Partiel 2020 - Fondamentaux Théoriques en Machine Learning

SCIA 2021

Exercice 1

a) Matrice de confusion pour EST_1 :

	FALSE	**TRUE**
FALSE	2	8
TRUE	1	6

Matrice de confusion pour EST_2 :

	FALSE	**TRUE**
FALSE	9	1
TRUE	3	4

Le risques empiriques pour les deux estimateurs sont donc :

- **EST_1** : $\frac{8+x}{17}$
- **EST_2** : $\frac{1+3x}{17}$

Si x < 3.5 l'estimateur 1 est avantagé et si x > 3.5 l'estimateur 2 est avantagé.

b)

c)

d) Tableau pour Y = TRUE avec X1 en ligne et X2 en colonne :

	1	**2**
1	1	4
2	1	1

Tableau pour Y = FALSE avec X1 en ligne et X2 en colonne :

	1	2
1	2	1
2	4	3

e) Calcul des probabilités en utilisant le théorème de Bayes (les divisions par P(X) sont retirés pour un soucis de lisibilité) :

- **X1 = 1 et X2 = 1 :**

$$P(Y = T \mid X1 = 1 \ \& \ X2 = 1) = \frac{7}{17} * \frac{1}{7} = \frac{7}{119} \approx 0{,}05882352941$$

$$P(Y = F \mid X1 = 1 \ \& \ X2 = 1) = \frac{10}{17} * \frac{2}{10} = \frac{20}{170} \approx 0{,}1176470588$$

- **X1 = 1 et X2 = 2 :**

$$P(Y = T \mid X1 = 1 \ \& \ X2 = 2) = \frac{7}{17} * \frac{4}{7} = \frac{28}{119} \approx 0{,}2352941176$$

$$P(Y = F \mid X1 = 1 \ \& \ X2 = 2) = \frac{10}{17} * \frac{1}{10} = \frac{10}{170} \approx 0{,}05882352941$$

- **X1 = 2 et X2 = 1 :**

$$P(Y = T \mid X1 = 2 \ \& \ X2 = 1) = \frac{7}{17} * \frac{1}{7} = \frac{7}{119} \approx 0{,}05882352941$$

$$P(Y = F \mid X1 = 2 \ \& \ X2 = 1) = \frac{10}{17} * \frac{4}{10} = \frac{40}{170} \approx 0{,}2352941176$$

- **X1 = 2 et X2 = 2 :**

$$P(Y = T \mid X1 = 2 \ \& \ X2 = 2) = \frac{7}{17} * \frac{1}{7} = \frac{7}{119} \approx 0{,}05882352941$$

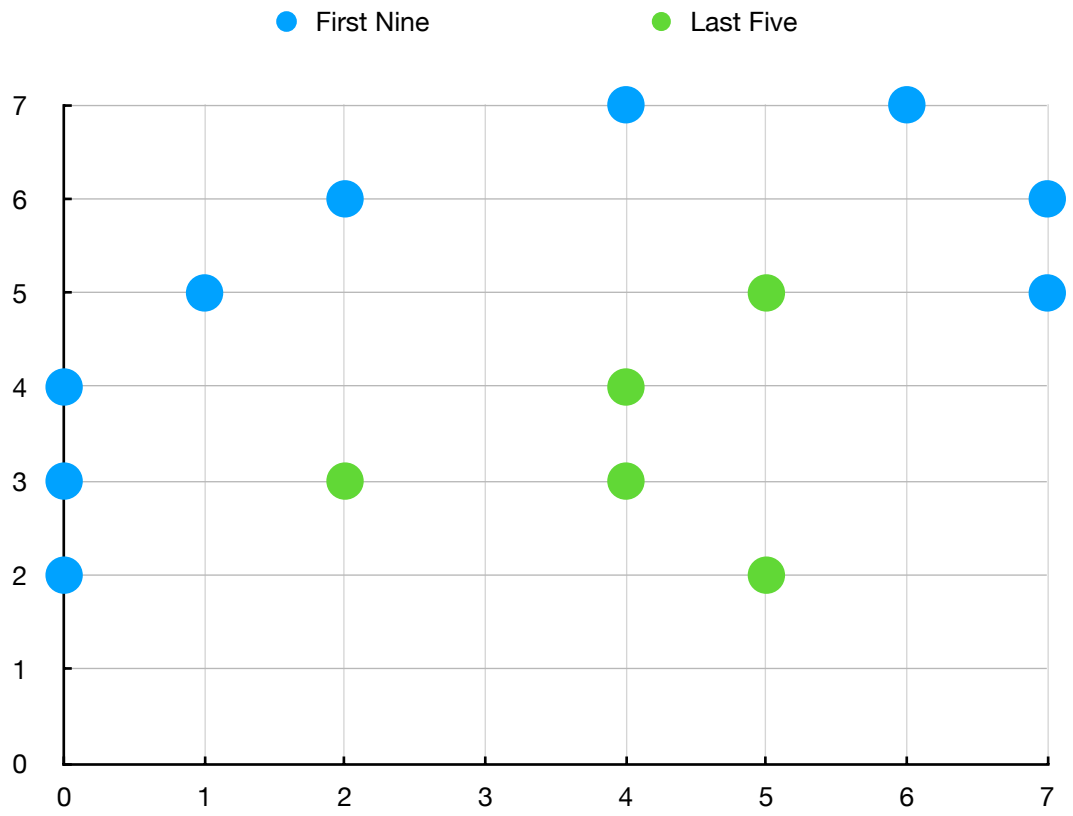
$$P(Y = F \mid X1 = 2 \ \& \ X2 = 2) = \frac{10}{17} * \frac{3}{10} = \frac{30}{170} \approx 0{,}1764705882$$

Nous obtenons donc ce tableau résumant l'estimateur bayesien naïf optimal :

	1	2
1	FALSE	TRUE
2	FALSE	FALSE

Exercice 2

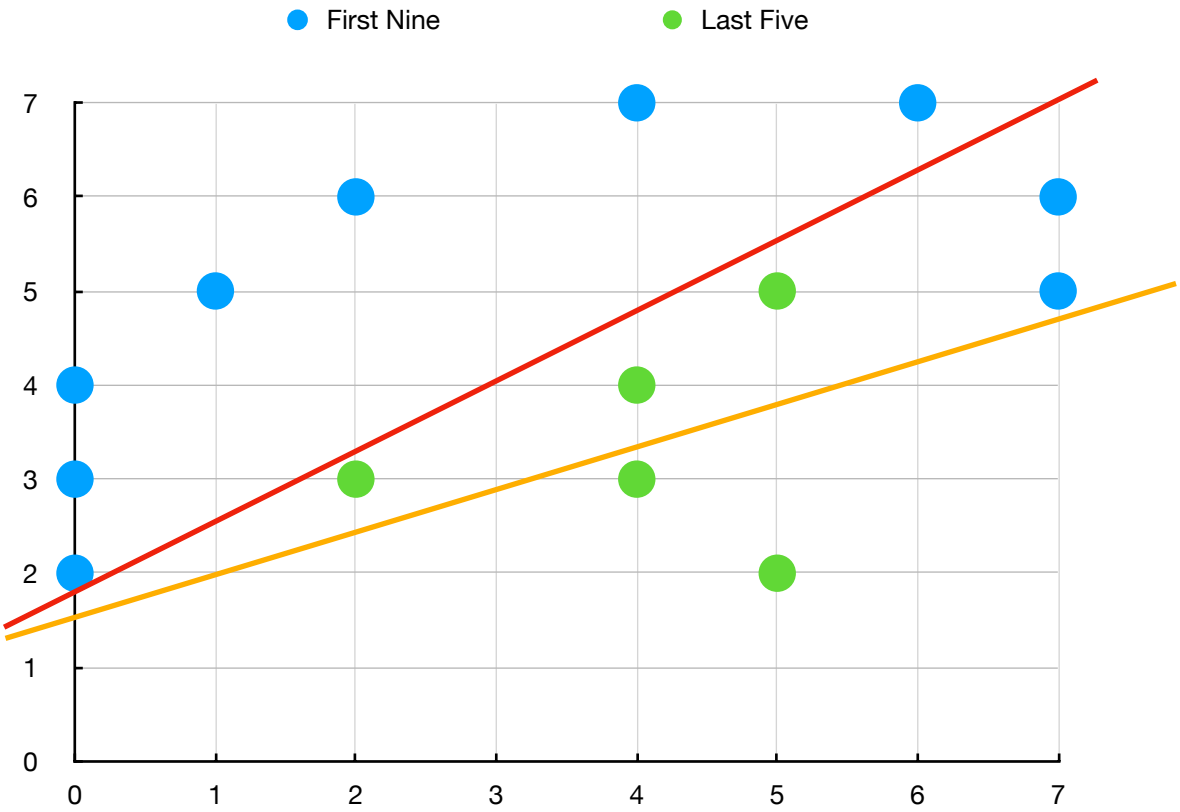
a)



b)

c) Il est possible de les séparer avec un SVM linéaire mais il ne sera pas optimal, dans le sens ou il ne pourra pas faire une séparation parfaite au vu de la disposition des points.

d)



Ici il est possible de voir en rouge un SVM avec une faible pénalisation pour les éléments négatifs (à l'extérieur) et une forte pénalisation pour les éléments positifs (verts). En orange c'est le contraire, un SVM avec une forte pénalisation pour les éléments négatifs et une faible pénalisation pour les éléments positifs.

e) Je propose la fonction suivante :

f(x1, x2) = {
- 0 si x1 > 1.5 ET x1 < 5.5 ET x2 <= 5
- 1 sinon
}

Exercice 3

a) Pour savoir si les deux variables concernées par un tableau croisé d'effectifs sont indépendantes il est possible d'utiliser le test du x^2.

On calcul d'abord les effectifs croisés espérés (avec # étant le nombre d'évènements) :

$$E_{i,j} = \frac{1}{n} \sum \sharp\{X = i\}\sharp\{Y = j\}$$

Ensuite il faut calculer les effectifs croisés observés :

$$O_{i,j} = \sharp\{X = i \& Y = j\}$$

Et finalement il est possible de calculer l'écart relatif entre les deux variables :

$$T = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Un T faible signifie que les variables sont indépendantes.

b) Pour trouver le nombre maximum de points pulvérisables il est possible de calculer la dimension de Vapnik-Chervonenkis. En effet cette dimension corresponds au nombre maximum d'éléments pulvérisable.

c) Intuitivement je dirais que la différence entre ambiguïté et risque, par exemple dans le contexte d'une prise de décision, marque la différence entre une décision prise en prenant conscience des informations externes (risque) et une décision prise sans prendre en compte ces mêmes informations (ambiguïtés) par exemple à cause de biais.