GENERAL ASSEMBLY

DATA SCIENCE

SYD DAT 8

Week 6 - Ensembles
Thursday 29th June

AGENDA

1. Review of Decision Trees
2. Ensembling
3. Bagging
4. Random Forest
5. Boosting
6. Lab
7. Other Uses
8. Discussion

DATA SCIENCE PART TIME COURSE

DECISION TREE REVISION

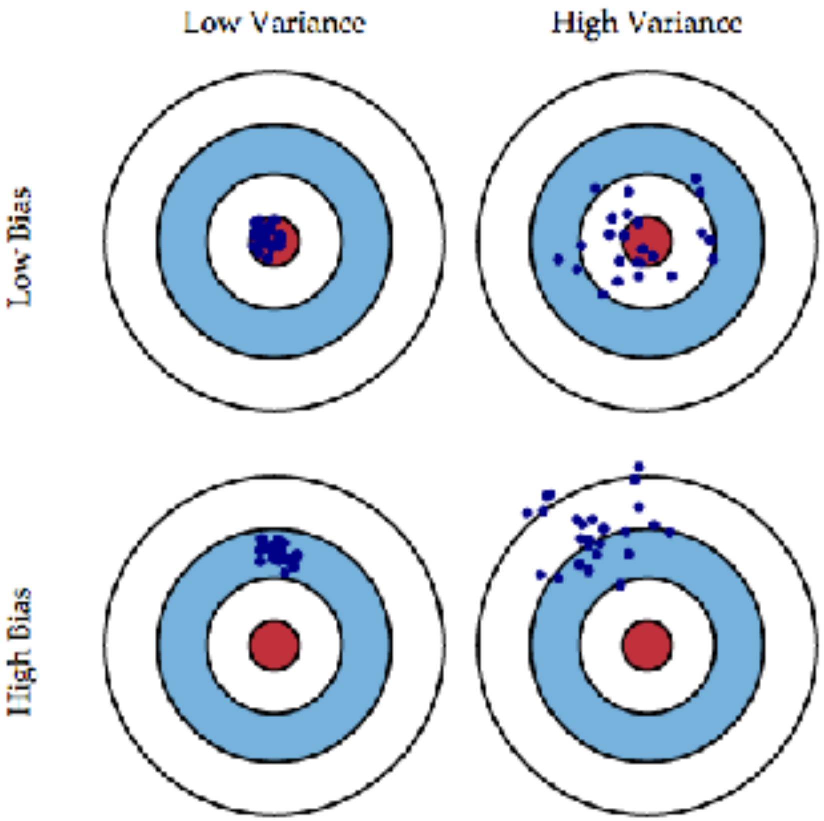
WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.
- Visually engaging and easy to interpret.
- Foundation for getting into very powerful techniques.
- Great for explaining to people!

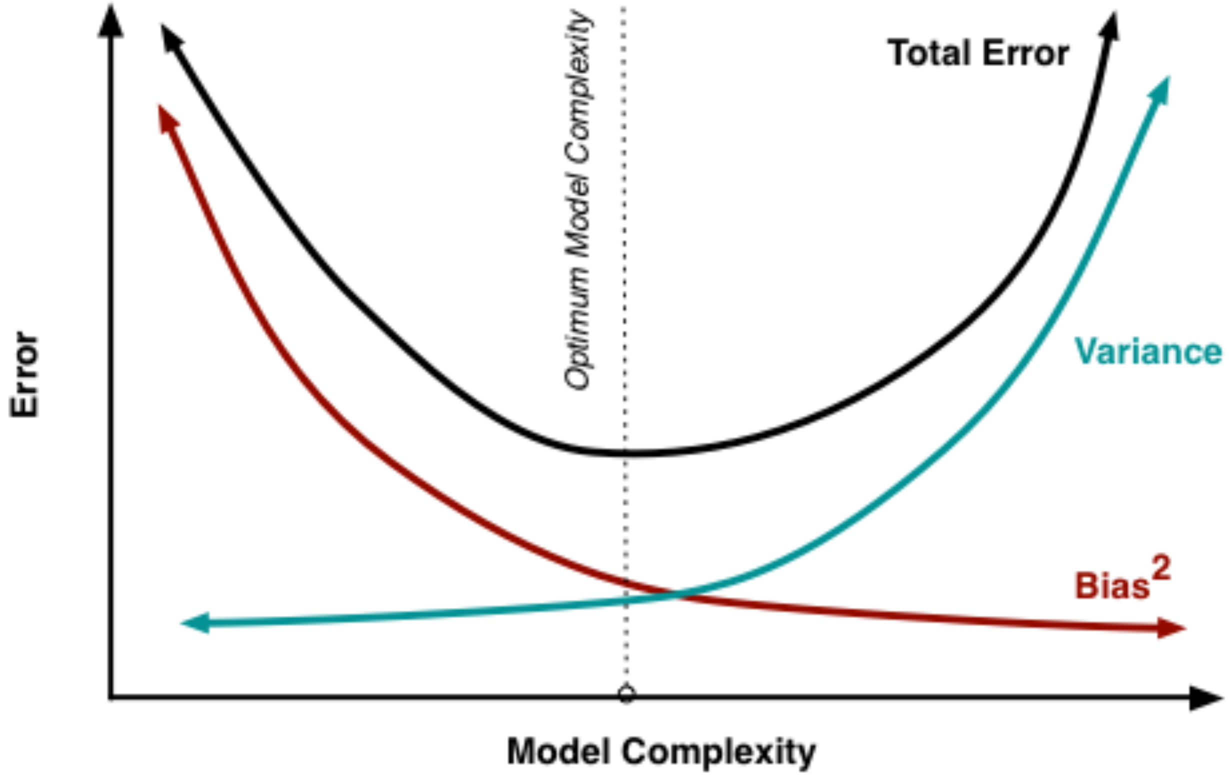
DECISION TREES - SHORTCOMINGS

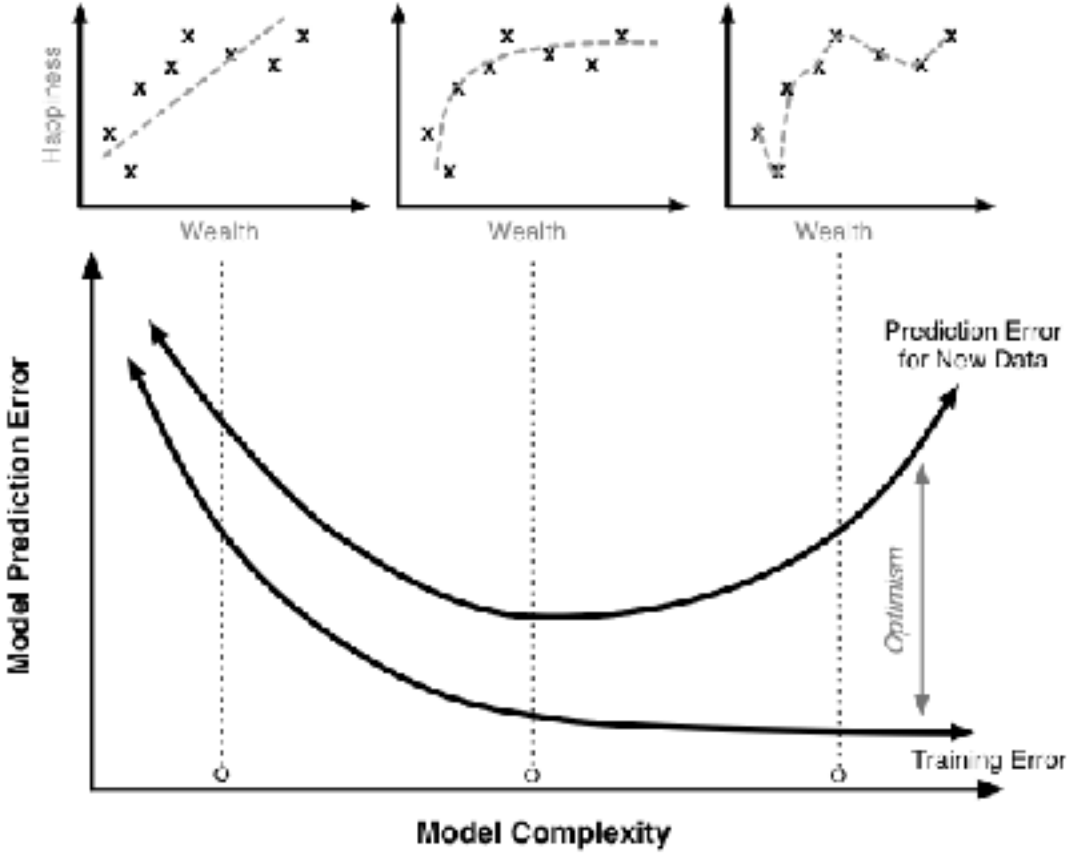
- Prone to overfitting.
- Predictive power is lower in comparison to many other modern techniques.

BIAS – VARIANCE TRADEOFF

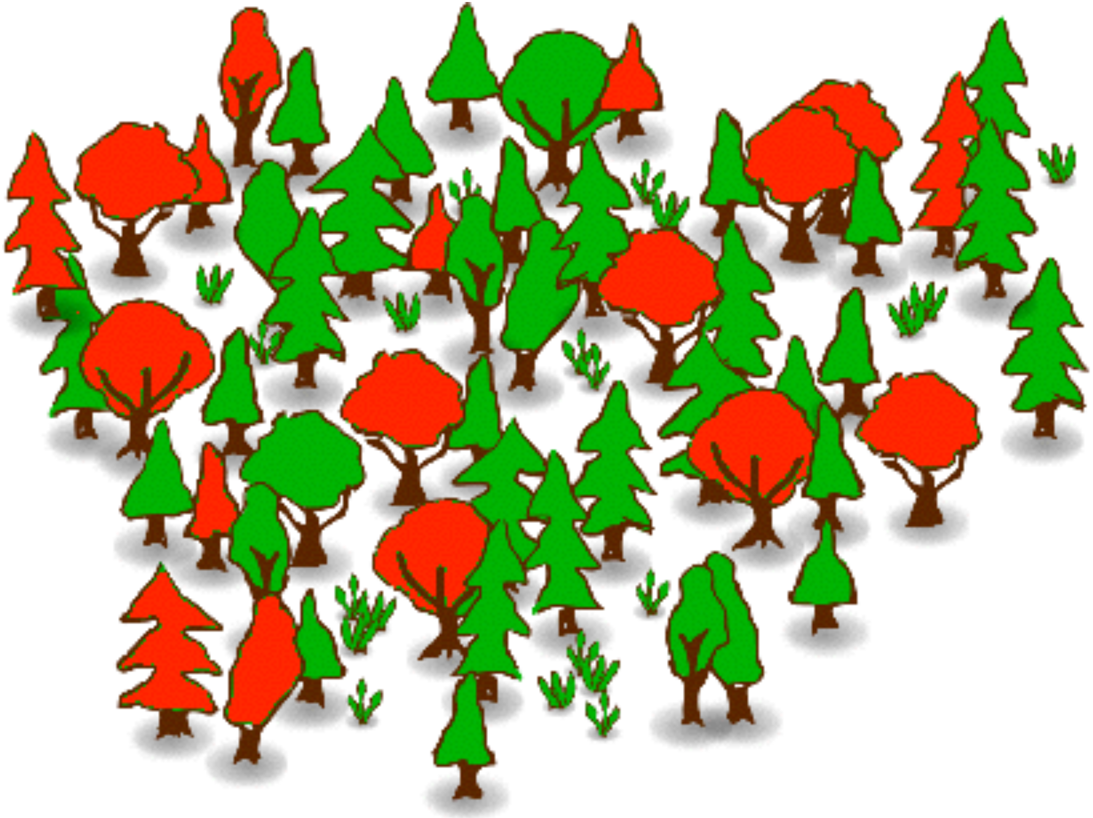


BIAS – VARIANCE TRADEOFF



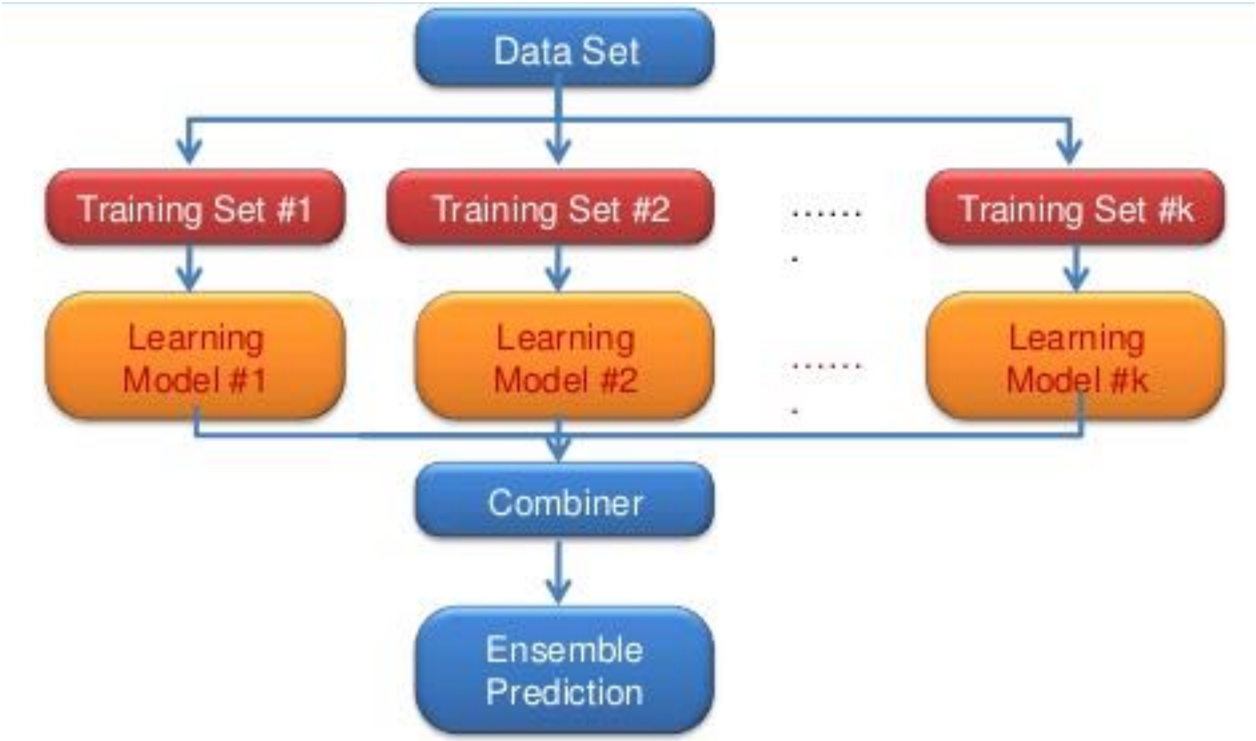


Happiness
Wealth
Wealth
Wealth
Prediction Error
for New Data
Optimism
Training Error
Model Prediction Error
Model Complexity



DATA SCIENCE PART TIME COURSE

ENSEMBLING


Data Set
Training Set #1
Training Set #2
Training Set #k
Learning Model #1
Learning Model #2
Learning Model #k
Combiner
Ensemble Prediction

Ensemble learning (or "ensembling") is the process of combining several models to solve a prediction problem, with the goal of producing a combined model that is more accurate than any individual model.

For classification problems, the combination is often done by majority vote.

For regression problems, the combination is often done by taking an average of the predictions.

ENSEMBLE MODELS – WHEN DO THEY WORK?

For ensembling to work well, the individual models must meet two conditions:

1. Models should be accurate (they must outperform random guessing)
2. Models should be independent (their predictions are not correlated with one another)

Typically Packaged Ensemble Methods

- Bagging
- Boosting

Typically Manually Applied Ensemble Methods

- Voting (classification)
- Averaging
- Rank Averaging
- Stacking
- Blending

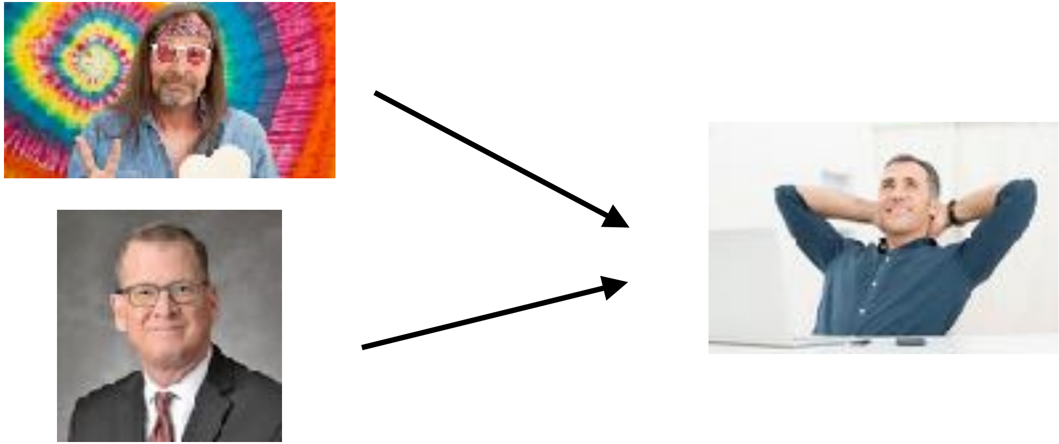
BAGGING

- Bootstrapped Aggregation = Bagging
- The bootstrap is a statistical technique that is used to quantify the uncertainty of a model
- Bootstrap samples are simply random samples with replacement

What is the bagging process?

1. Take repeated bootstrap samples (random samples with replacement) from the training data set
2. Train our method on each bootstrapped training set and make predictions
3. Average the predictions

This increases predictive accuracy by reducing the variance, similar to how cross-validation reduces the variance associated with the test set approach (for estimating out-of-sample error) by splitting many times and averaging the results.

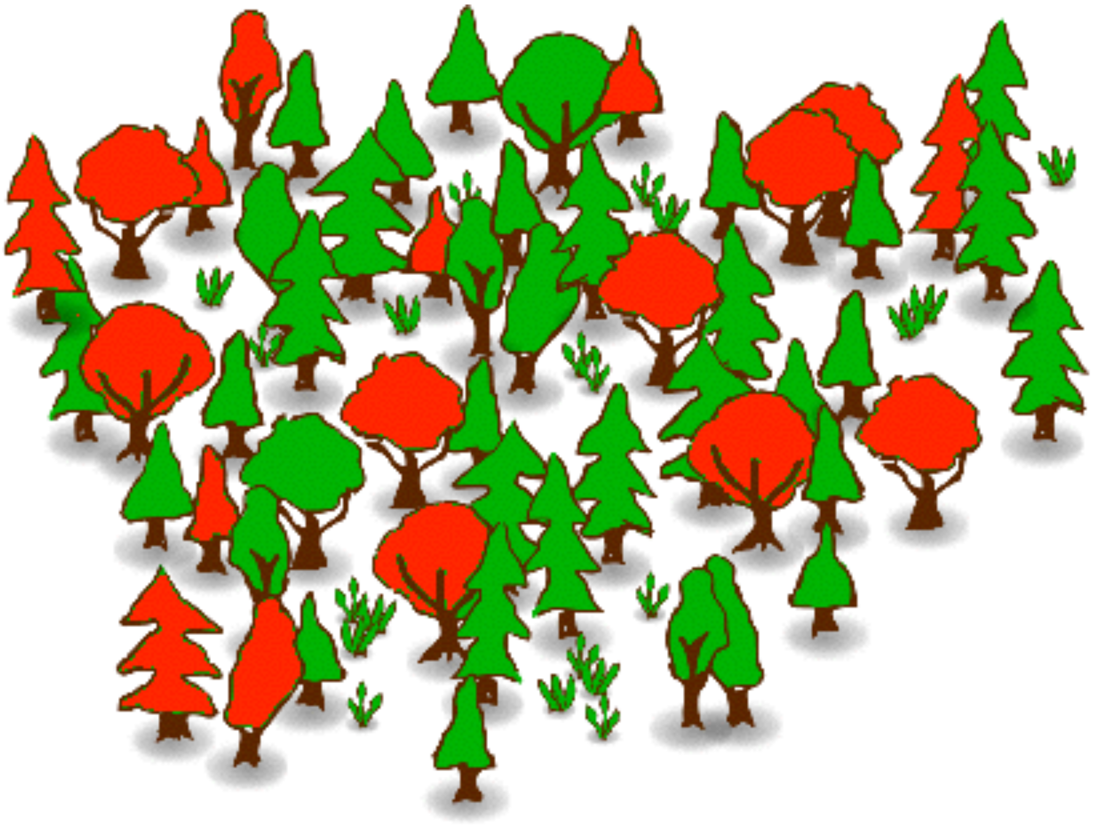


(for Decision Trees)

1. Grow B regression trees using B bootstrapped training sets
2. Grow each tree deep enough that each one has low bias
3. Every tree makes a numeric prediction, and the predictions are averaged (to reduce the variance)

DATA SCIENCE PART TIME COURSE

RANDOM FORESTS



Random Forests is a slight variation of bagged trees that has even better performance! Here's how it works:

- Exactly like bagging, we create an ensemble of decision trees using bootstrapped samples of the training set.
- However, when building each tree, each time a split is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is only allowed to use one of those m predictors.

However, when building each tree, each time a split is considered, a

random sample of m predictors

is chosen as split candidates from the

full set of p predictors.

The split is only allowed to use one of those

m predictors.

- A new random sample of predictors is chosen for every single tree at every single split.
- For classification, m is typically chosen to be the square root of p. For regression, m is typically chosen to be somewhere between p/3 and p.

What's the point?

- Suppose there is one very strong predictor in the data set. When using bagged trees, most of the trees will use that predictor as the top split, resulting in an ensemble of similar trees that are "highly correlated".
- Averaging highly correlated quantities does not significantly reduce variance (which is the entire goal of bagging).
- By randomly leaving out candidate predictors from each split, Random Forests "decorrelates" the trees, such that the averaging process can reduce the variance of the resulting model.

Although bagging increases predictive accuracy, it decreases model interpretability because it's no longer possible to visualize the tree to understand the importance of each variable.

- To compute variable importance for bagged regression trees, we can calculate the total amount that the mean squared error is decreased due to splits over a given predictor, averaged over all trees.
- A similar process is used for bagged classification trees, except we use the Gini index instead of the mean squared error.

OUT-OF-BAG ERROR

Bagged models have a very nice property: out-of-sample error can be estimated without using the test set approach or cross-validation. How it works:

- On average, each bagged tree uses about two-thirds of the observations. For each tree, the remaining observations are called "out-of-bag" observations.
- For the first observation in the training data, predict its response using only the trees in which that observation was out-of-bag. Average those predictions (for regression) or take a majority vote (for classification).
- Repeat this process for every observation in the training data.
- Compare all predictions to the actual responses in order to compute a mean squared error or classification error. This is known as the out-of-bag error.

BOOSTING

BOOSTING

Boosting involves the base learners being built sequentially (using previous models). The aim is to reduce the bias of the combined estimator.

Compare this to something like bagging where the models are built independently of one and other.

AdaBoost works by building sequential models on re-weighted versions of the data. Initially all data points have the same weighting = 1/N.

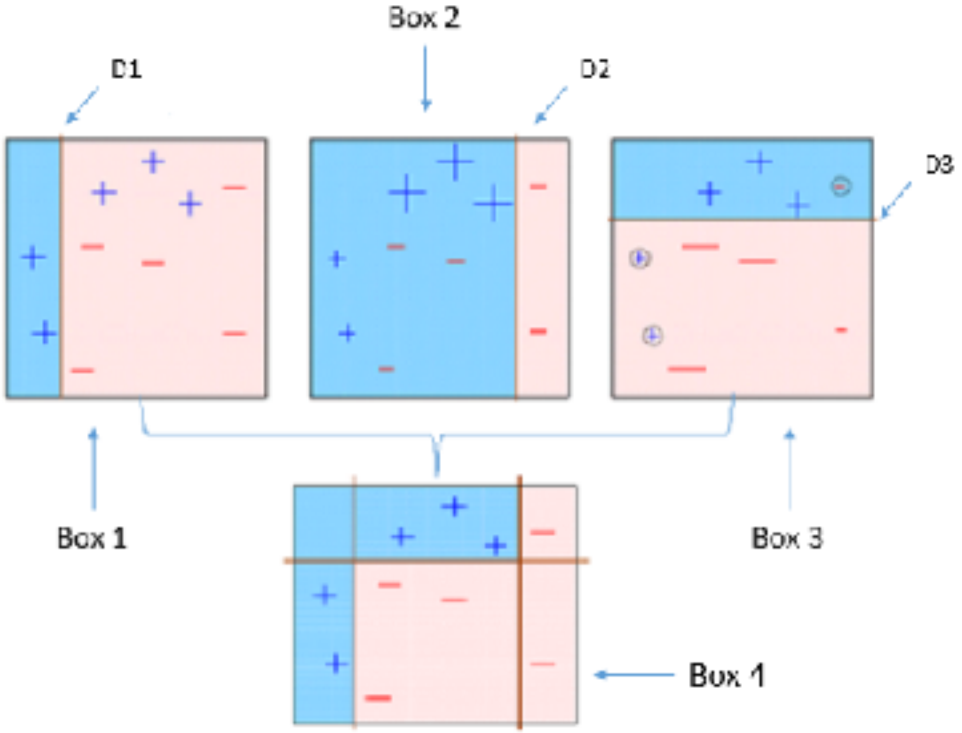
After the first model the error on the points is evaluated and the weighting for each observation is adjusted. So if we fit a model and an observation has a large difference between the observed and predicted we give that observation a higher weighting and run the model again.

This means observations that are difficult to predict become more influential in the model.

Accentuating the training labels for each sequential model.

See how box 2 (model 2) has the labels accentuated.

The simple models are combined for the final model.



LAB

SYNCHING YOUR FORK WITH THE COURSE REPO

1. re-name your labs with lab_name.<yourname>.ipynb (to prevent a conflict)
2. cd <path to the root of your SYD_DAT_6 local repo>
3. commit your changes ahead of sync
    - git status
    - git add .
    - git commit -m "descriptive label for the commit"
    - git status
4. download new material from official course repo (upstream) and merge it
    - git checkout master (ensures you are in the master branch)
    - git fetch upstream
    - git merge upstream/master

HOMEWORK

Homework

- Homework 2 - Due Friday 30th of June

Read the following

- Chapter 8.2 of Introduction to Statistical Learning - Bagging, Random Forests, Boosting