Module 3 Assignment on Classification

Jingwen Zhong // Graduate Student

3/4/2021

Module Assignment Questions

Q1) (***Bayes Classifier***)

`Bayes classifier` classifies an observation x_0 to the class k for which $p_k(x_0)$ is largest, where π_k is prior (proportion of k class in all classes over j):

$$p_k(x_0) = P(y = k|X = x_0) = \frac{\pi_k \cdot f_k(x_0)}{\sum \pi_j \cdot f_j(x_0)}.$$

Assume univariate (p=1) observation x in class k is iid from $N(\mu_k, \sigma_k^2)$, $f_k(x)$ is the density function of x with parameters μ_k, σ_k.

a. Show that the Bayes classifier in 2-class problem (so $k = 0, 1$) assigns the observation x_0 to the class k for which the discriminant score δ is largest when $\sigma_0 = \sigma_1$:

$$\delta_k(x_0) = x_0 \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Proof:

$$f_k(x_0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2}$$

Then

$$p_k(x_0) = \frac{\pi_k \cdot f_k(x_0)}{\sum \pi_j \cdot f_j(x_0)} = \frac{\pi_k \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2}}{\sum \pi_j \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2}}$$

$$log(p_k(x_0)) = log(\frac{\pi_k \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu_k)^2}}{\sum \pi_j \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu_j)^2}})$$

$$= log(\pi_k \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu_k)^2}) - log(\sum \pi_j \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu_j)^2})$$

$$= log(\pi_k) + log(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2\sigma^2}(x-\mu_k)^2 - log(\sum \pi_j \cdot e^{-\frac{1}{2\sigma^2}(x-\mu_j)^2}) + log(\frac{1}{\sqrt{2\pi}})$$

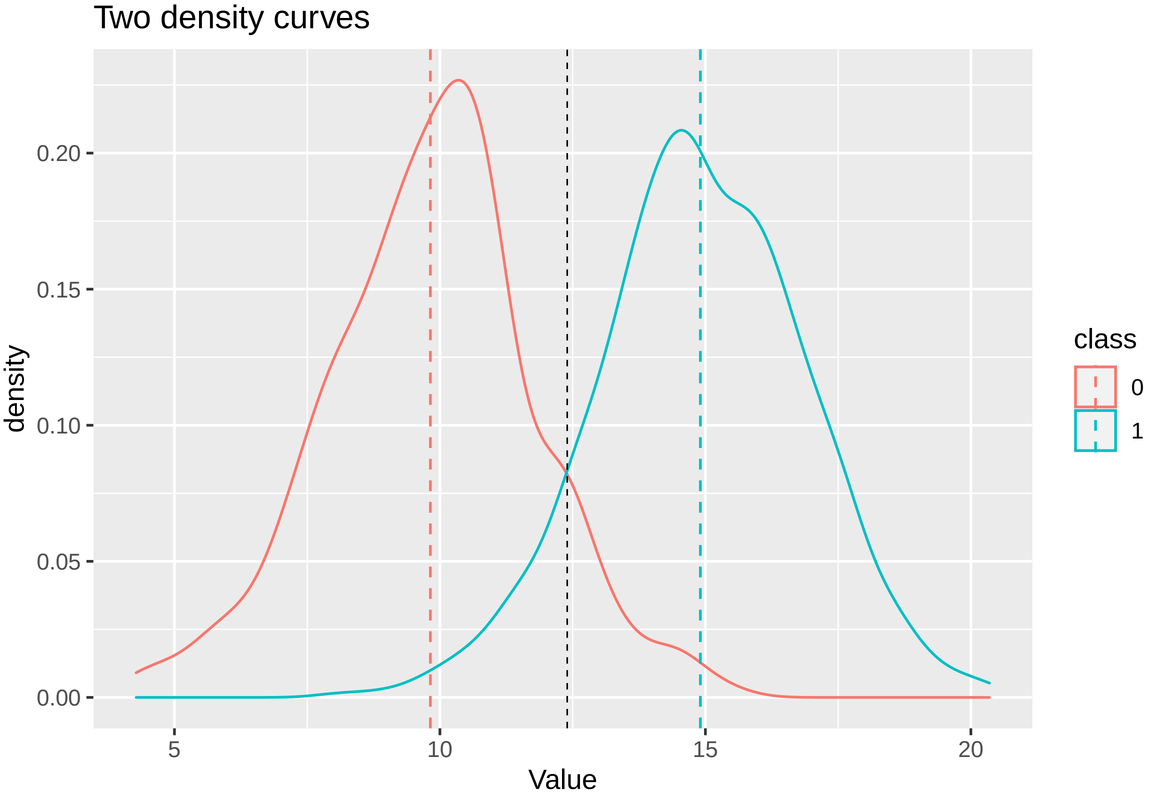
$$= log(\pi_k) - \frac{1}{2\sigma^2}(x^2 + \mu_k^2 - 2x\mu_k) - log(\sum \pi_j \cdot e^{-\frac{1}{2\sigma^2}(x^2+\mu_j^2-2x\mu_j)})$$

$$= log(\pi_k) - \frac{1}{2\sigma^2}(\mu_k^2 - 2x\mu_k) - log(\sum \pi_j \cdot e^{-\frac{1}{2\sigma^2}(\mu_j^2 - 2x\mu_j)})$$

Since the objective is to maximize $log(p_k(x_0))$, and $log(\sum \pi_j \cdot e^{-\frac{1}{2\sigma^2}(\mu_j^2-2x\mu_j)})$ do not depend on k, so we can remove it and then we obtain the discriminant score $log(\pi_k) - \frac{\mu_k^2}{2\sigma^2} + \frac{x_0\mu_k}{\sigma^2}$.

---

b. (Empirical Work) Verify `part a` with a simple empirical demonstration using normal densities in R with `dnorm()` or generated normal variables from `rnorm()` with $\mu_0 = 10, \mu_1 = 15, \sigma_0 = \sigma_1 = 2, \pi_0 = 0.3, \pi_1 = .7, mu2 = 15, pi2 = 0.7$. Plot the class densities or histograms in color, show the intersection between two class distributions (where the classification boundary starts), check one random value from each class by calculating the discriminant score so to confirm the class it belongs. How would you describe the misclassified values or regions? Calculate the error rate. What is the Bayes error rate?



```
##
## the intersection of this 2 classes is 12.39726
```

```
##
## A random number in class0 is: 7.749854
```

```
##
## discriminant score when assign x0 to class0 is: 5.670662
```

```
##
## discriminant score when assign x0 to class1 is: 0.5802766
## 7.749854 belongs to class0
```

```
##
## A random number in class1 is: 16.91708
```

```
##
## discriminant score when assign x0 to class0 is: 28.58873
```

```
##
## discriminant score when assign x0 to class1 is: 34.95738
```

```
## 16.91708 belongs to class1
```

```
## Bayes error rate is: 0.089
```

I descirbe misclassified values or regions as error.

---

c. Under `part a`, assume $\sigma_0 \neq \sigma_1$. Derive the Bayes classifier. Show work. Σ_k is a covariance matrix for the kth class

$$\sigma_k(x) = log(\pi_k) - \frac{\mu_k^2}{2\Sigma_k} - \frac{1}{2}log|\Sigma_k|$$

*** d. (BONUS) For p>1, derive the the Bayes classifier. Show work.

Q2) (*Four Models as Classifiers*)

The `Boston` data from `MASS` has 506 rows and 14 columns with the target variable `crim`, which is per capita crime rate by town. You will fit classification models (`KNN`, `logistic regression`, `LDA`, and `QDA`) in order to predict whether a given suburb has a crime rate above or below .3 per capita crime rate by town. Upper .3 may be labeled as `not really safe town` to raise a family. Use `80%-20% split validation test` approach.

```
## Warning in rm(Boston): object 'Boston' not found
```

a. Fit the `KNN`, `logistic regression`, `LDA`, and `QDA` models separately using all the predictors. Report `error rate` for each `train` and `test` data set. Use `error rate` = `1-accuracy`. Based on the test error rates, decide which model is best/better. Why?

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Table 1: The error rate of the train/test using 4 methods

	train error	test error
KNN	0.0000	0.0412
Logistic Regression	0.0244	0.0309
LDA	0.0489	0.0619
QDA	0.0293	0.0412

Based on the test error rates, Logistic regression model is best/better, because it has smallest test error rate.

---

b. Using the test data set, obtain confusion matrices and report only `recall`, `precision`, `f1` and `accuracy` metrics in a table. Comment on the findings. Based on this table, decide which model is best/better. Explain why some models do better than others. Which metric would be most important in this context? Why? Is this decision different from that of `part a`? Explain.

Table 2: The reports of confusion matrices

	Accuracy	Recall	Precision	F1
KNN	95.88	87.5	100	93.33
Logistic Regression	96.91	93.75	96.77	95.24
LDA	93.81	81.25	100	89.66
QDA	95.88	90.62	96.67	93.55

Based on this table, Logistic Regression model is best/better, since it has highest F1 score.

When the true decision boundaries are linear, then the LDA and logistic regression approaches will tend to perform well. When the boundaries are moderately non-linear, QDA may give better results. For much more complicated decision boundaries, a non-parametric approach such as KNN can be superior.

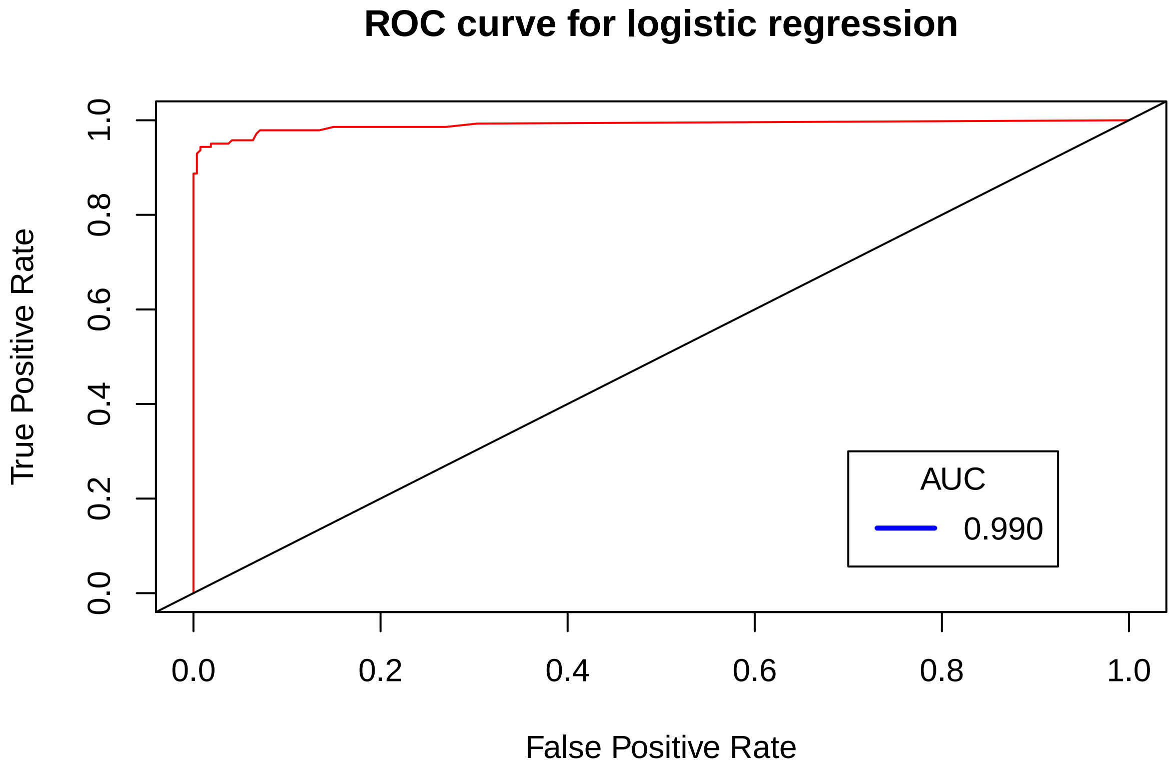
F1 is the most important in this context. Precision refers to how many of the data selected as positive by our algorithm are really positive. Recall refers to how many of the data that should actually be Positive are selected by us as Positive, and in this case, I can select everything as positive and get high recall rate but it won't help. F1 Score combines precision and recall, and it should be the best matrix. (Accuracy: It refers to how much of all the data we have correctly classified. However, if we only correctly classify no crime without making a judgment on the crime, we can also get a high accuracy rate, but it is not a good model for crime.)

This decision is not different from that of `part a`

---

c. Obtain the ROC curve for `logistic regression` based on train data set (plot of FPR vs TPR with classification threshold change). Plot it. Calculate the area under the curve. Explain what the curve and area tell about the model.

```
## plot:
```



An ideal ROC curve will hug the top left corner, and the larger area under the ROC curve the better the classifier, above graph tells that this logistic regression model works very well

---

d. How did you find the optimal k in the `KNN` classifier? Did you use `grid search` or `CV`? If not, use it and revise the results in part a and b. Did the results improve? I use the grid search to find the optimal k in knn:

```
## the best k is: 9
```

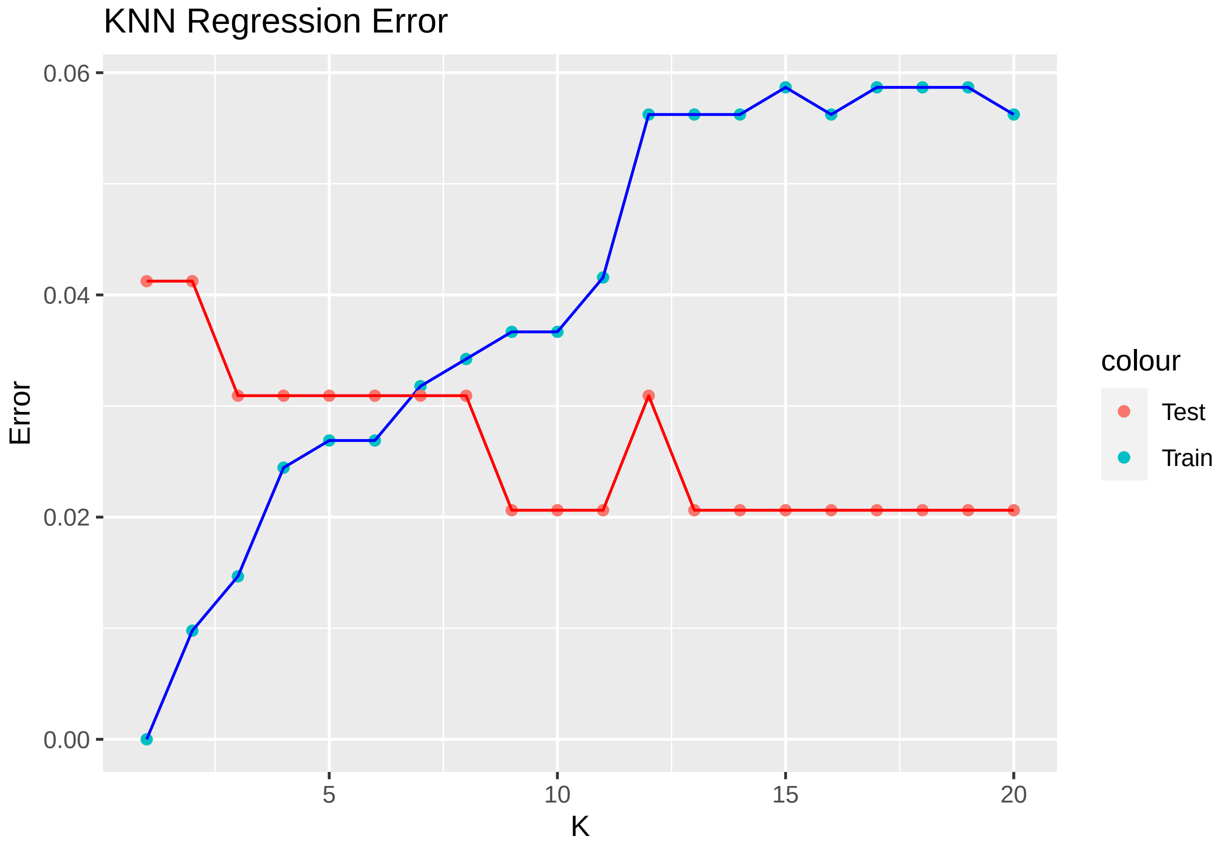


Table 3: update a): The error rate of the train/test using 4 methods

	train error	test error
old KNN	0.0000	0.0412
KNN	0.0367	0.0206
Logistic Regression	0.0244	0.0309
LDA	0.0489	0.0619
QDA	0.0293	0.0412

Table 4: update b): The reports of confusion matrices

	Accuracy	Recall	Precision	F1
KNN	97.94	93.75	100	96.77
Logistic Regression	96.91	93.75	96.77	95.24
LDA	93.81	81.25	100	89.66
QDA	95.88	90.62	96.67	93.55

The results improves, the test error decreased, and KNN is the best model based on the error rate.

e. What are the assumptions in each model? Do your best to describe each. Do your best to check these based on the fit. When you see assumption violation, what would you do to validate the fit?

- Assumptions for KNN:

  KNN is an non parametric lazy learning algorithm. For non parametric method, it means that it does not make any assumptions on the underlying data distribution.

- Assumptions for Logistic Regression:

  1. The observations are independent of each other
  2. Little or no multicollinearity among the independent variables
  3. The independent variables are linearly related to the log odds.
  4. A large sample size

- Assumptions for LDA:

  1. Equality of covariances among the predictor variables X across each all levels of Y
  2. Predictor variables X are drawn from a multivariate Gaussian (aka normal) distribution
  3. The number of predictor variables (p) to be less then the sample size (n)

- Assumptions for QDA:

  1. Predictor variables X are drawn from a multivariate Gaussian (aka normal) distribution
  2. The number of predictor variables (p) to be less then the sample size (n)
  3. Predictors shouldn't be highly correlated

No violation. When I see assumption violation, I can perform standardization to validate the fit.

Q3) (*Concepts*)

a. What would change if you perform `$k-$fold approach` instead of `validation set` approach for the model fits in Question 2? Just discuss conceptually. The k-fold CV estimate will be computed by averaging the error rate, the error rate will change, and it might be more percise.

---

b. To improve the test error rates in `part a` Q2, what strategies can be applied: list the ideas as many as possible. Try one of them and report the improved test error rate.

- increase the number of observations(n)
- select right features
- mix algorithms
- ensemble methods(bagging/boosting)
- tuning parameters
- cross validation

cross validation:

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Table 5: The error rate of the train/test using 4 methods

	train error	test error
KNN	0.0494	0.0220
Logistic Regression	0.0241	0.0328
LDA	0.0428	0.0503
QDA	0.0356	0.0356

---

c. Explain with less technical terms an estimation method employed in `binary logistic regression`. `MLE` and `gradient descent` are two of them. Gradient descent:

Suppose we are somewhere on a large mountain. Since we don't know how to go down the mountain, we decide to walk step by step, and every time we reach a position, we solve the gradient of the current position and follow the negative direction of the gradient. Going this way step by step, until we feel that we have reached the foot of the mountain (the gradient of that position is 0). However if we go on like this, it is possible that we don't go to the foot of the mountain, and stop at the certain lower part of the mountain.

---

d. (BONUS) Demonstrate with technical terms and numerically how `MLE` as estimation method employed in `binary logistic regression` works. Explain. The idea of maximum likelihood estimation: Ignore low probability, consider high probability events as true events, or use high probability to estimate true events

In binary logistic regression we define the prediction function returns a probability score between 0 and 1, and suppose our threshold was k:

$$p \leq k, class = 0$$

$$p \geq k, class = 1$$

which means when the probablity of the observation greater that k, we classify it as True(class = 1). When the probablity of the observation less that k, we classify it as False(class = 0)

I hereby write and submit my solutions without violating the academic honesty and integrity. If not, I accept the consequences.

How long did the assignment solutions take?: 15+ hrs

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Retrieved 1st Mar. 2021 from https://towardsdatascience.com/beginners-guide-to-k-nearest-neighbors-in-r-from-zero-to-hero-d92cd4074bdb

Retrieved 1st Mar. 2021 from https://towardsdatascience.com/all-the-annoying-assumptions-31b55df246c3

Retrieved 1st Mar. 2021 from https://www.statisticssolutions.com/assumptions-of-logistic-regression/

Retrieved 1st Mar. 2021 from http://uc-r.github.io/discriminant_analysis#linear

Retrieved 1st Mar. 2021 from https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/