Methodology: Assessment and Cross-Validation

Dan Sheldon

November 18, 2014
First story

- USPS uses a classifier to distinguish 4 from 9
- Pays $1 for every mistake
- How much money should it budget for 2015?
- **Model assessment (validation):** estimate prediction error on future unseen data (generalization)
Second story

- USPS uses regularized logistic regression to prevent overfitting in its classifier.
- What value of $\lambda$ will lead to the model with the least prediction error?
- **Model comparison (selection):** estimate prediction error for purpose of selecting the best model.
Two goals

**Model assessment**: estimate prediction error on future unseen data (generalization)

**Model comparison**: estimate prediction error for purpose of selecting the best model
Two goals

**Model assessment:** estimate prediction error on future unseen data (generalization)

**Model comparison:** estimate prediction error for purpose of selecting the best model

Can’t do either of these with data used to train the model
Data-Generating Mechanism

- Assumption: training data representative of future unseen data

- Formally, training examples and future test examples drawn independently from same probability distribution $\mathcal{P}$

$$\begin{align*}
(x^{(i)}, y^{(i)}) & \sim \mathcal{P} \\
(x, y) & \sim \mathcal{P}
\end{align*}$$

- How to think of this
  - huge bag of input-output pairs $(x, y)$ ("nature")
  - $m$ training examples pulled out randomly
  - future data drawn also pulled out randomly
  - (picture on board)
In an Ideal World

If we are “data rich”, this is what we would do:

- Validation set: labeled data reserved to compare models
- Test set: labeled data reserved to assess future performance

E.g., 50/25/25 split
In an Ideal World

If we are “data rich”, this is what we would do:

- **Validation set**: labeled data reserved to compare models
- **Test set**: labeled data reserved to assess future performance

E.g., 50/25/25 split

**Warning**: Terminology of validation/test not always consistently used
The Dilemma: Train vs. Test Size

What if you only have 100 training examples? 50? 10?

The dilemma

- More training data $\rightarrow$ more accurate classifier
- More test data $\rightarrow$ better estimate of generalization accuracy
Cross-Validation

(Assume assessment for now... how much will USPS pay?)

Beautiful and simple solution to train/test size dilemma:

- Split data in $k$ equal-sized “folds” (usually 2, 5, 10)
- For each fold, test on that fold while training on all others:

  1. Train
  2. Train
  3. Validation
  4. Train
  5. Train

- Estimate accuracy by averaging over all folds
Example

5-fold cross-validation

<table>
<thead>
<tr>
<th></th>
<th>Train folds</th>
<th>Test folds</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>12345</td>
<td>2,3,4,5</td>
<td>1</td>
<td>85%</td>
</tr>
<tr>
<td>12345</td>
<td>1,3,4,5</td>
<td>2</td>
<td>83%</td>
</tr>
<tr>
<td>12345</td>
<td>1,2,4,5</td>
<td>3</td>
<td>91%</td>
</tr>
<tr>
<td>12345</td>
<td>1,2,3,5</td>
<td>4</td>
<td>88%</td>
</tr>
<tr>
<td>12345</td>
<td>1,2,3,4</td>
<td>5</td>
<td>84%</td>
</tr>
</tbody>
</table>

Average accuracy = 88.2%
What if you need to do both model comparison and assessment?
What if you need to do both model comparison and assessment?

Fancier methods:

- One fold for validation (e.g. train/valid/test = 3/1/1)
- Nested cross-validation