Cross-Validation

Machine Learning

Model selection

Very broadly: Choosing the best model using given data

- What makes a model
  - Features
  - Hyper-parameters that control the hypothesis space
    - Example: depth of a decision tree, neural network architecture, etc.
  - The learning algorithm (which may have its own hyper-parameters)
  - Actual model itself
- The learning algorithms we see in this class only find the last one
  - What about the rest?

Model selection strategies

- Many, many different approaches out there
  - (Chapter 7 of Elements of Statistical Learning Theory)
  - Minimum description length
  - VC dimension and risk minimization
  - *Cross-validation*
  - Bayes factor, AIC, BIC, ….

Cross-validation

We want to train a classifier using a given dataset

We know how to train given features and hyper-parameters.

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

K-fold cross-validation

Given a particular feature set and hyper-parameter setting

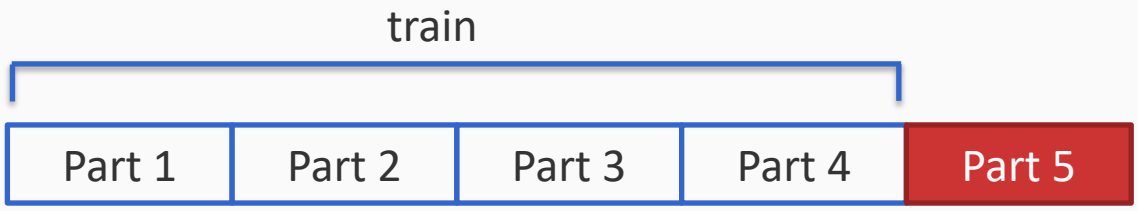
1. Split the data into K (say 5 or 10) equal sized parts

K-fold cross-validation

Given a particular feature set and hyper-parameter setting

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

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

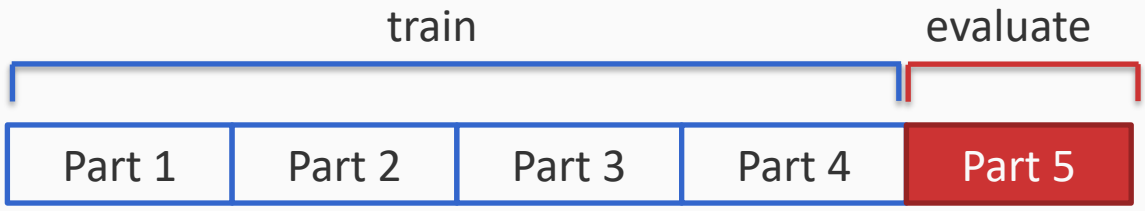


K-fold cross-validation

Given a particular feature set and hyper-parameter setting

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

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



$Accuracy_5$

K-fold cross-validation

Given a particular feature set and hyper-parameter setting

1. Split the data randomly 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*

Part 1	Part 2	Part 3	Part 4	**Part 5**	$Accuracy_5$
Part 1	Part 2	Part 3	**Part 4**	Part 5	$Accuracy_4$
Part 1	Part 2	**Part 3**	Part 4	Part 5	$Accuracy_3$
Part 1	**Part 2**	Part 3	Part 4	Part 5	$Accuracy_2$
Part 1	Part 2	Part 3	Part 4	Part 5	$Accuracy_1$

K-fold cross-validation

Given a particular feature set and hyper-parameter setting

1. Split the data randomly 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*

4. The quality of this feature set/hyper-parameter is the average of these K estimates

Performance = $(\text{accuracy}_1 + \text{accuracy}_2 + \text{accuracy}_3 + \text{accuracy}_4 + \text{accuracy}_5)/5$

K-fold cross-validation

Given a particular feature set and hyper-parameter setting

1. Split the data randomly 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*

4. The quality of this feature set/hyper-parameter is the average of these K estimates

$$\text{Performance} = (\text{accuracy}_1 + \text{accuracy}_2 + \text{accuracy}_3 + \text{accuracy}_4 + \text{accuracy}_5)/5$$

5. Repeat for every feature set/hyper parameter choice

Cross-validation

We want to train a classifier using a given dataset

We know how to train given features and hyper-parameters

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

Cross-validation

We want to train a classifier using a given dataset

We know how to train given features and hyper-parameters

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

1. Evaluate every feature set and 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