Biostat 202: Opportunities and Challenges of “Big Data”:

 **Machine Learning 4:** Cross-Validation, Ensembles, and Feature Importance

Aaron Wolfe Scheffler
Division of Biostatistics
Department of Epidemiology and Biostatistics
Outline

• Cross-validation
• Ensemble methods
  • Bagging: *Random Forests*
  • Boosting: *Adaboost*
  • Stacking
• Parameter tuning, continued
• Feature importance
Cross-validation
E.g. 5-fold cross-validation

1. Break up data into 5 “folds” of the same size

2. Set aside one fold for validation and use the rest to build model, repeat this process for each fold

- Estimate the predictive performance of our model on the combined validation data (i.e. the validation_1-5 sets)
Ensemble methods

• Train a bunch of predictive models and pool them to form a final prediction.

• **Advantage:**
  • improvement in predictive accuracy

• **Disadvantages:**
  • more computation
  • it is difficult to understand an ensemble of classifiers

• Examples: bagging, boosting, and stacking
Bagging
Boosting
Boosting vs Bagging

<table>
<thead>
<tr>
<th>Single Classifier</th>
<th>Bagging</th>
<th>Boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Iteration</td>
<td>Parallel</td>
<td>Sequential</td>
</tr>
<tr>
<td>Model</td>
<td>Model</td>
<td>Model</td>
</tr>
</tbody>
</table>

- Single Classifier: A single model is created.
- Bagging: Models are created in parallel, with each iteration adding new data to the existing models.
- Boosting: Models are created sequentially, with each iteration focusing on reducing the error of the previous model.
Stacking

- Initial dataset
- L weak learners (that can be non-homogeneous)
- Meta-model (trained to output predictions based on weak learners predictions)
Feature importance

• Even though we fool ourselves into only being concerned with prediction, we often want to understand which features impact our prediction.

• Given a target variable, which variables are important for prediction?

• The define this, you can develop measures of feature importance.

• Feature importance can either be pair wise or model based.
  • **Pairwise**: based directly on measures of association between target and individual features (*e.g.* correlation).
  • **Model based**: based on machine learning algorithms (*e.g.* regression coefficients, classification tree splits).
Orange Demo of Topics
Next time

• Clustering
  • k-means clustering
  • TwoStep clustering
  • Kohonen self-organizing map (SOM)

• Data reduction
  • Principal components analysis (PCA)
  • t-Distributed Stochastic Neighbor Embedding (t-SNE)

• Guidance for choosing machine learning algorithms
<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict time to readmission among diabetic patients</td>
<td></td>
</tr>
<tr>
<td>Prepare reproducible models</td>
<td></td>
</tr>
<tr>
<td>Collaborate as teams, submit work individually</td>
<td></td>
</tr>
<tr>
<td>You will have one hour (max) to perform this task</td>
<td></td>
</tr>
</tbody>
</table>