Ensemble Learning

(Bagging, Boosting, and Stacking)

Bagging

- Combine homogeneous “weak learners”
- Homogeneous = each weak learner is of the same type (e.g., all decision trees)
- Weak learners are trained independently
  - Parallelization is easy
- Aggregate predictions from the weak learners

So why is it called “bootstrap aggregating”?

Bootstrap part:

- Split into a training set and a test set
- Create B bootstrap samples (samples of size n_{train}, sample with replacement), train a weak learner for each of the B bootstrap samples

Aggregate part:

- Pass an observation through each of the B models
- For classification: Take a vote!
- As an example, say we have B = 50, for binary classification

Model 1	Model 2	Model 3	. . .	Model 50
Predicted class: B	Predicted class: A	Predicted class: B		Predicted class: B

- 42 votes for class B, 8 votes for class A => bagged model predicts class B

What’s the difference between bagging decision trees, and a random forest?

Bagged decision trees vs. random forests

- Random forest = a bagging algorithm, but one large difference
- With a random forest, we also select a random subset of our *p* variables when training any of the individual decision trees

Boosting

- Models are trained sequentially
- Weak learners are therefore not independent
  - Model at the current step depends on models at previous steps

AdaBoost (for binary classification)

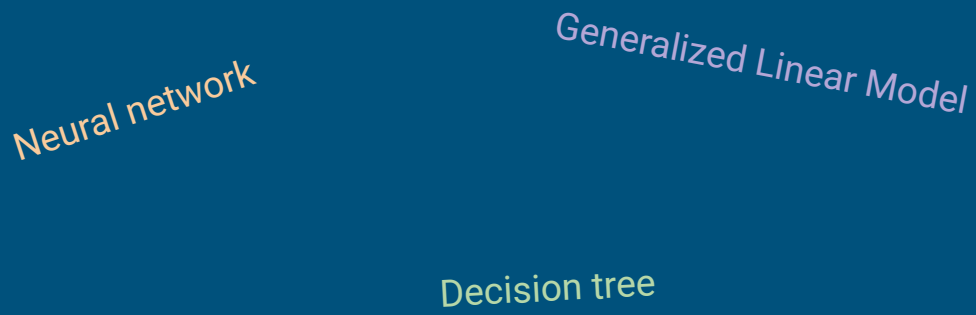
- “Adaptive Boosting”
- At each step, the weights of the observations that were previously misclassified are increased
- The “strong learner” (final model) is a weighted sum of the weak learners
    - The better the weaker learner, the higher its weight

Gradient Boosting

- Another boosting algorithm where the “strong learner” (final model) is a weighted sum of the weak learners
- Gradient descent is used to determine how to improve at each step in the sequence
- A generalization of boosting where optimization can be based off of any loss function (so long as it is differentiable)

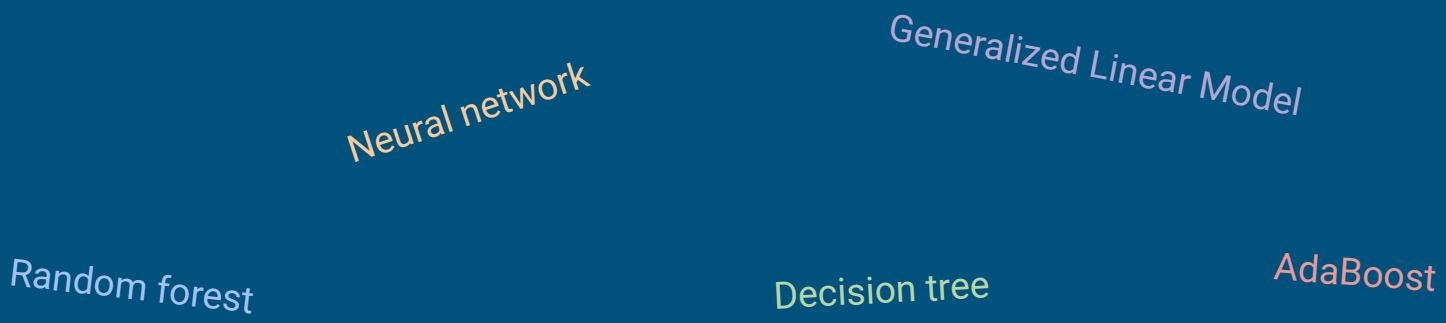
Stacking

- Ensemble method that combines heterogeneous weak learners
  - (They are combined using a “metalearning” algorithm)



Stacking

- Ensemble method that combines heterogeneous weak learners
  - (They are combined using a “metalearning” algorithm)



If ensemble methods are not very interpretable, why use them?

Why use ensemble methods?

- Often, interpretability is not our top priority
  - There may be situations where we want a model with high accuracy
  - In these cases, ensemble methods are highly desirable