Ch. 4 Classification - Problem Bank Questions

August 24, 2020

1. In 1846, the Donner party (Donner and Reed families) left Springfield, Illinois for California in covered wagons. After reaching Fort Bridger, Wyoming, the leaders decided to find a new route to Sacramento. They became stranded in the eastern Sierra Nevada mountains at a place now called Donner Pass (right) when the region was hit by heavy snows in late October. By the time the survivors were rescued on April 21, 1847, 40 out of 87 had died.

(a) What confounding variables might be present that could explain the significance of the sex indicator variable?

(b) From the Donner Party data, the log odds of survival were estimated to be 1.6 - 0.078 age+ 1.6 I_{female}, based on a binary response that takes the value 1 if an individual survived and I(female) is an indicator variable that takes the value 1 for females.

i. What would be the estimated equation for the log-odds of survival if the indicator variable for sex were 1 for males and 0 for females?

ii. What would be the estimated equation for the log-odds of perishing if the binary response were 1 for a person who perished and 0 for a person who survived?

(c) Given $\log\frac{p(x)}{1-p(x)} = B_0 + B_1x$, solve for p(x).

(d) What are the estimated probabilities of survival for men and women of ages 25 and 50?

(e) What is the age at which the estimated probability of survival is 50% for women and for men?

(f) Interpret 1.6, -0.078 and the second 1.6 in context.

2. Are these the possible reasons for selecting a metric other than misclassification error for evaluating a classifier. If yes, why using misclassification error would be a problem?

(a) The data set is imbalanced.

(b) False positives are much worse than false negatives for the given application.

(c) False negatives are much worse than false positives for the given application.

3. Suppose you perform a logistic regression of predicting whether an image is a cat ($Y = 1$) or a dog ($Y = 0$) based on the proportion of the image that is occupied by the animal (X, assume continuous between 0 and 1). The output of the regression is given below:

Coefficient	Estimate	Std. Error	Z value	p-value
(Intercept)	-4.3	1.1	-4	<2e-16
X	0.56	0.03	7.8	<2e-16

(a) What is the standard error of $\hat{\beta}_0$?
(b) What is the standard error of $\hat{\beta}_1$?
(c) What is the interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$?
(d) What is the conclusion about a hypothesis test for the value of $\hat{\beta}_1$?
(e) Given a new picture with $X = 0.77$, what is the probability that this image is a cat?

4. (a) Suppose you have a classification problem with more than two classes. How might you apply logistic regression (using only two classes at a time) to this data? What might be some advantages and/or drawbacks of such an approach?

(b) Suppose you are performing classification and your training set is very unbalanced: that is, 90% of the training set has $Y = 1$. What are some possible issues that could arise in this setting?

(c) Give an example that predicting a false negative is much more expensive than predicting a false positive.

(d) In some cases, it can be extremely important to look at a confusion matrix rather than just looking at the classification accuracy and/or error. Think of a scenario where this is the case.

5. (a) Why is logistic regression preferable to a linear regression where negative values become zero and values greater than 1 become 1?

(b) How can high collinearity "flip" the effect (sign of the coefficient) for a variable in multiple logistic regression (example in slide 11 of classification notes)

(c) Slide 18 of classification notes mentions that only K-1 linear functions are needed to solve for the probabilities of K classes. Explain why this is the case.

6. Bone density in adults is well known to be positively linked to both weight and activity level. Weight and activity level tend to be negatively linked.

(a) When you perform a simple regression to predict the risk for low bone density based on activity level, you see activity level has no effect on bone density. This directly goes against what you learning in Bone Density 101. Given what we know about the relationship between activity level and weight, justify why this might be the case.

(b) If you didn't know the relationship between weight and activity level, how could you figure this out?

(c) How could you change your regression to control for the relationship between activity level and weight?

(d) While performing a similar analysis, you notice that people who work more than 50 hours tend to have a higher risk for low bone density when compared to people who work less than 30 hours, even when controlling for weight and activity level. What could be the cause of this? (Hint: this question does not have one right answer and is just about making reasonable predictions).

7. While working at social media company InstaBook, you are tasked with determining how country of origin impacts the likelihood of deactivating your account. To figure this out, your lead gave you information about 1000 users - 500 of which deactivated their account and 500 who did not. You know that account deactivations are very rare.

(a) You are worried that because your sample dataset does not accurately reflect the overall population of account deactivations, your analysis may have problems. What can you accurately predict despite the imbalance?

(b) What would you not be able to accurately predict?

(c) If you knew the true proportion of users that deactivate their account, how could you potentially fix the faulty values you identified in (b)?

8. Suppose that we take a dataset, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use a logistic regression and get a training error of 20% and testing error of 30%. Next we use 1-nearest neighbors (K=1) and get an average error rate (averaged over both test and training datasets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

9. This problem has to do with odds.
a. On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?
b. Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?