Tree , Bagging , Random Forests and Boosting

Only used for predictions

not CI / Testing / significant variables.

- Classification trees ← Leo Breiman
- Boosted Random Forests ← run a lot of trees and find one that is not in the forest but better than all existing trees in the forests

Non-parametric

Arm Bended Problem

Two-class Classifications

If tree is small ⇒ good prediction properties.

⇒ how to determine when to stop } ⇒ in R package

huge ⇒ overfit

SPAM ⇒ R builtin

Sensitivity ⇒ proportion of true spam identified ↑ Type I ↑

Specificity ⇒ proportion of true email identified ↑ Power

want both to be high

overfitting ⇒ wont get high specificity & sensitivity

Freund → prof here

Decision Bondary : Tree

Model Averaging ⇒ Boosting > Random Forests > Bagging > Single Tree

Bagging

Bootstrap aggregation ⇒ bootstrap a thousand times

⇒ get a thousand trees

⇒ average the trees

smoother decision boundaries

Random Forests:

⇒ Randomly choose m features (refinement)

then ↳ bagging ⇒ average the trees

⇒ high dimensional

Boosting

⇒ Bootstrap, take features

⇒ which data was predicted well ⇒ not interesting ⇒ smaller weight

⇒ higher weight on misfit ⇒ weighted bootstrap ⇒ run trees.

⇒ which didn't predict well ⇒ reweight

↳ Loss function

run many forests

⇒ converged forests

pick the data you misclassified

⇒ won't be overfitting