ITCS 6156/8156 Fall 2023
Machine Learning

Ensemble Methods

Instructor: Hongfei Xue
Email: hongfei.xue@charlotte.edu
Class Meeting: Mon & Wed, 4:00 PM – 5:15 PM, CHHS 376

UNIVERSITY OF NORTH CAROLINA
CHARLOTTE

Some content in the slides is based on Dr. Raquel Urtasun’s lecture

Ensemble Methods

- Typical application: classification
- Ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way to classify new examples
- Simplest approach:
  1. Generate multiple classifiers
  2. Each votes on test instance
  3. Take majority as classification
- Classifiers are different due to different sampling of training data, or randomized parameters within the classification algorithm
- Aim: take simple mediocre algorithm and transform it into a super classifier without requiring any fancy new algorithm

Ensemble Methods

- Differ in training strategy, and combination method
  - Parallel training with different training sets
    1. Bagging (bootstrap aggregation) – train separate models on overlapping training sets, average their predictions
  - Sequential training, iteratively re-weighting training examples so current classifier focuses on hard examples: boosting
  - Parallel training with objective encouraging division of labor: mixture of experts

- Notes:
  - Also known as meta-learning
  - Typically applied to weak models, such as decision stumps (single-node decision trees), or linear classifiers

Bias & Variance

- Bias and Variance are two fundamental concepts in machine learning that pertain to the errors associated with predictive models.
- Bias: The differences between actual or expected values and the predicted values are known as bias error or error due to bias. Bias is a systematic error that occurs due to wrong assumptions in the machine learning process.
  - Low Bias: In this case, the model will closely match the training dataset.
  - High Bias: If a model has high bias, this means it can't capture the patterns in the data, no matter how much you train it. The model is too simplistic. This scenario is often referred to as underfitting.
- Variance: Variance is the amount by which the performance of a predictive model changes when it is trained on different subsets of the training data. More specifically, variance is the variability of the model that how much it is sensitive to another subset of the training dataset (i.e. how much it can adjust on the new subset of the training dataset).
  - Low Variance: Low variance means that the model is less sensitive to changes in the training data and can produce consistent estimates of the target function with different subsets of data from the same distribution.
  - High Variance: High variance means that the model is very sensitive to changes in the training data and can result in significant changes in the estimate of the target function when trained on different subsets of data from the same distribution.

Variance-bias Tradeoff

- Minimize two sets of errors:

  1. Variance: error from sensitivity to small fluctuations in the training set
  2. Bias: erroneous assumptions in the model

- Variance-bias decomposition is a way of analyzing the generalization error as a sum of 3 terms: variance, bias and irreducible error (resulting from the problem itself)

- Based on one of two basic observations:

  1. Variance reduction: if the training sets are completely independent, it will always help to average an ensemble because this will reduce variance without affecting bias (e.g., bagging)
     - reduce sensitivity to individual data points
  2. Bias reduction: for simple models, average of models has much greater capacity than single model (e.g., hyperplane classifiers, Gaussian densities).
     - Averaging models can reduce bias substantially by increasing capacity, and control variance by fitting one component at a time (e.g., boosting)

Ensemble Methods: Justification

- Ensemble methods more accurate than any individual members if:
  - Accurate (better than guessing)
  - Diverse (different errors on new examples)
- Why?
- Independent errors: prob k of N classifiers (independent error rate ϵ) wrong:

$$P(\text{num errors} = k) = \binom{N}{k} \epsilon^k (1 - \epsilon)^{N-k}$$

- Probability that majority vote wrong: error under distribution where more than $N/2$ wrong

Ensemble Methods: Justification

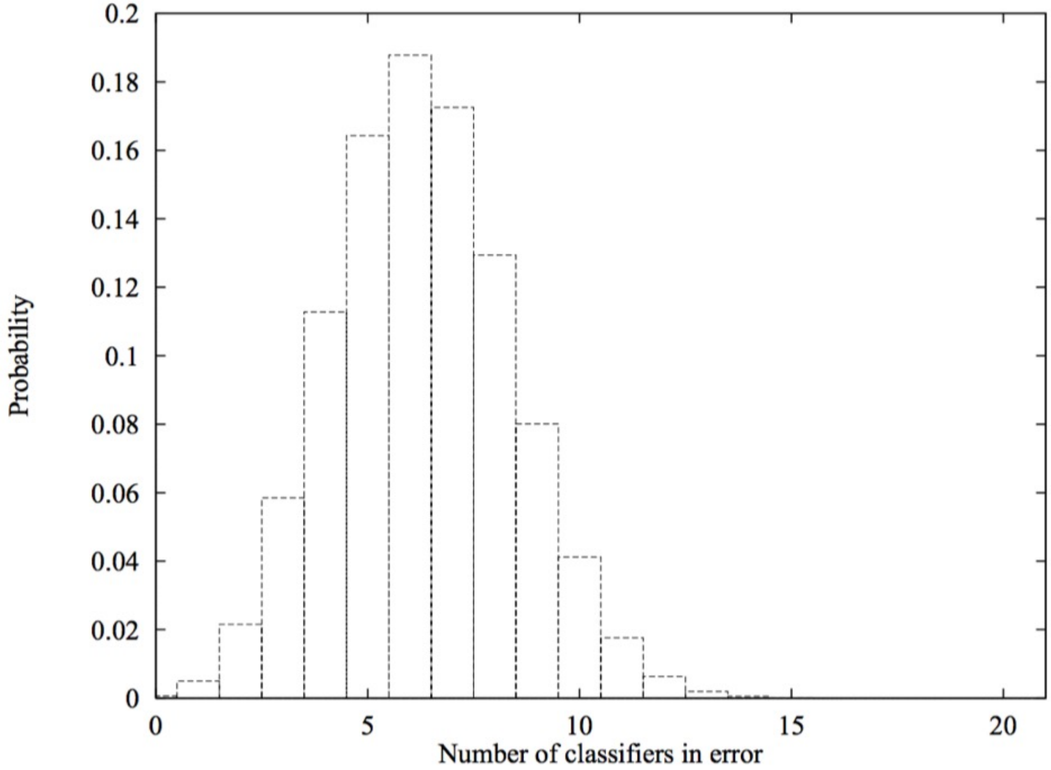


Figure : Example: The probability that exactly K (out of 21) classifiers will make an error assuming each classifier has an error rate of $\epsilon = 0.3$ and makes its errors independently of the other classifier. The area under the curve for 11 or more classifiers being simultaneously wrong is 0.026 (much less than ϵ).

Ensemble Methods: Justification

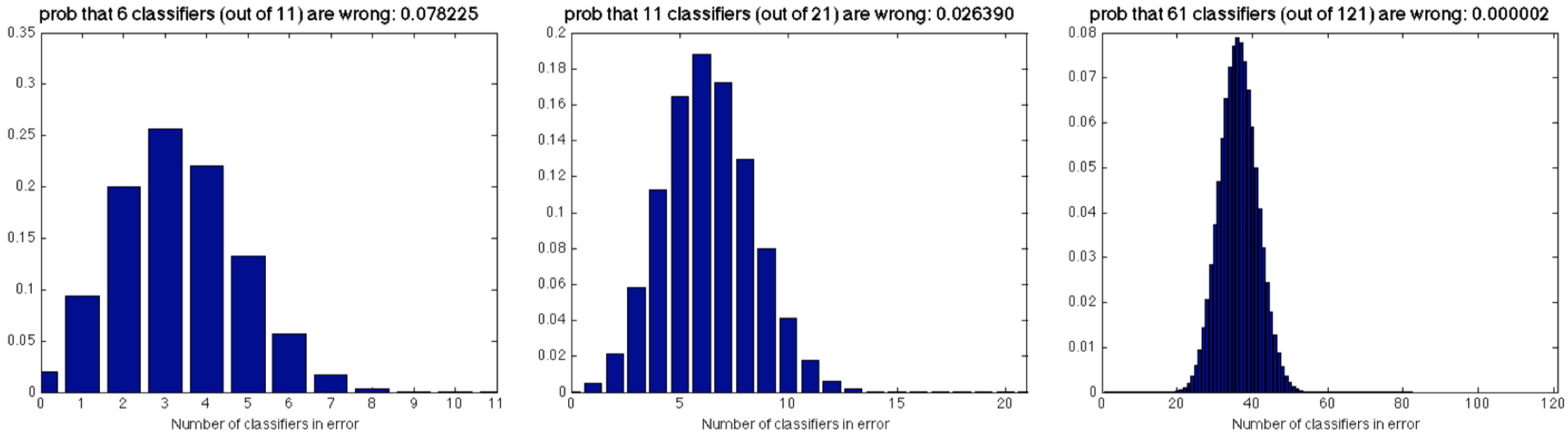


Figure : $\epsilon = 0.3$: **(left)** $N = 11$ classifiers, **(middle)** $N = 21$, **(right)** $N = 121$.

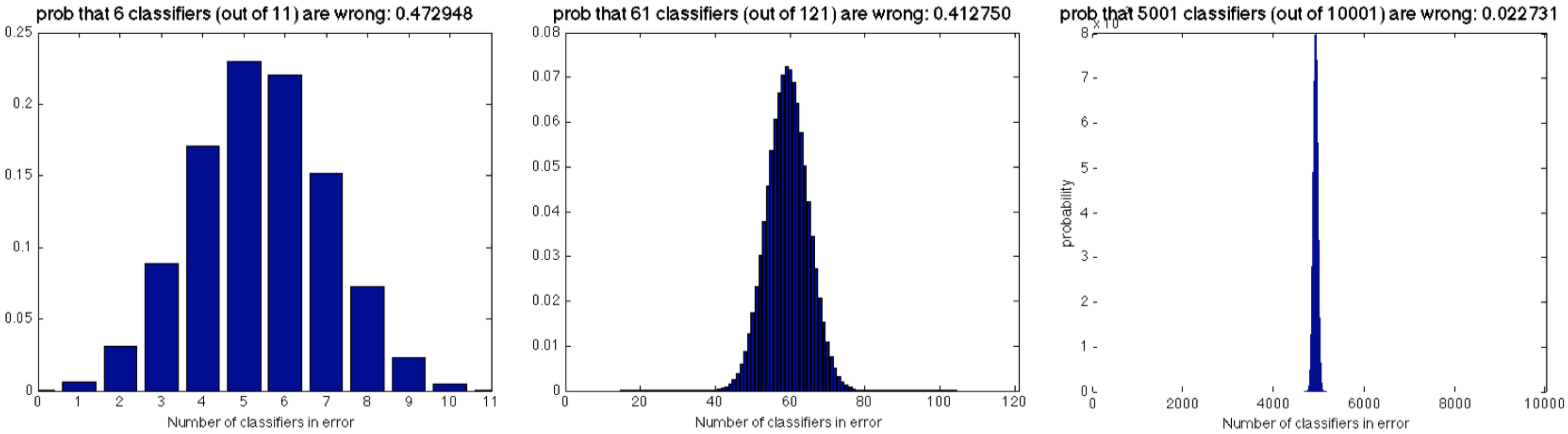
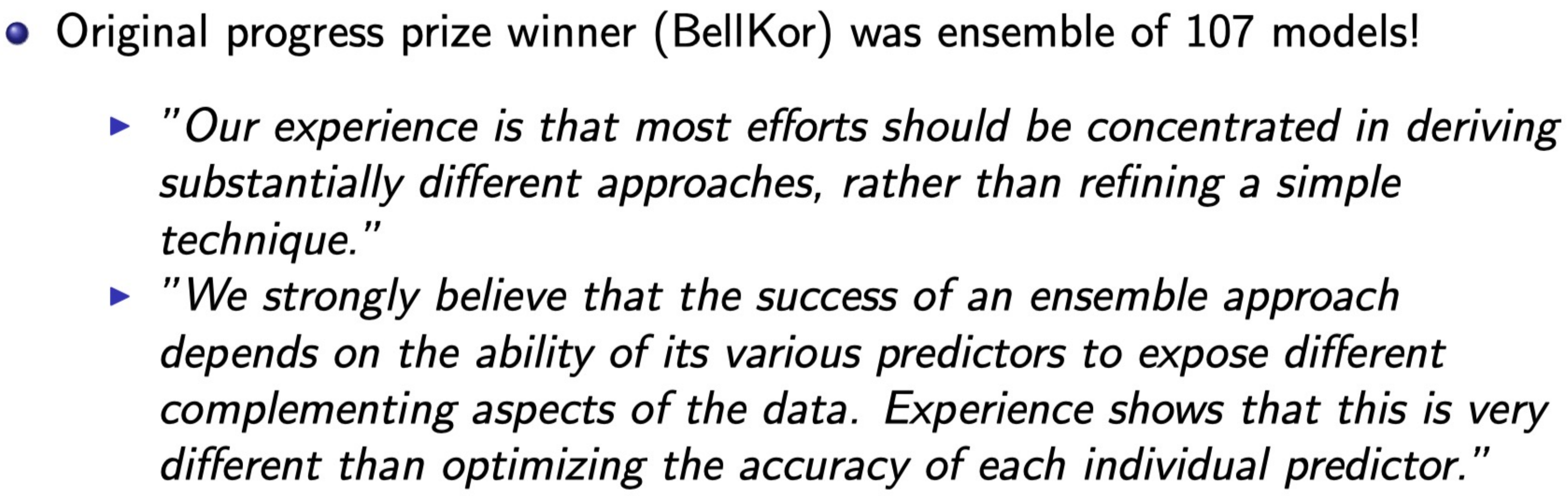


Figure : $\epsilon = 0.49$: **(left)** $N = 11$, **(middle)** $N = 121$, **(right)** $N = 10001$.

Netflix Prize 2007

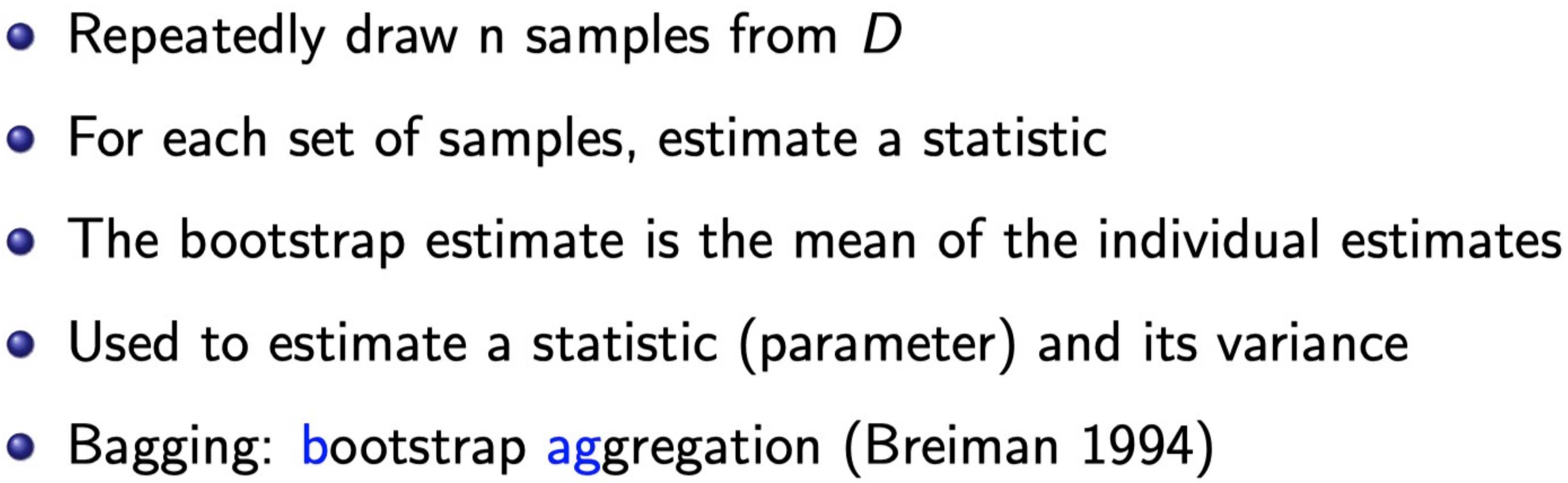
- The Netflix Prize was an open competition for the best collaborative filtering algorithm to predict user ratings for films, based on previous ratings without any other information about the users or films.

- Rewarded $50,000 in 2007.



- Original progress prize winner (BellKor) was ensemble of 107 models!
    - *”Our experience is that most efforts should be concentrated in deriving substantially different approaches, rather than refining a simple technique.”*
    - *”We strongly believe that the success of an ensemble approach depends on the ability of its various predictors to expose different complementing aspects of the data. Experience shows that this is very different than optimizing the accuracy of each individual predictor.”*

Bootstrap Estimation



- Repeatedly draw n samples from D
- For each set of samples, estimate a statistic
- The bootstrap estimate is the mean of the individual estimates
- Used to estimate a statistic (parameter) and its variance
- Bagging: bootstrap aggregation (Breiman 1994)

Bagging

- Simple idea: generate M bootstrap samples from your original training set. Train on each one to get y_m, and average them

$$y_{bag}^{M}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} y_m(\mathbf{x})$$

- For regression: average predictions
- For classification: average class probabilities (or take the majority vote if only hard outputs available)
- Bagging approximates the Bayesian posterior mean. The more bootstraps the better, so use as many as you have time for
- Each bootstrap sample is drawn with replacement, so each one contains some duplicates of certain training points and leaves out other training points completely

Random/Decision Forests

- Definition: Ensemble of decision trees
- Algorithm:
  - Divide training examples into multiple training sets (bagging)
  - Train a decision tree on each set (can randomly select subset of variables to consider)
  - Aggregate the predictions of each tree to make classification decision (e.g., can choose mode vote)

Questions?