Fiche d'auto-évaluation 03

November 17, 2020

Caractéristique = Feature

- Que veut dire qu'un modèle a une trop grande variance ? Et qu'est-ce que c'est la variance d'une caractéristique (écrivez la formule) ? La variance d'un modèle et la variance d'une caractéristique, sont elle la même chose ? Où sont elle des choses complètement différentes ?
- Quelles techniques de sélection de caractéristiques (feature selection) nous avons vu jusqu'à présent? Dis-en au moins 3.
- Écrivez les formules de variance (d'une colonne), écart type, coefficient de corrélation de Pearson
- Expliquez comment on construit un boxplot et que veulent dirent toutes les lignes qui le constitue.
- Si dans votre dataset il y des valeurs manquantes, qu'est-ce qu'il faut faire avant d'entraîner un modèle? (voir 01.introduction/b.preprocessing.ipynb)
- Si une des caractéristiques assume 4 valeurs catégoriques, vous pouvez les encoder de deux façons: (i) Vous les transformez en 0,1,2,3 ou (ii) vous appliquez le One-Hot Encoding. Dans quel est possible choisir (i)? Et dans quel cas (ii)?
- Expliquez comment on calcule le 35ème centile (percentile).
- Pourquoi, si on a N caractéristiques, nous représentons chaque exemple i avec un vecteur $\mathbf{x}^{(i)}$ de $N+1$ éléments (et pas seulement de N éléments?
- Dans les types de modèles vue jusqu'à présent, quels sont les méthodes et les hyper-paramètres qui permettent d'augmenter ou diminuer leur complexité?
- Comment peut-on se rendre compte si un modèle est en surapprentissage (overfitting), ou en sous-apprentissage (underfitting)? Comment peut-on le fixer dans les deux cas?
- A que sert la régularisation?

- Montrez une procédure correcte pour choisir le bon coefficient de régularisation.
- Montrez une procédure correcte pour choisir le bon pair coefficient de régularisation et dégrée de la régression polynomiale
- Est la mise à l'échelle (scaling) importante/utile quand on fait de la régression linéaire? Pourquoi? Et quand on fait de la régression de Ridge (Ridge Regression, c-à-d régression + régularisation)? Pourquoi?
- Par rapport à la descente de gradient dans la régression logistique, le gradient dont on parle est le gradient de quelle fonction? Définissez cette fonction exactement.
- La descente de gradient utilisée dans la régression logistique est utilisé pendant l'entraînement ou la prédiction?
- Par rapport la régression logistique, quel est le rapport entre la probabilité prédite par le modèle et la classe prédite?
- Est la régression logistique un classificateur linéaire ou non? Que veut cela dire?
- Que fait la fonction de softmax? Écrivez aussi sa formule.
- Écrivez la formule de la cross-entropie.
- Si vous avez un ensemble de données avec déséquilibre des classes et vous entraînez directement un classificateur, sans tenir compte du déséquilibre et vous obtenez une bonne justesse (accuracy), feriez-vous confiance à ce classificateur? Pourquoi?