

Apprentissage statistique

Classification

Stéphane Marchand-Maillet
Department of Computer Science



9 mai 2023

Résumé

Le but de ce programme de travail est de comprendre et mettre en pratique certains mécanismes de la classification automatique des données par apprentissage statistique. Le langage de support pour la programmation des algorithmes est **python**. On apprend d'abord à accéder aux données et leurs labels, puis on explore les algorithmes de classifications “k plus proches voisins”, “perceptron” et le “neurone logistique”

La **classification de données** consiste à apprendre et reconnaître automatiquement la catégorie (**classe**) d'un objet à partir de données exemples.

Exemple 1 (Classification d'image). Étant données des images des chiffres “0” et “1”, on veut automatiquement attribuer un **label** indiquant si l'image en question est un “1”. Les caractéristiques (**donnée**) d'une image seront les valeurs des pixels de l'image et le label sera 0 si l'image montre un “0” et le label sera 1 si l'image montre un “1”.

Exemple 2 (Classification de fruits). Étant données des prunes et des cerises, on veut automatiquement attribuer un label à chaque fruit indiquant si c'est une cerise (classe “cerise”, label 1) ou une prune (classe “prune”, label 0). La caractéristique d'un fruit sera son diamètre.

Remark 1 (Rapport de travail). Pour chacune des tâches de ce travail, il est demandé de rédiger un rapport détaillant et documentant les solutions trouvées aux questions posées. Les résultats des expériences seront décrits et commentés afin de tirer des conclusions argumentées de ces expériences.

Remark 2 (Difficulté). Les tâches sont marquées par difficulté estimée : ① : facile, ② : moyen, ③ : difficile

1 Classification de données

Un modèle de classification est un modèle de décision dépendant de paramètres. Il permet de prédire la classe (0 ou 1 pour la classification binaire) d'un objet inconnu en fonction des caractéristiques mesurées \mathbf{x} de cet objet.

Exemple 3 (Modèle de classification de fruits). Pour classifier des fruits (objets) selon leur diamètre (caractéristique), on définit une classe positive (classe 1), par exemple “cerise”, et les autres fruits seront de la classe négative (classe 0). Pour différencier la classe “cerise” (positive) de la classe “prune” (négative) on doit choisir un seuil (paramètre) sur le diamètre, en deçà duquel le fruit sera de la classe positive, sinon négative (décision).

Pour classifier les données, on distingue 3 phases. Les deux premières phases que sont l'apprentissage et le test visent à produire le modèle de classification (configurer les paramètres). Ces phases se font à l'aide de données connues (caractéristiques et labels connus). La finalité du modèle est son utilisation avec des données inconnues (caractéristiques connues, labels inconnus).

1.1 Apprentissage

Le but de la phase d'apprentissage (ou entraînement) est de configurer (à travers un algorithme d'apprentissage) les paramètres pour que les prédictions du modèle soient correctes sur l'ensemble d'entraînement.

Remark 3. Dans ce travail on verra 3 algorithmes d'apprentissage (kNN, perceptron, neurone logistique).

La qualité de l'apprentissage se mesure par une différence (erreur) pour chaque objet de l'ensemble d'entraînement, entre la classe prédite et la classe connue pour cet objet. La somme des erreurs forme l'erreur d'entraînement.

Remark 4. Dans le cas d'une classification binaire linéaire (2 classes séparées par une frontière linéaire, comme étudiée ici avec le perceptron ou le neurone logistique) on espère une erreur d'entraînement nulle pour des données séparables (par une frontière linéaire) et une erreur minimale pour des données non-séparables.

On voit donc que l'ensemble des données d'entraînement sera l'ensemble des caractéristiques des objets exemples pour l'entraînement avec leur label de classe (0 ou 1) que l'on fournit au modèle pour s'entraîner.

1.2 Test

Une fois les paramètres du modèle fixés, on simule son utilisation sur un cas réel en utilisant un ensemble de test. On estime la qualité du modèle lors de son utilisation en étudiant ses prédictions sur des objets dont on connaît les labels de classe (mais que l'on ne fournit pas au modèle). En comparant les erreurs faites par le modèle une fois ses paramètres fixés, on obtient l'erreur de test.

On voit donc que l'ensemble des données de test sera l'ensemble des caractéristiques des objets exemples pour l'entraînement avec leur label de classe (0 ou 1) que l'on ne fournira pas au modèle pour s'entraîner.

1.3 Utilisation

Une fois le modèle (et ses paramètres) validé(s) par une erreur de test acceptable, il peut être utilisé dans les conditions réelles. On reçoit un objet dont on mesure les caractéristiques que l'on fournit au modèle, qui produit une prédiction du label de classe de l'objet. Ne connaissant pas la classe réelle, on se fie à la qualité de l'entraînement et du test pour accepter cette prédiction.

Example 4 (Lecture de plaque minéralogique). Une société technologique entraîne le modèle pour reconnaître les chiffres de la plaque minéralogique (leur attribuer une classe 0, 1, ...). La société de parking utilisant ce modèle de prédiction fera des tests avec des plaques connues pour tester la fiabilité du système. Elle pourra alors utiliser ce système pour reconnaître et facturer automatiquement les véhicules entrant et sortant des parking.

On voit donc clairement que l'on a besoin de jeux de données d'entraînement et de test contenant les caractéristiques des objets et leurs labels. En pratique, on obtiendra une ensemble de données contenant les données (data) et les labels correspondants (labels) que l'on partagera en données d'entraînement (80% par exemple) et test (20%).

On s'attache donc d'abord à obtenir et préparer ces données pour l'entraînement et le test des modèles étudiés.

2 Préparer les données

Le but de cette partie est d'accéder à des données sur lesquelles appliquer des algorithmes de traitement. Les données seront soit chargées à partir de fichiers, soit générées aléatoirement selon un modèle de données.

Pour l'apprentissage statistique, une donnée est un vecteur $\mathbf{x}_i \in \mathbb{R}^D$ qui contient les caractéristiques de l'objet i . Son label sera noté $y_i \in \{0, 1\}$. On notera N le nombre de données disponibles pour l'entraînement.

Remark 5 (Dimension des données). Dans ce travail on se limitera aux valeurs de dimension $D = 1$ ou $D = 2$ afin de pouvoir visualiser les résultats (1D ou 2D) mais les procédures définies ici se généralisent sans difficulté aux cas $D > 2$.

2.1 Charger des données à partir de fichiers

Les données sont souvent accessibles à partir de fichiers informatiques qu'on lit pour obtenir les données. Les formats de ces fichiers peuvent être très variables (.txt, .ods, .xlsx, ...) et on utilisera l'outil de lecture adapté au format. Dans ce travail on utilise un format texte^(a) générique : le format CSV (comma separated values). Dans ce format, chaque ligne contient des valeurs séparées par des virgules. Dans notre cas, chaque ligne i contient les D données d'un vecteur \mathbf{x}_i séparées par des virgules.

(a). Les fichiers CSV peuvent être ouverts avec un éditeur de texte type **notepad** ou **vscode** ou être importés dans **open calc** ou **Excel**

Exercice 1 (① Lecture de fichier CSV). Étant donnés les fichiers fournis :

- (i) Écrire un programme qui, à partir d'un nom de fichier **xxx** lit les données dans le fichier **xxx-data.csv** et les labels correspondant dans le fichier **xxx-label.csv**. On stockera le résultat dans les variables (*numpy arrays*) **data** et **label**.

Hint. On pourra utiliser la fonction **numpy genfromtxt** pour lire ces fichiers

- (ii) Afficher le nombre **N** et la dimension **D** de ces données
- (iii) Sachant que pour ces données on a soit **D = 1**, soit **D = 2**, afficher graphiquement ces données en utilisant une couleur par label.

Hint. ② On pourra prévoir que les données peuvent éventuellement être de dimension **D > 2** et comporter plus que 2 classes. Dans ce cas, on proposera d'afficher 2 dimensions et les classes au choix.

Exercice 2 (① Utilisation de packages). Certaines librairies **python** offrent des facilités pour charger des ensembles de données standard.

- (i) Utilisez le package **datasets** de la librairie **sklearn** pour charger les ensembles **iris** et **wine**
- (ii) ② Testez l'affichage de ces données avec les outils de l'exercice 1

Exercice 3 (② Lecture des données MNIST).

- (i) Utilisez le package **datasets** de la librairie **keras** pour charger l'ensemble **MNIST**
- (ii) Transformez cet ensemble en un ensemble de données vectorielles en alignant les colonnes des images en un seul vecteur

Hint. On pourra utiliser la fonction **numpy reshape** qui réorganise les valeurs dans les tableaux

2.2 Générer des données automatiquement

Si on connaît les statistiques des données que l'on veut générer, on peut utiliser le générateur aléatoire (package **random**) pour tirer des valeurs au hasard tout en respectant les statistiques de l'étude.

Exercice 4 (Données de classification pour les fruits ①). On sait qu'une prune a un diamètre moyen égal à 6cm pour un poids moyen de 30g alors qu'une cerise a un diamètre moyen de 2cm pour un poids moyen de 5g.

- (i) En utilisant les fonctions **numpy randint** et **randn** du package **random** de **python**, générez des ensembles de taille (**N**) variable ou fixée de caractéristiques (diamètre et poids) et labels de cerises et de prunes
- (ii) Écrivez ces données dans des fichiers de format **CSV** de façon à pouvoir les lire et les afficher avec les programmes de l'exercice 1

Exercice 5 (② Données sur des cercles).

- (i) Grâce aux fonctions aléatoires **randint** et **rand**, générez des données se situant **autour de cercles séparés, s'intersectant ou concentriques** ^(b)
- (ii) Écrivez ces données dans des fichiers de format **CSV** de façon à pouvoir les lire et les afficher avec les programmes de l'exercice 1

Exercice 6 (① Utilisation de packages). Certaines librairies **python** offrent des facilités pour générer des ensembles de données de modèles standard.

- (i) Utilisez le package **datasets** de la librairie **sklearn** pour générer les différents types de données proposées.
- (ii) Écrivez ces données dans des fichiers de format **CSV** de façon à pouvoir les lire et les afficher avec les programmes de l'exercice 1

Exercice 7 (② Utilisation de Open Office ou Excel).

- (i) Utilisez **Open Calc** ou **Excel** pour générer des données de votre choix. Documentez
- (ii) Écrivez ces données dans des fichiers de format **CSV** de façon à pouvoir les lire et les afficher avec les programmes de l'exercice 1

(b). On rappelle que les points **x** d'un cercle de rayon **r** sont à une distance $d(\mathbf{x}, \mathbf{c}) = r$ du centre **c**. Les coordonnées $\mathbf{x} = [\mathbf{x}_{[0]}, \mathbf{x}_{[1]}]$ de ces points et du centre $\mathbf{c} = [\mathbf{c}_{[0]}, \mathbf{c}_{[1]}]$ sont donc tels que $d(\mathbf{x}, \mathbf{c}) = \sqrt{(\mathbf{x}_{[0]} - \mathbf{c}_{[0]})^2 + (\mathbf{x}_{[1]} - \mathbf{c}_{[1]})^2} = r$

2.3 Extraire les statistiques de données

Les données sont donc un tableau **numpy data** (*numpy array*) de dimension $N \times D$ dont chaque ligne est une donnée (les caractéristiques de l'objet correspondant) et chaque colonne les valeurs de cette caractéristique prises par les données^(c). On va faire des statistiques pour étudier la dispersion de ces données.

Exercice 8 (① Boîtes englobantes et intersection).

- (i) Pour un choix de données obtenues dans les étapes précédentes, calculez la **boîte englobante** des données (minimum, maximum) de chaque classe, ainsi que la **médiane** de chaque caractéristique. Pour ce faire, utilisez les fonctions **numpy median**, **min** et **max** sur les variables **data** et **label**
- (ii) Pour chacune des situations ci-dessus vérifiez si des données sont présentes dans l'intersection des boîtes englobantes des données. Quelle est votre interprétation ?

On peut enrichir ces statistiques par des statistiques plus globales.

Exercice 9 (① Moyenne et variance).

- (i) Pour un choix de données obtenues dans les étapes précédentes, calculez la moyenne (et la variance) de chacune des caractéristiques des données pour chaque classe.

Hint. On pourra utiliser simplement les fonctions **numpy** ou réimplémenter ces fonctions et comparer les résultats.

- (ii) Pour les données générées aléatoirement (section 2.2) vérifiez que les paramètres retrouvés correspondent bien aux paramètres de génération
- (iii) ② Variez le nombre de données considérées et comparez la qualité des l'estimation des paramètres de départ. Quelle est votre interprétation ?

On peut aussi visualiser la distribution des caractéristiques des données.

Exercice 10 (① Histogrammes).

- (i) Pour un choix de données obtenues dans les étapes précédentes, tracer les histogrammes de la distribution des caractéristiques des données, par classes.

Hint. On utilisera la fonction **hist** du package **pyplot** de la librairie **matplotlib**. On pourra varier le paramètre **bins** pour visualiser au mieux l'histogramme des données

- (ii) Quelle est votre interprétation de la lecture de ces histogrammes par rapport aux résultats des exercices 8 et 9 ?

3 Le classifieur “k-plus proches voisins”

Le classifieur des “k-plus proches voisins” kNN (*k-nearest neighbors*) utilise la similarité entre les objets pour effectuer la classification. Pour un objet inconnu \mathbf{x}_q , on cherche les k objets \mathbf{x}_i les plus proches (voisins) dans notre ensemble de données (pour lesquels on connaît les labels) et on choisit pour \mathbf{x}_q le label majoritaire (le plus fréquent) dans cet ensemble.

Remark 6. ▲ Ce classifieur ne nécessite pas d'entraînement, mis à part le choix de la valeur k . De plus, ce classifieur ne définit pas explicitement de frontière de classification globale mais attribue un label à chaque donnée selon le label de ses voisins (localement).

3.1 Proximité entre objets

La proximité $\text{prox}(\mathbf{x}_i, \mathbf{x}_j)$ entre les objets \mathbf{x}_i et \mathbf{x}_j est mesurée par la distance entre leurs caractéristiques. Plus la valeur est faible, plus les objets sont proches. On a plusieurs choix^(d) ^(e) :

- La différence : $\text{prox}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{d=1}^D |\mathbf{x}_i[d] - \mathbf{x}_j[d]|$
- La distance Euclidienne : $\text{prox}(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_i, \mathbf{x}_j)^2 = \sum_{d=0}^{D-1} (\mathbf{x}_i[d] - \mathbf{x}_j[d])^2$

Exercice 11 (Calcul de distance Euclidienne entre objets). Ecrivez un programme qui

- Tire aléatoirement les caractéristiques de dimension $D = 2$ pour N objets
- Calcule leur distance Euclidienne par la formule $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i[0] - \mathbf{x}_j[0])^2 + (\mathbf{x}_i[1] - \mathbf{x}_j[1])^2}$

(c). Attention : pour l'analyse de données, les vecteurs sont classiquement des colonnes. Le premier indice est l'indice de ligne

(d). la notation $\sum_{i=1}^N v_i$ se lit “somme pour i égal 1 à N des valeurs v_i ”, soit $\sum_{i=1}^N v_i = v_1 + v_2 + v_3 + \dots + v_N$

(e). la notation $\mathbf{x}[k]$ se lit “la coordonnée k du vecteur \mathbf{x} ”

- Fait la même chose pour $D > 0$ par la formule $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{d=0}^{D-1} (\mathbf{x}_i[d] - \mathbf{x}_j[d])^2}$

Exercice 12 (Voisinage). Écrivez un programme qui, pour un ensemble de données obtenu précédemment :

- Choisit une donnée aléatoirement (fonction `numpy.random.randint` pour le choix d'indice)
- Calcule la proximité avec chacune des données de l'ensemble (cf au choix ci-dessus)
- Trie les valeurs de proximité dans l'ordre croissant (fonction `numpy.argsort`)
- Imprime les labels des k plus proches voisins ($k < N$ au choix)
- Affiche les données sous forme d'un `scatter` plot (si 2D) avec la donnée choisie en rouge (`c='r'`), les k plus proches voisins en vert (`c='g'`), le reste en noir (`c='k'`). On pourra varier en affichant les données de la classe 0 avec des cercles (`marker='o'`) et les données de la classe 1 avec des croix (`marker='x'`)

3.2 Classification par “k-plus proches voisins” (kNN)

Exercice 13 (Classification kNN). En utilisant le programme précédent pour $k < N$ fixé, écrire un programme qui :

- Choisit une donnée \mathbf{x} aléatoirement
- Sélectionne les labels des k données $\mathcal{V}^k(\mathbf{x})$ les plus proches
- Attribue à (prédit pour) \mathbf{x} le label majoritaire dans cet ensemble
- Affiche le label prédit pour \mathbf{x} et le compare au label connu
- Comparer les résultats avec ceux de la fonction `neighbors.KNeighborsClassifier` du package `sklearn` [1]

3.3 Evaluation

4 Le perceptron

Le perceptron est un algorithme de classification binaire (2 classes) dont l'entraînement détermine les paramètres d'une frontière de classification. Dans ce cas précis la frontière de classification est linéaire et partage l'espace des données en 2 parties (de chaque côté de la frontière pour la classe 0 et la classe 1.

4.1 Cas 1D

Dans le cas le plus simple où chaque donnée n'a qu'une seule caractéristique (par exemple le diamètre pour des fruits), la classification revient à déterminer (par entraînement) une valeur d'un seuil (noté w_0) sur la caractéristique. Si la valeur de caractéristique de la donnée i est supérieure à ce seuil ($\mathbf{x}_i[0] > w_0$) on attribue le label 0 à cette donnée sinon 1.

4.2 Cas 2D

4.3 Cas général

5 Le neurone logistique

A Rappels mathématiques

A.1 Notation

Une variable $v \in \mathbb{R}$ représente un nombre réel (élément de l'ensemble \mathbb{R} des nombres réels). Si on doit énumérer des valeurs, on notera $v_i \in \mathbb{R}$ afin que v_0, v_1, \dots représentent les différentes valeurs à énumérer.

Exemple 5. (diamètre des cerises). Étant données N cerises, on mesure leur poids et on obtient $p_0 = 5.1g$ pour la première cerise, $p_1 = 4.9g$ pour la deuxième cerise, ..., $p_{N-1} = 5.3g$ pour la N ème cerise.

Remark 7 (Listes python). Les indices commencent avec l'indice 0 pour rester cohérent avec la notation des listes python.

```
x = [5, 4, 6, 2, 3, 7]
print(x[0])
# retourne 5
```

Cette notation permet d'écrire facilement la somme (\sum) ou le produit (\prod)

$$\sum_{i=0}^{N-1} p_i = p_0 + p_1 + \dots + p_{N-1} \qquad \prod_{i=0}^{N-1} p_i = p_0 \cdot p_1 \cdot \dots \cdot p_{N-1}$$

```
x = [5,4,6,2,3,7]
print(sum(x))      # écrit 27
```

A.2 Vecteurs et matrices

Dans ce travail, on définit un vecteur \mathbf{x} comme un élément de \mathbb{R}^D , c'est à dire une liste de D nombres réels. Par convention un vecteur s'écrit en colonnes

$$\mathbf{x} = \begin{bmatrix} x_{[0]} \\ \vdots \\ x_{[D-1]} \end{bmatrix}$$

$x_{[i]} \in \mathbb{R}$ est la $i^{\text{ème}}$ composante de \mathbf{x} .

```
import numpy as np
x = np.array([5,4,6,2,3,7])
print(x[3])      # écrit 2
print(x.sum())   # écrit 27
```

On peut faire les opérations d'addition et de soustraction entre des vecteurs de même taille.

On peut étendre la notion de vecteurs à celle de matrices comme tableau de chiffres en accolant des vecteurs les uns à coté des autres. Par exemple la matrice $\mathbf{A} \in \mathbb{R}^{N \times M}$ est une aggrégation de M vecteurs de taille N :

$$\mathbf{A} = \begin{bmatrix} a_{00} & \dots & a_{0M} \\ \vdots & & \vdots \\ a_{N0} & \dots & a_{NM} \end{bmatrix}$$

Pour python, une matrice est un vecteur (`numpy array`) de vecteurs :

```
import numpy as np
A = np.array([[5,4,6],
              [2,3,7],
              [2,8,4],
              [7,10,12]])
print(A.shape)    # donne la taille de la matrice soit (4,3)
print(A[2,1])     # écrit 8
```

On peut faire les opérations d'addition et de soustraction entre des matrices de même taille. On notera aussi a_{ij} le $j^{\text{ème}}$ élément de la $i^{\text{ème}}$ ligne de \mathbf{A} ($\mathbf{A}_{[i,j]}$).

A.3 Distance

Dans un repère Cartésien d'origine $\mathbf{o} = [0, 0]$, un vecteur $\mathbf{x} = [x_{[0]}, x_{[1]}]$ est équivalent au point de coordonnées $(x_{[0]}, x_{[1]})$. La distance entre les points $\mathbf{x} = [x_{[0]}, x_{[1]}]$ et $\mathbf{y} = [y_{[0]}, y_{[1]}]$ est donc la longueur du vecteur $\mathbf{x} - \mathbf{y}$, que l'on peut facilement calculer par le théorème de Pythagore.

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_{[0]} - y_{[0]})^2 + (x_{[1]} - y_{[1]})^2}$$

```
import numpy as np

def distance(x,y):
    return np.sqrt((x[0]-y[0])**2+(x[1]-y[1])**2 )

x = np.array([6,4])
y = np.array([2,3])
print(distance(x,y))    # retourne sqrt(17)
```

Que l'on peut généraliser (les deux vecteurs doivent avoir la même taille) :

```
import numpy as np

def distance(x,y):
    return np.sqrt(((x-y)**2).sum())

x = np.array([5,4])
y = np.array([2,3])
print(distance(x,y))    # écrit sqrt(10)
```

A.4 Aléa et Statistiques

Références

- [1] kNN classifier from the `sklearn` package. [sklearn.neighbors.KNeighborsClassifier](#).
- [2] Daniel Smilkov and Shan Carter. A neural network playground, 2018. (online : <https://playground.tensorflow.org>).