

Apprentissage sur un petit ensemble de données

Overfitting

- ▶ comment mesurer l'overfitting (mesures oop, performances à 1, ect)
 - ▶ Evaluer l'erreur du modèle pour les données d'entraînement et les données de validation et de test. Si l'erreur est faible pour les données d'entraînement par rapport au données de test, dans ce cas on peut dire qu'il y a overfitting.
- ▶ comment éviter l'overfitting (pourquoi la k-cross validation est pas suffisant, pourquoi un corpus de validation peut permettre de pallier à cela, ect.)

Variantes de corss-validation

- ▶ k-cross-validation
- ▶ k-fold cross-validation (un seul ensemble de données de test)
- ▶ k-fold cross-validation avec ensemble de test et de validation (plusieurs sous-ensembles comme données de test).

Application - Prédiction de l'activité cérébrale en fonction des signaux multimodaux

- ▶ k-fold cross-validation avec ensemble de test et de validation.
- ▶ Problème : la cross validation pose quelques problème pour les données séquentielles car elle tient pas en compte l'ordre chronologiques des données, mais on peut l'appliquer dans notre cas si on considère chaque sous-ensemble est une conversation sous l'hypothèse que l'ordre des conversations n'est pas important.
- ▶ Dans ce cas, on peut découper les données en 4 ensembles, chaque ensemble contient 6 conversations. Sur chaque ensemble on peut appliquer une k-fold cross-validation avec une seule conversations comme données de test, et pour les autres, on change aléatoirement l'ordre des conversations à chaque fois on considère une conversation comme données de validation, et on répète ce processus, jusqu'à ce que chaque conversation des données d'entraînement est utilisée un fois comme données de validation.

Techniques pour la génération de nouvelles données

- ▶ quelles sont les différentes techniques pour la génération de nouvelles données ?
- ▶ leurs limites ? peut-on les appliquées sur des données comportementales comme les nôtres ?