PS2: Classification

DS 5220: Supervised Machine Learning and Learning Theory
Omer Seyfeddin Koc

March 11, 2023

1 Classification Metrics (10%)

A new COVID-19 test-kit has been released and you need to assess its effectiveness. Results of the experiments performed with the kit are summarized in the following confusion matrix:

		Predictions (by the test)	
		+	-
Actual (covid status)	+	33	1
	-	3	72

1. How many false positives are there? how many false negatives?

   Solution : False Positive (FP) refers to the number of negative examples wrongly classified as positive. As stated in the table above, **the number of False Positive (FP) is 3.**

   False Negative (FN)refers to the number of positive examples wrongly classified as negative. As stated in the table above, **the number of False Negative (FN) is 1.**

2. What is the accuracy of the test?

   Solution : Accuracy is calculated as the number of all correct predictions divided by the total number of all predictions.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{33 + 72}{33 + 72 + 1 + 3} = \frac{105}{109} = \mathbf{0.9633}$$

3. What are the precision and recall of the test?

   Solution : **Precision** is the fraction of true positive samples in the group of samples declared as positive by the classifier.

$$\textbf{Precision} = \frac{TP}{TP + FP} = \frac{33}{33 + 3} = \frac{33}{36} = \mathbf{0.9167}$$

   Recall is the fraction of positive examples that are correctly classified by the model

$$\textbf{Recall} = \frac{TP}{TP + FN} = \frac{33}{33 + 1} = \frac{33}{34} = \mathbf{0.9706}$$

4. What is the F_1 score?

   Solution : Precision and recall can be combined into a single metric known as F_1 score. F_1 score is the harmonic mean of precision and recall.

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} = \frac{33}{33 + \frac{1}{2}(1 + 3)} = \frac{33}{35} = \mathbf{0.9429}$$

5. Would you rather have a higher precision or a higher recall in this case? Explain.

   Solution : In this scenario, it is important to detect the infected people. A negative result in a healthy person's test can create a false impression and cause unnecessary quarantine. This is undesirable situation, but it does not cause a health problem for us.

   However, the negative detection of a truly infected person by the test-kit causes that person to spread the virus to other people, so it is vital to accurately predict those who are actually infected.

   Recall is the fraction of positive examples that are correctly classified by the model. Therefore, **a higher recall value should be preferred for this scenario**.

2 Logistic Regression (20%)

In this exercise you will implement a logistic regression model to predict whether a student gets admitted into a university based on historical data from previous applicants.

1. Download the notebook LogisticRegressionEx.ipynb and the text file exams_data.txt from Canvas.
2. Follow the instructions in the notebook and complete the code for the functions written there.

Solution : The solution and all comments of the related problem are in the **OmerSKoc_PS2_2.ipyb** file.

3 k-Nearest Neighbors (KNN) (15%)

Consider the following data set comprised of three numerical features (f_1, f_2, and f_3) and one binary output y:

Example	f_1	f_2	f_3	y
$\mathbf{x}_1$	1	4	1	1
$\mathbf{x}_2$	1	2	3	1
$\mathbf{x}_3$	0	0	1	1
$\mathbf{x}_1$	-1	4	0	1
$\mathbf{x}_5$	1	0	-2	0
$\mathbf{x}_6$	-1	-1	1	0
$\mathbf{x}_7$	0	-1	0	0
$\mathbf{x}_8$	1	0	-3	0

Based on this data set, classify a new vector $x = (1, 0, 1)$ using KNN with $k = 3$ and Manhattan distance. Show your work.

Solution : Manhattan distance (L1) is the equation where the p value in Minkowski distance is 1. It is also known as city block or taxicab distance.

$$d(x,y) = \left(\sum_{j=1}^{d} |x_j - y_j|^p \right)^{\frac{1}{p}}$$

For Manhattan distance (L1), $p = 1$. The formula for Manhattan distance is as follows:

$$d_1(x,y) = \sum_{j=1}^{d} |x_j - y_j|$$

Since the vectors are in 3-dimensional space, our Manhattan distance formula will be as follows:

$$d_1(x,y) = |x_1 - x_2| + |y_1 - y_2| + |z_1 - z_2|$$

The Manhattan distances between each example $(x_1, x_2, ..., x_8)$ and new vector (x_{new}) will be calculated below:

$$\begin{aligned}
d_1(x_1, x_{new}) &= |1-1| + |4-0| + |1-1| = |0| + |4| + |0| = 4 \\
d_1(x_2, x_{new}) &= |1-1| + |2-0| + |3-1| = |0| + |2| + |2| = 4 \\
d_1(x_3, x_{new}) &= |0-1| + |0-0| + |1-1| = |-1| + |0| + |0| = 1 \\
d_1(x_4, x_{new}) &= |-1-1| + |4-0| + |0-1| = |-2| + |4| + |-1| = 7 \\
d_1(x_5, x_{new}) &= |1-1| + |0-0| + |-2-1| = |0| + |0| + |-3| = 3 \\
d_1(x_6, x_{new}) &= |-1-1| + |-1-0| + |1-1| = |-2| + |-1| + |0| = 3 \\
d_1(x_7, x_{new}) &= |0-1| + |-4-0| + |0-1| = |-1| + |-4| + |-1| = 6 \\
d_1(x_8, x_{new}) &= |1-1| + |0-0| + |-3-1| = |0| + |0| + |-4| = 4
\end{aligned}$$

According to Manhattan distance, k=3 nearest neighbor is shown with star($*$) in the table. For k=3, the 3-nearest neighbors are x_3, x_5 and x_6.

Example	f_1	f_2	f_3	y	$d_1(x_j, x_{new})$
$\mathbf{x}_1$	1	4	1	1	4
$\mathbf{x}_2$	1	2	3	1	4
$\mathbf{x}_3$	0	0	1	1	**1***
$\mathbf{x}_4$	-1	4	0	1	7
$\mathbf{x}_5$	1	0	-2	0	**3***
$\mathbf{x}_6$	-1	-1	1	0	**3***
$\mathbf{x}_7$	0	-1	0	0	6
$\mathbf{x}_8$	1	0	-3	0	4

Finally, the y value for x_{new} will be determined by looking at the y values of the 3-nearest neighbors. Two of the three closest neighbors belong to the y=0 class and one to the y=1 class. Therefore, the y-value for x_{new} will be 0 since the majority has y=0.

Based on this dataset, we classified the x_{new} = (1,0,1) as y=0 using KNN with k = 3 and Manhattan distance.

4 Naive Bayes Classification (15%)

The following table shows a part of a customer database of an electronic store:

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Use Naive Bayes to predict whether the following customer will buy a computer:

$$x = (age = youth, income = medium, student = yes, credit\_rating = fair)$$

Don't use Laplace smoothing. Show your work.

Solution : We first estimate the class prior probabilities based on their frequency in the data:

$$P(buys\_computer = yes) = 9/14 = 0.6429$$
$$P(buys\_computer = no) = 5/14 = 0.3571$$

We now estimate the conditional probabilities of the features in the new sample:

$$\begin{aligned}
Age : & P(age = youth|buys\_computer = yes) = 2/9 = 0.2222 \\
& P(age = youth|buys\_computer = no) = 3/5 = 0.6000 \\
\\
Income : & P(income = medium|buys\_computer = yes) = 4/9 = 0.4444 \\
& P(income = medium|buys\_computer = no) = 2/5 = 0.4000 \\
\\
Student : & P(student = yes|buys\_computer = yes) = 6/9 = 0.6667 \\
& P(student = yes|buys\_computer = no) = 1/5 = 0.2000 \\
\\
Credit\ Rating : & P(credit\_rating = fair|buys\_computer = yes) = 6/9 = 0.6667 \\
& P(credit\_rating = fair|buys\_computer = no) = 2/5 = 0.4000
\end{aligned}$$

Therefore, the class posterior probabilities are (α = 1/P(x) is a constant term):

$$\begin{aligned}
P(Yes|X) &= \alpha \cdot P(Yes) \cdot P(A = Youth|Yes) \cdot P(I = Medium|Yes) \cdot P(S = Yes|Yes) \cdot P(CR = Fair|Yes) \\
&= \alpha \cdot 0.6429 \cdot 0.2222 \cdot 0.4444 \cdot 0.6667 \cdot 0.6667 \\
&= 0.0282\alpha \\
P(No|X) &= \alpha \cdot P(No) \cdot P(A = Youth|No) \cdot P(I = Medium|No) \cdot P(S = Yes|No) \cdot P(CR = Fair|No) \\
&= \alpha \cdot 0.3571 \cdot 0.6000 \cdot 0.4000 \cdot 0.2000 \cdot 0.4000 \\
&= 0.0069\alpha
\end{aligned}$$

Since $P(Yes|X) > P(No|X)$, the sample is classified as *buys_computer = yes.*

5 Spam Filter (40%)

A spam filter is a computer program that classifies e-mail messages as either spam (unwanted) or ham (wanted). The spam filter maintains a set of features. A feature is a characteristic of messages that the spam filter can use to distinguish ham from spam. A feature could be simply the presence of a word in the message, for example, "the message contains the word 'lottery'", and it could be much more complex/powerful, for example, "the subject line is all capitals" or "the message body mentions many Internet domains".

Given (1) a set of features, (2) a set of messages known to be ham, and (3) a set of messages known to be spam, the spam filter learns how to classify incoming messages as ham or spam.

In this exercise we are going to train a spam filter using the Apache SpamAssassin public mail corpus.

Solution : The solution and all comments of the related problem are in the **OmerSKoc_PS2_5.ipyb** file.