Examen – Traitements avancés d'informations SHS

Durée : 1h

1. Méthodes de régression et classification

(1) Soit D un ensemble d'apprentissage composé d'un sous ensemble de n^+ exemples positifs $x_i \in D^+$ et d'un sous ensemble de n^- exemples négatifs $x_i \in D^-$. Que fait le classifieur suivant ? Quel est le principal défaut de cet algorithme d'apprentissage ?

$$\hat{y} = f(x) = sign\left(\left\|x - \frac{1}{n^-}\sum_{D^-} x_i\right\|_1 - \left\|x - \frac{1}{n^+}\sum_{D^+} x_i\right\|_1\right)$$

Et celui-ci ?

$$\hat{y} = f(x) = sign\left(\max_{x_i \in D^-} \langle x, x_i\rangle - \max_{x_i \in D^+} \langle x, x_i\rangle\right)$$

(2) A quoi sert-il de faire de la validation croisée plutôt que de faire un simple *split* en un *training set* et un *validation set* ?

(3) Que fait la fonction *sklearn.linear_model.lasso_path(X, y, n_alphas=1000)* ?

Que contiendront les vecteurs colonnes `coefs[:,0]` et `coefs[:,999]` retournés par cette fonction ?

(4) Etant donné un ensemble de B estimateurs $\widehat{f_b}(x)$ et un ensemble de test X composé de 9 échantillons, on observe les prédictions en figure 4. D'après vous, avec quelle type de méthodes a été appris cet ensemble d'estimateurs (gradient boosting ou bagging) ? Pourquoi ?

```
f_0=[ 0.21 -0.62 -0.62  0.21  0.21  0.21  0.21  0.21  0.21]
f_1=[ 0.15  0.15 -0.73  0.15  0.15  0.15  0.15  0.15  0.15]
f_2=[ 0.17 -0.52 -0.52  0.17  0.17  0.17  0.17  0.17  0.17]
f_3=[ 0.12  0.12 -0.6   0.12  0.12  0.12  0.12  0.12  0.12]
f_4=[ 0.14 -0.43 -0.43  0.14  0.14  0.14  0.14  0.14  0.14]
f_5=[ 0.1   0.1 -0.5   0.1   0.1   0.1   0.1   0.1   0.1]
f_6=[ 0.12 -0.36 -0.36  0.12  0.12  0.12  0.12  0.12  0.12]
f_7=[ 0.12 -0.3  -0.3   0.12  0.12 -0.3   0.12  0.12  0.12]
f_8=[ 0.07 -0.46  0.07  0.07  0.07  0.07  0.07  0.07  0.07]
f_9=[ 0.07  0.07 -0.37  0.07  0.07  0.07  0.07  0.07  0.07]
f_10=[ 0.1  -0.23 -0.23  0.1   0.1  -0.23  0.1   0.1   0.1 ]
f_11=[ 0.05 -0.37  0.05  0.05  0.05  0.05  0.05  0.05  0.05]
f_12=[ 0.06  0.06 -0.29  0.06  0.06  0.06  0.06  0.06  0.06]
f_13=[ 0.07 -0.21 -0.21  0.07  0.07  0.07  0.07  0.07  0.07]
f_14=[ 0.06  0.06 -0.19  0.06  0.06 -0.19  0.06  0.06  0.06]
```

(5) Quelle est l'expression de la règle de décision d'un SVM ? Qu'appelle-t-on les vecteurs supports ?

(6) Qu'est-ce que l'astuce du noyau ? Quel est son intérêt ?

2. L'inférence Bayésienne

Nous souhaitons mettre en place une plateforme permettant de prédire le vainqueur d'une élection en temps réel au fur et à mesure des dépouillements. Nous avons dès le départ (i.e. avant le premier dépouillement) une idée du résultat grâce aux sondages qui ont pu être réalisé au préalable. Nous supposons donc que le pourcentage d'individus qui voteraient pour le candidat A est décrit par la variable aléatoire p suivant une distribution beta de paramètres α et β. Ces paramètres forment notre *a priori* et dépendent donc des sondages. Leur amplitude donne notre confiance en ce dernier. Dit autrement, nous avons :

$$p \sim \mathcal{B}(\alpha, \beta),\ \ \alpha, \beta \in \mathbb{R}^{+*}$$

Nous connaissons les moments de la loi beta :

$$\mathbb{E}[p] = \frac{\alpha}{\alpha+\beta} \text{ et } \mathbb{V}\mathrm{ar}[p] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Contrairement, à un sondage, nous ne recevons par les votes un par un, mais agrégés par bureau de vote. Le bureau de votes 45 pourrait donc nous dire que parmi ses 1267 votants, 856 ont voté pour le candidat A. On modélise donc le résultat d'un bureau de vote par une variable aléatoire notée $x^{(i)}$ indiquant le nombre de votant pour le candidat A, sachant que le i^{eme} bureau de vote possède $n^{(i)}$ électeurs. De plus, $x^{(i)}$ suit une loi binomiale de paramètres $n^{(i)}$ et p. Dit autrement, nous avons :

$$x^{(i)} \sim Bi(n^{(i)}, p)$$

Nous connaissons également les moments de la binomiale :

$$\mathbb{E}[x^{(i)}] = n^{(i)}p \text{ et } \mathbb{V}\mathrm{ar}[p] = n^{(i)}p(1-p)$$

On rappelle que la loi beta possède la densité de probabilité suivante :

$$f_{\mathrm{beta}}(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ x \in [0,1], \alpha, \beta \in \mathbb{R}^{+*}$$

Et la binomiale la suivante :

$$f_{\mathrm{binom}}(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k},\ k \in \{0, \dots, n\},\ n \in \mathbb{N},\ p \in [0,1]$$

Remarque 1 : La densité de la loi beta est parfois définie de la manière suivante :

$$f_{\mathrm{beta}}(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}_{[0,1]}(x),\ \alpha, \beta \in \mathbb{R}^{+*},\ x \in \mathbb{R}$$

où $\mathbb{1}_{[0,1]}(x)$ vaut 1 si $x \in [0,1]$ et 0 sinon. Cela permet d'avoir une densité définie sur $\mathbb{R}$ et non pas uniquement sur $[0,1]$.

Questions

(1) Qu'est-ce qu'une loi conjuguée ?
(2) À quoi servent les méthodes de Monte Carlo en général et les méthodes de Markov Chain Monte Carlo en particulier ?
(3) Dans le modèle ci-dessus, quels sont les paramètres ? Et les hyper-paramètres ?
(4) Quel est la loi a posteriori du paramètre p (la proportion de gens qui votent pour le candidat A) ?
(5) Notons $F(x; \alpha, \beta)$, la fonction de répartition de la loi beta. Quelle est la probabilité que le paramètre p soit compris entre 0.2 et 0.45 (donnez la forme analytique de la solution) ?
(6) Gardons $F(x; \alpha, \beta)$, la fonction de répartition de la loi beta. De quelle manière celle-ci est-elle reliée à la densité de probabilité de la loi beta $f_{\text{beta}}(x; \alpha, \beta)$?

Remarque 2 : Les bureaux de vote donnant leur résultat en premier n'ont généralement pas les mêmes dynamiques électorales que ceux donnant leur résultat en dernier. Ainsi, il serait préférable d'avoir une modélisation beaucoup plus fine en pratique.
