Cross-Validation

Machine Learning
Spring 2021
Model selection

Very broadly: Choosing the best model using given data

– What makes a model
  • Training data
  • Hyper-parameters
  • Loss
Cross-validation

We want to train a classifier using a given dataset

We know how to train given hyper-parameters.

How do we know what the best hyper-parameters are?

Examples of hyperparameters: C in SVM, Number of epochs in Perceptron, Kernel parameters in nonlinear SVM, ....
Question

• If the hyper-parameters are present in the cost function, can we directly optimize it? Think about SVM
K-fold cross-validation

*Given a particular feature set and hyper-parameter setting*

1. Split the data into \( K \) (say 5 or 10) equal sized parts

| Part 1 | Part 2 | Part 3 | Part 4 | Part 5 |
K-fold cross-validation

*Given a particular feature set and hyper-parameter setting*

1. Split the data into K (say 5 or 10) equal sized parts

2. Train a classifier on four parts and evaluate it on the fifth one
K-fold cross-validation

*Given a particular feature set and hyper-parameter setting*

1. Split the data into K (say 5 or 10) equal sized parts

2. Train a classifier on four parts and evaluate it on the fifth one

![Diagram showing train and evaluate process with parts 1 to 5]

Accuracy$_5$
K-fold cross-validation

Given a particular feature set and hyper-parameter setting

1. Split the data into K (say 5 or 10) equal sized parts

2. Train a classifier on four parts and evaluate it on the fifth one

3. Repeat this using each of the K parts as the validation set
K-fold cross-validation

*Given a particular feature set and hyper-parameter setting*

1. Split the data into K (say 5 or 10) equal sized parts

2. Train a classifier on K-1 parts and evaluate it on the remaining one

3. Repeat this using each of the K parts as the *validation set*

4. The quality of this hyper-parameter is the average of these K estimates
   
   Performance = \( \frac{\text{accuracy}_1 + \text{accuracy}_2 + \text{accuracy}_3 + \text{accuracy}_4 + \text{accuracy}_5}{5} \)
K-fold cross-validation

*Given a particular feature set and hyper-parameter setting*

1. Split the data into K (say 5 or 10) equal sized parts

2. Train a classifier on K-1 parts and evaluate it on the remaining one

3. Repeat this using each of the K parts as the *validation set*

4. The quality of this hyper-parameter is the average of these K estimates
   
   Performance = \( \frac{\text{accuracy}_1 + \text{accuracy}_2 + \text{accuracy}_3 + \text{accuracy}_4 + \text{accuracy}_5}{5} \)

5. Repeat for every hyper parameter choice
Cross-validation

We want to train a classifier using a given dataset
We know how to train given the hyper-parameters

How do we know what the best hyper-parameters are?
Cross-validation

We want to train a classifier using a given dataset
We know how to train given the hyper-parameters

How do we know what the best hyper-parameters are?
1. Evaluate every hyper-parameter using cross-validation (could be computationally expensive)
2. Pick the best according to cross-validation performance
3. Train on full data using this setting