Linear Data Analysis
Classification - Assesment With Confusion Matrix

Cain Susko

Queen's University
School of Computing

March 16, 2022

a Data Labels as Dependent Variables

this lecture will cover data which has labels. labesl are categorical dependent data. Classification is the predicion of data given previoys data. A confusion matrix is used to assess the classification.

Dependent Variables Consider observation i which has a label which is provided by a different system (human, AI, ...) . i is a categorical variable in a finite set. An independent variable cannot be categorical. Additionally, a category cannot be scaled. For new data, we want to *predict* the new label. This implies that **a label is a dependent variable** y_i.

Binary Classification We mostly do supervised binary classification. the supervised means that we are given labels; the binary means there are only 2 labels, and the classification means we want to classify new data with a label.
Given m data:

- Each observation is a row $\underline{\mathrm{x}}_i$ of n variables.
- each binary label $y_i \in \{-1, +1\}$

To score a classification, we map $z : \mathbb{R}^n \to \mathbb{R}$. Prediction / classification / quantization is a map:

$$q : \mathbb{R} \to \{-1, +1\} \text{ or } q(z(\vec{u})) = \pm 1$$

such that for each data vector there are 2 labels and 2 predictions, thus there are 4 contingencies possible.

Colouring in MatLab we can use the command `gscatter` to colour our categorical data.

b Confusion Matrix for Binary Labels

in binary calssification we have 4 contingencies, 2 from the system that labeled the data, and 2 from a quantization of our labelling algorithm.

	+1	-1
+1	TP	FN
-1	FP	TN

We use this Confusion Matirx as a truth table in order to know what is or is not correct within the contingencies of data. the sum of the first row gives the number of positive vectors that are provided by the user. the sum of the second row is the number of all negative data vectors provided by the user. the rows represent the provided labels and the columns represent the predicted labels. each space is filled with the number of True Positives, False Negatives, False Positives, and True Negatives.

Sensitivity The True Positive Rate (TPR) can be calculated from the confusion matrix as:

$$TPR = \frac{TP}{P}$$

where P is the number of positives and TP is the number of true positives

Specificity The True Negative Rate (TNR) can be calculated as:

$$TNR = \frac{TN}{N}$$

where the variables are the same for TPR accept 'negative' instead of 'positive'

Accuracy The Accuracy can be calculated as:

$$ACC = \frac{TP + TN}{P + N}$$

Errors There are 2 types of errors in classification:

I False Positive (FP)

II False Negative (FN)

c Example of Confusion Matrix

Consider Data with the labels:

$$y_1 = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \qquad q_1 = \begin{bmatrix} 1 \\ -1 \\ 1 \\ 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}$$

where y_1 are the user provided labels and q_1 is the classification of said data using a labeling algorithm. Where the rows represents the provided labels

	+1	-1
+1	5	1
-1	2	4

and the columns represent the predicted class (label).

Given Scores Instead of calculating the predicted label, we shall use the score from some other classification algorithm:

$$\vec{z_1} = \begin{bmatrix} -2.5 \\ -2 \\ -1.5 \\ -1 \\ -0.5 \\ 0 \\ 0.5 \\ 1 \\ 1.5 \\ 2 \\ 2.5 \\ 3 \end{bmatrix}$$

We can then create a confusion table using the threshold Θ where:

- class -1 iff $z_i < \Theta$
- class +1 iff $z_i \geq \Theta$

We must first find the proper hyperparameter Θ for best classification

Learning Summary

Students should now be able to:

- visualize labebled data
- threshold data scores to predict classes
- compute a confusion matrix for a threshold
- measure the FNR and FPR for a given threshold.