Cross validation

COMS 4721

The model selection problem

Objective

▶ Often necessary to consider many different models (e.g., types of classifiers) for a given problem.
▶ Sometimes “model” simply means particular setting of hyper-parameters (e.g., \(k\) in \(k\)-NN, number of nodes in decision tree).

Terminology

The problem of choosing a good model is called model selection.

Model selection by hold-out validation

(Henceforth, use \(h\) to denote particular setting of hyper-parameters / model choice.)

Hold-out validation

Model selection:

1. Randomly split data into three sets: training, validation, and test data.

2. Train classifier \(\hat{f}_h\) on training data for different values of \(h\).
3. Compute Validation (“hold-out”) error for each \(\hat{f}_h\): \(\text{err}(\hat{f}_h, \text{Validation})\).

4. Selection: \(\hat{h} = \text{value of } h \text{ with lowest Validation error.}\)
5. Train classifier \(\hat{f}\) using \(\hat{h}\) with training and validation data.

Model assessment:

6. Finally: estimate true error rate of \(\hat{f}\) using test data.

Main idea behind hold-out validation

Classifier \(\hat{f}_h\) trained on training data \(\longrightarrow\) \(\text{err}(\hat{f}_h, \text{Validation})\).

Classifier \(\hat{f}_h\) trained on training and validation data \(\longrightarrow\) \(\text{err}(\hat{f}_h, \text{Test})\).

The hope is that these quantities are similar!

(Making this rigorous is actually rather tricky.)
Beyond simple hold-out validation

Standard hold-out validation:

| Training | Validation | Test |

Classifier \( \hat{f}_h \) trained on Training data \( \rightarrow \) err(\( \hat{f}_h \), Validation).

Could also swap roles of Validation and Training:

- train \( \hat{f}_h \) using Validation data, and
- evaluate \( \hat{f}_h \) using Training data.

Classifier \( \hat{f}_h \) trained on Validation data \( \rightarrow \) err(\( \hat{f}_h \), Training).

Idea: Do both, and average results as overall validation error rate for \( h \).

Model selection by \( K \)-fold cross validation

Model selection:
1. Set aside some test data.
2. Of remaining data, split into \( K \) parts ("folds") \( S_1, S_2, \ldots, S_K \).
3. For each value of \( h \):
   - For each \( k \in \{1, 2, \ldots, K\} \):
     - Train classifier \( \hat{f}_{h,k} \) using all \( S_i \) except \( S_k \).
     - Evaluate classifier \( \hat{f}_{h,k} \) using \( S_k \): err(\( \hat{f}_{h,k} \), \( S_k \)).

   Example: \( K = 5 \) and \( k = 4 \)

   | Training | Training | Training | Validation | Training |

   - \( K \)-fold cross-validation error rate for \( h \): \( \frac{1}{K} \sum_{k=1}^{K} \text{err}(\hat{f}_{h,k}, S_k) \).

4. Set \( \hat{h} \) to the value \( h \) with lowest \( K \)-fold cross-validation error rate.

5. Train classifier \( \hat{f} \) using selected \( \hat{h} \) with all \( S_1, S_2, \ldots, S_K \).

Model assessment:
6. Finally: estimate true error rate of \( \hat{f} \) using test data.

How to choose \( K \)?

**Argument for small \( K \)**
Better simulates "variation" between different training samples drawn from underlying distribution.

| Training | Validation |

\( K = 2 \)

| Validation | Training |

\( K = 4 \)

| Validation | Training | Training | Training |
| Training | Validation | Training | Training |
| Training | Training | Validation | Training |
| Training | Training | Training | Validation |

**Argument for large \( K \)**
Some learning algorithms exhibit phase transition behavior (e.g., output is complete rubbish until sample size sufficiently large).

Using large \( K \) best simulates training on all data (except test, of course).

In practice: usually \( K = 5 \) or \( K = 10 \).

Recap

- **Model selection**: goal is to pick best model (e.g., hyper-parameter settings) to achieve low true error.

  - **Two common methods**: hold-out validation and \( K \)-fold cross validation (with \( K = 5 \) or \( K = 10 \)).

  - **Caution**: considering too many different models can lead to overfitting, even with hold-out / cross-validation.

  (Sometimes "averaging" the models in some way can help.)