GENERAL ASSEMBLY

INTRO TO ENSEMBLES, BAGGING, AND BOOSTING

Ritika Bhasker, Data Science Immersive

AGENDA

- What is an ensemble method?
- What is bagging?
- What is boosting?
- Use bagging with SKLearn

GENERAL ASSEMBLY

WHAT IS AN ENSEMBLE

ENSEMBLE METHODS

- Ensemble techniques are supervised learning methods to improve model performance by *combining* several base models in order to enlarge the space of possible hypotheses to represent our data.

- When is this useful?

- What is the hypothesis space?

THE HYPOTHESIS SPACE

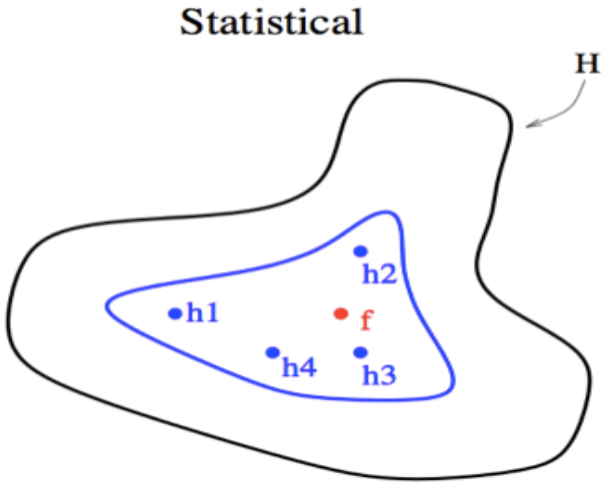
- The hypothesis space is all hypotheses that could explain parts of the observed data. In any supervised learning task, we're essentially looking within the hypothesis space for the most appropriate function to describe the relationship between our features and our target

WHY WOULD OUR BASE CLASSIFIER PERFORM BADLY

- There are a few reasons why our base classifier may not perform well when it's trying to approximate the ~true~ classification function/hypothesis.

THE STATISTICAL PROBLEM

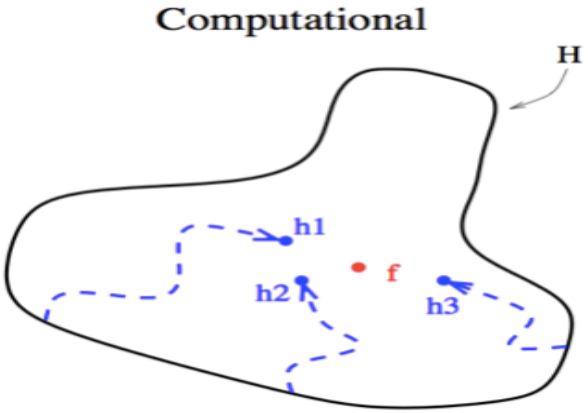
- If the training data you have is small compared to the size of the hypothesis space, your algorithm may find many different hypotheses in the hypothesis space that all give the same accuracy on the data.



An ensemble method will let you 'average out' base classifier predictions to find a good approximation of that true hypothesis

THE COMPUTATIONAL PROBLEM

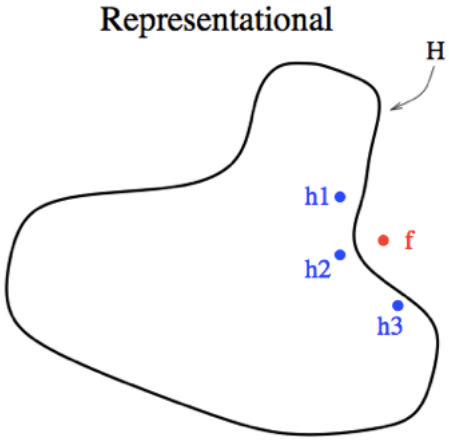
- Many learning algorithms work by performing some form a local search and may get stuck in a local optima. With decision trees for example, an exhaustive search of the entire hypothesis space is extremely complex.



An ensemble method in this case allows you to run local searches from many different starting points. This would be more effective than depending on just one classifier.

THE REPRESENTATIONAL PROBLEM

- Sometimes our true function cannot be expressed in terms of our hypothesis at all. For example, we learned that decision trees work by forming rectilinear partitions of feature space. But what if our true function is a diagonal line?



An ensemble method in this case allows us to expand the space of representable functions and better approximate our true function.

ENSEMBLES? ENSEMBLES!

- **Characteristics of Ensemble methods**
- In order for an ensemble classifier to outperform a single base classifier, the following conditions must be met:
- **Accuracy:** The base classifiers must outperform random guessing
- **Diversity:** There must be some misclassification on different training examples

INTRODUCTION: BAGGING

- Bagging (or bootstrap aggregating) is a method that involves manipulating the training set by resampling with replacement.
- Samples are independently created by resampling training data using uniform weights. In other words, *each model in the ensemble votes with equal weight.*
- Bagging helps reduce overfitting (high variance) by aggregating multiple base classifiers together
- Example of bagging: random forests

INTRODUCTION: BOOSTING

- Boosting involves building out base estimators sequentially, where each new estimator tries to reduce the bias of the combined estimator. Boosting is particularly useful where we're trying to combine several weak models (shallow trees, for example) to build a powerful ensemble.
- Example of boosting: AdaBoost, Gradient Tree Boosting, XGBoost

GENERAL ASSEMBLY

BAGGING IN SKLEARN