Math 407

Model Selection and Assessment

Chapter 5 in ISL, Chapter 7 in ESLII

Overview of the process of choose a model:

1. Select a few methods to try that are appropriate for the type of data you have and your (client's) needs.
2. Train models using the methods from 1. on a set of data (training set)
3. Choose the model with the best performance on a different set of data (test set 1)
4. Get an unbiased estimate of your chosen model's performance using a third set of data (test set 2)

Overview of the process of choose a model:

1. Select a few methods to try that are appropriate for the type of data you have and your (client's) needs.
2. Train models using the methods from 1. on a set of data (training set)
3. Choose the model with the best performance on a different set of data (test set) *This step is called "model selection"*
4. Get an unbiased estimate of your chosen model's performance using a third set of data (validation)

This step is called "model assessment"

How should we proceed?

- If **lots** of data is available, split it into three parts – training, test and validation.
- If a limited amount of data is available, use a **resampling** method
  - K-fold cross-validation
  - Bootstrap

Resampling methods

- Can be used for **model selection** or **model assessment**

- Can be *computationally intense* (could take days or *weeks* to run if you have a large dataset and complex method)

- The method of **Cross Validation** is very popular for estimating a model's performance
- **Bootstrapping** is a more general tool for estimating "anything" about your model, including it's performance

Leave One Out Cross Validation

Suppose we are trying to estimate a **model's performance** and we have a dataset of **n** data points (n observations of **X** and **Y**)

1. Train the model using all the data points except one

(training set size = n-1)

2. Test the model using the left out data point

(test set size = 1)

*Do steps 1 and 2 a total of **n times**, each time leaving out a different point to be the test set.*

Average over the prediction errors from step 2 to get an estimate of MSE, MAE, or the misclassification rate, etc.

10-fold Cross Validation

Suppose we are trying to estimate a **model's performance** and we have a dataset of **n** data points (n observations of **X** and **Y**)

1. Train the model using 90% of the data points

(training set size = 0.9n)

2. Test the model using the remaining 10% of the data points

(test set size = 0.1n)

*Do steps 1 and 2 a total of **10 times**, each time leaving out a different 10% of the data as the test set.*

Average over the prediction errors from step 2 to get an estimate of MSE, MAE, or the misclassification rate, etc.

k-fold Cross Validation

Suppose we are trying to estimate a **model's performance** and we have a dataset of **n** data points (n observations of **X** and **Y**)

1. Train the model using 100(1-1/k)% of the data points

   (training set size = n-n/k)

2. Test the model using the remaining (100/k)% of the data points

   (test set size = n/k)

Do steps 1 and 2 a total of ***k times****, each time leaving out different fold of the data as the test set.*

Average over the prediction errors from step 2 to get an estimate of MSE, MAE, or the misclassification rate, etc.

How should we choose k=number of folds?

5 fold CV – test set size is 20% of n, MSE ≈ average over 5 errors

10 fold CV – test set size is 10% of n, MSE ≈ average over 10 errors

20 fold CV – test set size is 5% of n, MSE ≈ average over 20 errors

LOO CV – test set size is 1, MSE ≈ average over n errors

How does the choice of k affect your estimate of MSE?

How should we choose k=number of folds?

5 fold CV – test set size is 20% of n, MSE ≈ average over 5 errors
10 fold CV – test set size is 10% of n, MSE ≈ average over 10 errors
20 fold CV – test set size is 5% of n, MSE ≈ average over 20 errors
LOO CV – test set size is 1, MSE ≈ average over n errors

How does the choice of k affect your estimate of MSE?

- Averaging over more errors means more stability so a *lower variance*
- Having a larger test size, means a better estimate of the error so *less bias*

The choice of k means you are choosing the tradeoff between bias and variance in the estimate of the MSE of the MSE!!!

Reporting your model’s performance

- Keep in mind that whether you use CV or a single test set, the MSE or misclassification rate you obtain is only an **estimate** of your model’s performance.

- For the bank note data, we found kNNs with k=10 had an overall misclassification rate of 6.6% on the test set of 137 bills.

- It would be more informative to know the misclassification rate of the model on **all bank notes.**

Some terminology from statistics

Population = *all objects for which you are interested in predicting Y*

Sample = *a subset of the population*

- For the bank note data, we found kNNs with k=10 had an overall misclassification rate of 6.6% on the test set of 137 bills.

a sample

- It would be more informative to know the misclassification rate of the model on **all bank notes**.

The population of interest

Some more terminology from statistics

Parameter = *a numerical summary of your* ***population***

Statistic = *a numerical summary of your* ***sample***

- For the bank note data, we found kNNs with k=10 had an overall **misclassification rate of 6.6%** on the test set of 137 bills.

a statistic

- It would be more informative to know the **misclassification rate** of the model on all bank notes.

The parameter (generally unknown)

Inferential Statistics

Inferential Statistics = inferring something about the population from a sample

What are useful tools from inferential statistics for machine learning?

- Estimation techniques (bootstrap, confidence intervals, etc)
- Sample size calculations
- Sampling methods

Inferential Statistics

Inferential Statistics = inferring something about the population from a sample

What are useful tools from inferential statistics for machine learning?

- Estimation techniques (bootstrap, confidence intervals, etc)

Ex: estimate the population MSE based on test set MSE

- Sample size calculations

Ex: how large a test set do I need? Training set?

- Sampling methods

Ex: how can I collect data?

how should I divide my data into test and training sets?

what limitations does the way my data was collected place on when my model should be used?

Estimating a parameter from a sample

For the bank note data, we found kNNs with k=10 had an overall misclassification rate of 6.6% on the test set of 137 bills.

So our single best guess of the *misclassification rate of our model on all bills* is 6.6%.

One option: we could compute the statistic on the sample and use it as our best guess of the value of the parameter

Estimating a parameter from a sample

For the bank note data, we found kNNs with k=10 had an overall misclassification rate of 6.6% on the test set of 137 bills.

So I am 95% confident that the *misclassification rate of our model on all bills* is between 2.4% and 10.8%.

A better option: we could compute a range of plausible values for the parameter given the sample

Confidence Intervals

A **95% confidence interval for the parameter $\boldsymbol{\alpha}$** is an interval [L, U] computed from a sample so that $P(L< \boldsymbol{\alpha} <U) = 0.95$

95% CI for a population proportion

IF

a sample was collected following a Binomial Process and has at least 10 successes and 10 failures,

THEN

a 95% CI for a population proportion π is given by

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the sample proportion computed from a sample of size n

95% CI for a population proportion

If

we assume the bills in our test set are ***independent***, and there are at least 10 genuine bills and 10 fake bills in the test set,

THEN

a 95% CI for *misclassification rate of our model on all bills* is given by

$$0.066 \pm 1.96\sqrt{\frac{0.066(1-0.066)}{137}}$$

where $\hat{p}$ = 0.066 is the proportion of bills misclassified in our test set of size n = 137.

Interpretation: I am 95% confident that the *misclassification rate of our model on all bills* is between 2.4% and 10.8%.

How large should my test set be?

Suppose we want to know the population misclassification rate to within 1% with 95% confidence.

This means we want the half-width of the CI to be 0.01:

$$0.01 = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Solve for n?

How large should my test set be?

Suppose we want to know the population misclassification rate to within 1% with 95% confidence.

This means we want the half-width of the CI to be 0.01:

$$0.01 = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Worse case is that $\hat{p} = 0.5$. Solving for n in

$$0.01 = 1.96\sqrt{\frac{1}{4n}}$$

yields n = $\frac{1.96^2}{4(0.01^2)}$ *= 9604*

In order to estimate the misclassification rate to within 1%, I need to have at least 9604 bills in my test set.

Sample size calculations:
How should large should my test and training sets be?

1. Calculate the test set size by specifying the degree of accuracy you want to have in estimating the misclassification rate.
2. Put the rest of the data in your training set.
   - Some methods have rules of thumb: e.g. need at least 10 datapoints per predictor variable in regression
   - Some methods have sample size formulas you can compute. Many methods don’t.