Meta-Learning

Predictive models are like potato chips

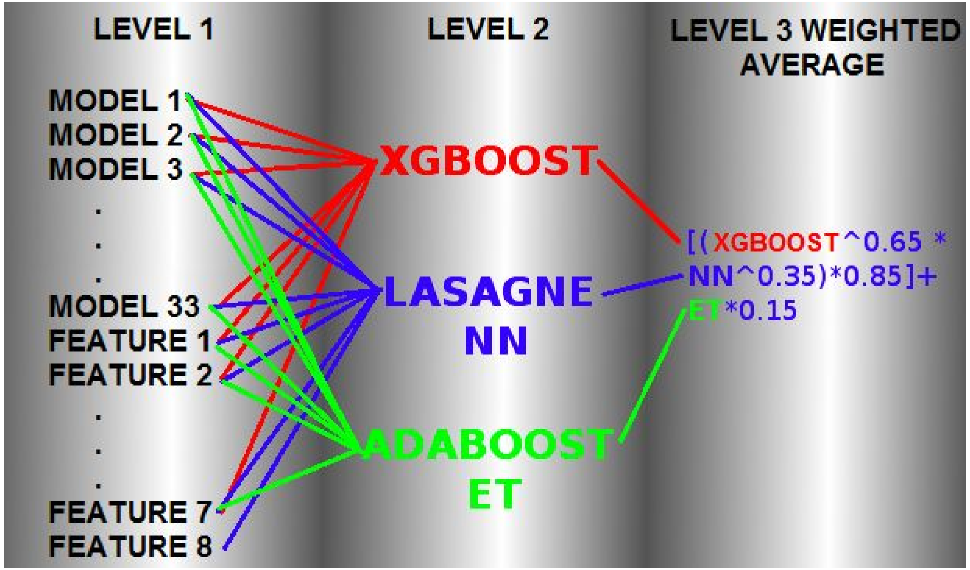
Sometimes you can't have just one.

Need a combination of methods (**ensemble**) for certain situations.

http://www.julescatering.com/blog/wp-content/uploads/2012/11/Bert-Lahr-Frito-Lay-05-1967.jpg

Layers of Learning

Gilberto Titericz Junior (top-ranked user on Kaggle.com) used this setup to win the $10,000 Otto Group Product Classification Challenge.



33 models???

3 levels???

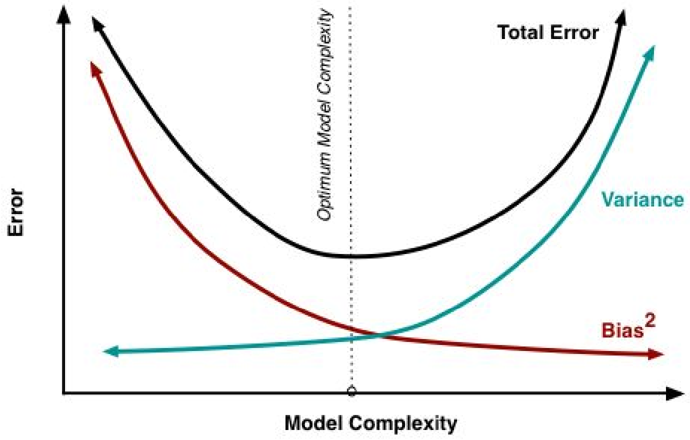
LASAGNE???

Why so many models?

Recall: A single model on its own is often prone to bias and/or variance.

- **Bias** - Systematic or "consistent" error. Associated with underfitting.
- **Variance** - Random or "deviating" error. Associated with overfitting.

A tradeoff exists. We want to minimize both as much as we can.



http://scott.fortmann-roe.com/docs/docs/BiasVariance/biasvariance.png

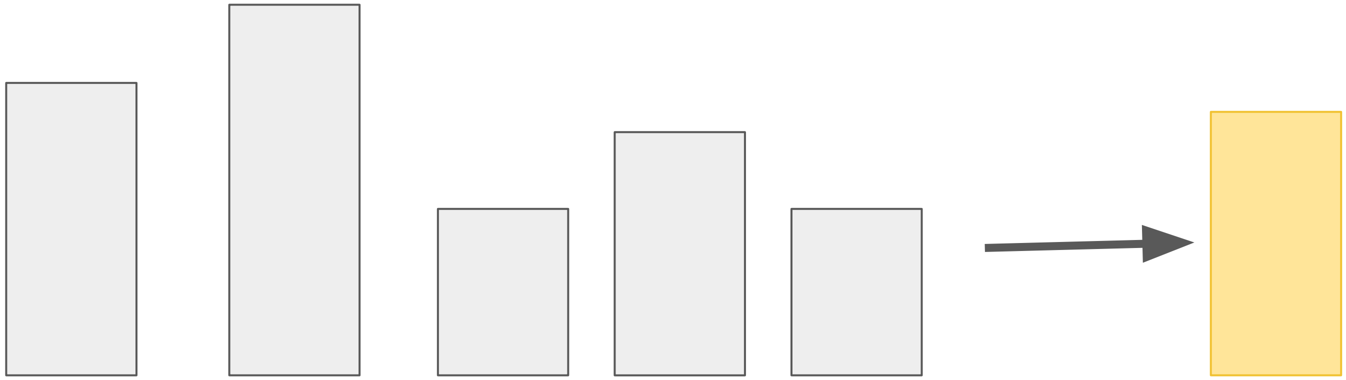
Ensembles and Hypotheses

- Recall the definition of “hypothesis.”
- Machine learning algorithms search the **hypothesis space** for hypotheses.
    - Set of mathematical functions on real numbers
    - Set of possible classification boundaries in feature space
- More searchers are more likely to find a “good” hypothesis that minimizes bias.
- We can then combine the searchers’ results in a way that minimizes variance.

Introduction: Ensemble Averaging

Basic ensemble composed of a **committee** of learning algorithms.

Results from each algorithm are averaged into a final result, reducing variance.



More Sophisticated Ensembles

Three important ensembles to know:

Boosting	Bagging	Stacking

http://ep.yimg.com/ca/I/yhst-96751435660117_2272_27759083

https://wattsupwiththat.files.wordpress.com/2014/02/plastic_bag.jpg

http://www.clipartkid.com/images/220/stack-of-money-clipart-shorttermloanscenter-com-u5k9gG-clipart.jpg

Boosting

Boosting decreases **bias** and prevents **underfitting.**

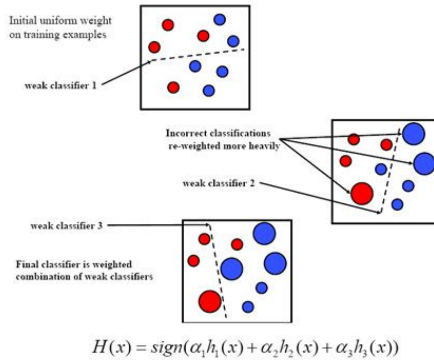
A **sequential ensemble**. Models are applied one-by-one based on how previous models have done.

- Apply a model on a subset of data.
- Check to see where the model has badly classified data.
- Apply another model on a new subset of data, giving preference to data badly classified by the model.

http://pad3.whstatic.com/images/thumb/7/7f/Carry-Someone-Who's-Bigger-Than-You-Step-6.jpg/aid166374-728px-Carry-Someone-Who's-Bigger-Than-You-Step-6.jpg

AdaBoost

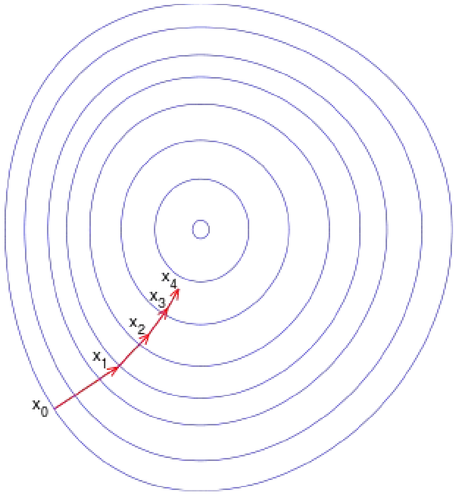
- Short for **ada**ptive **boost**ing
- Uses "weak learners" - simple models that do slightly better than random guessing
    - Example: **decision stump** (decision tree with one level)
- Sequentially generates weak learners, adjusting newer learners based on mistakes of older learners
- Combines output of all learners into weighted sum



http://www.cc.gatech.edu/~kihwan23/imageCV/Final2005/images/bestrong.JPG

XGBoost

- Short for eXtreme Gradient Boosting
- Sequentially generates “weak learners” like Adaboost
- Updates model by computing cost function
  - Computes gradient of cost function
  - Direction of greatest decrease = negative of gradient
  - Creates new learner with parameters adjusted in this direction



https://upload.wikimedia.org/wikipedia/commons/thumb/f/ff/Gradient_descent.svg/350px-Gradient_descent.svg.png

Boosting Demo

We'll be comparing the predictive power of the `xgboost` package in R to the standard logistic regression model (`glm`) on the same dataset.

Demo time!

Bagging

Bagging decreases **variance** and prevents **overfitting.**

Short for **b**ootstrap **agg**regat**ing**.

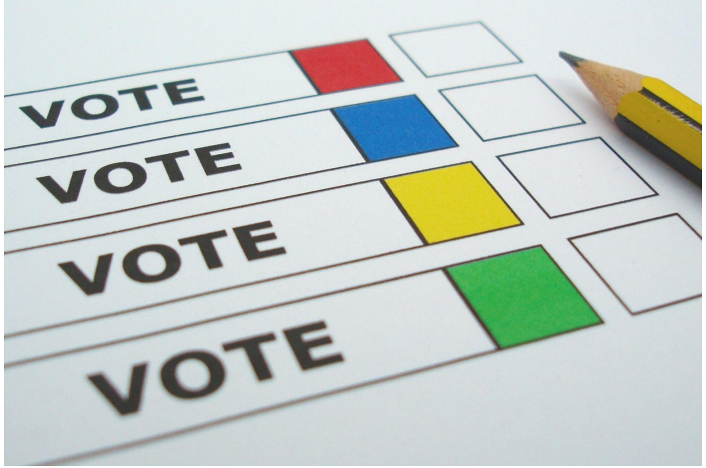
A **parallel ensemble**. Models are applied without knowledge of each other.

- Apply each model on a random subset of data.
- Combine the output by averaging (for regression) or by majority vote (for classification)
- A more sophisticated version of ensemble averaging.

https://cdn.shopify.com/s/files/1/0043/9252/products/the-vintage-tote-bag-1.jpg?v=1470499389

Random Forests

- Designed to improve accuracy over CART
- Much more difficult to overfit
- Works by building a large number of CART trees
    - Disadvantage: Makes model harder to understand and follow
    - Each tree in the forest “votes” on outcome
    - Outcome with the most votes becomes our prediction



https://nssa-nsca.org/wp-content/uploads/2016/05/Vote.jpg

Random Forests

- Wouldn't each CART tree be identical?
  - Right! So random forest changes each tree's training data a bit

  - Each tree is trained on a random subset of the data
  - Example - original data: 1 2 3 4 5
  - New data:

    2 4 5 2 1 ----> first tree

    4 1 3 2 1 ----> second tree

    3 5 1 5 2 ----> third tree

Random Forest Parameters

- Minimum number of observations in a branch
  - `nodesize` parameter, similar to `minbucket` in CART
  - Smaller the node size, more branches, longer the computation

- Number of trees
  - `ntree` parameter
  - Fewer trees means *less accurate* prediction
  - More trees means *longer computation* time
  - Diminishing returns after a couple hundred trees

Bagging Demo Time!

How Stacking Works

Assumption: can improve performance by taking a **weighted average** of the predictions of models.

- Apply models on subsets of your data (how you choose them is up to you)
- Obtain predictions and perform linear regression on the predictions
    - This gives you the coefficients of the weighted average
- Result: a massive blend of potentially hundreds of models!

https://upload.wikimedia.org/wikipedia/commons/thumb/9/95/Gojira_1954_Japanese_poster.jpg/220px-Gojira_1954_Japanese_poster.jpg

Stacking

Linear regression...

...on models.

http://akns-images.eonline.com/eol_images/Entire_Site/2015725/rs_634x920-150825112125-634-mccauley-culkin-home-alone-2-08255.jpg

https://upload.wikimedia.org/wikipedia/en/1/18/Inception_OST.jpg

Stacking Example

We'll use the `lm` function in R on the results of three different models:

- (finish this later)

Demo time!

Where to Learn More

- The course notes (available on the course website)
- http://stats.stackexchange.com/questions/18891/bagging-boosting-and-stacking-in-machine-learning
- http://scott.fortmann-roe.com/docs/BiasVariance.html
- http://mlwave.com/kaggle-ensembling-guide/

Coming Up

Your problem set:

Next week: Learning how to analyze text

See you then!

Cornell Data Science