Validation croisée / cross-validation

Toby Dylan Hocking
toby.dylan.hocking@usherbrooke.ca

September 17, 2024

La validation croisée

Les données sont divisées en :

- train = entraînement, utilisé pour apprendre la fonction de prévision f.
- test, utilisé pour évaluer la qualité de la fonction de prévision f.

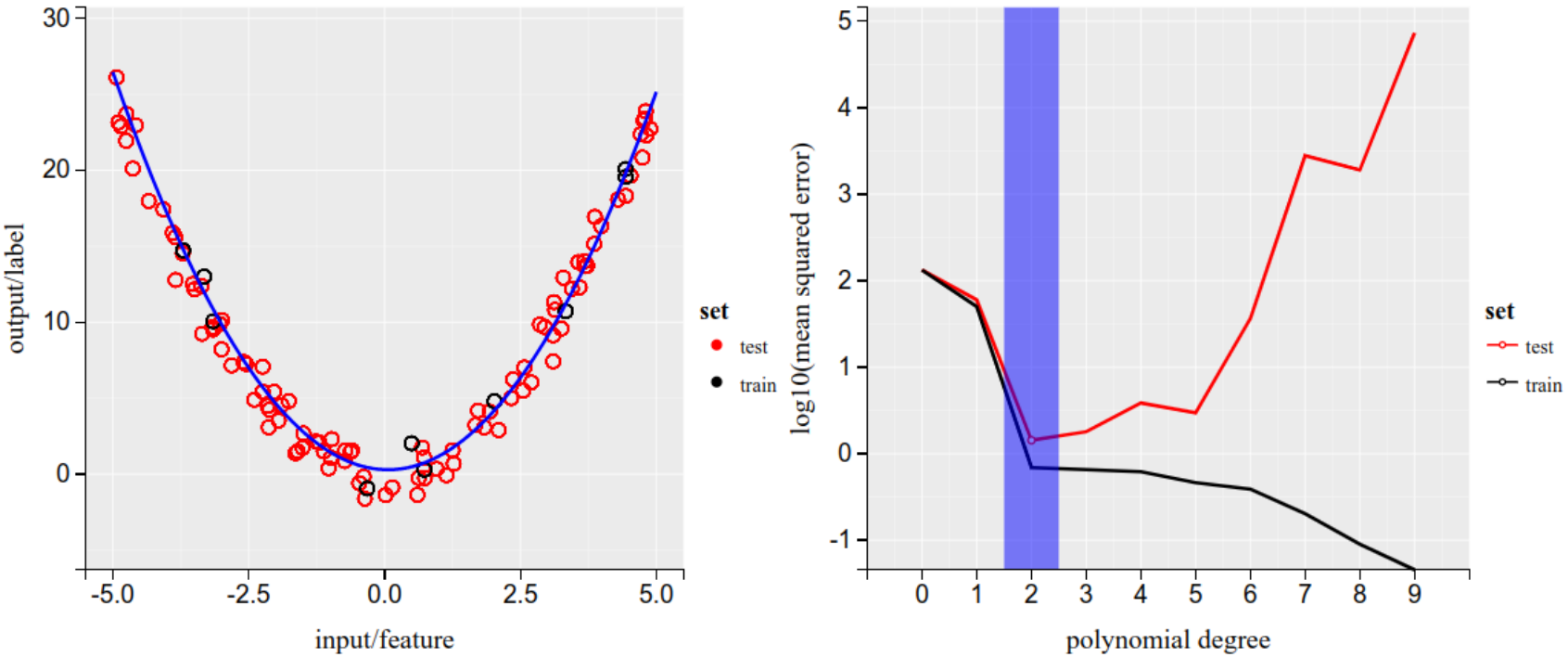
Sur-apprentissage et Sous-apprentissage

Jeux de données standards

Simulations : quand est-ce que l'apprentissage est possible ?

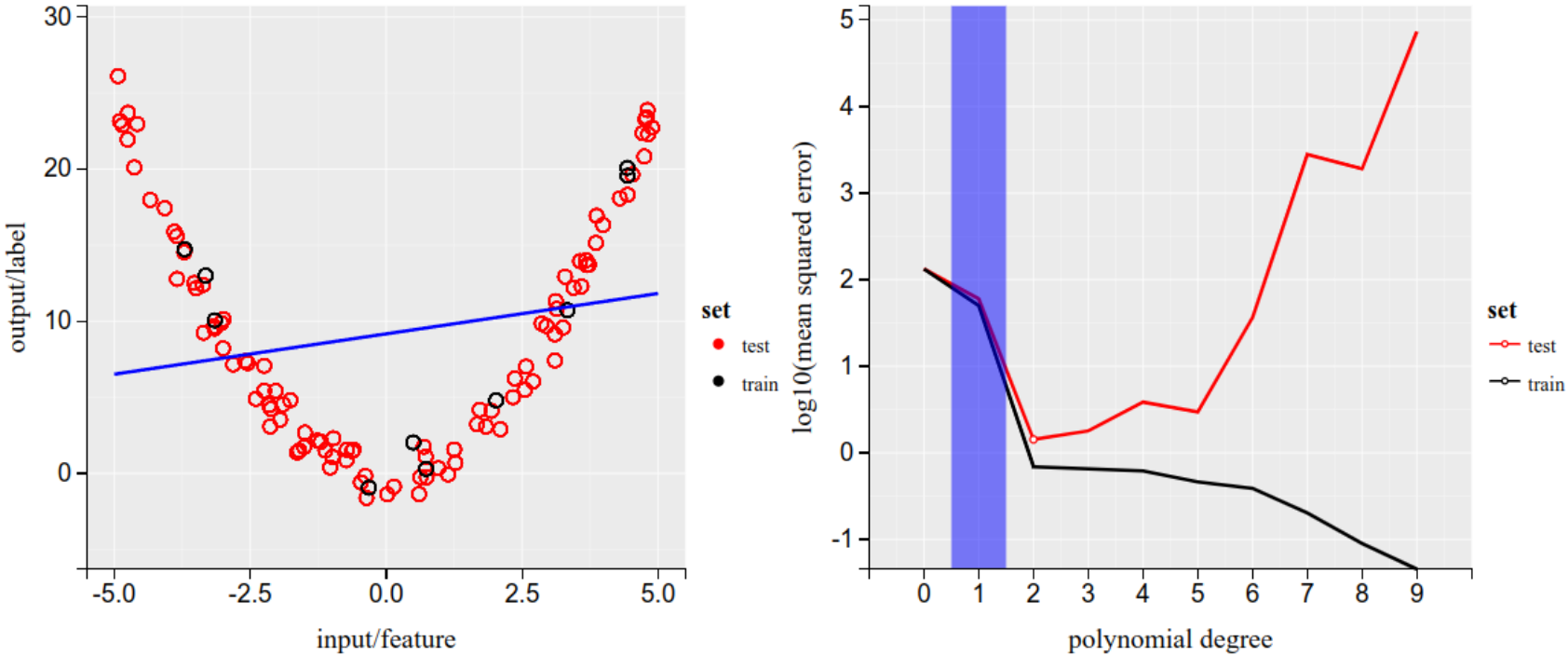
Bon apprentissage

- Fonction de bonne complexité (polynome degré 2)
- Bon régularité, bon variabilité.
- `https://tdhock.github.io/2020-02-03-capacity-polynomial-degree/`



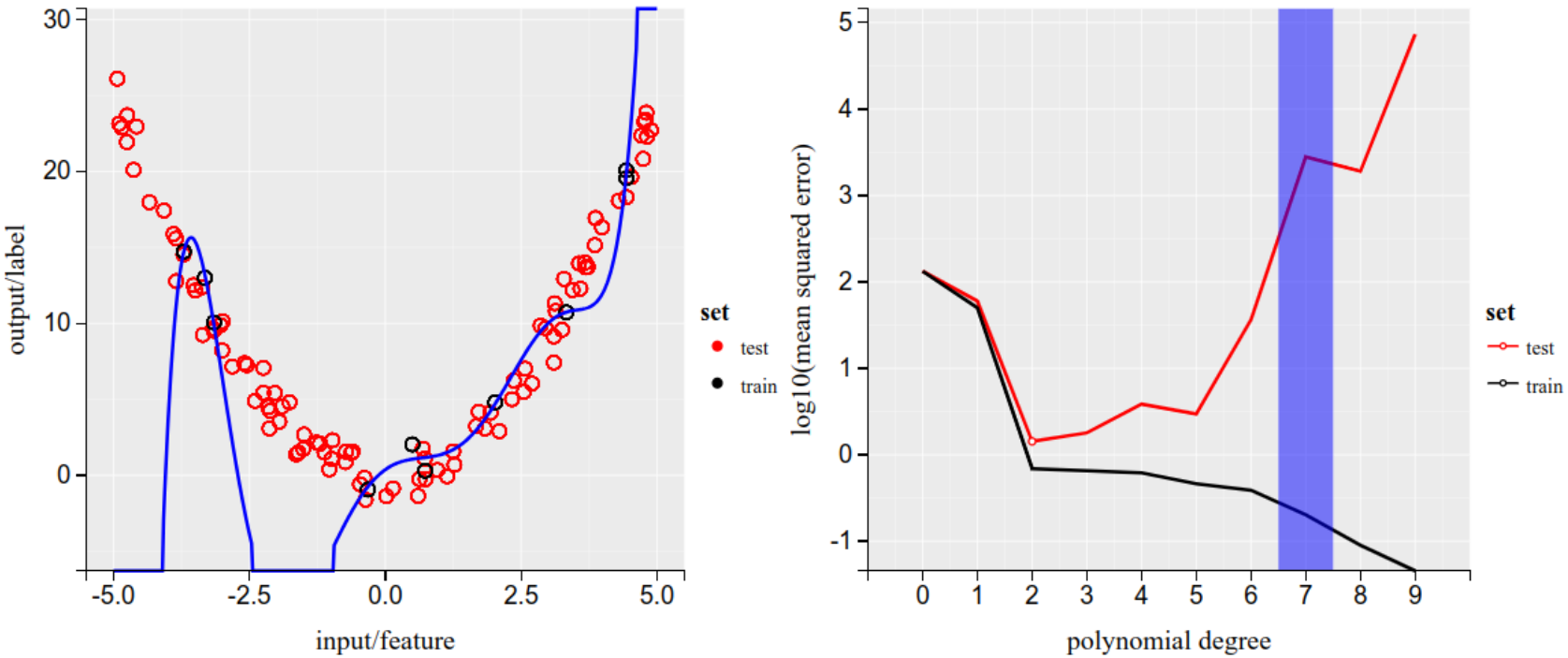
Sous-apprentissage

- Fonction trop simple (polynome degré 1)
- Trop régulier, pas assez variable.
- `https://tdhock.github.io/2020-02-03-capacity-polynomial-degree/`



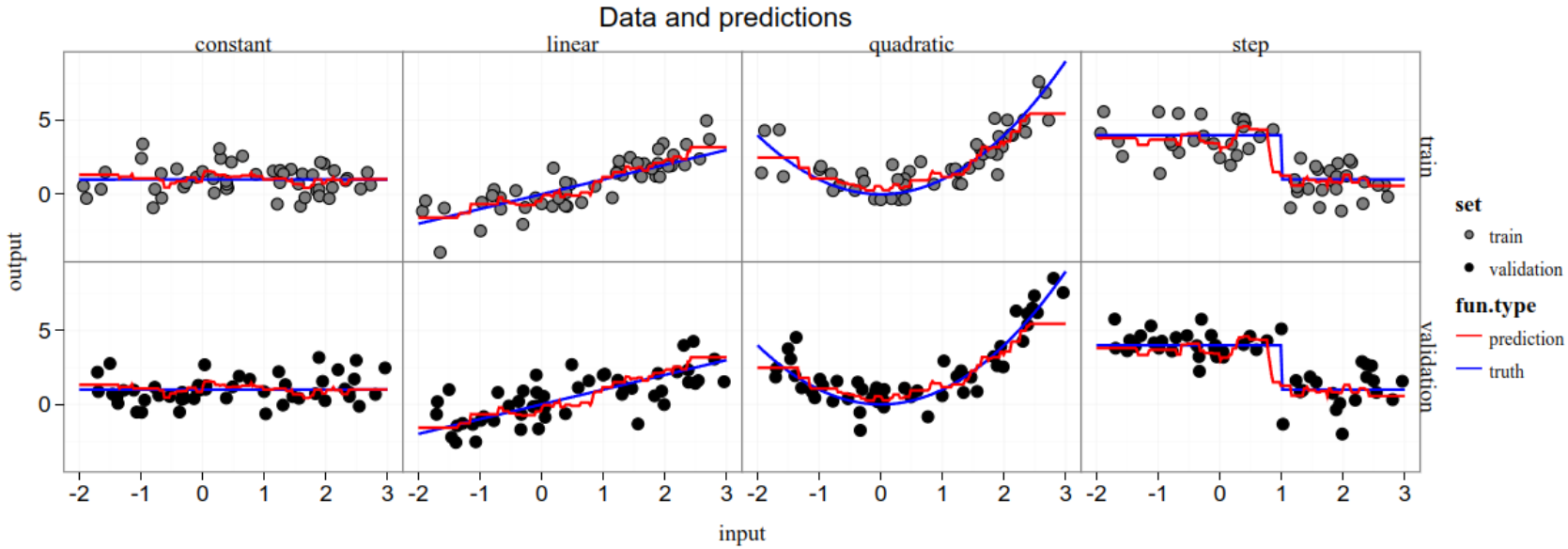
Sur-apprentissage

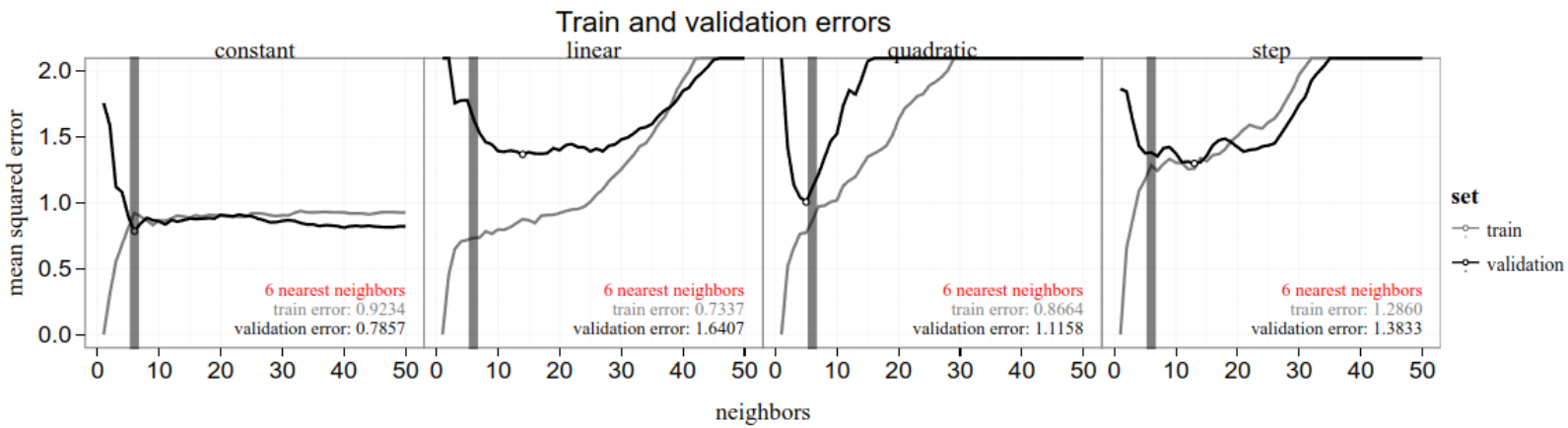
- Fonction trop complexe (polynome degré 7)
- Pas assez régulier, trop variable.
- `https://tdhock.github.io/2020-02-03-capacity-polynomial-degree/`



Sur- et Sous-apprentissage

- https://tdhock.github.io/2019-01-nearest-neighbor-regression-one-split/





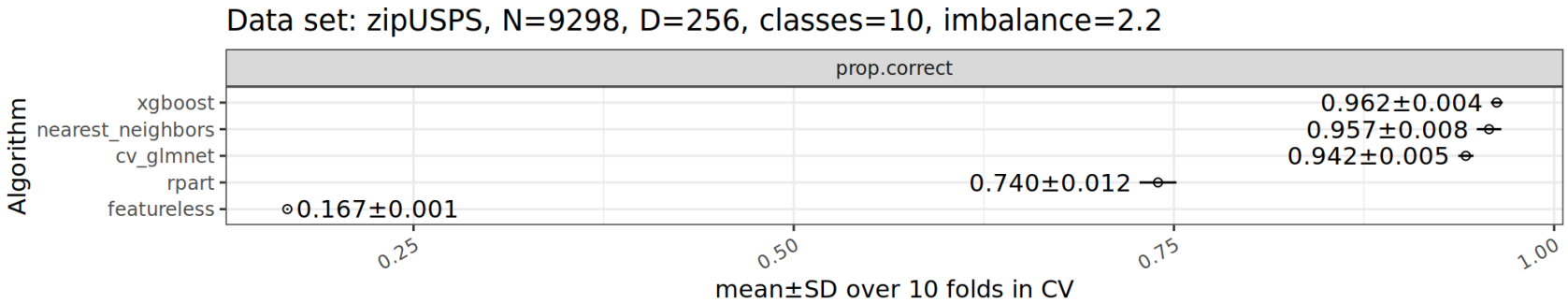
Sur-apprentissage et Sous-apprentissage

Jeux de données standards

Simulations : quand est-ce que l'apprentissage est possible ?

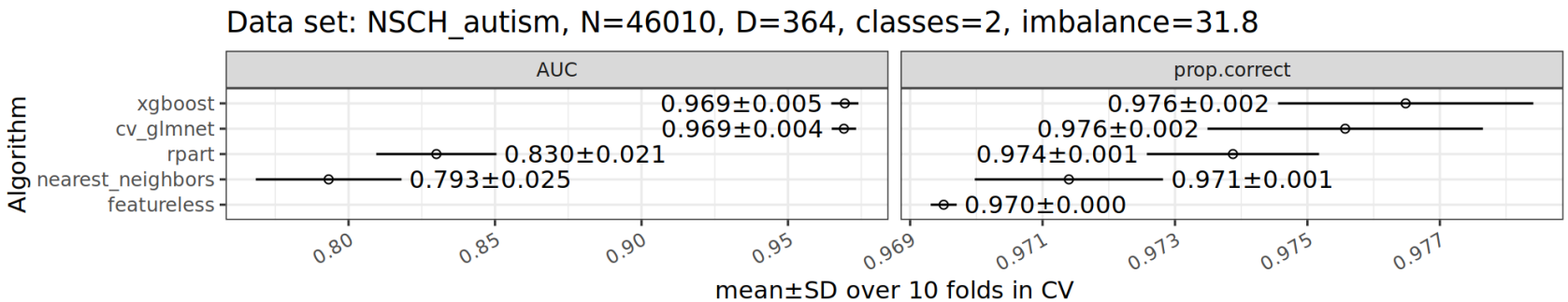
Classification d'images de chiffres (zipUSPS)

- prop.correct = proportion correcte
- xgboost meilleur, mais voisins (nearest_neighbors) est très proche... est-ce que la différence est significative ?

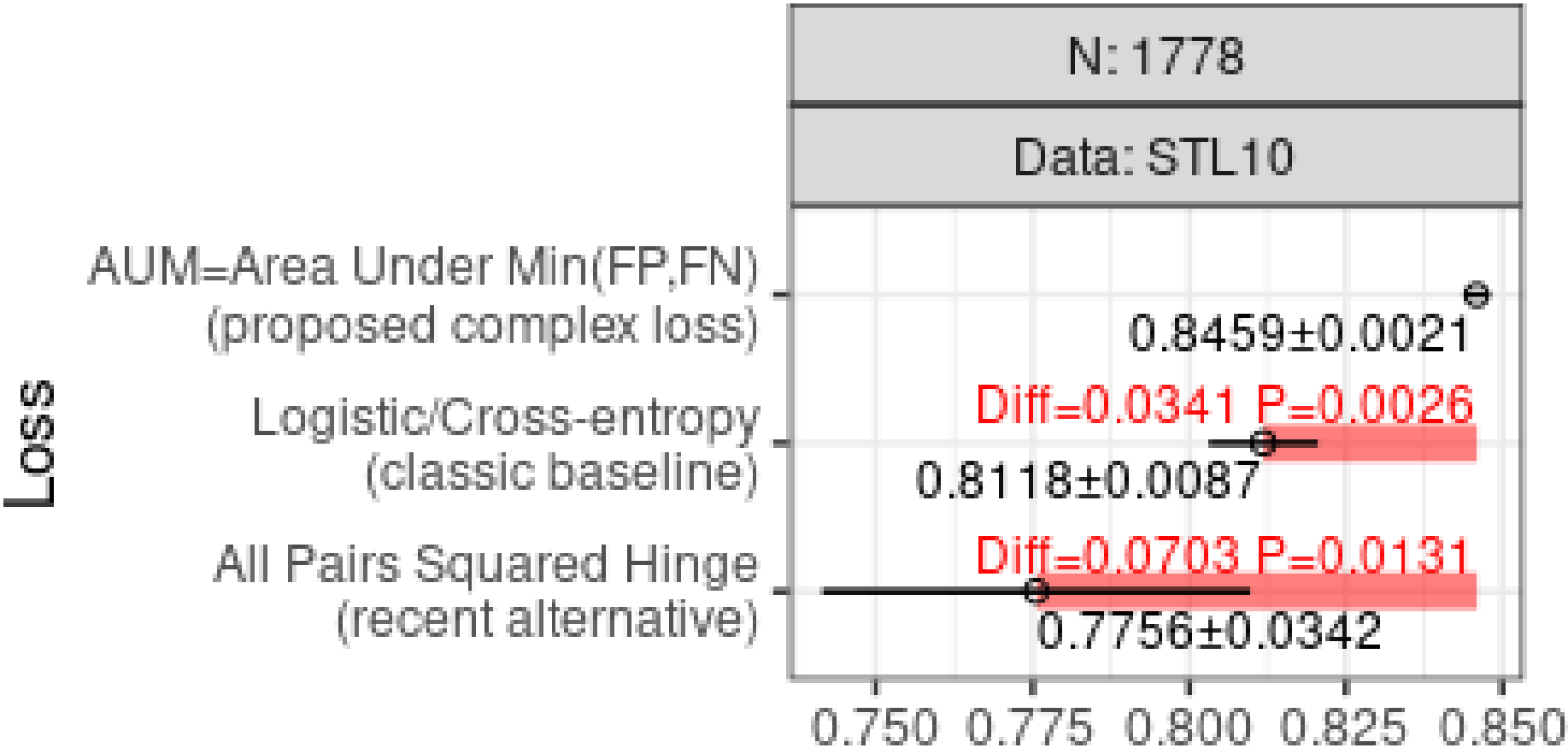


Classification d'autisme

- prop.correct = proportion correcte
- AUC = Area Under ROC Curve / Aire sous la courbe ROC (Taux de Vrai Positive vs. Taux de Faux Positive)
- xgboost meilleur, mais modèle linéaire (cv_glmnet) est très proche... est-ce que la différence est significative ?

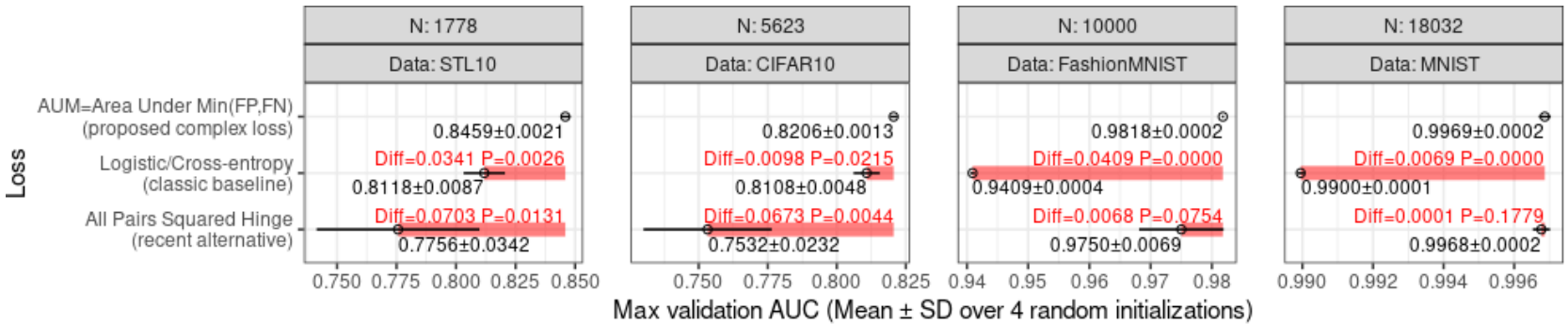


Visualisation du test de Student



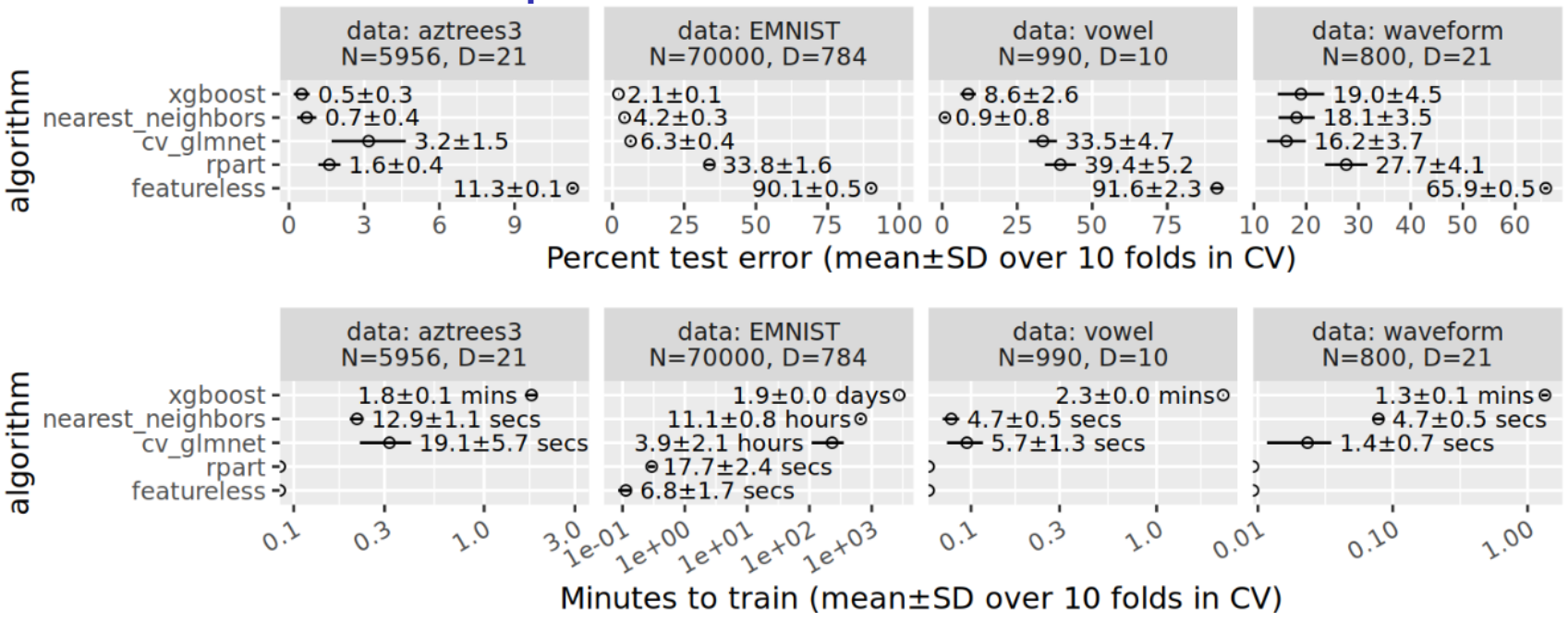
- X = Aire sous la courbe ROC.
- $p < 0.05 \Rightarrow$ différence significative.
- https://tdhock.github.io/blog/2024/viz-pred-err/

Visualisation du test de Student



- $p < 0.05 \Rightarrow$ différence significative.
- $p > 0.05 \Rightarrow$ différence non-significative.
- `https://tdhock.github.io/blog/2024/viz-pred-err/`

Taux d'erreur et temps de calcul



- featureless est le plus rapide, et toujours le plus erroné.
- xgboost est le plus lent, le moins erroné dans aztrees3, EMNIST.
- Les plus proches voisins (nearest_neighbors) est le meilleur dans vowel.
- Modèle linéaire (cv_glmnet) est le meilleur dans waveform.
- Pour chaque jeu de données, on ne sait pas quel algo est préferable, jusqu'au moment de voir le résultat de la V-C.

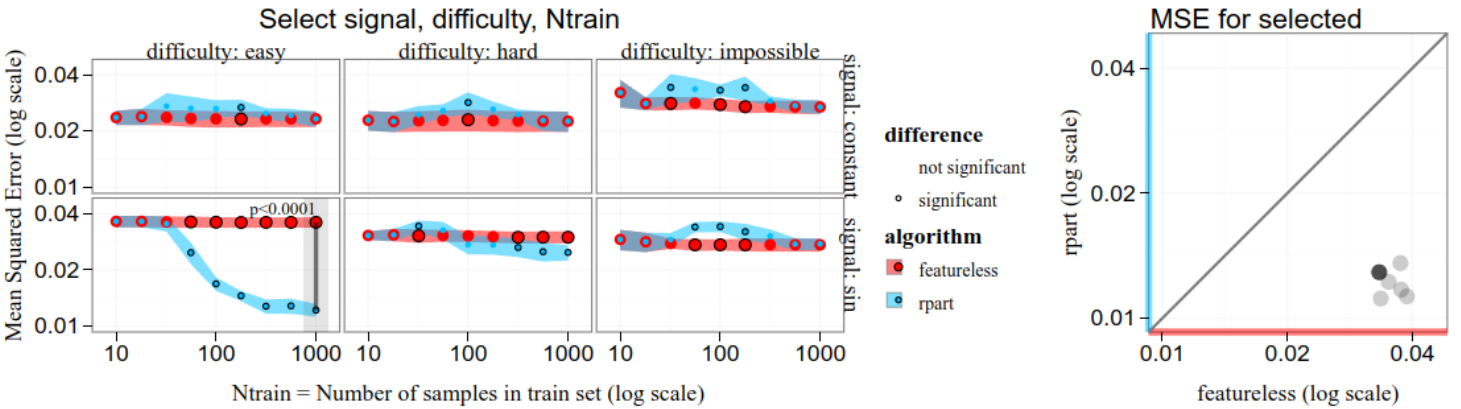
Sur-apprentissage et Sous-apprentissage

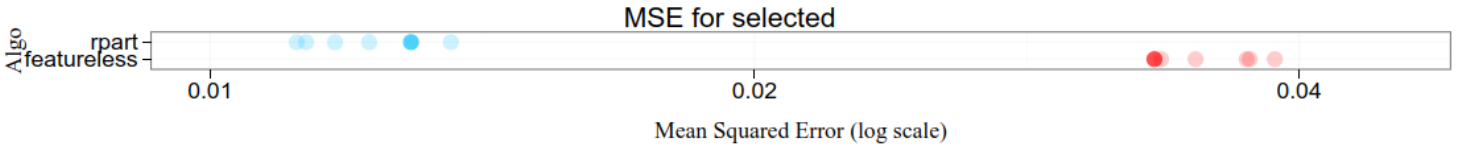
Jeux de données standards

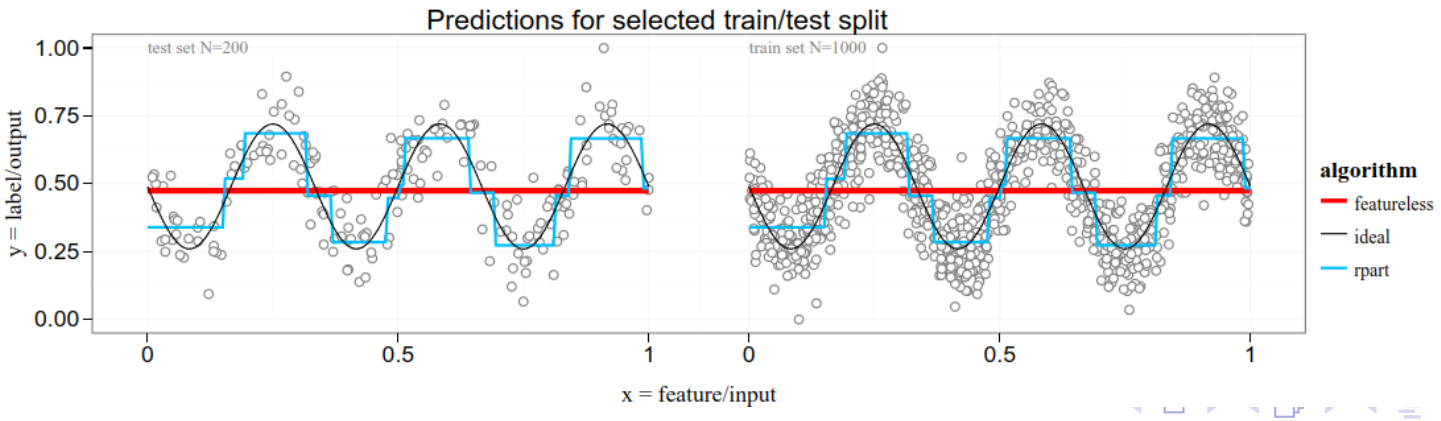
Simulations : quand est-ce que l'apprentissage est possible ?

Apprentissage facile / Easy learning

- N=1000, bruit : facile, signal : sin.
- Erreur de rpart plus petit que featureless, $p < 0.0001$
- `https://tdhock.github.io/2024-09-16-K-fold-CV-train-sizes-regression/`

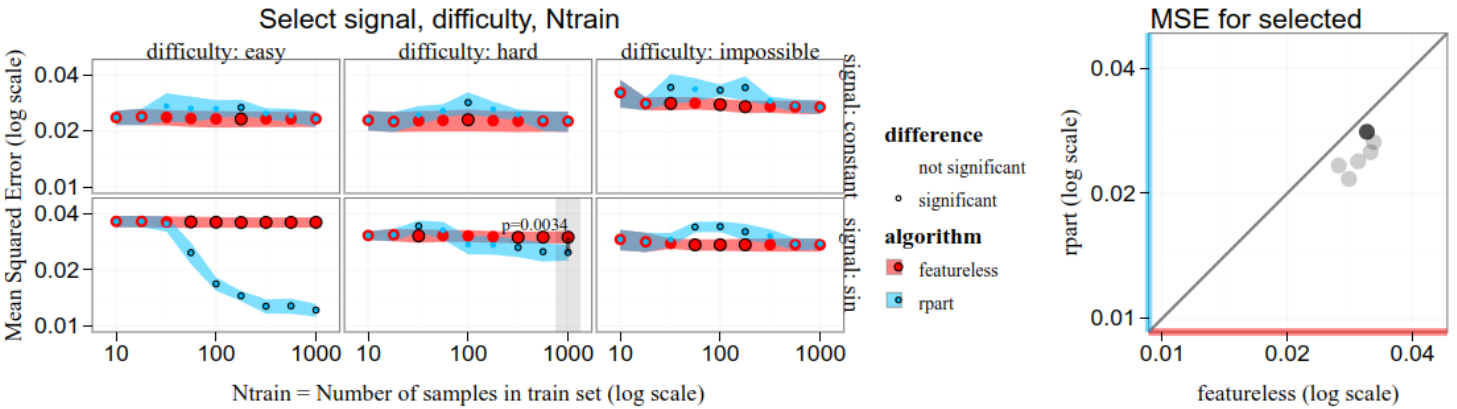


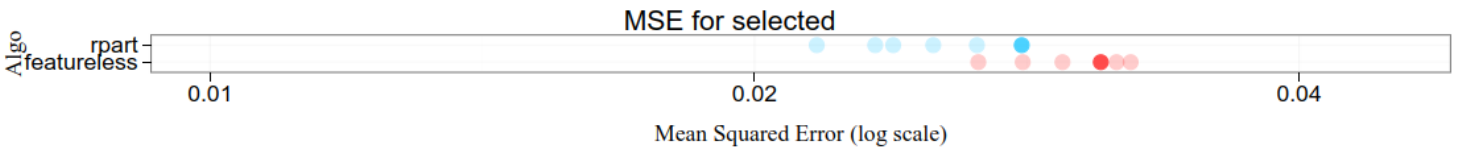


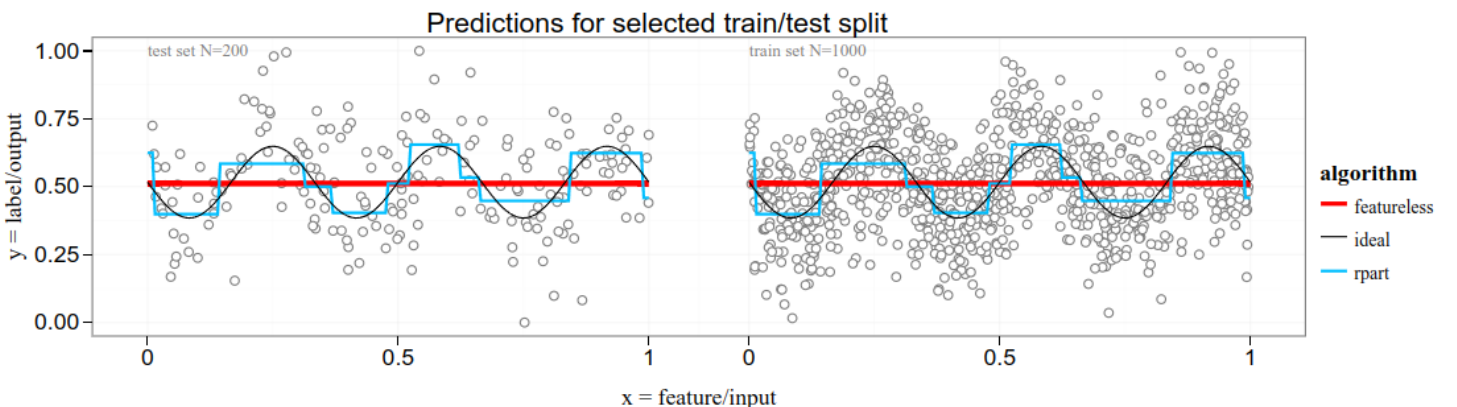


Apprentissage possible / Possible learning

- N=1000, bruit : dificile, signal : sin.
- Erreur de rpart plus petit que featureless, $p = 0.0034$
- `https://tdhock.github.io/2024-09-16-K-fold-CV-train-sizes-regression/`

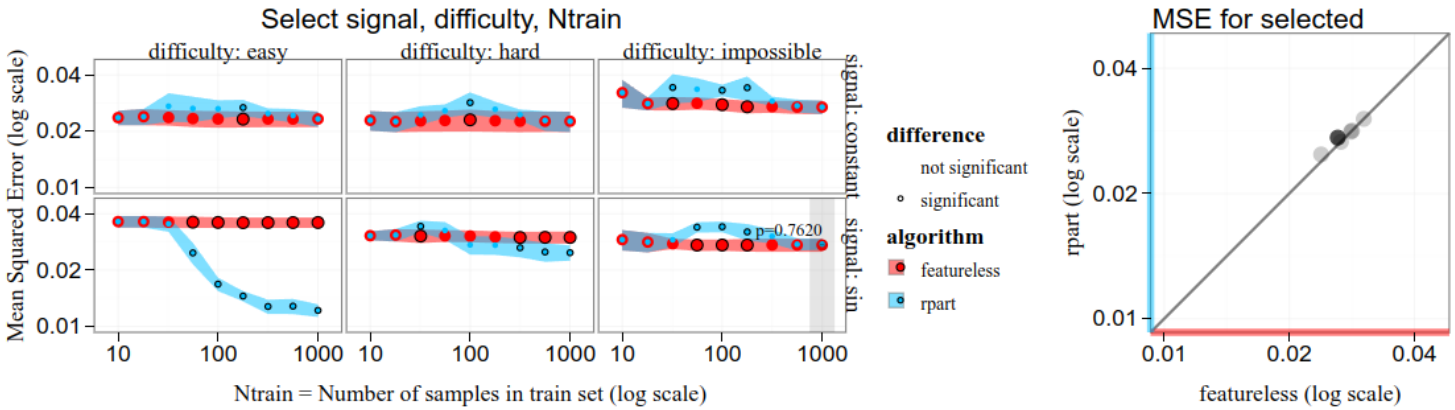


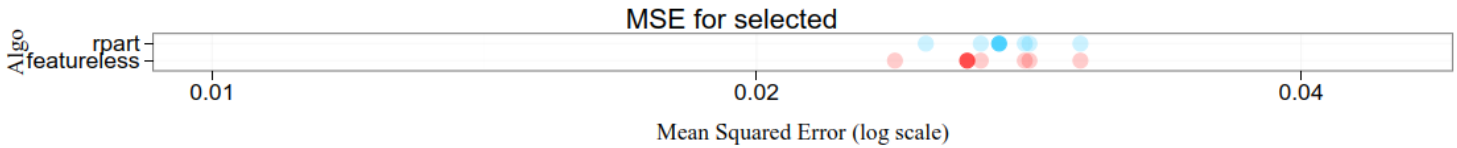


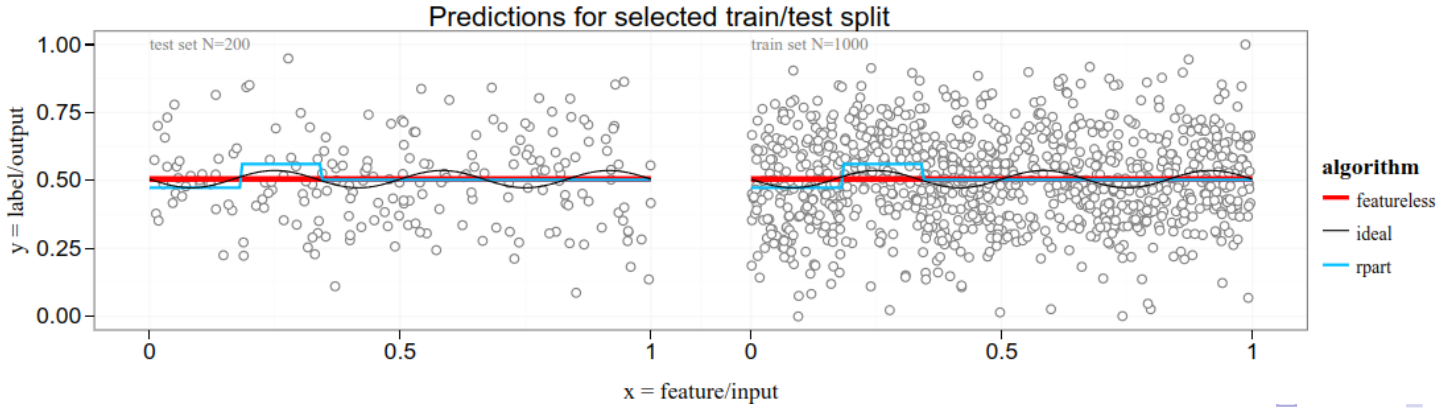


Apprentissage impossible 1 / Impossible learning 1

- N=1000, bruit : impossible, signal : sin.
- Erreur de rpart équivalent à featureless, $p = 0.7620$.
- Mode d'échec 1 : trop de bruit / signal pas assez fort.

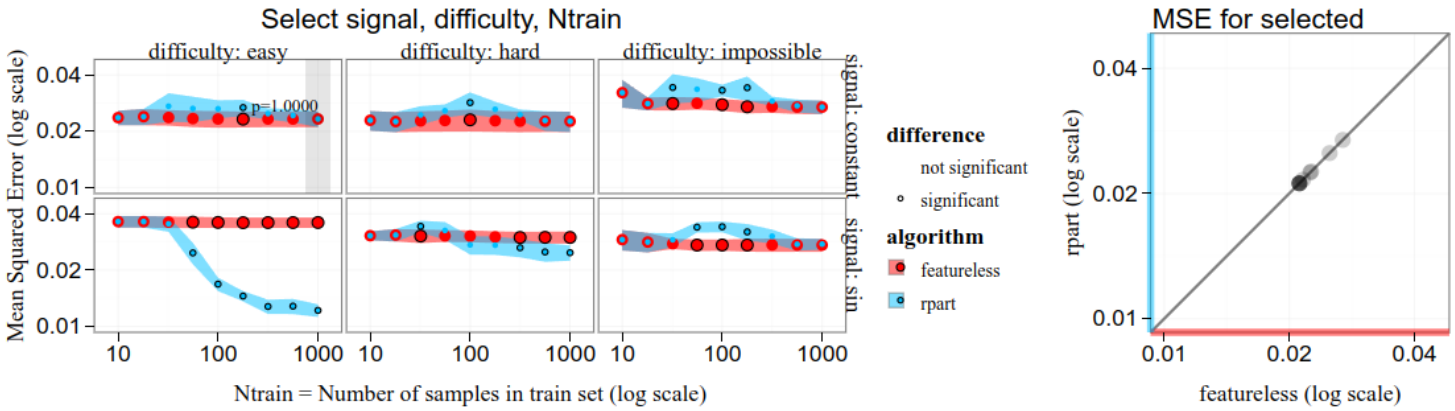


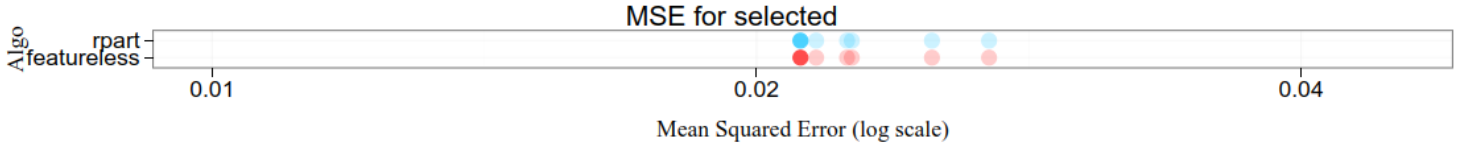




Apprentissage impossible 2 / Impossible learning 2

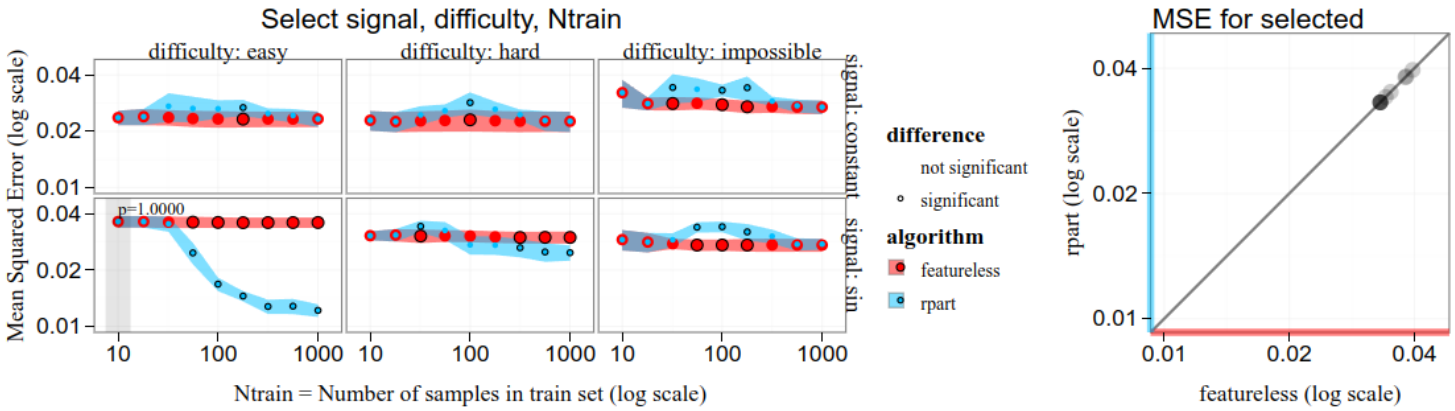
- N=1000, bruit : facile, signal : constant.
- Erreur de rpart équivalent à featureless, $p = 1$.
- Mode d'échec 2 : pas de rélation entre entrée x et sortie y.

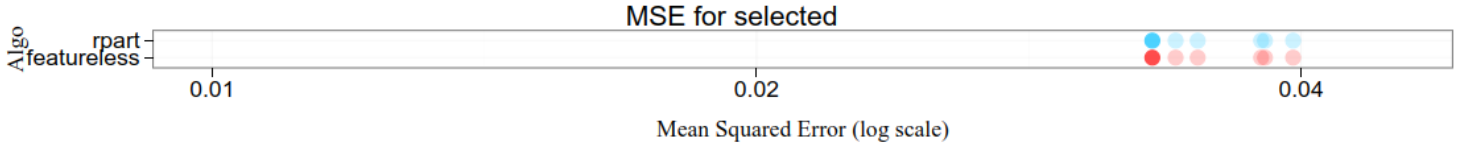


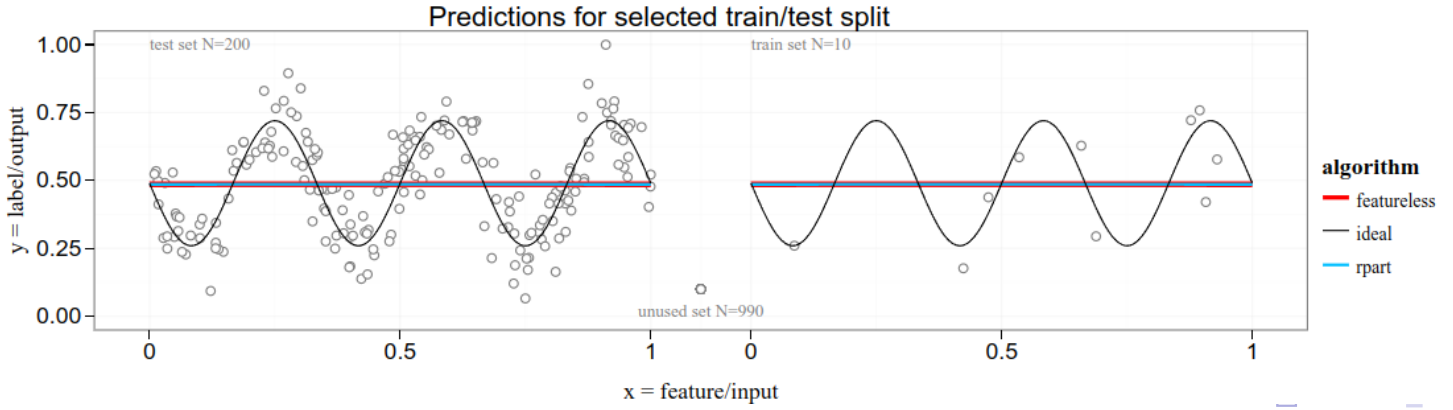


Apprentissage impossible 3 / Impossible learning 3

- N=10, bruit : facile, signal : sin.
- Erreur de rpart équivalent à featureless, $p = 1$.
- Mode d'échec 3 : pas assez de données.







Sur-apprentissage / Overfitting

- N=178, bruit : facile, signal : constant.
- Erreur de rpart plus grand featureless, $p = 0.0325$.
- Mode d'échec 3 : pas assez de données.