hw_02

lisa liubovich

2024-05-28

1. Ex. 2.4.7. KNN Prediction

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

obs.	X_1	X_2	X_3	Y	Distance to X = (1,1,1)	Neighbor for k = 1	Neighbor for k = 3
1	0	3	0	Red	$\sqrt{6}$		
2	2	0	0	Red	$\sqrt{3}$		
3	0	1	3	Red	$\sqrt{5}$		
4	0	1	2	Green	$\sqrt{2}$		
5	-1	0	1	Green	$\sqrt{5}$		
6	1	1	1	Red	0		

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 1$ using K-nearest neighbors. Do this manually (no code) and show your work.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 1$.

for observation 1: square root of: $(0\text{-}1)^2 + (3\text{-}1)^2 + (0\text{-}1)^2 = \text{-}1^2 + 2^2 + 1^2 = 1 + 4 + 1 = 6 \rightarrow \sqrt{6}$

for observation 2: square root of: $(2\text{-}1)^2 + (0\text{-}1)^2 + (0\text{-}1)^2 = 1^2 + \text{-}1^2 + \text{-}1^2 = 1 + 1 + 1 = 3 \rightarrow \sqrt{3}$

for observation 3: square root of: $(0\text{-}1)^2 + (1\text{-}1)^2 + (3\text{-}1)^2 = \text{-}1^2 + 0^2 + 2^2 = 1 + 0 + 4 = 5 \rightarrow \sqrt{5}$

for observation 4: square root of: $(0\text{-}1)^2 + (1\text{-}1)^2 + (2\text{-}1)^2 = \text{-}1^2 + 0^2 + 1^2 = 1 + 0 + 1 = 2 \rightarrow \sqrt{2}$

for observation 5: square root of: $(\text{-}1\ \text{-}1)^2 + (0\text{-}1)^2 + (1\text{-}1)^2 = \text{-}2^2 + \text{-}1^2 + 0 = 4 + 1 + 1 = 5 \rightarrow \sqrt{5}$

for observation 5: square root of: $(1\text{-}1)^2 + (1\text{-}1)^2 + (1\text{-}1)^2 = 0 + 0 + 0 = 0$

(b) What is our prediction with $k = 1$? Why?

When k = 1, our prediction is Red. This is because the nearest neighbor to X1 = X2 = X3 =1 is observation 6, which has a distance of 0 between each observation and the test point, and the color of this observation is Red.

(c) What is our prediction with $k = 3$? Why?

When k = 3, our prediction is Red. This is because the three nearest neighbors to X1 = X2 = X3 = 1 are observations 2, 4, and 6, which are $\sqrt{3}$, $\sqrt{2}$, and 0 respectively. Their respective classifications are Red, Green, and Red. Red then wins by popular majority.

(d) If the decision boundary in this problem is highly nonlinear, would we expect the best value for k to be large or small? Why?

If the decision boundary is highly nonlinear, this indicates a flexible model which is associated with a small value of k.

In classification problems, decision boundaries separate different response categories in the X space of independent variables.

2 Ex. 4.8.13 (part) Stock Market Prediction, part 1. KNN

We want to predict the behavior of the stock market in the following week. The Weekly data set, (from the {ISLR2} package), contains 1,089 observations with the following 9 variables.

- Year: The year that the observation was recorded
- Lag1: Percentage return for previous week
- Lag2: Percentage return for 2 weeks previous
- Lag3: Percentage return for 3 weeks previous
- Lag4: Percentage return for 4 weeks previous
- Lag5: Percentage return for 5 weeks previous
- Volume: Volume of shares traded (average number of daily shares traded in billions)
- Today: Percentage return for this week
- Direction: A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week

(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)

# data(Weekly)

# summary(Weekly)
# ggpairs(Weekly)
```

(b) Split the data set into training and testing data and use the KNN method with $K = 9$ to predict Direction as the response based on the five lag variables plus Volume as predictors. Use split percentage of 70%. Why or why not is this a good split percentage? Use seed 1235.

(c) Compute the confusion matrix, showing cross-tabulation of the actual and predicted responses.

(d) Compute the classification rate, which is the overall fraction of **correct** predictions.

(e) Tuning. Try different values of the tuning parameter K and select the optimal one; the one which minimizes the prediction *error* rate. Report the optimal K and the associated **error rate**.

3 Ex. 4.8.4. Curse of Dimensionality

When the number of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are *near* the test observation for which a prediction must be made. This phenomenon is known as the *curse of dimensionality*, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this curse.

(a) Suppose we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly (evenly) distributed on [0,1]. Associated with each observation is a response value. We wish to predict a test observation's response using only observations that are **within 10%** of the range of X closest to that test observation. For instance, to predict the response for a test observation with $X = 0.6$, we will use observations in the range [0.55,0.65].

- On average, what fraction of the available observations will we use to make the prediction?

(b) Now suppose we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1,X_2) are uniformly distributed on [0,1] $\times$ [0,1] and are independent. We wish to predict a test observation's response using only observations that are within 10% of the range of X_1 **and** within 10% of the range of X_2 closest to that test observation. For instance, to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range [0.55,0.65] for X_1 and in the range [0.3,0.4] for X_2.

- On average, what fraction of the available observations will we use to make the prediction?

(c) Now suppose we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation.

- What fraction of the available observations will we use to make the prediction?

(d) Using your answers to parts (a) - (c), argue that a drawback of KNN when p is large is there are very few training observations "near" any given test observation.

(e) We wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations (for our variables that are i.i.d. Uniform (0, 1)). For $p = 1$, 2, and 100, What is the length of each side of the hypercube? Comment on your answer. *Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$, a hypercube is simply a line segment, when $p = 2$ it is a square, and when $p = 100$ it is a 100-dimensional cube.*

4 Ex. 4.8.9. What are the Odds?

(a) On average, what fraction of people with an odds of 0.33 of defaulting on their credit card payment will in fact default?

(b) Suppose an individual has a 10% chance of defaulting on her credit card payment. What are the odds they will default?

5 Ex. 4.8.6. What does it take to get an A?

Suppose we collect data on a group of students in a undergraduate statistics class using the following variables:

- X_1 : hours studied
- X_2 : undergrad GPA, and
- Y : receive an A (1) or Not an A (0).

We fit a logistic regression and produce estimated coefficients,

- β_0 = -6, β_1 = 0.05, and β_2 = 1.

(a) Estimate the probability a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

(b) How many hours would the student in part (a) need to study to have a 50% (predicted) chance of getting an A in the class?

6 Ex. 4.8.13.(part) Stock Market Prediction, Part 2. Logistic Regression

We want to predict the behavior of the stock market in the following week. The Weekly data set, (from the {ISLR2} package), contains 1,089 observations with the following 9 variables.

- Year: The year that the observation was recorded
- Lag1: Percentage return for previous week
- Lag2: Percentage return for 2 weeks previous
- Lag3: Percentage return for 3 weeks previous
- Lag4: Percentage return for 4 weeks previous
- Lag5: Percentage return for 5 weeks previous
- Volume: Volume of shares traded (average number of daily shares traded in billions)

• Today: Percentage return for this week

• Direction: A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week

(a) Perform a logistic regression with Directionas the response and the five lag variables plus Volume as predictors. Do any of the predictors appear to be statistically significant? If so, which ones?

(b) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by the logistic regression model.

(c) Fit a logistic regression model with Lag2 as the only predictor using as training data the period from 1990 to 2008. Compute the confusion matrix and the overall fraction of correct predictions for the held out test data (that is the data from 2009, and 2010). How does it compare to before? What does that suggest?

(d) Plot an ROC curve for the logistic regression on the test data from (c), using different probability thresholds and add an diagonal line for FPR = TPR. Interpret the plot in terms of how useful the model might be?

7 Ex. 4.8.5, LDA v QDA

(a) If the Bayes decision boundary is *linear*, do we expect LDA or QDA to perform better on the training set? On the test set?

(b) Compare the expected performance of LDA and QDA on the training set and then on the test set if the Bayes decision boundary is *non-linear*.

(c) In general, as the sample size n increases, do we expect the *test prediction accuracy* of QDA relative to LDA to improve, decline, or be unchanged? Why?

(d) **True or False**: If the Bayes decision boundary for a given problem is linear, we will probably achieve a superior *test error rate* using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. **Justify your answer.**

8 Ex. 4.8.7. Non-uniform Prior. Predicting Issuance of a Stock Dividend

Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover:

• The mean value of percent profit X for companies that issued a dividend was $\bar{x}_1 = 10$, while the mean for those that didn't was $\bar{x}_2 = 0$.

• The variance of X for these two sets of companies was $\sigma^2 = 36$.

• 80% of companies issued dividends.

Assuming that X follows a Normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Hint: Recall that the density function for a Normal random variable is f(x) = $1/\ \sigma\sqrt{2}\pi\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. You will need to use the Bayes theorem. Also, use R function dnorm(x, mean, sd) to compute the pdf ($f(x)$) of Normal distribution.

We can also use the probability calculator in R.

9 Ex. 4.8.13.(part) Stock Market Prediction, Part 3. LDA, QDA and Summary

We want to continue with trying to predict the behavior of the stock market in the following week. We computed a KNN estimate and a Logistic regression estimate already. We will now look at LDA and QDA using the {MASS} package and compare across these prediction methods.

The Weekly data set, (from the {ISLR2} package), contains 1,089 observations with the following 9 variables.

- Year: The year that the observation was recorded
- Lag1: Percentage return for previous week
- Lag2: Percentage return for 2 weeks previous
- Lag3: Percentage return for 3 weeks previous
- Lag4: Percentage return for 4 weeks previous
- Lag5: Percentage return for 5 weeks previous
- Volume: Volume of shares traded (average number of daily shares traded in billions)
- Today: Percentage return for this week
- Direction: A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week

(a) Use LDA with a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of **correct predictions** for the held-out data (that is, the data from 2009 and 2010).

(b) Repeat (a) using QDA.

(c) Using the results from the previous questions (Stock Market Prediction. Part 1, 2), compare the correct classification rates on the testing data obatined from the 4 methods: KNN, Logistic regression, LDA and QDA. Recommend one or more methods. Explain your rationale.