Apprentissage statistique

Synthèse

Vincent Lefieux

Machine learning

- On souhaite estimer une fonction de lien, en régression ou en classification supervisée, sans hypothèse de loi.
- Lorsqu'on estime, on fait face à deux erreurs distinctes :
  - L'erreur d'approximation : on recherche une solution dans un espace restreint.
  - L'erreur d'estimation : dans l'espace considéré, il existe une différence entre l'estimateur et la solution optimale dans cet espace.
- Plus l'espace dans lequel on recherche l'estimateur est complexe (grand), plus le biais (l'erreur d'approximation) est faible et la variance (l'erreur d'estimation) élevée. Et vice-versa.
- On souhaite minimiser le risque (erreur de généralisation), basé sur une fonction de perte (fonction de coût). On l'estime par le risque empirique.
- Le risque empirique sous-estime le risque. Pour pallier cela, et éviter du sur-apprentissage, on utilise la validation croisée.

CART

- Le principe des arbres de décision et de régression est de scinder l'espace en deux de manière récursive.
- On trouve une racine, des branches, des nœuds et des feuilles.
- A chaque étape on recherche la covariable et le seuil associé optimaux qui permettent de minimiser l'hétérogénéité des deux sous-échantillons obtenus (ex : erreur de classification dans le cas de la classification supervisée, variance dans le cas de la régression).
- La règle de prévision considérée dans une feuille est le vote majoritaire dans le cas de la classification supervisée et la moyenne dans le cas de la régression.
- On peut élaguer un arbre de manière à éviter du sur-apprentissage.
- Les arbres souffrent d'une variance élevée (une petite variation dans l'échantillon peut conduire à des résultats très différents).

Bagging

- Le bagging fait partie des méthodes d'agrégation.
- Le bagging agrège des modèles présentant un biais faible et une variance forte estimés sur des échantillons bootstrappés (de même taille que l'échantillon initial, issus d'un tirage avec remise).
- Les forêts aléatoires (random forests) sont basées sur des « grands » arbres (donc avec une forte variabilité) et diminuent la dépendance entre les modèles en présentant à chaque arbre un nombre fixé, plus petit, de covariables tirées au sort.
- Il n'y a pas de risque de sur-apprentissage mais il faut néanmoins limiter le nombre d'arbres par souci de parcimonie numérique.
- Ces méthodes sont réputées efficaces numériquement.

Boosting

- Le boosting fait partie des méthodes d'agrégation.
- AdaBoost agrège de manière récursive des règles faibles en sur-pondérant à chaque étape les points mal prévus.
- AdaBoost est équivalent à une méthode de gradient boosting qui consiste à minimiser un risque empirique « convexifié » de manière récursive.
- Il faut limiter le nombre d'itérations car il existe un risque de sur-apprentissage.
- On doit choisir le pas de descente de l'algorithme d'optimisation.
- On trouve parmi les modèles classique AdaBoost et LogitBoost pour la classification supervisée et L^2-Boosting pour la régression.
- Ces méthodes sont réputées efficaces numériquement.

SVM

- Les séparateurs à vaste marge, SVM, ont été introduits dans le cadre de la classification supervisée binaire.
- Dans le cas linéairement séparable, on cherche l'hyperplan optimal qui scinde l'échantillon en deux en maximisant la marge.
- Dans le cas non linéairement séparable, on relaxe la contrainte en ajoutant des variables ressorts (*slack variables*). Il faut alors choisir, par validation croisée, l'hyperparamètre qui contrôle le compromis entre le nombre d'erreurs de classification et le niveau de la marge.
- Pour faciliter la séparation, on peut utiliser l'« astuce du noyau » qui envoie les observations dans un espace de représentation (*feature space*). On choisit numériquement le noyau le plus adapté.
- Dans le cas de la régression, on peut utiliser les SVR.
- Ces méthodes sont réputées efficaces numériquement.

Réseaux de neurones denses (DNN)

- Un neurone formel lie des entrées avec des poids associés et à une sortie via plusieurs opérations : somme pondérée des entrées, ajout d'un biais puis application d'une fonction d'activation.
- Il faut ajouter des couches aux réseaux de neurones afin de traiter des problèmes non-linéairement séparables.
- Chaque neurone d'une couche d'un perceptron multicouche est relié aux neurones de la couche adjacente.
- On parle de réseau de neurones profond ou dense (DNN) dès qu'on a plus de 2 couches.
- L'estimation d'un réseau de neurones est classiquement réalisée à l'aide de l'algorithme de rétro-propagation des erreurs.
- Ces méthodes sont réputées efficaces numériquement mais requièrent de gros volumes de données et des capacités de calcul importantes.