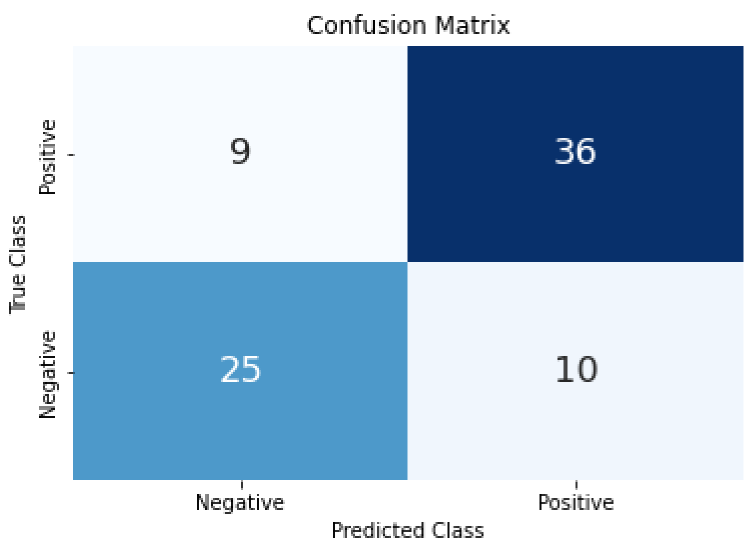
ECE M148 | Homework 4
Introduction to Data Science | Due: May 10, 12:00 p.m
Instructor: Lara Dolecek | TAs: Harish G V, Jayanth Shreekumar

Please upload your homework to Gradescope by May 10, 12:00 p.m.
You can access Gradescope directly or using the link provided on BruinLearn.
You may type your homework or scan your handwritten version. Make sure all the work is discernible.

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?

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?

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

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

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

Prove that

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

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

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.

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.