Travaux pratiques d'IA

Série 5: Naive Bayes & Logistic Regression

Données

1. Les données à utiliser pour ce TP se trouvent dans les fichiers data_train.csv et data_test.csv. Il s'agit de prédire l'achat d'un produit en fonction du sexe, de l'âge et du salaire d'un individu.
2. Les 3 premières colonnes spécifient les covariables tandis que la dernière colonne correspond aux labels.
3. Le fichier data_test.csv sert à évaluer les performances des modèles développés à partir des données du fichier data_train.csv.
4. Il est recommandé d'utiliser pandas et/ou NumPy pour manipuler les données. Notamment la méthode *get_dummies* de pandas vous permet de convertir la première covariable en deux covariables binaires.

1 Naive Bayes

Le but de cette section est d'implémenter Naive Bayes. Voici les différentes étapes à accomplir:

1. Calculer la distribution empirique des labels.
2. Pour chaque valeur des labels, estimer les paramètres des distributions des covariables.
3. Implémenter la fonction de densité gausienne et la fonction de probabilité de Bernoulli.
4. Étant donné de nouvelles covariables, prédire les labels correspondants.

2 Logistic Regression

Le but de cette section est d'implémenter Logistic Regression. On suppose donc le modèle suivant:

$y_i \overset{ind}{\sim} Bernoulli(p_i), p_i = \sigma(w^T x_i + b), \sigma(z) = \frac{1}{1+\exp^{-z}}$

1. Sur papier, dériver:
   (a) $p(y_i|x_i; w, b)$
   (b) $\log(p(y_i|x_i; w, b))$
   (c) $\frac{\mathrm{d}\sigma(z)}{\mathrm{d}z}$ comme une fonction de σ
   (d) $\frac{\mathrm{d}\log(p(y_i|x_i;w,b))}{\mathrm{d}w_j}$
   (e) $\frac{\mathrm{d}\log(p(y_i|x_i;w,b))}{\mathrm{d}b}$
2. Implémenter une fonction *train_logistic_regression* qui prend en arguments:

(a) Une matrice de covariables X

(b) Un vecteur de labels y

(c) Un vecteur de poids initial w

(d) Une valeur de biais initiale

(e) Un nombre d'itérations num_iters

(f) Un taux d'apprentissage learning_rate

et qui renvoit les poids et le biais entraînés par descente de gradient à minimiser

$$-\sum_{i=1}^{N} \log(p(y_i|x_i; w, b))$$

avec N le nombre d'exemples d'entraînement.

3 Evaluation

Comparer les performances des modèles développés en les évaluant sur les données de data_train.csv et data_test.csv selon les métriques suivantes: accuracy, precision, recall, F1 score.