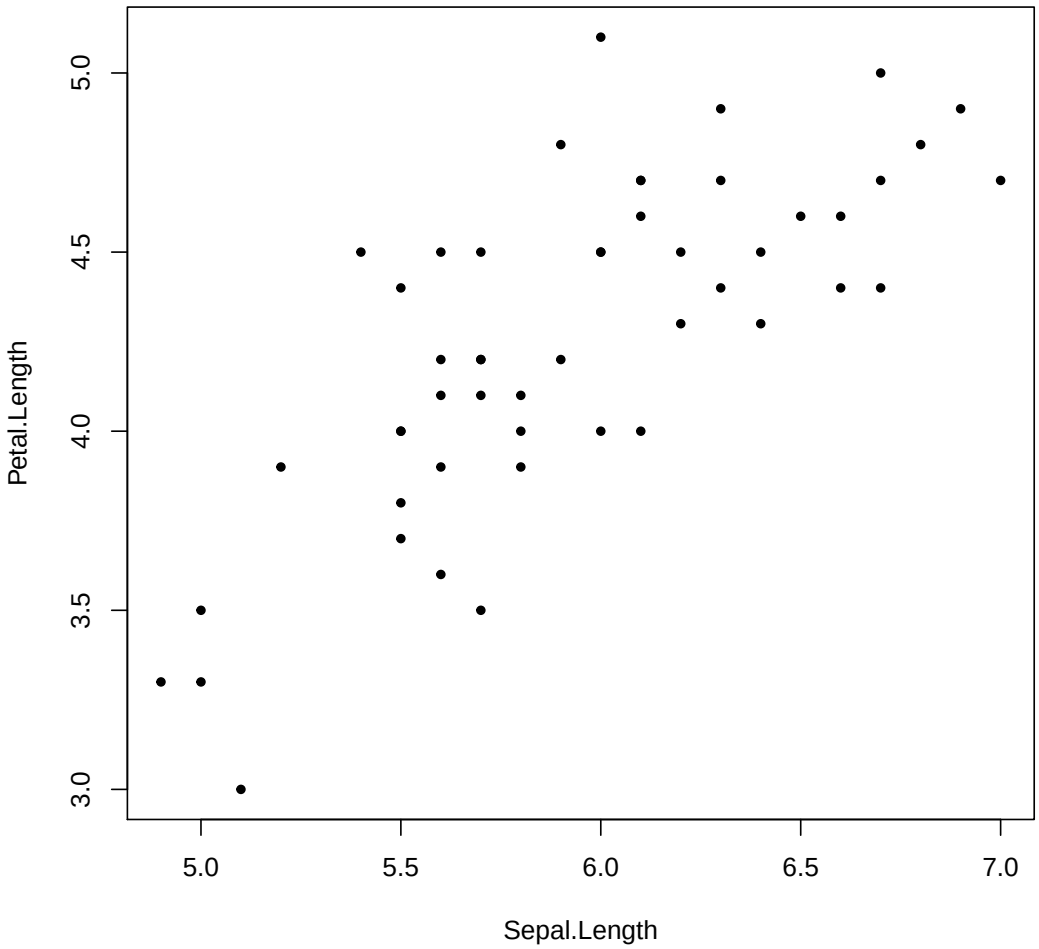

FIGURE 1 – Les éléments de l'une des classes pour deux des variables