HW4-605840292

May 9, 2023

1 Question 1

1.1 (a)

[4]:

```
FPR = 10/(10+25)
print("False Positive Rate is", round(FPR,2))
```

```
False Positive Rate is 0.29
```

[5]:

```
FNR = 9/(9+36)
print("False Negative Rate is", round(FNR,2))
```

```
False Negative Rate is 0.2
```

[7]:

```
25+9+46-59
```

[7]:

```
21
```

1.2 (b)

If we increase the probability threshold t in logistic regression, the model becomes **more conservative in making positive predictions**. As a result, fewer observations are classified as positive, which means that the false positive rate (FPR) **tends to** decreases and the false negative rate (FNR) tends to increases.

2 Question 2

2.1 (a)

[12]:

```
Pro_fire = 0.01*0.99
Pro_no_fre = 0.99*(1-0.9)
total_prob = Pro_fire+Pro_no_fre
print("the probability that the smoke detector sounds an alarm is␣
 ↪approximately", round(total_prob,2))
```

```
the probability that the smoke detector sounds an alarm is approximately 0.11
```

2.2 (b)

A: Fire occurrence

B: Alarm sounds

P(B|A) is the probability that the alarm sounds given that there is a fire, which is 0.99.

P(A) is the probability of a fire occurring, which is 0.01.

P(B) is the probability that the alarm sounds, which is 0.11.

[16]:

```
P_B_A = 0.99
P_A = 0.01
P_B = total_prob
P_A_B = (P_B_A*P_A)/P_B
print("Given that you heard the alarm sound, the probability that there was␣
 ↪actually a fire is", round(P_A_B,2))
```

```
Given that you heard the alarm sound, the probability that there was actually a
fire is 0.09
```

2.3 (c)

Detection of hazardous fire: The smoke detector has a high accuracy rate of 99% in detecting hazardous fires. This indicates that it is reliable in alerting you to potential fire incidents, which is crucial for the safety of the kitchen and the restaurant.

False alarms due to cooking smoke: The smoke detector has a 10% false alarm rate when there is no hazardous fire but smoke from cooking is detected.

While false alarms can be inconvenient and disruptive, it is common for smoke detectors to trigger false alarms in such situations. The provided accuracy rate of 90% under non-fire conditions suggests that the detector performs relatively well in distinguishing between actual fires and cooking smoke. Considering these factors, the smoke detector appears to be quite useful overall.

3 Question 3

We could set

$$\frac{1}{1+e^{-\beta^T x_i}} = \sigma(\beta^T x)$$

To start, the probability of one data point is:

$$P(Y=y|X=x) = \sigma(\beta^T x)^y * [1-\sigma(\beta^T x)]^{(1-y)}$$

Since each datapoint is independent, the probability of all data is:

$$L(\beta) = \prod_{i=1}^{n} P(Y = y^{(i)}|X = x^{(i)})$$

$$L(\beta) = \prod_{i=1}^{n} \sigma(\beta^T x^{(i)})^{y^{(i)}} * [1-\sigma(\beta^T x^{(i)})]^{(1-y^{(i)})}$$

Before we start take derivative, here is the side note of chain rule:

$$\frac{\partial \sigma(z)}{\partial z} = \sigma(z)[1-\sigma(z)]$$

Derivative of gradient for one datapoint(x,y):

$$\frac{\partial L(\beta)}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} y log \sigma(\beta^T x) + \frac{\partial}{\partial \beta_j}(1-y) log[1-\sigma(\beta^T x)]$$

$$= [\frac{y}{\sigma(\beta^T x)} - \frac{1-y}{1-\sigma(\beta^T x)}]\frac{\partial}{\partial \beta_j}\sigma(\beta^T x)$$

$$= [\frac{y}{\sigma(\beta^T x)} - \frac{1-y}{1-\sigma(\beta^T x)}]\sigma(\beta^T x)[1-\sigma(\beta^T x)]x_j$$

$$= \frac{y-\sigma(\beta^T x)}{\sigma(\beta^T x)[1-\sigma(\beta^T x)]}\sigma(\beta^T x)[1-\sigma(\beta^T x)]x_j$$

$$= [y-\sigma(\beta^T x)]x_j$$

Because the derivative of sums is the sum of derivatives, the gradient of theta is simply the sum of this term for each training datapoint:

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

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

4 Question 4

4.1 (a) One vs All

In the One vs All approach for multi-class classification, we train multiple binary classifiers, where each classifier is designed to distinguish one class from all the other classes. For example, if we have N classes, we would train N classifiers. During prediction, we apply each classifier to the input data, and the class with the highest probability or confidence score from the individual classifiers is assigned as the final predicted class.

4.1.1 Advantages of One vs All:

Simplicity: The One vs All approach is conceptually straightforward and easy to implement.

4.1.2 Disadvantages of One vs All:

Class Overlap: The One vs All approach assumes that the classes are mutually exclusive and independent. However, in some cases, classes may overlap or exhibit complex relationships, which can lead to suboptimal results. And in some cases, no class chose this point.

4.2 (b) All vs All

In the All vs All approach for multi-class classification, we train binary classifiers for every possible pair of classes. For N classes, we need $\binom{N}{2}$ classifiers. During prediction, each classifier votes for its predicted class, and the class with the most votes is assigned as the final predicted class.

4.2.1 Advantages of All vs All:

Handling Class Overlap: The All vs All approach can handle situations where classes overlap or have complex relationships since each classifier is trained specifically for a pair of classes.

Balanced Training Data: With pairwise classifiers, each classifier is trained on a balanced subset of the data, which can help in achieving better performance.

4.2.2 Disadvantages of All vs All:

Complexity: The All vs All approach requires training a large number of classifiers, which can be computationally expensive, especially for a large number of classes.

Decision Boundary Ambiguity: In cases where classes have overlapping regions, the decision boundaries learned by different classifiers may conflict or create ambiguous regions, leading to potential misclassifications.

5 Question 5

5.1 (a)

True:

Positive predictive value (PPV), precision, is the probability that a sample is truly positive given that the model classifies it as positive. It is calculated as the number of true positive predictions divided by the sum of true positive and false positive predictions:$PPV = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$

5.2 (b)

False:

total should be **2(k-1)** parameters which is 6

5.3 (c)

False:

Logistic regression models the log-odds as a linear function of the input features. This linear relationship between the log-odds and the input features results in a linear decision boundary in the feature space. The decision boundary is a hyperplane that separates the classes.

If the log-odds function is modeled as a quadratic function, it would deviate from the linear relationship and would not represent logistic regression. Instead, it would be a form of non-linear regression.

To achieve a non-linear decision boundary in logistic regression, one common approach is to introduce non-linear transformations of the input features. This can be done by adding polynomial terms, interaction terms, or applying other non-linear transformations to the features. By incorporating these non-linear transformations, logistic regression can capture more complex relationships and produce a non-linear decision boundary.

5.4 (d)

False:

While a 97% test accuracy may initially seem like the model is performing well, it does not provide a comprehensive evaluation of its effectiveness for detecting fraudulent credit card transactions. Accuracy alone is not sufficient for assessing the performance of a classifier, especially in scenarios with imbalanced classes such as fraudulent transactions.

In this case, detecting fraudulent transactions is typically a critical task, and false negatives (fraudulent transactions classified as non-fraudulent) can have significant consequences. Simply achieving a high overall accuracy does not necessarily imply that the model is performing well in identifying fraudulent transactions.

5.5 (e)

True:

For a very good classification model, we expect the confusion matrix to be dominated by diagonal entries. The diagonal entries of the confusion matrix represent the correctly classified instances, where the predicted class matches the true class. These are the true positives and true negatives.