Comp6940 Worksheet with Solutions

Performance Metrics:

Introduction:

Emergency physicians, like other specialists, are faced with different patients and various situations every day. They have to use ancillary diagnostic tools like laboratory tests and imaging studies to be able to manage them (1-8). In most cases, numerous tests are available. Tests with the least error and the most accuracy are more desirable. The power of a test to separate patients from healthy people determines its accuracy and diagnostic value (9). Therefore, a test with 100% accuracy should be the first choice. This does not happen in reality as the accuracy of a test varies for different diseases and in different situations. For example, the value for diagnosing cancer varies based on pre-test probability. It shows high accuracy in low risk patient and low accuracy in high risk ones. The characteristics of a test that reflects the aforementioned abilities are accuracy, sensitivity, specificity, positive and negative predictive values.

Definitions:

- **Patient**: positive for disease
- **Healthy**: negative for disease
- True positive (TP) = the number of cases correctly identified as patient
- False positive (FP) = the number of cases incorrectly identified as patient
- True negative (TN) = the number of cases correctly identified as healthy
- False negative (FN) = the number of cases incorrectly identified as healthy

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN /TP + TN + FP + FN

Sensitivity (Recall): The sensitivity of a test is its ability to determine the patient cases correctly. To estimate it, we should calculate the proportion of true positive in patient cases. Mathematically, this can be stated as:

Sensitivity = TP / TP + FN

Precision: The Precision of a test is its ability to determine how many of the patients cases were relevant.

Precision = TP / TP + FP

Specificity: The specificity of a test is its ability to determine the healthy cases correctly. To estimate it, we should calculate the proportion of true negative in healthy cases. Mathematically, this can be stated as:

Specificity = TN / TN + FP

Examples:

1. Imagine we have a sample of 100 cases, 50 healthy and the others patient. If a test can be positive for all patients and be negative for all the healthy ones. it is 100% accurate.

   Accuracy 50+50/50+50+0+0 = 100%

   Sensitivity 50/50+0 = 100%

   Specificity 50/50+0 = 100%

   Precision 50/50+0 = 100%

2. If the test can only diagnose 25 out of the 50 patients and has reported the others as healthy.

   Accuracy: Of the 100 cases that have been tested, the test could determine 25 patients and 50 healthy cases correctly. Therefore, the accuracy of the test is equal to 75 divided by 100 or 75%.

   Sensitivity: From the 50 patients, the test has only diagnosed 25. Therefore, its sensitivity is 25 divided by 50 or 50%.

   Specificity: From the 50 healthy people, the test has correctly pointed out all 50. Therefore, its specificity is 50 divided by 50 or 100%.

   Precision: From the 50 patients, the test has correctly pointed out all 25. Therefore, its precision is 25 divided by 25 or 100%.

3. This time we will assume that the test has been able to identify 25 of the 50 healthy cases and has reported the others as patients.

   Accuracy: Of the 100 cases that have been tested, the test could identify 25 healthy cases and 50 patients correctly. Therefore, the accuracy of the test is equal to 75 divided by 100 or 75%.

   Sensitivity: From the 50 patients, the test has diagnosed all 50. Therefore, its sensitivity is 50 divided by 50 or 100%.

   Specificity: From the 50 healthy cases, the test has correctly pointed out only 25. Therefore, its specificity is 25 divided by 50 or 50%.

Precision: From the 50 patients, the test has correctly pointed out all 50 and incorrectly classed 25 health people as patients. Therefore, its precision is 50 divided by 75 or 66%.

Imbalanced data: Undersampling & Oversampling

Introduction

Many medical applications, we have datasets where we have two classes for the main outcome; normal samples and relevant samples. For example in a cancer detection application we might have a small percentages of patients with cancer (relevant samples) while the majority of samples might be healthy individuals.

The main motivation behind the need to preprocess imbalanced data before we feed them into a classifier is that typically classifiers are more sensitive to detecting the majority class and less sensitive to the minority class. Thus, if we don't take care of the issue, the classification output will be biased, in many cases resulting in always predicting the majority class.

Definitions

- **Undersampling**: This method works with majority class. It reduces the number of observations from majority class to make the data set balanced. Random undersampling method randomly chooses observations from majority class which are eliminated until the data set gets balanced. Informative undersampling follows a pre-specified selection criterion to remove the observations from majority class.

- **OverSampling**: This method works with minority class. It replicates the observations from minority class to balance the data. It is also known as upsampling. Random oversampling balances the data by randomly oversampling the minority class. Informative oversampling uses a pre-specified criterion and synthetically generates minority class observations.

- **Synthetic Data Generation**: Instead of replicating and adding the observations from the minority class, it overcome imbalances by generates artificial data. It is also a type of oversampling technique. In regards to synthetic data generation, synthetic minority oversampling technique (SMOTE) is a powerful and widely used method. SMOTE algorithm creates artificial data based on feature space (rather than data space) similarities from minority samples.

- **Cost Sensitive Learning (CSL)**: Cost-Sensitive Learning is a type of learning in data mining that takes the misclassification costs (and possibly other types of cost) into consideration. The goal of this type of learning is to minimize the total cost.

Forecasting with Single Exponential Smoothing

Definition:

Forecasting Formula

Forecasting the next point The forecasting formula is the basic equation:

$S_{t+1} = \alpha y_t+(1-\alpha)S_t, 0<\alpha\leq 1, t>0.$

In other words, the new forecast is the old one plus an adjustment for the error that occurred in the last forecast.

Example:

The table below shows the movement of the price of a commodity over 12 months.

Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Price	25	30	32	33	32	31	30	29	28	28	29	31

1. Calculate a 6 month moving average for each month. What is the forecast for month 13?
2. Apply exponential smoothing with smoothing constants of 0.7 and 0.8 to derive forecasts for month 13.

Solution

1. Now we cannot calculate a 6 month moving average until we have at least 6 observations - i.e. we can only calculate such an average from month 6 onward. Hence we have:

   m6 = (25 + 30 + 32 + 33 + 32 + 31)/6 = 30.50

   m7 = (30 + 32 + 33 + 32 + 31 + 30)/6 = 31.33

   m8 = (32 + 33 + 32 + 31 + 30 + 29)/6 = 31.17

   m9 = (33 + 32 + 31 + 30 + 29 + 28)/6 = 30.50

   m10 = (32 + 31 + 30 + 29 + 28 + 28)/6 = 29.67

   m11 = (31 + 30 + 29 + 28 + 28 + 29)/6 = 29.17

   m12 = (30 + 29 + 28 + 28 + 29 + 31)/6 = 29.17

   The forecast for month 13 is just the moving average for the month before that i.e. the moving average for month 12 = m12 = 29.17.

2. Applying exponential smoothing with a smoothing constant of 0.7 we get:

   M1 = Y1 = 25

M2 = 0.7Y2 + 0.3M1 = 0.7(30) + 0.3(25) = 28.50

M3 = 0.7Y3 + 0.3M2 = 0.7(32) + 0.3(28.50) = 30.95

M4 = 0.7Y4 + 0.3M3 = 0.7(33) + 0.3(30.95) = 32.39

M5 = 0.7Y5 + 0.3M4 = 0.7(32) + 0.3(32.39) = 32.12

M6 = 0.7Y6 + 0.3M5 = 0.7(31) + 0.3(32.12) = 31.34

M7 = 0.7Y7 + 0.3M6 = 0.7(30) + 0.3(31.34) = 30.40

M8 = 0.7Y8 + 0.3M7 = 0.7(29) + 0.3(30.40) = 29.42

M9 = 0.7Y9 + 0.3M8 = 0.7(28) + 0.3(29.42) = 28.43

M10 = 0.7Y10 + 0.3M9 = 0.7(28) + 0.3(28.43) = 28.13

M11 = 0.7Y11 + 0.3M10 = 0.7(29) + 0.3(28.13) = 28.74

M12 = 0.7Y12 + 0.3M11 = 0.7(31) + 0.3(28.74) = 30.32

As before the forecast for month 13 is just the average for month 12 = M12 = 30.32.

3. Applying exponential smoothing with a smoothing constant of 0.8 we get:

M1 = Y1 = 25

M2 = 0.8Y2 + 0.2M1 = 0.8(30) + 0.2(25) = 29.00

M3 = 0.8Y3 + 0.2M2 = 0.8(32) + 0.2(29.00) = 31.40

M4 = 0.8Y4 + 0.2M3 = 0.8(33) + 0.2(31.40) = 32.68

M5 = 0.8Y5 + 0.2M4 = 0.8(32) + 0.2(32.68) = 32.14

M6 = 0.8Y6 + 0.2M5 = 0.8(31) + 0.2(32.14) = 31.23

M7 = 0.8Y7 + 0.2M6 = 0.8(30) + 0.2(31.23) = 30.25

M8 = 0.8Y8 + 0.2M7 = 0.8(29) + 0.2(30.25) = 29.25

M9 = 0.8Y9 + 0.2M8 = 0.8(28) + 0.2(29.25) = 28.25

M10 = 0.8Y10 + 0.2M9 = 0.8(28) + 0.2(28.25) = 28.05

M11 = 0.8Y11 + 0.2M10 = 0.8(29) + 0.2(28.05) = 28.81

M12 = 0.8Y12 + 0.2M11 = 0.8(31) + 0.2(28.81) = 30.56

As before the forecast for month 13 is just the average for month 12 = M12 = 30.56.

Betweenness centrality

Definition:

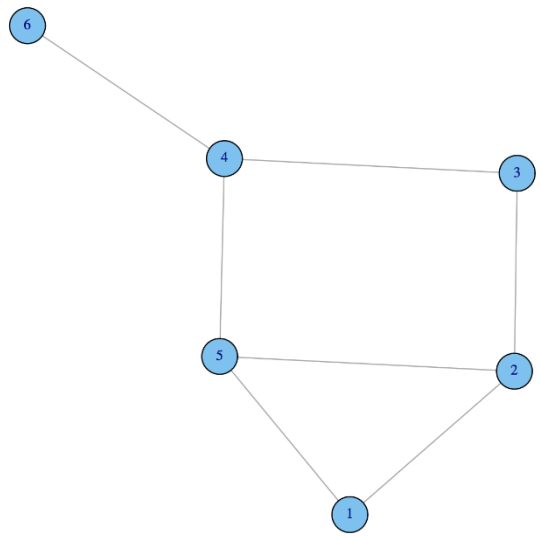
The betweenness of a vertex v in a graph G := (V, E) with V vertices is computed as follows:

- For each pair of vertices (s, t), compute the shortest paths between them.
- For each pair of vertices (s, t), determine the fraction of shortest paths that pass through the vertex in question (here, vertex v).
- Sum this fraction over all pairs of vertices (s, t).
- More compactly the betweenness can be represented as:

$$Betweenness\,(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

- where σst is total number of shortest paths from node s to node t and σst(v) is the number of those paths that pass through v.

Example:



	σst	σst(v)	σst(v)/σst
BC for 1			
2 to 3	1	0	0
2 to 4	1	0	0
2 to 5	1	0	0

	σst	σst(v)	σst(v)/σst
2 to 6	1	0	0
3 to 4	1	0	0
3 to 5	1	0	0
3 to 6	1	0	0
4 to 5	1	0	0
4 to 6	1	0	0
5 to 6	1	0	0
			0
BC for 2			
1 to 3	1	1	1
1 to 4	1	0	0
1 to 5	1	0	0
1 to 6	1	0	0
3 to 4	1	0	0
3 to 5	2	1	0.5
3 to 6	1	0	0
4 to 5	1	0	0
4 to 6	1	0	0
5 to 6	1	0	0
			1.5
BC for 3			
1 to 2	1	0	0
1 to 4	1	0	0
1 to 5	1	0	0
1 to 6	1	0	0
2 to 4	2	1	0.5
2 to 5	1	0	0

	σst	σst(v)	σst(v)/σst
2 to 6	2	1	0.5
4 to 5	1	0	0
4 to 6	1	0	0
5 to 6	1	0	0
			1
BC for 4			
1 to 2	1	0	0
1 to 3	1	0	0
1 to 5	1	0	0
1 to 6	1	1	1
2 to 3	1	0	0
2 to 5	1	0	0
2 to 6	2	1	0.5
3 to 5	2	1	0.5
3 to 6	1	1	1
5 to 6	1	1	1
			4
BC for 5			
1 to 2	1	0	0
1 to 3	1	0	0
1 to 4	1	1	1
1 to 6	1	1	1
2 to 3	1	0	0
2 to 4	2	1	0.5
2 to 6	2	1	0.5
3 to 4	1	0	0
			3

	σst	σst(v)	σst(v)/σst
BC for 6			
1 to 2	1	0	0
1 to 3	1	0	0
1 to 4	1	0	0
1 to 5	1	0	0
2 to 3	1	0	0
2 to 4	1	0	0
2 to 5	1	0	0
3 to 4	1	0	0
3 to 5	1	0	0
			0