Midterm III

2022-12-12

Your name:

Questions

Q1: Logistic Regression for Classification

We are interested in predicting the diabetes status of patients depending on their Plasma glucose concentration using the `PimaIndiansDiabetes2` dataset from the R package `mlbench`. The diabetes status is stored in the variable `diabetes` and the Plasma glucose concentration is stored in the variable `glucose`. Given below are the data preparation steps.

```
set.seed(123)
data(PimaIndiansDiabetes2)
db <- PimaIndiansDiabetes2 %>% drop_na() %>% select(glucose, diabetes)
db_single <- db %>% select(diabetes, glucose) %>%
  mutate(diabetes = fct_relevel(diabetes, ref = "neg"))
glimpse(db_single)
## Rows: 392
## Columns: 2
## $ diabetes <fct> neg, pos, pos, pos, pos, pos, pos, neg, pos, neg, pos, pos, n~
## $ glucose  <dbl> 89, 137, 78, 197, 189, 166, 118, 103, 115, 126, 143, 125, 97,~
```

```
db_split <- initial_split(db_single, prop = 0.75)
# Create training data
db_train <- db_split %>% training()
# Create testing data
db_test <- db_split %>% testing()
```

(a) What is the reference level of the factor `diabetes`? How many observations are in the test and train datasets? *Answer:* 98 and 294, respectively.

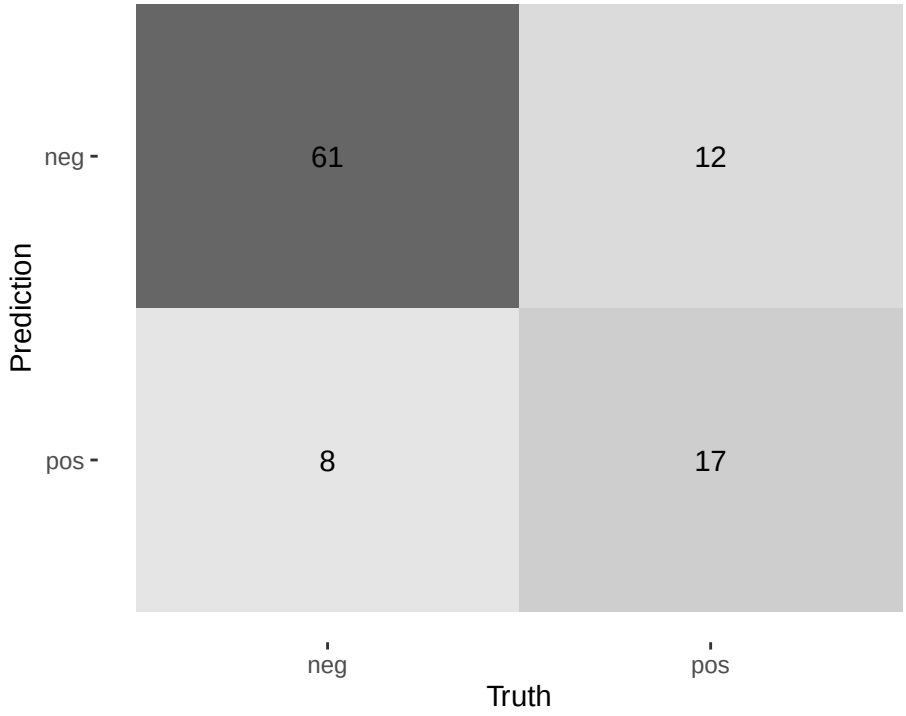
We fit a logistic regression model to the training dataset to predict the diabetes status using Plasma glucose level. The summary of the logistic regression model is given below.

```
tidy(fitted_logistic_model)
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -6.30     0.759       -8.30 1.06e-16
## 2 glucose       0.0443   0.00575      7.71 1.22e-14
```

(b). For what Plasma glucose level, the probability of having diabetes is 1/2? *Answer:*

(c). What is the odds of getting diabetes, if one has a Plasma glucose level of 150? *Answer:*

(d). Now, lets predict the diabetes status of patients in the test set using our model fitted using the training set and classify a patient as positive if the predicted probability is at least 0.5, and negative otherwise. Answer the following questions using the resulting confusion



matrix.

(i) Calculate the accuracy of this classifier.

(ii) Calculate the specificity of this classifier.

(iii) Calculate the sensitivity of this classifier.

(e). Which one of False Positive (FP) or False Negative (FN) is more detrimental in this example? You may assume positive diabetes as a positive case in answering this question.

Q2 Clustering using k-NN

The dataset for this question contains data on 400 students regarding their admission status, GRE score, GPA and rank stored under the variables `admit`, `gre`, `gpa`, and `rank`, respectively. We would like to use k-nearest neighbor algorithm to cluster the students.

```
gpa_data <- read_csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/GPA.csv")
gpa_data <- gpa_data %>% mutate(admit = as.factor(admit),
                                admit = fct_relevel(admit, ref ="Yes"))
glimpse(gpa_data)
## Rows: 400
## Columns: 4
## $ admit <fct> No, Yes, Yes, Yes, No, Yes, Yes, No, Yes, No, No, No, Yes, No, Y~
## $ gre   <dbl> 380, 660, 800, 640, 520, 760, 560, 400, 540, 700, 800, 440, 760,~
## $ gpa   <dbl> 3.61, 3.67, 4.00, 3.19, 2.93, 3.00, 2.98, 3.08, 3.39, 3.92, 4.00~
## $ rank  <dbl> 3, 3, 1, 4, 4, 2, 1, 2, 3, 2, 4, 1, 1, 2, 1, 3, 4, 3, 2, 1, 3, 2~
```

(a). Why do we need to split the data into training and testing set? Explain. *Answer:*

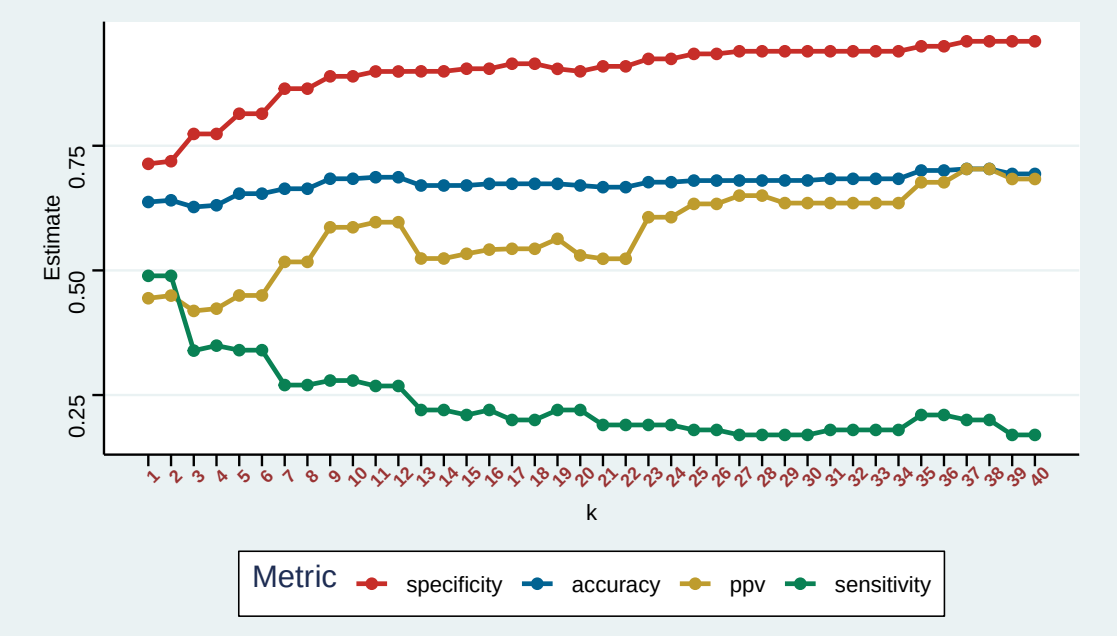
(b). Why do we need to standardize data before fitting a k-NN model? *Answer:* Standardization is required in k-NN because the calculation of Euclidean distance is influenced by the feature that has more variability.

```
gpa_recipe <- recipe(admit ~. , data = gpa_train) %>%
  step_scale(all_predictors()) %>% # scale the predictors
  step_center(all_predictors()) %>% # center the predictors
  prep

gpa_knn_spec <- nearest_neighbor(mode = "classification",
                                 engine = "kknn",
                                 weight_func = "rectangular",
                                 neighbors = tune())
```

(c). Briefly explain why do we need to do cross validation when fitting a machine learning model. *Answer:*

Let's do a 10-fold cross validation across a grid of number of neighbors. The following is a plot of how the various metrics change when we vary `k`, the number of neighbors. A corresponding table with the optimal number of neighbors for the different metric is also given.



neighbors	metric
37	Accuracy
1	Sensitivity
37	Specificity
37	PPV

(d). What optimal number of neighbors would you pick based on the plot and the table? Explain your reasoning. *Answer:*

(e). Why do you think the sensitivity and the specificity grow in opposite pattern as the number of neighbors increases? *Answer:*

The sensitivity and specificity of the k-Nearest Neighbor algorithm grow in opposing patterns as the number of neighbors increases because the algorithm uses a voting system to determine the classification of a data point. As the number of neighbors increases, it becomes increasingly likely that the majority vote will be correct, leading to higher sensitivity. However, due to the increased number of neighbors, the algorithm is also more likely to be misled by noisy data points, leading to lower specificity.

Q4 Miscellaneous

(a) The k-NN method of classification will yield different accuracies as the value of k changes. As k approaches n, the sample size, what value will the accuracy of the method approach? Explain. *Answer:*

(b) Which of the following can act as possible termination conditions in k-Means? Explain.

1. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
2. Centroids do not change between successive iterations.
3. All of the above

Answer:

(c) Explain how you can use total within cluster sum of squares to find the "best" choice of k in a k-means clustering algorithm. *Answer:*

Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.

(d) (True/False) A model suffering from overfitting will most likely have high bias. *Answer:*

(e) (Multiple Choice) Given the following models trained using k-NN, the model which could result in overfitting will most likely have the value of k as

1. 2
2. 10
3. 20

Answer: