BAGGING AND BOOSTING

A Review of Ensemble Methods

Miranda Gibbons, DSI-DC-4

Agenda

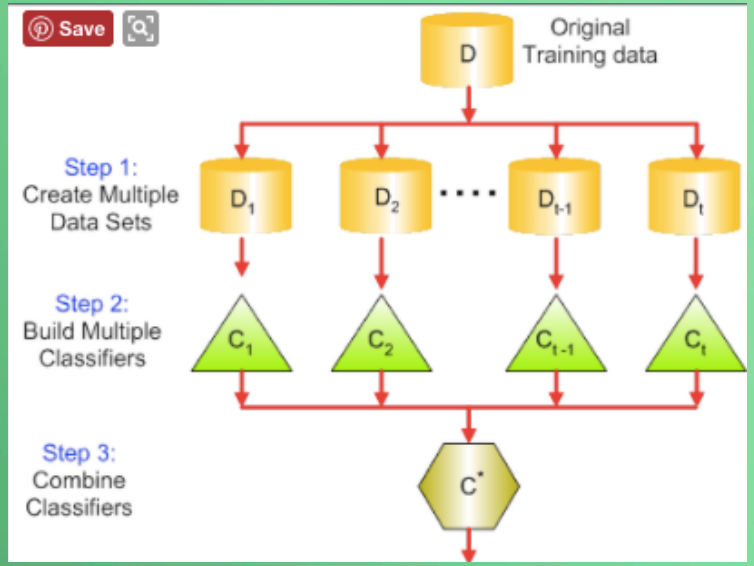
- Define Bagging, Boosting
- When & why might we use these methods?
- Specific examples

What is Bagging?

- Bagging, or Bootstrap Aggregating, is defined as:
  - An ensemble method that creates multiple predictors by resampling training data with replacement iteratively, and aggregates those predictors into a single model
- Resampling our training data with replacement
- Each sample has a uniform weight
- Reduces overfitting (decreases VARIANCE)
- Creates a single model based on independent, aggregated classifiers

Different Bagging Techniques

- Bagging Classifier/Regressor
- Extra Trees (Extremely Randomized Trees)
- Random Forest

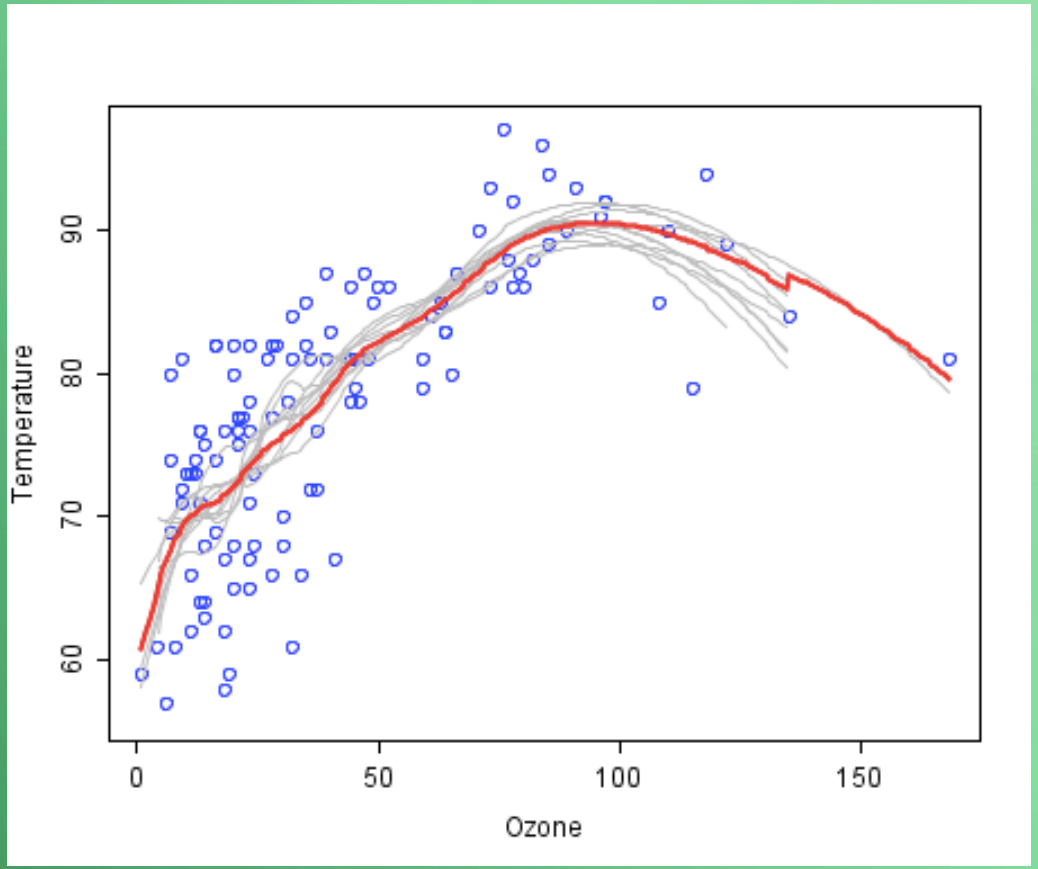


When is Bagging Used?

- Aim to decrease the error due to variance of our complex model
- Not recommended for models with high levels of error due to bias
- E.g. decision trees are prone to overfitting – good candidate for implementing bagging metaheuristic
- Can be used for any classification or regression model

Ex: Bagging

- Ozone data

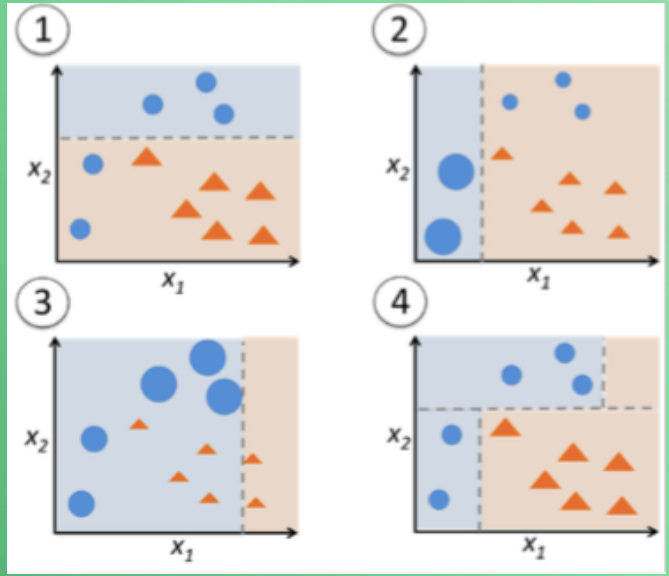


What is Boosting?

- Ensemble method where models/base estimators are built out sequentially
- First iteration: samples have uniform weight
    - Subsequently, misclassified data is weighted heavier, while correctly predicted training data is decreased in weight
- In this way, each subsequent model corrects for misclassified data
- Reduces error due to BIAS

Different Boosting Techniques

- AdaBoost
- Gradient Tree Boosting/ Gradient Boosted Regression Trees
- XGBoost (Extra Gradient Boost)
  - (NB: not in sklearn)

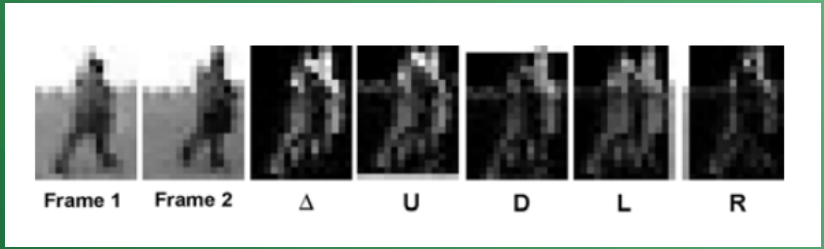


When is Boosting Used?

- Aim to decrease the error due to bias in our model
  - Subsequently increases error due to variance
- Best used with weak models (e.g. shallow decision trees)
- Kaggle competitions (Joking? Maybe?)

Ex: Boosting

- Object Categorization in Images



http://www.merl.com/publications/docs/TR2003-90.pdf

Resources

- http://www.mit.edu/~9.520/spring06/Classes/class10.pdf
- https://stats.stackexchange.com/questions/18891/bagging-boosting-and-stacking-in-machine-learning
- https://www.analyticsvidhya.com/blog/2015/09/questions-ensemble-modeling/
- https://sebastianraschka.com/faq/docs/bagging-boosting-rf.html
- http://scikit-learn.org/stable/modules/classes.html#module-sklearn.ensemble
- BAGGING:
  - http://rdcu.be/rlVM (Breiman article)
  - http://scikit-learn.org/stable/modules/ensemble.html#bagging
- BOOSTING:
  - http://www.cs.princeton.edu/courses/archive/spr08/cos424/readings/Schapire2003.pdf
  - http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf
  - http://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
  - https://www.youtube.com/watch?v=wPqtzj5VZus (Trevor Hastie talk)
  - http://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/
  - http://xgboost.readthedocs.io/en/latest/
  - http://www.hpl.hp.com/techreports/Compaq-DEC/CRL-2001-1.pdf