Séance 7 : apprentissage statistique

Critères de construction d'un classifieur

	longueur pétales	largeur des pétales	variété
apprentissage	4,7	1,4	versicolor
	4,5	1,5	versicolor
	4,9	1,5	versicolor
	6	2,5	virginica
	5,1	1,9	virginica
	5,9	2,1	virginica
test	3,9	1,4	??
	5,1	2,4	??
	3	4,2	??

Une tâche de classification consiste à construire, à partir d'un ensemble d'apprentissage, un modèle permettant de proposer des inférences.

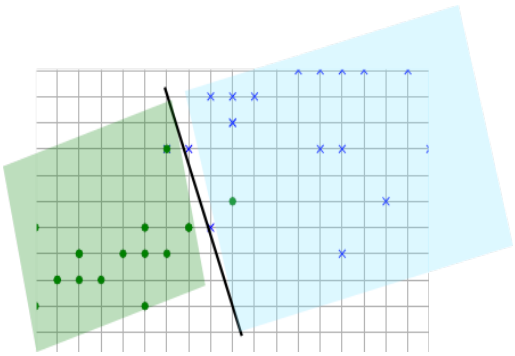
A ce modèle est associée une *complexité* et une *erreur de généralisation*.

Nous allons illustrer chacune de ces notions avec l'exemple des iris de Fisher.

Ensemble d'apprentissage : On dispose à la fois des variables prédictives et des étiquettes pour chaque donnée.

Exemple : chaque donnée correspond à un iris, les variables prédictives sont longueur/largeur de pétale/sépale, l'étiquette est virginica, versicolor ou setosa.

Modèle : c'est une fonction qui aux variables prédictives associe la catégorie prédite. Cette fonction détermine une frontière de décision qui la caractérise



Exemple : on choisit de modéliser la frontière de décision par une droite :

$f(x,y) = 1 \quad \text{si} \quad a \times x + b \times y - c > 0$

$f(x,y) = 0 \quad \text{si} \quad a \times x + b \times y - c < 0$

Reste ensuite à décider ce qui se passe pour les points situés sur la frontière de décision.

Les paramètres du modèle sont les coefficients a, b et c qui déterminent l'orientation et la position de la droite dans le plan (cf séance sur les équations de droite).

Les coefficients du modèle sont estimés grâce aux données d'apprentissage, de façon à minimiser l'erreur de classification.

Ensemble de test : il sert à estimer l'erreur de généralisation : la capacité du modèle à classifier de nouveaux échantillons. En pratique, les données d'apprentissage et de test sont au choix du modélisateur : l'erreur de test sert à mesurer le risque associé à l'utilisation du modèle.

Travaux pratiques : un classifieur virginica / versicolor

L'échantillon d'apprentissage et l'échantillon de test sont représentées pages 3 et 4

1. Etape d'apprentissage :
   (a) Est-il possible de construire un classifieur sans erreur ?

   (b) Montrer que la plus petite erreur possible est 1/60 et décrire la droite qui permet d'obtenir cette erreur.

   (c) Quelle est l'équation $y = m\,x + p$ associée à cette droite ?

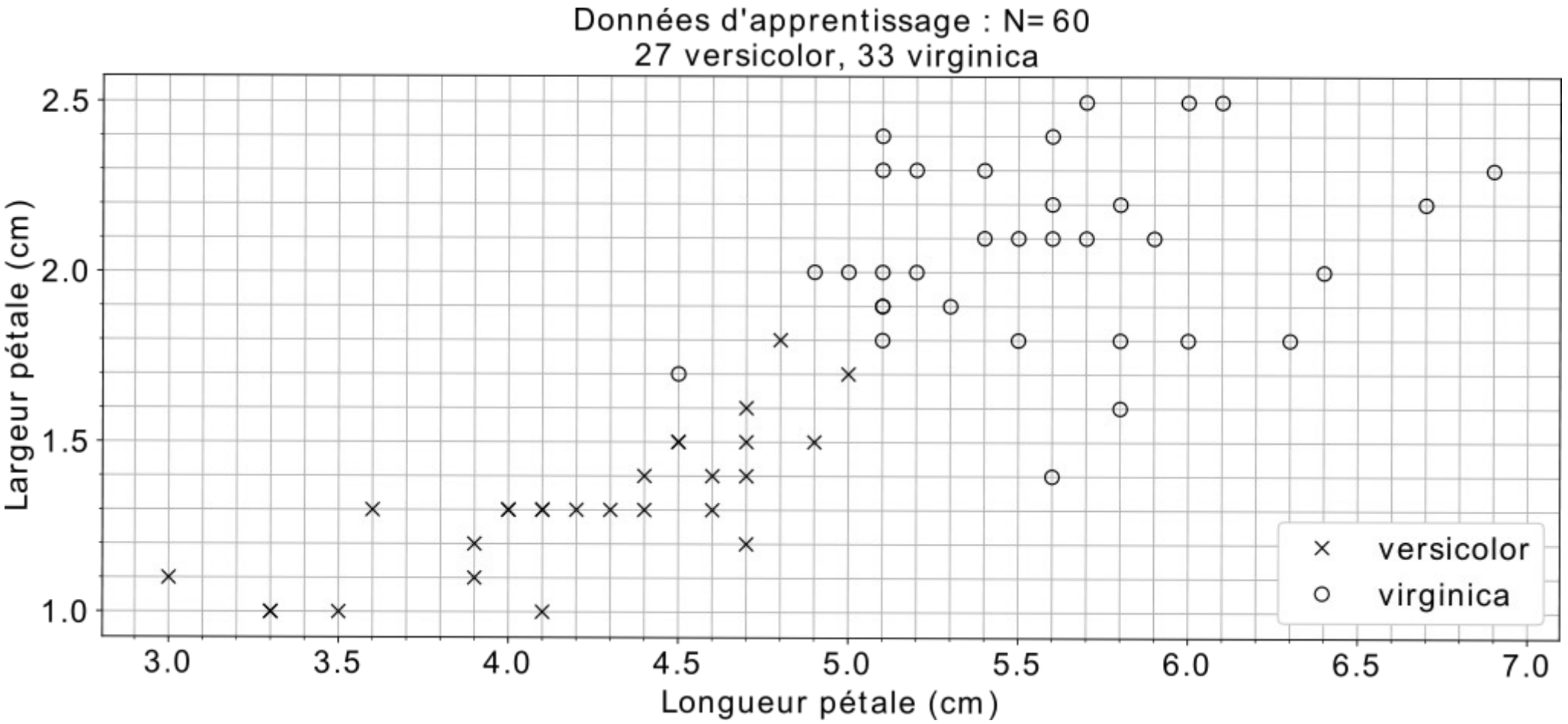
   (d) Vérifier que :

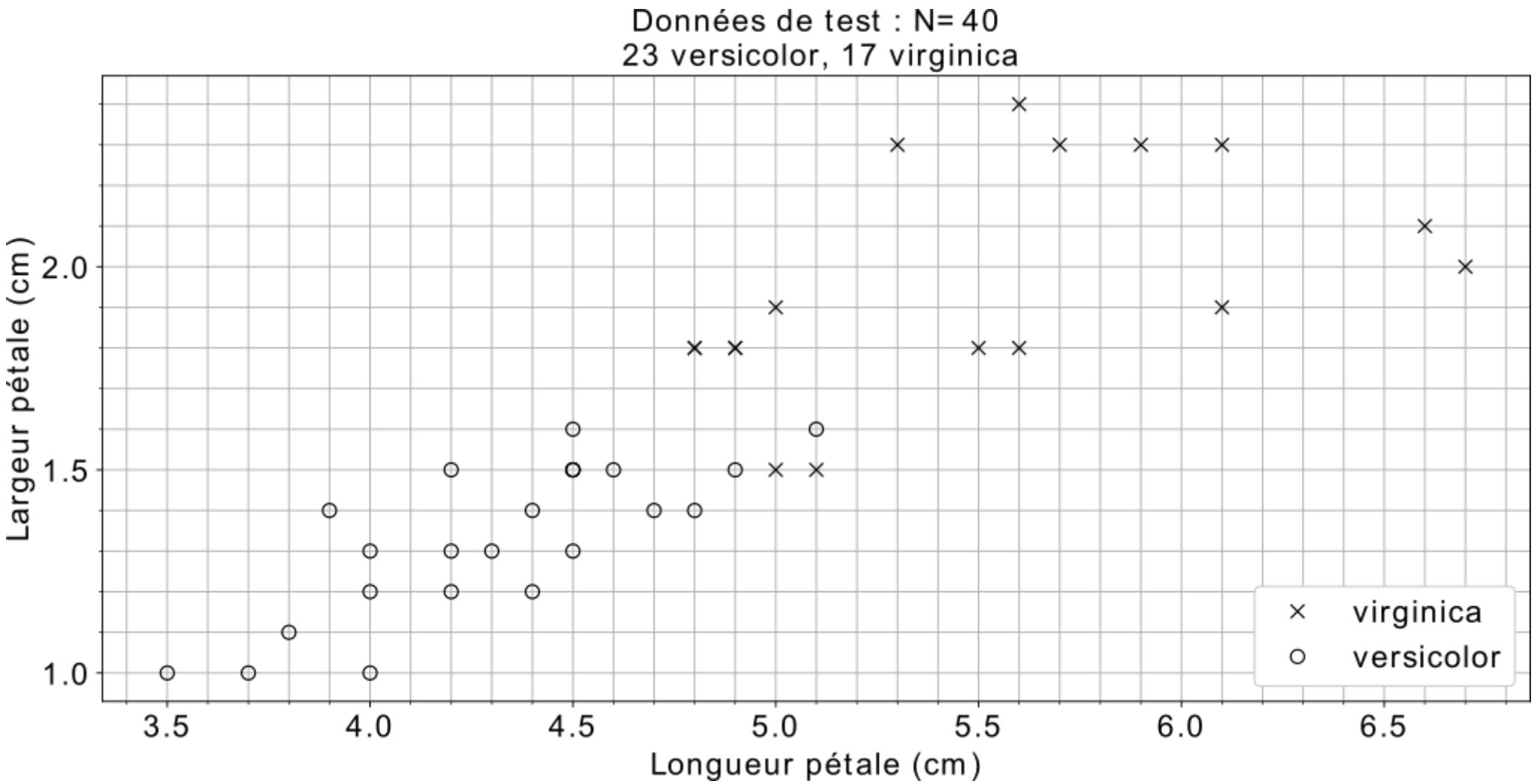
      1. pour le versicolor situé en (3;1,1), on a bien $y - (m\,x + p) < 0$
      2. pour le virginica situé en (5;2), on a bien $y - (m\,x + p) > 0$

2. Etape de test :

   (a) Tracer la frontière de décision estimée à l'étape précédente sur la figure « Données test »

   (b) Comparer l'erreur commise à l'apprentissage, et l'erreur commise à l'étape de test


Données d'apprentissage : N= 60
27 versicolor, 33 virginica
Largeur pétale (cm)
2.5
2.0
1.5
1.0
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
Longueur pétale (cm)
versicolor
virginica


Données de test : N= 40
23 versicolor, 17 virginica
Largeur pétale (cm)
1.0
1.5
2.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
Longueur pétale (cm)
virginica
versicolor