Examen - Fondamentaux théoriques en Machine Learning

Nicolas Bourgeois

SCIA, S9, 2019-2020

L'examen dure 2 heures. Tous les documents sont autorisés. Le rendu s'effectue au format pdf dans un seul fichier. Les graphiques peuvent être réalisés avec n'importe quel logiciel/librairie ou même scannés. Les exercices sont indépendants et les questions du dernier exercice sont indépendantes.

Exercice 1

Dans tout cet exercice, on considère le tableau de données suivant, contenant les prédictions de deux estimateurs sur un échantillon de test.

X1	X2	Y	Y_EST_1	Y_EST_1
1	2	T	T	T
2	2	F	T	F
1	1	T	F	F
1	2	F	T	T
1	2	T	T	T
2	2	F	T	F
2	1	F	T	F
2	2	T	T	F
2	1	F	T	F
2	1	F	T	F
1	2	T	T	T
1	1	F	F	F
2	1	T	T	F
1	1	F	F	F
2	2	F	T	F
2	1	F	T	F
1	2	T	T	T

On suppose une fonction de perte qui pénalise de 1 en cas de faux positif (T est prédit alors que F est attendu) et de x en cas de faux négatif.

a) Quel est le risque empirique, sur l'échantillon considéré, associé respectivement aux estimateurs EST_1 et EST_2 ? Comparez-les en fonction des valeurs de x.

On suppose à partir de maintenant que x = 2.

b) Construisez un estimateur qui minimise le risque empirique (parmi l'ensemble des estimateurs possibles, sans hypothèse).

c) Expliquez pourquoi il n'était pas nécessaire ici de se placer dans le cadre bayésien naïf. (Au moins deux raisons)

d) Produisez maintenant les tableaux agrégés où la corrélation n'apparaît plus entre X_1 et X_2 (selon le modèle de ceux utilisés dans les exercices du cours).

e) Utilisez ces tableaux pour produire l'estimateur bayesien naïf optimal. Comparez avec l'estimateur optimal trouvé plus haut.

Exercice 2

Dans tous cet exercice on considère le jeu de données suivant :

x_1	x_2
0	2
0	3
0	4
1	5
2	6
4	7
6	7
7	6
7	5
2	3
4	3
4	4
5	2
5	5

a) Dessinez les points et identifiez deux composantes probables (les 9 premiers points contre les 5 derniers). Quelles méthodes non supervisées vous paraissent à même de les distinguer ?

b) Représentez le dendogramme d'une classification hiérarchique ascendante. N'oubliez pas de préciser quelle fonction vous avez choisie pour évaluer le coût de l'intégration (l'axe des ordonnées) mais pas besoin de faire les calculs exacts ; par contre l'ordre des associations doit être exact.

c) Il s'avère que les deux composantes correspondent effectivement à deux classes qu'on appellera 1 (à l'intérieur) et -1 (à l'extérieur). Est-il possible de les séparer avec un SVM linéaire ?

d) Fixez deux niveaux de pénalisation arbitraires et dessinez approximativement les SVM linéaires optimaux associés.

e) Proposez une kernelisation qui permette de séparer les deux classes, par exemple en rajoutant une fonction $x_3 = f(x_1, x_2)$.

Exercice 3

Pour chacun des problèmes suivants, expliquez comment vous procéderiez :

a) Si, étant donné un tableau croisé d'effectifs, je voulais savoir si les deux variables concernées sont indépendantes ?

b) Si je voulais trouver le nombre maximum de points pulvérisables à la surface d'une sphère ?

c) Si je voulais expliquer à une personne non-spécialiste la différence entre risque et ambiguité ?