PART 1. LARGE SCALE DATA ANALYTICS
5. EVALUATION AND VALIDATION TECHNIQUES

Sangmi Lee Pallickara
Computer Science, Colorado State University
http://www.cs.colostate.edu/~cs435

Today’s topics

• Evaluation and Validation Techniques

Why evaluation/validation?                                   (1/2)

• Process for model selection and performance estimation

  • Model selection (fitting the model)
    • Most of the models have one or more free parameters
      \[ h(x) = \theta_0 + \theta_1 x_1 \]
      
    • How do we select the "optimal" parameter(s) or model for a given classification problem/predictive analytics?

Why evaluation/validation?                                   (2/2)

• Performance estimation
  • Once we have chosen a model, how do we estimate its performance (accuracy)?

• Performance is typically measured by the true error rate
  • e.g. the classifier’s error rate on the entire population

Challenges                                   (1/2)

• If we had access to an unlimited number of examples these questions have a straightforward answer
  • Choose the model that provides the lowest error rate on the entire population
    • Of course, that error rate is the true error rate

• In real applications we only have access to a subset of examples, usually smaller than we wanted
  • What if we use the entire available data to fit our model and estimate the error rate?
    • The final model will normally overfit the training data
    • We already used the test dataset to train the data
Challenges

- This problem is more pronounced with models that have a large number of parameters
  - The error rate estimate will be overly optimistic (lower than the true error rate)
    - In fact, it is not uncommon to have 100% correct classification on training data

- A much better approach is to split the training data into disjoint subsets: the holdout method

The Holdout Method

- Split dataset into two groups
  - Training set
    - Used to train the model
    - Used to estimate the error rate of the trained model
  - Test set
  - Simplest cross validation method

Drawbacks of the holdout method

- For a sparse dataset, we may not be able to set aside a portion of the dataset for testing
- Based on the where "split" happens, the estimate of error can be misleading
- Sample might not be representative

- The limitations of the holdout can be overcome with a family of resampling methods
  - More computational expense
  - Stratified sampling
  - Cross Validation
  - Random subsampling
  - K-fold cross validation
  - Leave-one-out Cross Validation

Random Subsampling

- K data splits of the dataset
  - Each split randomly selects a (fixed) number of examples without replacement
  - For each data split, retain the classifier from scratch with the training examples and estimate E with the test examples
True Error Estimate

- The true error estimate is obtained as the **average of the separate estimates** $E_i$.

- This estimate is significantly better than the holdout estimate

$$E = \frac{1}{K} \sum_{i=1}^{K} E_i$$

---

$k$-Fold Cross-validation

- Create a $k$-fold partition of the dataset
  - For each of the $k$ experiments use $K-1$ folds for training
    - The remaining one for testing

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
<th>Experiment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total number of examples

---

Leave-one-out Cross Validation

- Leave-one-out is the degenerate case of $k$-Fold Cross validation
  - $k$ is chosen as the total number of examples
  - For a dataset with $N$ examples, perform $N$ experiments
  - Use $N-1$ examples for training, the remaining example for testing

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>...</th>
<th>Experiment $N$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total number of examples
How many folds are needed?

- The choice of the number of folds depends on the size of the dataset
  - For large datasets, even 3-Fold Cross Validation will be quite accurate
  - For very sparse datasets, you may have to consider leave-one-out
    - To get maximum number of experiments.

- A common choice for $k$-fold Cross Validation is $k=10$

Three-way data splits

- If model selection and true error estimates are computed simultaneously
  - The data needs to be divided into three disjoint sets
    - Training set
    - E.g. to find the optimal weights
    - Validation set
      - A set of examples used to tune the parameters of a model
      - To find the “optimal” number of hidden units or determine a stopping point for the back propagation algorithm
    - Test set
      - Used only to assess the performance of a fully-trained model

- After assessing the final model with the test set, you must not further tune the model.

Plain Accuracy

- Classifier accuracy
  - General measure of classifier performance

$$\text{Accuracy} = \frac{\text{(Number of correct decisions made)}}{\text{(Total number of decisions made)}}$$

- Pros
  - Very easy to measure

- Cons
  - Cannot consider realistic cases

The Confusion Matrix

- A type of contingency table
  - $n \times n$ matrix
    - The columns labeled with actual classes
    - The rows with predicted classes

- Separates out the decisions made by the classifier
  - How one class is being confused for another
  - Different sorts of errors may be dealt with separately
Problems with Unbalanced Classes

- Consider a classification problem where one class is rare
  - Sifting through a large population of normal entities to find a relatively small number of unusual ones
  - Looking for defrauded customers, or defective parts
  - The class distribution is unbalanced or skewed

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 500 & 200 \\
\bar{\text{Y}} & 0 & 300 \\
\end{array}
\]

Confusion Matrix of \( A \): evaluated with 1,000 datapoints

Confusion Matrix of \( B \): evaluated with 1,000 datapoints

Which model is better?

Why accuracy is misleading

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 500 & 0 \\
\bar{\text{Y}} & 100 & 500 \\
\end{array}
\]

Confusion Matrix of \( A \)

Confusion Matrix of \( B \)

Balanced Population

True Population

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \text{Y} & \bar{\text{Y}} \\
\hline
\text{Y} & 100 & 100 \\
\bar{\text{Y}} & 500 & 500 \\
\end{array}