Week 2

Paper 1

Random Forests

Leo Breiman

Machine Learning 45, 5-32 (2001)

[Random Forests | SpringerLink](https://link.springer.com/article/10.1023/a:1010933404324)

* Shows the math behind random forests
* Explains out of the bag error estimation and how using that estimate removed the need to a set aside test set
* Compares using random forest to using adaptive boosting (AdaBoost)
* Random forest is less susceptible to noise, since it is not concentrating weight
* Mentions difficulty of working with many weak inputs, which can be common with medical data

Summary:

This paper is attributed to be the first mention and description of the random forest technique. Random forest classifiers are based on decision trees which have been generated from random subsets of the features (columns of the data, considering that rows are observations of the data). The classification is determined by majority vote of the decision trees by row. Random Forest can be used for classification (predicting class label) or regression (predicting continuous value). An advantage of random forest is it’s performance with noisy data, since it is not weighting the data and hence potentially weighting noise. Random forests do not overfit data due to the randomness of the selecting variables to use for the prediction. The “out of bag” estimate for a class label is the estimate using trees that do not contain the class label, and it is a good estimate of overall error rate. Random forests can have similar error rate as another machine learning algorithm, AdaBoost. The math behind random forests are covered.

Paper 2

PCA Embedded Random Forest

Charles Gardener, Dan Chia-Tien Lo

IEEE SoutheastCon 2021

[PCA Embedded Random Forest | IEEE Conference Publication | IEEE Xplore (uwf.edu)](https://ieeexplore-ieee-org.ezproxy.lib.uwf.edu/document/9401949)

Study involves combining Random Forest methodology with PCA to improve precision. PCA is used to create new combination vectors for dependent features in this case. It is ideal to create a new feature when a dependency is found, but this methodology would take the guesswork out of it and automate the creation of the new features. The random forest technique would then incorporate the new features. This allows the single split of a random forest / decision tree to split on a variable that combines multiple features. Another gain is the fact that each tree has unique features due to the pca being applied on each tree, so the trees are less correlated, which has been shown to increase accuracy.

Week 3

Paper 1

[A random forest guided tour | SpringerLink](https://link.springer.com/article/10.1007/s11749-016-0481-7)

Literature review. Lots of math covered and the fact that the behavior of random forests is really not very well mathematically described. Lots of references for further materials to review such as additional papers by Leo Breiman. Section 1: intro. Section 2: review of mathematical background. Section 3: theoretical math from niche scenarios that are mathematically describable. Section 4: math on splits and other related math. Section 5: variable selection impacts and math. Section 6: further related usage of random forests such as survival analysis and streaming data with random forests (starting with incomplete data rather than full dataset).

Paper 2

CONSISTENCY FOR A SIMPLE MODEL OF RANDOM FORESTS

Leo Breiman, 2004

<https://www.stat.berkeley.edu/~breiman/RandomForests/consistencyRFA.pdf>

Someone pointed out that the random forest method was not consistent. This paper was an attempt to create some simple random forest models where mathematical consistency could be proven. Refers to a paper by Yi Lin & Yongho Jeon below that showed that the random forest method was a form of adaptive nearest neighbors. Introduces concept of strong and weak variables to help explain. Attempts to create equations to bound variance and bias for this simple model to prove consistency. My personal thought is that the problems themselves aren’t always consistent so the solutions couldn’t be either.

Paper 3

Random Forests and Adaptive Nearest Neighbors

Yi Lin & Yongho Jeon, 2002

<https://www.tandfonline.com/doi/epdf/10.1198/016214505000001230>

Shows that random forest can be viewed as adaptive nearest neighbors. I didn’t read much of this and mostly skimmed it. Mentions loss functions and gradient descent. Mentions that the tree or split is a form of distance measure and therefore it is a nearest neighbor approach. Excluded categorical variables from their test datasets.

Paper 4

Bagging Predictors

Leo Breiman, 1996

<https://link.springer.com/content/pdf/10.1007/BF00058655.pdf>

Prelude to random forest method. Before there was random forest there was bagging predictors. Discusses bootstrap aggregating and the math behind it. Discusses limitations of bagging predictors such as not being able to improve on a stable procedure. Mentions other multi-tree solutions that came before the random forest method, noted as voting over multiple trees and using split methods of alternative splits and oblique splits. Calculations of the improvement of the bagging method over the standard tree method. Covers classification and regression.

Week 4

Paper 1

An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization

T.G. Dietterich, 1999

[An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization | SpringerLink](https://link.springer.com/article/10.1023/a:1007607513941)

Compared a method of randomizing split of C4.5 decision tree method with AdaBoost and Bagging methods. Used 33 test datasets and compared the performance of each. Considered an influence on Leo Breiman and the random forest method. Results printed with confidence interval “Kohavi” plots. Discusses the effect of randomization leading to a diverse set of decision trees. Comparison of performance with “noise” or changed class labels for percentages of the data also tested since it was noted that adaboost did not do well with such noise. The reason being that adaboost amplifies the effect of the noise. Mentions a K error (kappa-error) metric for calculating the differences between classifiers for samples and discusses the math for calculating it. Has to do with diversity of what they called hypotheses. This article was referenced in the original Random Forests paper by Leo Breiman.

Paper 2

Rotation Forest: A New Classifier Ensemble Method

J.J. Rodriguez; L.I. Kuncheva; C.J. Alonso, 2006

[Rotation Forest: A New Classifier Ensemble Method | IEEE Journals & Magazine | IEEE Xplore (uwf.edu)](https://ieeexplore-ieee-org.ezproxy.lib.uwf.edu/document/1677518)

Attempt to improve on random forest method. Discusses bagging, boosting, and random forest methods. Mentions that adaboost is “equivalent to fitting an additive logistic regression model by a stage-wise estimation procedure,” and mentions that some of the math behind bounds and generalization error have been proven, providing references. Mentions that AdaBoost was successful because of imparting diversity to the ensemble, but there was a tradeoff in accuracy. Uses PCA to create new features in a “rotation” method to classify the rotated data. Mentions math behind rotation, that rotation may not be optimal for feature extraction, and that there are other linear transformations/projections that could be used instead of PCA. Mentions using disjoint sets of data to create different classifiers and that it could use intersecting. Uses K sets. Mentions that PCA yields some zero eigenvectors and therefore that the number of components varies. Kept all the components rather than discard low ranking ones. Finds that the rotation forest method has more accurate but less diverse (but still reasonably diverse) individual classifiers and that contrasts with random forests having more diverse but less accurate classifiers. Main finding is that the rotation method tested provided superior results.