Rasik Kane **Text Analytics COMP47600** **21/11/2020**
19200172 **Practical 8**

Concept of Clustering

In statistics, clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). Example: By looking at sample space [2, 8, 344, 338, 4, 350, 7], we can naively observe that we have 2 categories of numbers (2,4,7, 8) and (344, 338, 350) as they are "closer" on number line (1-D space). Similarly, N-D space observation sets like tweets, social media posts, inventory list of supermarkets can be classified into classes.

To distinguish if a data sample belongs to class A or Class B [Or class C Or……], we need to mathematically quantify sample in "Vector Feature Space". Threshold of similarity is used as a baseline to determine if 2 samples or documents are dissimilar. If a given sample pair is more similar than threshold, they are part of a same cluster. Similarity is often calculated using distance based metrices like cosine similarity, Manhattan distance etc.

Question 1 A system that detects tweets about different candidates in a general election have classified a gold-standard set of 200 tweets, 100 of which are identified to be about the election and 100 classed as being about non-election things. When the similarity threshold of the system is varied from 1 – 50; different numbers of correct and incorrect answers are obtained , that is different numbers of True Positives (TP), False Negative (FN), False Positives (FP) and True Negatives (TN) tweets.

Given problem data is inference drawn by comparing the ground truth and the predictions provided by algorithm which assesses pool of tweets and assign binary labels "Related to election" and "Not related to election". This data is nothing but "BINARY CONFUSION MATRIX" for tweet clustering algorithm.

		Prediction	
		Positive	Negative
Ground Truth	Positive	TRUE +	FALSE -
	Negative	FALSE +	TRUE -

With respect to pool of tweets and classification criteria as 'relating to topic ELECTION': label 1 represents : "Yes - Belongs to…" and label 0 represents : "No - Does not belong to.."

- **TP** : Algorithm correctly recognizes that Sample "belongs to.."
- **TN** : Algorithm correctly recognizes that Sample "Does not belong to.."
- **FP** : Algorithm erroneously classify sample as "belongs to.." class though Sample "Does not belong to..".
  It is usually called Type 1 error.
- **FN** : Algorithm erroneously classify sample as "Does not belong to.." class though Sample "Belong to..".
  It is usually called Type 2 error.

1(a) Taking this data, compute the Precision and Recall for the system at each threshold and identify the threshold values at which it does best, according to the F1 measure.

Confusion matrix gives summary of a classifier / clustering model. To quantify performance on scale, typically following metrics are used.

$$\textbf{Accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP + FP + TN + FN}}$$

Accuracy gives count of correctly labelled data. When data is balanced alike given data, classification gives good idea about model performance. But this metric is often confusing as it tells little about Boundary of clustering. From accuracy, one cannot answer the question that "samples Which cluster label are influencing the result."

Precision $= \frac{\mathbf{TP}}{\mathbf{TP + FP}}$

Precision captures effect of Type 1 error, where labels belonging to class 0 are misclassified as class 1 labels. For common NLP classifications like Spam mail detection[IS SPAM(1) Or Not)(0)], good precision is necessary. Otherwise, user shall have her/ his important mails into spam folder.

Recall $= \frac{\mathbf{TP}}{\mathbf{TP + FN}}$

Recall captures effect of Type 2 error, where labels belonging to class 1 are misclassified as class 0 labels. Recall should be high for classifier segregating instigating twitter handles [Is Offensive(1) or not(0)]; otherwise malicious handles would be considered harmless and shall continue on agenda.

$$\mathbf{f}_{\beta} = \frac{(1+\beta^2)*Precision*Recall}{\beta^2 * \mathbf{Precision + Recall}} = \frac{(1+\beta^2)*TP}{[(1+\beta^2)*TP] + [\beta^2 * \mathbf{FN}] + \mathbf{FP}}$$

f_β score combines precision and recall into one metric by calculating harmonic mean. F-score is a reliable metric when DATA IS UNBALANCED i.e. class distribution is skewed towards a class. This could be a case if our given sample had 150 Election related tweets and only 50 Unrelated tweets. In this situation, Accuracy would not be a good metric; as training data corpus would have more feature values aligned with "YES" class. Hence, an untrained classifier would be inherently BIASED towards majority class. Here, F-score is more effective as it has higher coefficients associated with TRUE classifications.

f_1 (β=1) treats precision and recall equally important. F1 score is a better metric when Precision and Recall cannot be balanced.
As can be seen, β boosts the weight of recall value in f-score. For example :

- For given data, IF TWEET IS ABOUT ELECTION OR NOT CAN BE IMPORTANT ISSUE TO SECURITY AGENCIES. Agencies would not want to miss a tweet into category "Not related to election" (i.e. False Negative); but would be tolerant about non-related tweets classified as election tweet (False positives).
- **So, client shall need model with Higher emphasis on recall → which means critical analysis of True negatives. So, a HIGHER value of β should be used i.e. f2,f3 would give better model for client requirements than f1.**

When metrics are calculated from provided confusion table, following results are obtained

threshold	accuracy	recall	precision	f1	TP_rate	FP_rate
1	0.59	0.2	0.909091	0.3279	0.2	0.02
5	0.725	0.5	0.909091	0.6452	0.5	0.05
10	0.75	0.6	0.857143	0.7059	0.6	0.1
15	0.8	0.8	0.8	0.8	0.8	0.2
20	0.79	0.88	0.745763	0.8073	0.88	0.3
25	0.75	0.9	0.692308	0.7826	0.9	0.4
30	0.725	0.95	0.655172	0.7755	0.95	0.5
35	0.68	0.96	0.615385	0.75	0.96	0.6
40	0.635	0.97	0.580838	0.7266	0.97	0.7
50	0.59	0.98	0.550562	0.7050	0.98	0.8

Observations

- Given data is balanced that is, number of samples for Yes and No class as equal. So, Accuracy can be a good metric to assess quality of model.
- F1 score is not decisive indicator alone since precision and recall give clear picture about miscalssification individually.
- As data is balanced, it is more likely that F1 and Accuracy scores are covariant.

As noticed, Highest F1 score is obtained for Threshold = 20. We can confirm above claim with reference to following graph.

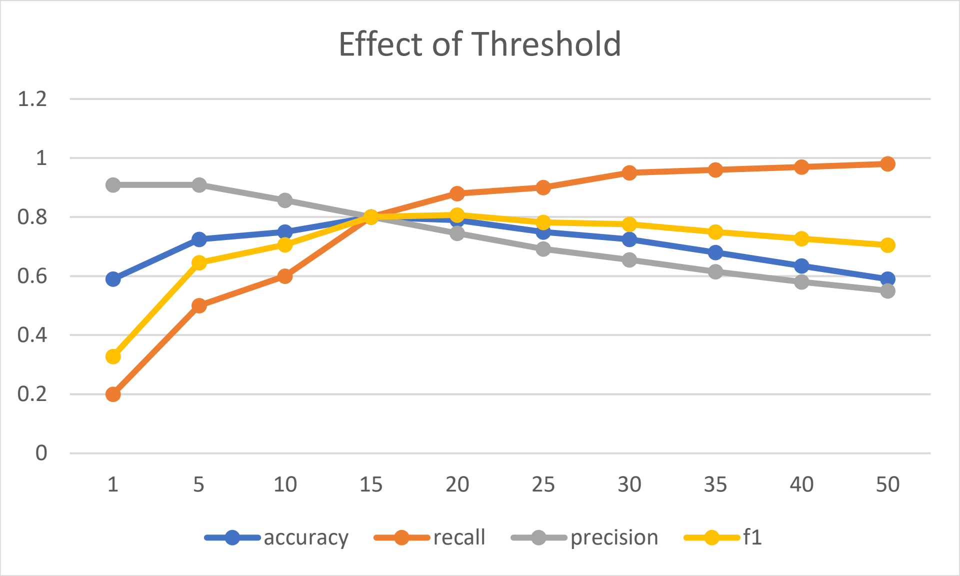

Figure 1Threshold vs classification metrics

Inference

- Accuracy and F1 scores are covarient and both have ~ highet value at threshold = 20
- We notice a knot of lines at threshold =15. This indicates sweet spot for hyperparameter “threshold”.
- Depending upon expectation from model [if effect of RECALL(FP) IS IMPORTANT or PRECISION(FP)]; we can choose to tune threshold about value 15.

1(b) plot the ROC for this data

Receiver operating characteristic curve is a chart that visualizes the tradeoff between true positive rate (TPR) and false positive rate (FPR) for each threshold value.

- Performance of model with respect to majority class detection is measured using ROC curve. Higher the TP_rate and lower the FP_rate, better is classifier at assessing a positive class.
- From the graph, we can see that distance between false positive and True positive sample count is better for threshold range (15→20) and it is highest at threshold =20

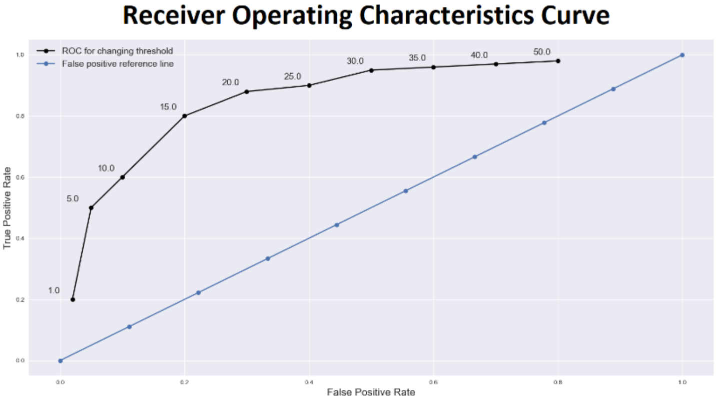

Figure 2 ROC curve with best threshold value ~ 20

- Indicator of a better classifier is **Area under the curve**. With normalized units, It indicates probability that a positive sample would be ranked higher than negative sample and pushed away from False positive reference line [reference]. We can visually confirm that if lines passing trough thresholds which are parallel to false positive reference line are drawn, area between line for threshold 20 shall be highest.

1(c) plot the DET curve for the same data

As noticed from figure 1, true action of deciding best threshold happens in “knot area” around 15. Same can be seen in figure 2 ROC curve, where direction of curve rapidly changes. But, due to the linear scale used for plotting ROC curves, different classifiers usually only differ in the top left corner of the graph and appear similar for a large part of the plot.
With DET graph, one can take a zoomed in look in this high derivative area using **Logarithmic scale**. DET curves represent straight lines in normal deviate scale.

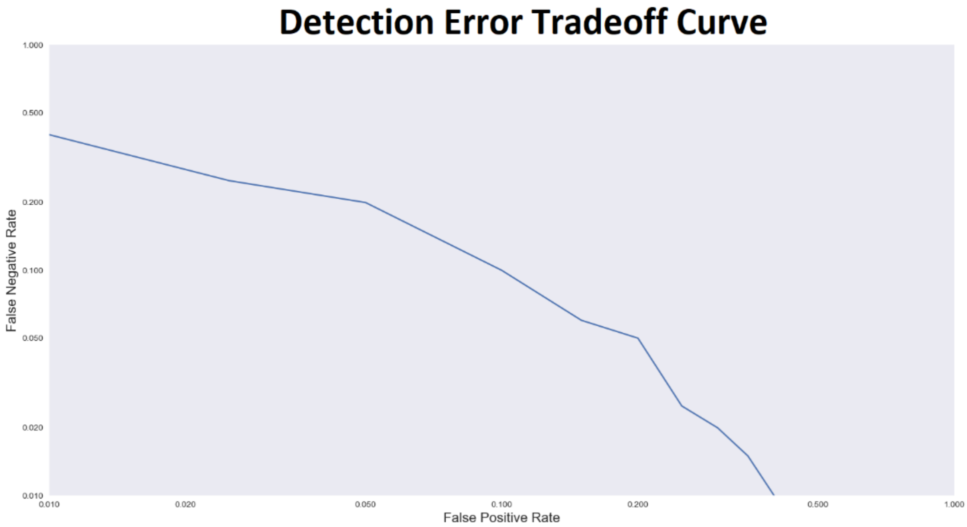

Figure 3 DET curve

Inference

DET curve focusses on Missed detection [False Negatives] as well as Wrong suggestions [False Negatives]. But importantly, **rather than plotting the probabilities themselves alike ROC curve, DET graph plots the normal deviates that correspond to the probabilities of classes.** [reference]. So, it is a better visual measure for analyzing performance results of threshold tradeoff.