Statistical Machine Learning

Day 39

Review: Training, Selecting and Assessing a Model; Reporting

Steps in a Machine Learning Project

1. Reframe a prediction question in terms of math and statistics
2. Find, clean and transform an appropriate dataset
3. Understand the dataset (visualize, summarize)
4. Train, select and assess a prediction model
5. Report results

Types of Learning

- **Unsupervised Learning**

  dataset only has examples of the predictors X, no examples of Y

  Example: K-means

- **Semi-supervised Learning**

  dataset has some examples of both X and Y, some examples with only X

  Example: K-means with majority vote

- **Supervised Learning**

  all examples in the dataset have both X and Y

  Examples: LDA, QDA, Bayes Classifier, Linear, Logistic, GLMs, ridge, LASSO, Splines, SVM, Neural Networks, Trees, kNN

- **Reinforcement Learning**

  given a set of constraints, find a course of actions that maximizes a reward

Step 4: Training, selecting and assessing

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

This step is called "model assessment"

Types of Supervised Learning

Discriminative Models

Learn a **boundary** between classes by either

- Estimating P(Y|X) ("probabilistic algorithm")

or

- Assuming the boundary takes a specific form ("non-probabilistic algorithm")

Generative Models

Model the entire data-generating process, both the X's and Y's, P(Y, X)

The additional structure works well when the dataset is relatively small and/or is close to the truth data-generating process

Classify each method as **Discriminative** or **Generative**: if Discriminative, classify it as a **probabilistic** or **non-probabilistic** algorithm

LDA,

QDA,

Bayes Classifier,

Linear,

Logistic,

GLMs,

Ridge,

LASSO,

Splines,

SVM,

Neural Networks,

Trees,

kNN

Rank the methods by their **flexibility** (i.e. restrictive assumptions about X and Y)

LDA,

QDA,

Bayes Classifier,

Linear,

Logistic,

GLMs,

Ridge,

LASSO,

Splines,

SVM,

Neural Networks,

Trees,

kNN

Rank the methods by their **interpretability,** that is, how easy it is to see how X is connected to the predicted Y

LDA,

QDA,

Bayes Classifier,

Linear,

Logistic,

GLMs,

Ridge,

LASSO,

Splines,

SVM,

Neural Networks,

Trees,

kNN

Step 4: Training, selecting and assessing

a. Choose a few methods to try that are appropriate for the type of data you have and your (client's) needs.

b. Train models using the methods from a. on a set of data (training set)

c. Select the model with the best performance on a different set of data (test set) *This step is called "model selection"*

d. Assess the model's actual performance with an unbiased estimate obtained from a third set of data (validation)

This step is called "model assessment"

Which methods train by **plugging the data into a formula**?

Which methods train by using a **numerical algorithm** like Newton's method or gradient descent to maximize a measure of prediction quality?

LDA,

QDA,

Bayes Classifier,

Linear,

Logistic,

GLMs,

Ridge,

LASSO,

Splines,

SVM,

Neural Networks,

Trees,

kNN

Which methods should I try for the following prediction questions?

How old is someone?

How many pairs of shoes does someone own?

Does a person have heart disease?

Is there a stop light in front of my car? What does the signal say?

Step 4: Training, selecting and assessing

a. Choose a few methods to try that are appropriate for the type of data you have and your (client's) needs.
b. **Train models using the methods from a. on a set of data (training set)**
c. **Select the model with the best performance on a different set of data (test set)** ***This step is called "model selection"***
d. **Assess the model's actual performance with an unbiased estimate obtained from 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 such as **K-fold cross-validation** (K = 5 or 10 is common)

- Can be used for **model selection** and/or **model assessment**
- Can be *computationally intense so* K = 5 or 10 is common

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 example, with 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.**

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.