Synthèse - Machine Learning

Léa Calem - Fatima Layla - Laureline Martin

18 octobre 2019

1 Description du jeu de données

1.1 Quel est le type de problème ?

Dans ce projet, nous devons classifier des images. les images reprensentent soit des t-shirt soit des robes.
Le problème est de type supervisé car nos données sont déjà annotées :

$$S = (x_i, y_i)$$

Tel que : x_i = ensemble des images, y_i = ensemble des étiquettes des classes C_1 et C_2.
Classification : $y_i = C_1, C_2$. Avec les classes : $C_1 = \{0$ T-shirt/top$\}$ et $C_2 = \{3$ Dress$\}$.

À l'aide des méthodes d'apprentissage, on recherche la fonction $h(x)$ qui à toutes images x associées une étiquette $\hat{(}y)$. Le but est d'obtenir des valeurs $\hat{(}y)_i$ proches des y_i, pour tout (x_i, y_i) appartenant à S.
(cours 02_methodo_etu.pdf/ slide 9)

1.2 quelles sont les données ?

1.2.1 Modélisation des données sous forme de matrice

Nombre d'observations : 14 000 :
— 7 000 de la classe C_1
— 7 000 de la classe C_2

1.2.2 Séparation des jeux de données

1. Données d'entraînement : sous-ensemble de données destiné à l'apprentissage du modèle. On utilise 75% des données, soit 10 500 .
2. Données de test : sous-ensemble de données destiné à l'évaluation du modèle (ce jeu de données ne doit en aucun cas être utilisé lors de la conception du modèle). On utilise 25% des données, soit 3 500.

1.3 Description des données

Les données sont des images de taille 28x28 pixels (784 pixels) composées de niveau de gris (valeur allant de 0 à 255). Sur ces images, seul l'objet est coloré donc le reste de l'image est en blanc, la valeur des pixels à 0.

1.4 Description statistique

Graphe du nombre de pixels différents de 0 en fonction de la classe. On divise l'image en 3, on cherche la moyenne du blanc de la 1ère et 3éme partie.

1.5 Paramètres

Étant donné que les images x_i sont bruitées, on suppose que les pixels ayant une valeur < 25 sont blanc.
Comme paramètre : écart entre blanc et gris tous les 28 pixels.

2 Méthodologie

2.1 Méthodes d'apprentissage utilisées

— Les K-NN (K plus proches voisins) : Fatima
— La régression : Lauréline
— SVM : Léa

2.2 Protocole de comparaison

Pour comparer les résultats des différentes méthodes d'apprentissages utilisées, nous évaluons le taux d'erreurs sur des jeux identiques de données.