STAT 154: Study Guide/Practice Problems for Midterm 2

Johnny Hong

November 8, 2017

1 True/False

Indicate True or False for each of the following statements. Justify your answers.

1. In logistic regression, there are closed-form solutions for the maximum likelihood estimates of the parameters.
2. Tree-based methods can only be applied to classification problems.
3. For a two-class classificaiton problem, the (theoretical) AUC of a random classifier (a classifier that assigns the class label at random) is 0.5.
4. QDA is always better than LDA, since QDA allows for more flexible decision boundaries.
5. k-NN yields linear decision boundaries.
6. Logistic regression requires a Gaussian assumption on the predictors.
7. In (probabilistic) LDA, the prior probabilities are typically chosen via cross-validation.
8. True positive rate and false positive rate must add up to 1.

2 Concept questions

1. Explain the similarities and differences between LDA and QDA from the probabilistic modeling standpoint.
2. Explain why logistic regression is more appropriate than linear regression in a two-class classification problem.
3. Describe an advantage and a disadvantage of decision trees over LDA.

3 Mathematical Questions

1. Let $x_{ik} \in \mathbb{R}$ be the value of the ith observation in class k for $i = 1, ..., n_k$, $j = 1, ..., p$, and $k = 1, ..., K$.

$$TSS = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x})^2, \tag{1}$$

$$BSS = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2, \tag{2}$$

and

$$WSS = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2. \tag{3}$$

(a) Prove that $TSS = BSS + WSS$.

(b) Recall that the correlation ratio is defined as follows:

$$\eta^2 = \frac{BSS}{TSS}. \tag{4}$$

Use part (a) to prove that $0 \leq \eta^2 \leq 1$. When does $\eta^2 = 0$? When does $\eta^2 = 1$? What is the interpretation in each case?

(c) Recall that the F ratio is defined as

$$F = \frac{BSS/(K-1)}{WSS/(n-K)}, \tag{5}$$

where $n = n_1 + \cdots + n_K$. Show that there is a one-to-one relationship between the F ratio and the correlation ratio.

4 Coding questions

1. Write an R function called `performance_metrics()` that takes a confusion matrix as an input and returns a vector of length six, containing true positive rate, true negative rate, false positive rate, false negative rate, specificity and sensitivity. Assume that the confusion matrix is a 2×2 matrix of the following form:

	Actual $Y = 1$	Actual $Y = 0$
Predicted $Y = 1$	a	b
Predicted $Y = 0$	c	d

where a, b, c, d are nonnegative integers.