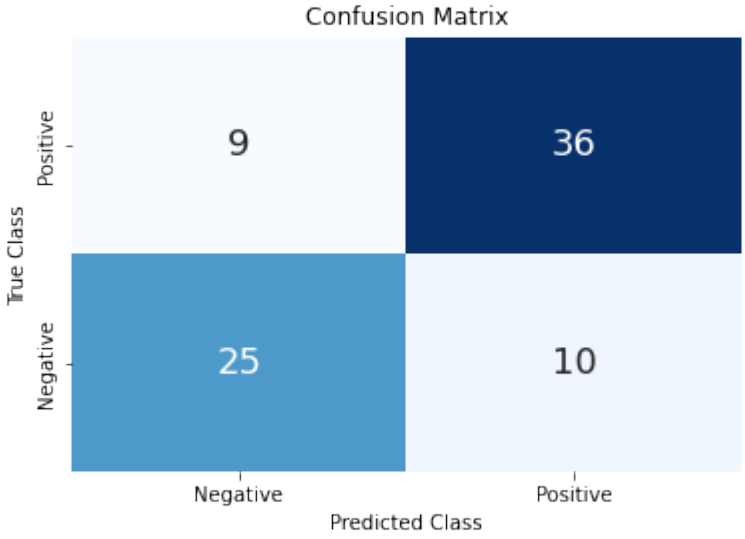
ECE M148 Homework 4 Solutions
Introduction to Data Science Due: May 10, 12:00 p.m
Instructor: Lara Dolecek TAs: Harish G V, Jayanth Shreekumar

1. Suppose we have the following confusion matrix outputted from a logistic regression using the probability threshold $P(Y = Positive) \geq t$, i.e. we classify the sample as Positive if $P(Y = Positive)$ is greater than t otherwise we classify as Negative.



(a) Compute the false positive and false negative rates.

(b) How would you expect the confusion matrix to change if we increased t?

Solution:

(a)

$$\text{False Positive Rate} = \frac{\text{Number of Negative Classes predicted as Positive}}{\text{Number of True Negative Classes}}$$
$$= \frac{10}{35} \approx 0.2857$$

$$\text{False Negative Rate} = \frac{\text{Number of Positive Classes predicted as Negative}}{\text{Number of True Positive Classes}}$$
$$= \frac{9}{45} = 0.2$$

(b) As t increase, it reduces the number of samples that are classified as Positive. Thus, the right column of the confusion matrix would decrease and the left column would increase.

2. *Bayes Theorem.* Consider that you own a small restaurant. You have a smoke detector in your kitchen. The chances that a hazardous fire occurs in the kitchen is pretty rare, say 1%. The smoke alarm is pretty accurate in detecting such fire and it sounds the alarm 99% of the time. However, the alarm is poorly calibrated and it also sounds an alarm sometimes when there is no fire, due to smoke detected from cooking. The accuracy of the smoke alarm under non-fire condition is 90%.

   (a) What is the probability that the smoke detector sounds an alarm?

   (b) Given that you heard the alarm sound, what is the probability that there was actually a fire?

   (c) Comment on how useful the smoke detector is and would you consider replacing it?

   Solution:

   (a) Let S be the event that the smoke detector sounds an alarm and F be the event that there actually was a fire.

$$\begin{aligned} P(S) &= P(S|F)P(F) + P(S|F^c)P(F^c) \\ &= 0.99 \cdot 0.01 + 0.1 \cdot 0.99 = 0.1089 \end{aligned}$$

   (b)

$$\begin{aligned} P(F|S) &= \frac{P(S|F)P(F)}{P(S)} \\ &= \frac{0.99 \cdot 0.01}{0.1089} \approx 0.0909 \end{aligned}$$

   (c) This example shows us that although the accuracy of the smoke detector seems pretty good on the first look, having an accuracy of 99% when there is a fire and 90% when there is no fire, in reality when the alarm sounds, there is only a 9% chance that there really was a fire. This makes it practically unusable and needs to be replaced or recalibrated.

3. Logistic regression is minimizing the following cross-entropy loss function:

$$L(\beta) = -\sum_{i=1}^{n} y_i \log(\frac{1}{1+e^{-(\beta^T x_i)}}) + (1-y_i)\log(1-\frac{1}{1+e^{-(\beta^T x_i)}})$$

where β is a vector of parameters, n is the number of samples,x_i is a k dimensional data sample, and $y_i \in \{0,1\}$ is a binary variable that represents the class of sample i.

Logistic regression is generally solved using iterative methods. One such method is the gradient descent method where we start with random values for $\{\beta_j^1 : 1 \leq j \leq k\}$ and we update them using the gradient rule

$$\beta_j^{t+1} = \beta_j^t - \eta \frac{\partial(\beta)}{\partial \beta_j^t}$$

for all j such that $1 \leq j \leq k$ where η is the step-size.

Prove that

$$\frac{\partial(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} (\frac{1}{1+e^{-(\beta^T x_i)}} - y_i) x_i^j$$

where x_i^j is the jth element of the ith sample.

Solution: Note that we can re-write the loss function as

$$\begin{aligned}
L(\beta) &= -\sum_{i=1}^{n} y_i \log(\frac{1}{1+e^{-(\beta^T x_i)}}) + (1-y_i)\log(1-\frac{1}{1+e^{-(\beta^T x_i)}}) \\
&= -\sum_{i=1}^{n} y_i \log(\frac{1}{1+e^{-(\beta^T x_i)}}) + (1-y_i)\log(\frac{e^{-(\beta^T x_i)}}{1+e^{-(\beta^T x_i)}}) \\
&= -\sum_{i=1}^{n} y_i \log(\frac{1}{1+e^{-(\beta^T x_i)}}) + (1-y_i)\log(\frac{1}{1+e^{(\beta^T x_i)}}) \\
&= \sum_{i=1}^{n} y_i \log(1+e^{-(\beta^T x_i)}) + (1-y_i)\log(1+e^{(\beta^T x_i)})
\end{aligned}$$

Now, we perform the derivative using the standard rules of derivatives:

$$
\begin{aligned}
\frac{\partial(\beta)}{\partial\beta_j} &= \sum_{i=1}^{n} y_i \frac{\partial \log(1+e^{-(\beta^T x_i)})}{\partial\beta_j} + (1-y_i)\frac{\partial \log(1+e^{(\beta^T x_i)})}{\partial\beta_j} \\
&= \sum_{i=1}^{n} -y_i \frac{e^{-(\beta^T x_i)} x_i^j}{1+e^{-(\beta^T x_i)}} + (1-y_i)\frac{e^{(\beta^T x_i)} x_i^j}{1+e^{(\beta^T x_i)}} \\
&= \sum_{i=1}^{n} -y_i \frac{e^{-(\beta^T x_i)} x_i^j}{1+e^{-(\beta^T x_i)}} + (1-y_i)(\frac{e^{(\beta^T x_i)} x_i^j}{1+e^{(\beta^T x_i)}} \times \frac{e^{-(\beta^T x_i)}}{e^{-(\beta^T x_i)}}) \\
&= \sum_{i=1}^{n} -y_i \frac{e^{-(\beta^T x_i)} x_i^j}{1+e^{-(\beta^T x_i)}} + (1-y_i)\frac{x_i^j}{1+e^{-(\beta^T x_i)}} \\
&= \sum_{i=1}^{n} -y_i x_1^j \frac{1+e^{-(\beta^T x_i)}}{1+e^{-(\beta^T x_i)}} + \frac{x_i^j}{1+e^{-(\beta^T x_i)}} \\
&= \sum_{i=1}^{n} -y_i x_1^j + \frac{x_i^j}{1+e^{-(\beta^T x_i)}} \\
&= \sum_{i=1}^{n} (\frac{1}{1+e^{-(\beta^T x_i)}} - y_i) x_i^j
\end{aligned}
$$

4. In your own words, explain the following types of multi-class classification methods:

   (a) One vs All

   (b) All vs All

   Provide the advantages and disadvantages of each method.

   Solution:

   Assume that we have K classes.

   (a) One vs All

   For each class, we make a binary classifier that determines whether the data point belongs to this class or not. Each classifier will provide a score for that class and we select the class that has the highest score. This creates K classifiers.

   Advantages:

   - Scales well with the number of classes

   Disadvantages:

   - Can result in an imbalance of data for each classifier. For example, if there are 3 classes that make up $\frac{1}{3}$ of the data, then each classifier is trained with $\frac{1}{3}$ positive samples and $\frac{2}{3}$ negative samples.

   (b) All vs All

   For every pair of classes, we create a binary classifier that specifies which among these two classes is more likely. We select the class that has the highest score/votes among the binary classifiers. This creates $K(K-1)/2$ binary classifiers.

   Advantages:

   - Avoids the issue of imbalance in the data

   Disadvantages:

   - The number of classifiers grows quadratically with the number of classes which can get very large
   - Reduces the number of data points used in each classifier

5. True or False questions. For each statement, decide whether the statement is True or False and provide justification (full credit for the correct justification).

   (a) For a classification model, positive predictive value is the probability that a model classifies a sample as positive given that the true label of the sample is positive.

   (b) Assume we are working with a multinomial logistic regression such that $P(Y = i|X) = e^{\beta_{0,i}+\beta_{1,i}X}P(Y = K|X)$ for $1 \leq i \leq K-1$. For a dataset with 1 feature and 4 possible class labels, the number of learnable parameters $\beta_{j,i}$ is 8.

   (c) If the log-odds function is modeled as a quadratic, logistic regression can provide a non-linear decision boundary.

   (d) You are building a classifier to detect fraudulent credit card transactions. Your employer states that a 90% success in detection of fraudulent transactions is good enough. You test your model on the next 1000 transactions and get a 97% test accuracy. Therefore, your model is doing much better than what is required.

   (e) For a very good classification model, we expect the confusion table to be dominated by diagonal entries.

   Solution:

   (a) **False**. Positive predictive value measures how confident we are that a predicted positive label is actually positive.

   (b) **False**. The answer is 6 because we only need to learn the parameters for $P(Y = i|X), 1 \leq i \leq K-1$ since $P(Y = K|X)$ can be inferred from the others. Thus, there are $3 \times 2 = 6$ parameters.

   (c) **True**. Since we classify based on the value of the log-odds function using a threshold, a quadratic log-odds function would result in a quadratic decision boundary.

   (d) **False**. This is a classic case of class imbalance. The number of fraudulent transactions is very very small compared to valid transactions. As such, a prediction of "valid transaction" always will also yield a very good result. Accuracy is not a good metric to measure the usefulness of the model in this case.

   (e) **True**. The diagonal entries represent the points from a particular class correctly classified as that class. All other entries represent misclassification of some kind.