Data Mining - Assignment-4

Anirudh K M - anikamal

9th April, 2016

1 Problem 1

First inorder the plot the ROC curve, we sort the given set of predictions in decreasing order. From, here we will analyze the cases where we need minimum and maximum number of thresholds.

It seems that a very good ideal classifier will have an area of 1.0 below the ROC curve.

For an ideal classifier, we can observe that initially as we pick each threshold from the top, the true positive rate increases to 1 and the 1-FPR(False Positive Rate) remains at 0. At a particular threshold we can see that the (TPR, FPR) is (1,0) and after this the we can see that (TPR, FPR) moves towards the point (1,1).

The ideal threshold would be the point where we would have the TP(True Positive) = n+, TN(True Negative) = n-, FP(False Positive) = 0, FN(False Negative) = 0.

So, from this observation we can see that for a very good ideal classifier, we need only three points to calculate the curve under the ROC curve.

These three points should give us the value of (TPR, FPR) to be (0,0), (1,0) and (1,1).

Next, in the case of a poor classifier we can see that the trace of (TPR,FPR) is very wavy. so in this case, we will b required to calculate the (TPR,FPR) for all the threshold points to calculate the area under the curve.

So the number of threshold points needed would be "n" points.

Generally for a poor classifier the area under the curve is observed to be 0.5.

2 Problem 2

Ward's method can be implemente in order to merge two clusters so that the sum of squared errors is reduced.

So in this method, with the distances given we combine two clusters and compute the SSE and compare with the previous value, we keep on repeating this step for all the clusters and finally decide on which cluster to be combined.

3 Problem 3

a

Both the $F_{K-1} \times F_{K-1}$ and $F_{K-1} \times F_1$ methods are being implemented.

b

The binarization of the input file is done using the code binarize.py.

Dataset	Support	Candidata count	Frequent count
Mushroom	0.01	1098	851
Mushroom	0.2	57	27
Mushroom	0.4	34	8
Solar	0.2	348	335
Solar	0.4	110	99
Solar	0.6	39	31
Nursery	0.2	207	120
Nursery	0.3	116	45
Nursery	0.5	45	9

Resutls obtained using the $F_{K-1} \times F_{K-1}$ method.

method.

Dataset	Support	Candidata count	Frequent count
Mushroom	0.01	1263	851
Mushroom	0.2	72	27
Mushroom	0.4	37	8
Solar	0.2	357	335
Solar	0.4	114	99
Solar	0.6	39	31
Nursery	0.2	211	120
Nursery	0.3	118	45
Nursery	0.5	45	9

Results obtained using $F_{K-1} \times F_1$ method.

From these two table we can observe that $F_{K-1} \times F_{K-1}$ method has better performance beacuse of less generation of candidate itemset compared to $F_{K-1} \times F_1$ method.

c

Dataset	Support	Frequent count	Maximal frequent items	closed frequent items
Mushroom	0.01	851	35	132
Mushroom	0.2	27	6	25
Mushroom	0.4	8	5	8
Solar	0.2	335	7	131
Solar	0.4	99	3	67
Solar	0.6	31	1	31
Nursery	0.2	120	58	120
Nursery	0.3	52	28	52
Nursery	0.5	9	9	9

Comparison of closed frequent items and maximal frequent itemset with frequent itemset for various support and datasets.

From this table we can see that as the support increases the count of frequent itemset, maximal frequent itemset and closed frequent itemset almost tends to be the same and keeps on decreasing.

d

Dataset	Support	confidence	No of rules:Brute Force	No of rules:Pruning
Mushroom	0.01	0.5	22228	10476
Mushroom	0.01	0.6	22228	5071
Mushroom	0.01	0.2	22228	16214
Mushroom	0.2	0.5	86	81
Mushroom	0.2	0.6	86	65
Mushroom	0.2	0.2	86	86
Mushroom	0.4	0.5	6	6
Mushroom	0.4	0.6	6	6
Mushroom	0.4	0.2	6	6
Solar	0.2	0.5	7456	5171
Solar	0.2	0.6	7456	4372
Solar	0.2	0.7	7456	3063
Solar	0.3	0.5	2160	1798
Solar	0.3	0.6	2160	1439
Nursery	0.2	0.5	602	535
Nursery	0.2	0.3	602	602
Nursery	0.2	0.6	602	457

Table showing the number of rules generated based on brute force and confidence pruning for various dataset, support and confidence.

From this we can see that as the confidence increases the pruning is proven to be more effective and also when we see there is an increases in support the brute force and the confidence pruning method doesn't seems to make any difference.

e

The association rules derived for various dataset, support, confidence are written to the file F1RulesByConfidenceMeasure.t and FK1RulesByConfidenceMeasure.txt.

From these rules generated we can observe that as the support and confidence increases we see that less rules are being generated.

So the rules are highly dependent on these parameters.

f

Dataset	Support	Lift	No of rules:Brute Force	No of rules:Pruning
Mushroom	0.01	1.2	22228	13389
Mushroom	0.01	3	22228	4454
Mushroom	0.01	1.75	22228	4572
Mushroom	0.2	1.2	86	59
Mushroom	0.2	3	86	52
Mushroom	0.2	1.75	86	52
Mushroom	0.4	1.2	6	6
Mushroom	0.4	3	6	6
Mushroom	0.4	1.75	6	6
Solar	0.2	1	7456	6665
Solar	0.2	1.5	7456	1390
Solar	0.2	2	7456	1327
Solar	0.3	1	2160	1883
Solar	0.3	1.5	2160	598
Nursery	0.2	1	602	515
Nursery	0.2	1.5	602	307
Nursery	0.2	1.2	602	309

Table showing the number of rules generated based on brute force and lift pruning for various dataset, support and lift.

Similar to the confidence pruning, in the measure of lift also we can see that as lift increases, the pruning shows its effect.

Also, when the support goes beyond a particular value, irrespective of the lift the no of rules remain the same.

The association rules derived for various dataset, support, lift are written to the file F1RulesByLiftMeasure.txt and FK1RulesByLiftMeasure.txt.

4 Problem 4

Paper: An Impossibility Theorem for Clustering
Author: Jon Kleinberg, Department of Computer Science, Cornell University

Summary:
Subsection 1:
Kleinberg points out the goals of clustering is to group the heterogeneous objects together by representing them as a set S of n points and defining distance function to measure the similarity between the points as a measure of the pairwise distances $d(i, j)$. And he denies the possibility of having a unified framework for the clustering process at the technical level and explains how it backs this claim with the help of impossibility theory and the tradeoffs it causes in some well known clustering techniques.
Although this goal is an intuitive approach to the clustering process and seem conceptually unified, it is rather not clearly defined at the implementation level is what is the claim made by Kleinberg and the paper attempts to prove this with the help of axiomatic frameworks. So it defines three axioms on various clustering techniques, each of which results in the impossibility theorem of not all axioms/properties being satisfied by any clustering technique. Then raises the question – how to interpret the impossibility results and discusses a priori function that generates a fixed number of clusters and shows other axiomatic approaches followed by researchers.

Subsection 2:
This section defines a clustering function conceptually for a set S and distance function d, the cluster

function f computes the partition Γ , so set of Γ will be the cluster. Further, it states the impossibility theorem by introducing distance metric properties such as Positivity, Symmetry and Triangle Inequality that should be satisfied by the pairwise distances and explains three properties on the clustering such as:
[1] **Scale Invariance**: function will not change with the units of distance measure. Axiomatically, stated as $f(d) = f(\alpha * d)$ where α is the constant and $f(d)$ is clustering function.
[2] **Richness**: Each point by itself is a cluster i.e. Range of the clustering function $f(d)$ is equal to the number of Γ in the set S.
[3] **Consistency**: For any clustering function f, if we transform the existing cluster distances within the cluster and between the cluster, the resulting function f' arises from the transformation or f' is the consistent variance of transformation.

Finally the impossibility theorem is defined as there is no clustering function f that satisfies all these three properties and before this is proved, there is a need to examine the relation between there properties to show that at least two of these properties are satisfied by the cluster. For this purpose the single-linkage procedure is introduced that creates a set of related functions.

Single-linkage procedure treats each point as a cluster and keeps on combining clusters until a stopping condition is reached. Again to represent axiomatically, from a set S it creates a weighted graph G_d of edge weight $d(I, j)$ for the node i and j. Now this weighted graph G_d creates linkages to the neighboring nodes until a stopping condition is reached. And there are three stopping conditions that are defined such as
$*$ **k-cluster stopping condition** – A parameter k is passed if the number of data points is at least k, for the function to stop adding more nodes to the subgraph to form a cluster.
$*$ **distance-r stopping condition**: Restrict the subgraph to include elements with edge distance r.
$*$ **distance-r stopping condition scale- stopping condition**: For a maximum pairwise distance p', we only add edges with weights $\alpha * p'$.

Kleinberg, now defines the theorem that only TWO of the three properties of clustering function discussed above can be held and it is summarized as:
[a] Scale-Invariance and Consistency is satisfied by K-cluster stpping condition with a $k >= 1$, and $n >= k$.
[b] Scale-Invariance and Richness is satisfied by scale- stopping condition for any positive $\alpha < 1$, and any $n >= 3$
[c] Richness and Consistency is satisfied by distance-r stopping condition for any $r > 0$, and any $n >= 2$

Subsection 3:
To prove the impossibility property, a concept of *antichain* is introduced and showed that any clustering function under discussion is eventually an *antichain* hence proving impossibility. It states that for a partition Γ' to be a refinement of Γ only if there is a $C \subseteq \Gamma$ so that $C' \subseteq C$.
Now Theorem 3 states a similar set of properties on the abstract function f inline with the above discussed Kleinberg's stopping condition summaries and both the theorems are proved to hold the impossibility property true:
Theorem 3.1 states that the $Range(f)$ is an antichain, if a clustering function f satisfies Scale-Invariance and Consistency.
Theorem 3.2 states that there is a clustering function f satisfying Scale-Invariance and Consistency if the $Range(f) = A$.

Subsection 4:
This section attempts to prove that the centroid based clustering technique will not satisfy the Consistency property. There by it defines the centroid based clustering techniques such as k-means and k-median clustering techniques as the one that selects k of the input points as spatial centroids, and then defines clusters by assigning each point in S to its nearest centroid.

Subsection 5:

As we saw the in the section 3, the relaxation of the Richness property when encountered with the Scale-Invariance and Consistency property together and the relaxation of the Scale-Invariance property when encountered with the Richness and Consistency property together, this section introduces the concept of relaxations of Consistency

Critical Assessment:
So the approach towards clustering problems is to gauge the axiomatic correctness of various clustering methods rather than optimizing the clustering problem itself, so as a result Kleinberg's axioms are subjected to various inconsistencies in terms of the formalization framework is concerned.
And I view this as a successful attempt to prove the impossibility condition but with a lot of assumptions in terms of formulating the axiomatic proofs. Also regarding the quality of the paper, I feel this paper presents ideas in an abstract way that can be applied to a wide function set which is a very good property of a high quality paper, for example the graph and subgraph notations for the clusters can be applied to any novel graph data structure in classical statistics.
The Strength of this paper being Kleinberg presented the proofs for the impossibility theorem with the help of well defined theorems and axiomatic reasoning and the probably the weakness also lies in the similar lines where he tries to write of the possibility of having a clustering function which satisfies all the properties.

Paper: Measures of Clustering Quality: A Working Set of Axioms for Clustering
Author: Margareta Ackerman and Shai Ben-David from School of Computer Science, University of Waterloo, Canada

Summary:
Subsection 1:
As the abstract clearly states that this is an attempt to formulate a quality measure for the clustering technique in hand rather that making axiomatic reasoning and rule generation on clustering methods.

So this starts off with argument that it is due to excessive adherence to the prescribed form of clustering and is not a permanent characteristics of it. The crux of the paper is a clustering quality measure(CQM) it proposes with the help of the three properties discussed in the Kleinberg's paper of Impossibility theorem. CQM is defined as a function which provides non-negative real number showing the offers a certain confidence interval for the clustering technique at hand and can be extensively be used for comparing the results between various clustering techniques based on its parameters.

To strengthen the CQM methods, the author says that clustering is not mere grouping of the data according to the distance measures but it is rather an extraction of meaningful patterns out of data in an un-supervised fashion and demonstrate that clustering can be done in polynomial time. It explains further that rather than working on cluster functions like Kleinberg they manage the quality measures that encompass those functions. Also, Kleinberg's axioms can be made into a system with no inconsistency. They present a feature called Cluster Quality Measure(CQM) which when given an information set and it is shaped into any number of groups , CQM gives a measure of how solid or weak the cluster depends on the score it gives. CQM deals with the same standards gave by Kleinberg. Also, numerous CQM's have been suggested that work on the same principles or axioms provided.

Subsection 2:
This section defines the notations used for the clustering-quality measure (CQM) such as the distance d and the cluster denoted C with k partition as $C = C1, C2, ..., Ck$.and since we are dealing with the pairwise distances for the computation of cluster's quality, it also discusses the distance metric properties

such as [1]Positivity, [2]Symmetry and [3]Triangle Inequality.

Subsection 3:
This section reiterates over the previous paper of Impossibility and the three major concepts such as:
[1] **Scale Invariance**: function will not change with the units of distance measure. Axiomatically, stated as $f(d) = f(\alpha * d)$ where α is the constant and $f(d)$ is clustering function.
[2] **Richness**: Each point by itself is a cluster i.e. Range of the clustering function $f(d)$ is equal to the number of Γ in the set S.
[3] **Consistency**: For any clustering function f, if we transform the existing cluster distances within the cluster and between the cluster, the resulting function f' arises from the transformation or f' is the consistent variance of transformation.
And subsequently defines the impossibility theorem that only TWO of the three properties of clustering function discussed above will be held true by any clustering method.

Subsection 4:
These axioms are similar to the Kleinberg's axioms but are more from focused towards the clustering quality of the clustering method itself. Author compares these CQM axioms on the scale of soundness and completeness. However, author points out that it is not easy to define soundness and completeness for the clustering as we do not have clear definition for clustering.

Soundness and Completeness of Axioms:
Soundness implies that each component of the class fulfills all axioms and completeness implies that each property shared by all objects of the class is inferred by the axioms. **Isomorphism Invariance**:
This axiom of clustering quality measures does not have a relating Kleinberg property. The arrangement of properties comprising of Isomorphism Invariance, Scale Invariance, Consistency, and richness is a reliable arrangement of axioms. They say that Relative margin quality measure fulfills every one of the four axioms.
Now we are acquainted with two novel QCMs; a measure that infers and discusses linkage based clustering, and a measure for center based clustering. Linkage-based clustering implies that at whatever point a couple of points offer the same cluster they are associated by means of a tight chain of points in that cluster. The time taken for a clustering measure to figure the quality is polynomial time is n.

Subsection 5:
To muddle over this issue, paper recommends to replace soundness by the requirement that all three axioms should be satisfied by the set of clustering functions in study.

Using these axioms, author points out that Kleinberg's axioms fails to successfully prove that there are some naturally occurring algorithms which satisfies two axioms. Furthermore, author expounds equations on the quality measures to incorporate Clustering Isomorphism and Isomorphism Invariance.

In the next section of the paper, author tries to validate these axioms to formulate a mathematical formula for computing the CQM value. These axioms are validated using examples of Weakest Link, Additive Margin, and Computation complexity. Author also states that, time needed to compute the clustering quality measure is of very high importance in terms of the clustering technique as this will be very useful in comparing various set of clustering techniques. And when set of center is given time required by relative or additive Margin is $Orderof : O(n^k + 1)$ however this complexity comes down to $Orderof : O(nk)$ and $Orderof : O(n^2)$ for Relative and additive models respectively.

Finally the author advocates on the clustering quality measures which are depends on the number of clusters for the centroid based clustering methods like kmeans and kmedians. Author also mentions in this section that the CQM that violates scale invariance and richness axioms can be avoided by using the scaling method known as Lnormalisation.

5 Reference

1. Introduction to Data Mining by Pang-Ning Tan (Author), Michael Steinbach (Author), Vipin Kumar (Author).