Assignment 2

Benny Chen

November 7, 2023

1 Logistic Regression

A country has two political parties, denoted by A and B here for simplicity. The probability that a voter votes for party A or party B is found to be modeled well by a 2-class logistic discrimination model with the following linear function:

$$g(x) = 0.2x_1 + 0.1x_2 - 0.3x_3 - 3 \tag{1}$$

where x_1 is the family income (in \$10,000; if you are given \$20,000, $x_1 = 2$), x_2 is the number of years of education, and x_3 is the gender (1 for male and 0 for female). Note that the sigmoid functionc an be written as $\frac{1}{1+\exp(-g(x))}$

The output y of the logistic discrimination model represents the probability that a column voter with attributes $x = (x_1; x_2; x_3)^T$ will vote for party A.

(a) Consider a male voter with a family income of \$50,000 and 20 years of education. According to the model, what is the *probability* that he will vote for party A?

(b) Consider a female voter with a family income of \$30,000 and 12 years of education. According to the model, what is the *log odds (logit)* of the probability that she will vote for party A?

Answer:

(a) We first calculate the linear function $g(x)$:

$$g(x) = 0.2(5) + 0.1(20) - 0.3(1) - 3 = 0.3 \tag{2}$$

Then we plug $g(x)$ into the sigmoid function:

$$\frac{1}{1+\exp(-g(x))} = \frac{1}{1+\exp(-0.3)} = 0.574 \tag{3}$$

(b) We first calculate the linear function $g(x)$:

$$g(x) = 0.2(3) + 0.1(12) - 0.3(0) - 3 = -1.2 \tag{4}$$

We then plug $g(x)$ into the sigmoid function:

$$\frac{1}{1+\exp(-g(x))} = \frac{1}{1+\exp(-(-1.2))} = 0.231 \tag{5}$$

We then calculate the log odds:

$$\ln(\frac{p}{1-p}) = \ln(\frac{0.231}{1-0.231}) = -1.2 \tag{6}$$

2 Classification Evaluation Measures

Consider a test data of 1000 samples with two classes: + class (100 samples) and − class (900 samples). We have two random classifiers C1 and C2.
We note that classifier C1 classifies test data to + class randomly with a probability p ($0 \leq p \leq 1$) and classifier C2 classifies test data to + class randomly with a probability 2p.

(a) What is the expected TPR and FPR for C1 and C2?

(b) Is C2 a better classifier than C1? *Hint: The random guess line in an ROC curve corresponds to TPR = FPR.*

Answer:

(a) Equations: $TRP = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$
C1: $TP = \frac{100}{1000}p$ $FN = \frac{100}{1000}(1-p)$ $FP = \frac{900}{1000}p$ $TN = \frac{900}{1000}(1-p)$

$$TPR = \frac{TP}{TP+FN} = \frac{p}{p+1-p} = \frac{p}{1} = p \tag{7}$$

$$FPR = \frac{FP}{FP+TN} = \frac{9p}{9p+9-9p} = \frac{9p}{9} = p \tag{8}$$

C2: $TP = \frac{100}{1000}2p$ $FN = \frac{100}{1000}(1-2p)$ $FP = \frac{900}{1000}2p$ $TN = \frac{900}{1000}(1-2p)$

$$TPR = fracTPTP + FN = \frac{2p}{2p+1-2p} = \frac{2p}{1} = 2p \tag{9}$$

$$FPR = \frac{FP}{FP+TN} = \frac{18p}{18p+9-18p} = \frac{18p}{9} = 2p \tag{10}$$

(b) C2 is not a better classifier than C1 because the TPR and FPR are the same for both classifiers.

3 Theory of Support Vector Machines

(a) The following is the primal formualtion of L2 SVM (with squared slack variables), a variant of the standard SVM .

$$\min_{w,b,\xi} \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^{l} \xi_i^2 \tag{11}$$

$$s.t. \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i \in \{1, \ldots, l\}, \quad \xi_i \geq 0, \quad i \in \{1, \ldots, l\} \tag{12}$$

If we remove the last constraint ($\xi_i \geq 0$), we might get a simpler problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^{l} \xi_i^2 \tag{13}$$

$$s.t. \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i \in \{1, \ldots, l\} \tag{14}$$

Please provide the Lagrangian of the above simplified formulation.

(b) Please find the partial derivative of the Lagrangian in (a) with respect to w, b, and ξ_i.

Answer:

1. The Lagrangian of the above simplified formulation is:

$$L(w, b, \xi, \alpha) = \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^{l} \xi_i^2 - \sum_{i=1}^{l} \alpha_i (y_i(w^T x_i + b) - 1 + \xi_i) \tag{15}$$

2. The partial derivative of the Lagrangian in (a) with respect to w, b, and ξ_i is:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{l} \alpha_i y_i x_i = 0 \tag{16}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{17}$$

$$\frac{\partial L}{\partial \xi_i} = \sum_{i=1}^{l} C\xi_i - \alpha_i = 0 \tag{18}$$

4 Machine Learning Evaluation (ROC Curve)

You have been asked to develop a classification model for diagnosing whether a patient is infected with a certain disease. To help you construct the models, your collaborator has provided you with a small training set (N = 10) with

equal number of positive and negative examples. You tried several approaches and found two most promising models, C1 and C2. The outputs of the models in terms of predicting whether each of the training examples belong to the "positive (+)" class are summarized in the table below. The first row shows the probability a training example belongs to the positive class according to classifier C1, while the second row shows the same information for classifier C2. The last row indicates the true class label of the 10 training examples.

$P(y = +\|C_1)$	0.1	0.15	0.2	0.3	0.31	0.4	0.62	0.77	0.81	0.95
$P(y = +\|C_2)$	0.25	0.49	0.05	0.35	0.66	0.6	0.7	0.65	0.55	0.99
y	−	+	−	−	+	−	+	+	−	+

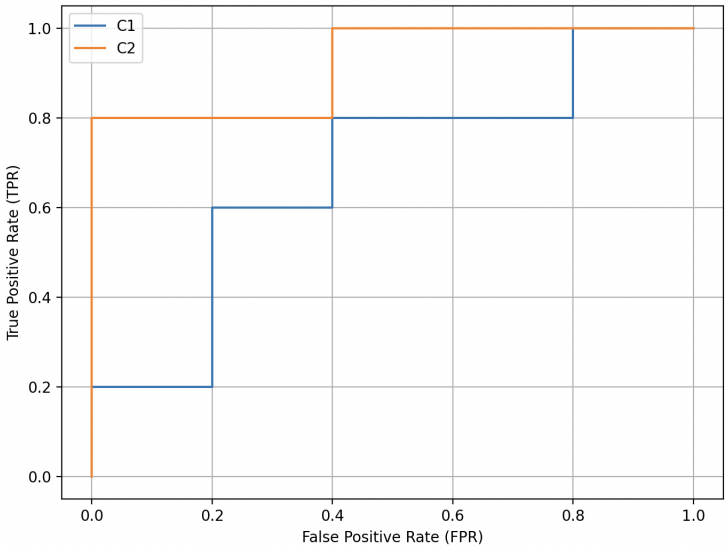
For each model, we will evaluate different thresholds within the range of [0, 1], and a sample with probability $P(y = +|Cx)$ (x is either 1 or 2) that is lower than this threshold will be estimated as -, or + if greater than this threshold. By varying the thresholds (referred to the lecture slide on ROC), you can study the model performance and draw the ROC.

(a) Draw the corresponding ROC curves for both classifiers on the same plot.

(b) Which classifier can be considered better? Why?

Answer:

(a)



(b) C2 is better because it has a higher AUC of 0.92 compared to C1's AUC of 0.68.

5 Clustering

Consider the following set of one-dimensional data points: $\{0.1, 0.25, 0.45, 0.55, 0.8, 0.9\}$. All the points are located in the range between $[0, 1]$.

(a) Suppose we apply k-means clustering to obtain three clusters, A, B, and C. If the initial centroids are located at $\{0, 0.4, 1\}$, respectively, we have the following cluster assignment of the data points.

Iter	0.10	0.25	0.45	0.55	0.80	0.90	A	B	C
0	A	B	B	B	C	C	0.00	0.40	1.00
1	A	A	B	B	C	C	0.1	0.42	0.85
2	A	A	B	B	C	C	0.18	0.5	0.85
3	A	A	B	B	C	C	0.18	0.5	0.85

Find the sum-of-squared error (SSE) of the clustering after the third iteration. And find the silhouette coefficient of data point 0.25 after the third iteration.

(b) For the dataset given in part (a), is it possible to obtain empty clusters? Why?

Answer:

1. To calculate the SSE, we first calculate the distance between each point and its centroid:

$$d(0.1, 0.18) = 0.0064 \tag{19}$$

$$d(0.25, 0.18) = 0.0049 \tag{20}$$

$$d(0.45, 0.5) = 0.0025 \tag{21}$$

$$d(0.55, 0.5) = 0.0025 \tag{22}$$

$$d(0.8, 0.85) = 0.0025 \tag{23}$$

$$d(0.9, 0.85) = 0.0025 \tag{24}$$

Then we sum them up to get the SSE:

$$SSE = 0.0064 + 0.0049 + 0.0025 + 0.0025 + 0.0025 + 0.0025 = 0.0213 \tag{25}$$

The formula for silhouette coefficient is:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{26}$$

We first calculate $a(i)$:

$$a(i) = a(.25) = \frac{|.1 - .25|}{1} = 0.15 \tag{27}$$

We then calculate $b(i)$:

$$b(i) = b(.25) = \min\{\frac{.2+.3}{2}, \frac{.55+.65}{2}\} = 0.25 \tag{28}$$

We can now calculate the silhouette coefficient:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} = \frac{0.25 - 0.15}{\max\{0.15, 0.25\}} = 0.4 \tag{29}$$

2. No, it's not possible to have a empty cluster since the centroids will always have a point in the clusters.