Quiz #5 – Solutions

1. If we have enough data to partition the dataset into training, validation, and test samples, which one of the following classification models is most likely to be the best when applied to the test sample?

a) A model with 18% training error, 22% validation error, and 75% sensitivity.
b) A model with 17% training error, 20% validation error, and 74% sensitivity.
c) A model with 21% training error, 21% validation error, and 75% sensitivity.
d) A model with 19% training error, 20% validation error, and 75% sensitivity.

Solution: D

The best model is the one with the lowest validation error and highest sensitivity. Better sensitivity is preferable to lower training error. We can make more complex models to better fit our training data, but they do not perform well in any other data.

2. Suppose we have 100 individuals, some of whom have type 2 diabetes. We use a classification model based on health and lifestyle profiles of the individuals to estimate the probability that each individual is diabetic. If we use a 0.5 probability cut-off, the model predicts that 60 individuals are diabetic, of whom 54 individuals actually are diabetic, while of the other 40 individuals, only 5 are diabetic. Which of the following are possible results that could result from increasing the cut-off probability to 0.6? (Select all that apply.)

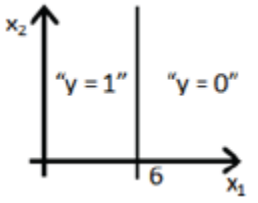
a) 47 actual diabetics out of 49 predicted diabetics; 12 actual diabetics out of 51 predicted non-diabetics.
b) 50 actual diabetics out of 53 predicted diabetics; 2 actual diabetics out of 47 predicted non-diabetics.
c) 55 actual diabetics out of 67 predicted diabetics; 6 actual diabetics out of 33 predicted non-diabetics.
d) 58 actual diabetics out of 69 predicted diabetics; 1 actual diabetic out of 31 predicted non-diabetics.

Solution: A

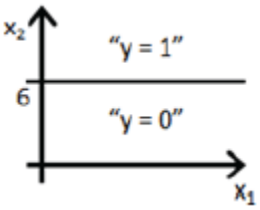
The number of diabetics in the sample is fixed at 54+5=59 and the number of non-diabetics is fixed at 100-59=41. This rules out the second and third possibilities, which have 50+8=58 and 55+6=61 diabetics, respectively. Increasing the probability cut-off means that fewer individuals' posterior probabilities will exceed this threshold, and so fewer than 60 individuals will be predicted diabetics. This rules out the fourth possibility, which has 69 predicted diabetics. The first possibility satisfies both these criteria (47+12=59 actual diabetics, 49 predicted diabetics) and is the only correct possibility.

3. Suppose you train a logistic regression classifier, $p(x) = 1/(1+\exp(-\beta_0-\beta_1 x_1-\beta_2 x_2))$. Suppose $\beta_0=-6$, $\beta_1=0$, $\beta_2=1$. Which of the following figures represents the decision boundary found by your classifier? (Note that the letters for the answers are at the bottom-left of the figures.)

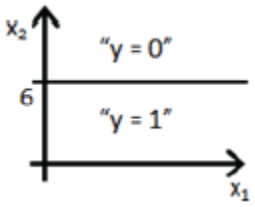
a)



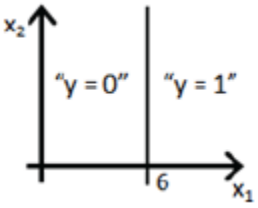
b)



c)



d)



Solution: B

Since $\beta_0=-6$, $\beta_1=0$, $\beta_2=1$ and the classifier is $p(x) = 1/(1+\exp(6-0x_1-1x_2)) = 1/(1+\exp(6-x_2))$. Thus, only x_2 matters and the decision boundary must be horizontal. Since this classifier is an increasing function of x_2, the "y=1" region must be at the top of the graph.

4. Which of the following are correct statements? (Select all that apply).

a) Logistic regression and linear discriminant analysis are similar but logistic regression relies on fewer assumptions.
b) Linear and quadratic discriminant analysis assume Gaussian distributions for the predictor variables.
c) Linear discriminant analysis assumes each class has the same covariance matrix.
d) Quadratic discriminant analysis assumes each class has a different covariance matrix.

Solution: A, B, C, D

5. Arrange the following steps in a *k*-nearest neighbors analysis of a classification problem in the most appropriate order (assume *k* is fixed):

a) For each point to be classified, find the *k* closest training observations and classify using the most common class.
b) For each point to be classified, calculate the distance from each training observation using an appropriate distance metric.
c) Standardize the predictor/feature variables if appropriate to do so.

Solution: A:3, B:2, C:1

The steps are standardize, calculate distances, classify.

6. Which of the following are true statements about nearest neighbors classifiers? (Select all that apply.)

a) Nearest neighbors classifiers work best in applications in which the decision boundaries are highly regular and can be well approximated by intersecting hyperplanes.
b) Increasing the number of neighbors, *k*, increases the smoothness of the decision boundary found by a nearest neighbors classifier.
c) Since nearest neighbors classifiers require no model to be fit, the computational load is relatively light compared to model-based methods.

Solution: B

Increasing *k* makes for a smoother decision boundary in a *k*-nearest neighbors analysis. However, nearest neighbors classifiers do not work best in applications in which the decision boundary is highly regular and can be well approximated by intersecting hyperplanes – this is the kind of situation in which *k*-nearest neighbors do poorly relative to linear parametric methods. The other statement about computational load is also false because *k*-nearest neighbors is actually very resource intensive because of the number of calculations that have to be done and the storage requirements.

Use for Questions 7 – 9:

Consider using a logistic regression model to predict whether a potential customer will respond to an offer that costs $1 to mail and is expected to yield $5 revenue if the contact responds. A pilot study has been conducted in which offers were made to people on whom three predictor variables were also measured: R (recency), F (frequency), M (monetary). The resulting logistic regression model estimated from the training data leads to the equation:

Probability of response = $1/(1+\exp(0.5–0.2R–0.1F–0.3M))$

Six customers in the validation sample are then sorted in order of the probability of a response, and it is possible to calculate how expected responses and net profits change as the number of offers increases, both for sending offers using the ordered customers from the model and for sending offers at random. Note that there were a total of 3 responses from 6 offers in this sample, and total net profit from making all 6 offers is 3×5–6×1=$9 (remember it costs $1 to mail and is expected to yield $5 revenue if the contact responds). So, for example, making 1 offer to the top person on the list yields 1 response, whereas making 1 offer at random yields (3/6)=0.5 expected responses. Net profit for making 1 offer to the top person on the list yields 1×5–1×1=$4, whereas making 1 offer at random yields an expected profit of (1/6)×9=$1.50. The next three questions will ask you to complete the following entries in the table for the 5th customer in the validation sample.

R	F	M	Prob. of Response	Cumulative Offers	Respond (1: Yes, 0: No)	Model: Responses	Random: Expected Responses	Model: Net Profit	Random: Expected Net Profit
1	2	3	0.690	1	1	1	0.5	4	1.5
2	3	1	0.622	2	1	2	1.0	8	3.0
2	2	1	0.599	3	0	2	1.5	7	4.5
1	2	0	0.475	4	1	3	2.0	11	6.0
0	0	1		5	0	3	2.5		
0	0	0	0.378	6	0	3	3.0	9	9.0

7. What is the "Probability of Response" for the 5th customer? (Round to two decimal places.)

Solution: 0.45

Probability of response = 1/(1+exp(0.5–0.2R–0.1F–0.3M)) = 1/(1+exp(0.5–0.2(0)–0.1(0)–0.3(1))) = 1/(1+exp(0.2)) = 0.45.

8. What is the "Model: Net Profit" for the 5th customer? (Round to nearest whole number.)

Solution: 10

3 x 5 – 5 x 1 = 10

9. What is the "Random: Expected Net Profit" for the 5th customer? (Round to one decimal place.)

Solution: 7.5

(5 / 6) x 9 = 7.5