**Tentative schedule**

|  |  |  |
| --- | --- | --- |
| Dates | Section | Status |
| March 13-16th | ~~Introduction/Outline/Thesis~~ | done |
| March 18th -19th | Background | Bootstrap/bagging done |
| March 26th-27th | Body |  |
| April 1st-2nd | Conclusion/References/Edit |  |

Outline for Big Data term paper:

Ensemble Methods

1. *Introduction; why am I writing about this topic*
   1. Talk about predictive modeling and data science, my current job, use of trees in building models, popularity of random forests.
   2. Related to on the job modeling; uplift modeling used for marketing campaign (<http://link.springer.com.proxygw.wrlc.org/article/10.1007%2Fs10618-014-0383-9>) (2014)[[1]](#footnote-1)
   3. Have worked quite intimately with Random Forests, would like to get to know Ensemble methods well as Random Forests always perform well. This would allow me to scaled use to other algorithms and come up with ideas to improve modeling techniques using ensemble method approach with multiple models.

*Example of a good introduction:* from Ensemble-based classifiers PDF

Given the potential usefulness of ensemble methods, it is not surprising that a vast number of methods are now available to researchers and practitioners. The aim of this paper is to provide an introductory yet extensive tutorial for the practitioners who are interested in building ensemble-based classification systems.

The rest of this paper is organized as follows: In Sect. 2we present the ingredients of ensemble-based systems. In Sect. 3 we present the most popular methods for combining the base classifiers outputs. We discuss diversity generation approaches in Sect. 4. Section 5 presents selection methods for making the ensemble compact. Hybridization of several ensemble strategies is discussed in Sect. 6. Section 7 suggests criteria for selecting an ensemble method from the practitioner point of view. Finally, Sect. 8 concludes the paper.

1. *Background*
   1. Bootstrapping:

Where the ability to use bagging and boosting comes from. If it were not for bootstrapping, we would need multiple training sets. Instead we are able to create multiple training sets from one actual training set but resampling the training set multiple times; this process is called bootstrapping.

* 1. Bagging

Bagging, random forests, and boosting use trees as building blocks to construct more powerful prediction models. (PG. 316 ISLR)

Hence a natural way to reduce the variance and hence increase the prediction accuracy of a statistical learning method is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions. (PG. 316 ISLR)

<http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf> (ISLR – 2013)[[2]](#footnote-2)

* 1. Boosting

Another approach for improving the predictions resulting from a decision tree. Instead of independent bootstrapped data sets, boosting is grown sequentially, using information from previously grown trees. Boosting does not involve bootstrap sampling. Instead each tree is fit on modified version of the original data set (ISLR pg. 321-322) (can add algorithm on pg. 322). Has three tuning parameters; number of trees (*B)* which we use cross validation to select *B*  or overfit can occur; the shrinkage parameter *lambda* , number of *d* splits in each tree, which controls the complexity of the boosted ensemble.

* 1. Random Forests

Uses Bagging, improvement on bagging with ‘small tweak that decorrelates the trees’ (PG. 320 ISLR)

Improves prediction at expense of interpretability . Variable importance is gained. Main difference between bagging and RF is the choice of predictor subset sixe *m.* Bagging : *m=p,* Random Forests: *m=sqrt(p)*

This *sqrt(p)*  is what leads to reduction in both test error and OOB error over bagging. (can add Figure 8.8 from ISLR)

* 1. Other Ensemble methods
     1. Hidden Markov Model (multiple models employed to obtain the target model) <http://ieeexplore.ieee.org.proxygw.wrlc.org/xpls/abs_all.jsp?arnumber=6587849&tag=1> (29 August 2013)[[3]](#footnote-3)
     2. Ensemble-based classifiers Lior Rokach (2010) – in term paper folder: history and background here… pg.3 (The ensemble ide in supervised learning has been…)[[4]](#footnote-4)
     3. Important to distinguish between dependent frameworks and independent frameworks for building ensembles.
     4. Stacking? (mentioned in ‘Ensemble Methods: Foundations and Algorithms”)
     5. Naïve Bayes, Support Vector Machines and Kernel Methods (Ensemble Methods: foundations..)
  2. Newer ideas/additional areas of interest
     1. Rotation Forest; a new classifier ensemble method: <http://ieeexplore.ieee.org.proxygw.wrlc.org/xpls/abs_all.jsp?arnumber=1677518&tag=1> (2006)
     2. Ensemble learning from multiple information sources:

<http://search.proquest.com.proxygw.wrlc.org/docview/1537156612?pq-origsite=summon&accountid=11243&selectids=10000008,1006323,1006324,1006102,1006985>

* + 1. Semi-supervised learning using ensemble methods (Ensemble Methods: Foundations and Algorithms, 2012) [[5]](#footnote-5)

1. Body
   * 1. Models I will focus on:
        1. Random Forest
        2. Naïve Bayes
        3. Independent Ensemble methods
     2. Ensemble method using multiple models:

Independent ensemble methods: Each classifier is built independently and their results are combined in some fashion. This is sometimes referred to as “cooperative”

<http://arxiv.org/pdf/1106.0257.pdf> (popular ensemble methods PDF)[[6]](#footnote-6)

* + 1. Ensemble method using the same model over and over:

Dependent framework ensemble methods: outcome of a certain classifier affects the creation of the next classifier. Sometimes referred to as “successive”.

Pattern Classification Using Ensemble Methods (pg.42)

*GOOD BOOK GOES OVER IND. AND DEP. METHODS*: <https://books.google.com/books?id=4qnUwdoaVbsC&pg=PA41&lpg=PA41&dq=independent+vs.+dependent+framework+ensemble+methods&source=bl&ots=imFR2SWTnH&sig=TIXXFUdcMK5D_-oklAkLZ1yvlDs&hl=en&sa=X&ved=0ahUKEwiI6KrJx77LAhUHsIMKHaKcAPgQ6AEINTAD#v=onepage&q=independent%20vs.%20dependent%20framework%20ensemble%20methods&f=false>[[7]](#footnote-7)

* + 1. Bagging
    2. Boosting
    3. How bagging and boosting are used respectively within models. Problems and benefits with both. Other ideas around both methods. When to use bagging vs. boosting.
    4. Run through sample data set (maybe Titanic data on Kaggle) to illustrate all ensemble methods discussed here.

1. Conclusion

1. First reference: Ensemble methods for uplift modeling [↑](#footnote-ref-1)
2. Second reference- ISLR [↑](#footnote-ref-2)
3. Third reference – Hidden Markov Models [↑](#footnote-ref-3)
4. Fourth reference – ensemble based classifiers white paper [↑](#footnote-ref-4)
5. Fifth reference: Ensemble Methods: Foundations and Algorithms [↑](#footnote-ref-5)
6. Sixth reference: Popular Ensemble Methods (ONE BEFORE 2006, 1999) [↑](#footnote-ref-6)
7. Seventh reference: Pattern Classification Using Ensemble Methods (2010) [↑](#footnote-ref-7)