Vishvesvaraya Technological University

Interpreting students Feedback

Paridhi Gupta(1PE16CS107)
paridhi-gupta1998

paridhigupta555@gmail.com
Sanjana K Rao(1PE16CS142)

Booyah1898

sanjkrao@gmail.com
Naveen Naidu(1PE16CS106)
Supervisor: Gowri Srinivasa

Report contains the analysis of different clustering algorithms on the dataset

Department of Computer Science

May 3, 2019

Declaration

All sentences or passages quoted in this document from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure.

Name: _____

Signature: _____

Date: _____

Abstract

This project report analyzes the survey on student's feedback on current scenerio of teaching.This report is based on the dataset taken from
http://archive.ics.uci.edu/ml/datasets/turkiyestudentevaluation
The analysis was found necessary to answer questions like which teacher is not giving their 100 percent and how many students are actually satisfied with what they are getting in institution.

Chapter 1

Understanding of data

1.1 About Dataset

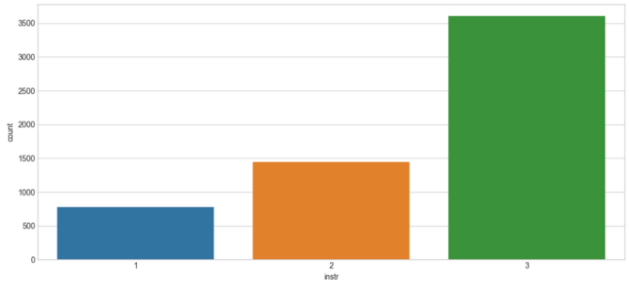
5 head attributes and 28 questions asked in survey for assesment of quality of classes. instr: Instructor's identifier; values taken from 1,2,3
class: Course code (descriptor); values taken from 1-13
repeat: Number of times the student is taking this course; values taken from 0,1,2,3,...
attendance: Code of the level of attendance; values from 0, 1, 2, 3, 4
difficulty: Level of difficulty of the course as perceived by the student; values taken from 1,2,3,4,5
Q1: The semester course content, teaching method and evaluation system were provided at the start.
Q2: The course aims and objectives were clearly stated at the beginning of the period.
Q3: The course was worth the amount of credit assigned to it.
Q4: The course was taught according to the syllabus announced on the first day of class.
Q5: The class discussions, homework assignments, applications and studies were satisfactory.
Q6: The textbook and other courses resources were sufficient and up to date.
Q7: The course allowed field work, applications, laboratory, discussion and other studies.
Q8: The quizzes, assignments, projects and exams contributed to helping the learning.
Q9: I greatly enjoyed the class and was eager to actively participate during the lectures.
Q10: My initial expectations about the course were met at the end of the period or year.
Q11: The course was relevant and beneficial to my professional development.
Q12: The course helped me look at life and the world with a new perspective.
Q13: The Instructor's knowledge was relevant and up to date.
Q14: The Instructor came prepared for classes.
Q15: The Instructor taught in accordance with the announced lesson plan.
Q16: The Instructor was committed to the course and was understandable.
Q17: The Instructor arrived on time for classes.
Q18: The Instructor has a smooth and easy to follow delivery/speech.
Q19: The Instructor made effective use of class hours.
Q20: The Instructor explained the course and was eager to be helpful to students.
Q21: The Instructor demonstrated a positive approach to students.
Q22: The Instructor was open and respectful of the views of students about the course.

Q23: The Instructor encouraged participation in the course.
Q24: The Instructor gave relevant homework assignments/projects, and helped/guided students.
Q25: The Instructor responded to questions about the course inside and outside of the course.
Q26: The Instructor's evaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives.
Q27: The Instructor provided solutions to exams and discussed them with students.
Q28: The Instructor treated all students in a right and objective manner.

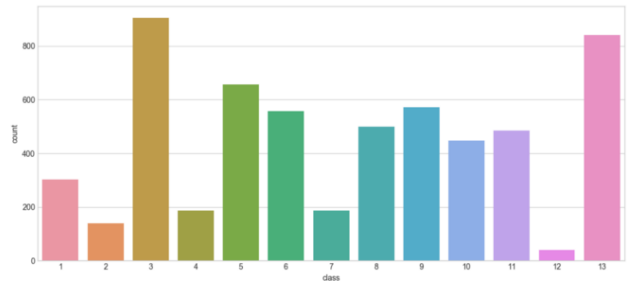
Q1-Q28 are all Likert-type, meaning that the values are taken from 1,2,3,4,5
Dataset link-http://archive.ics.uci.edu/ml/datasets/turkiye+student+evaluation

1.1.1 What can we infer from the data?

There are three instructors,each instructor teaches some specific number of students. According to the graph instr 3 teaches max number of the students.



But there can be variation of students in diffrent classes so,we cannot say that instr 3 teaches max number of classes . So,lets see graph between classes and count of student.



Calculation of means of each question asked is done and then mean of those means is taken which comes up to 3.1861.Each question's mean is near to 3 . Standard deviation is 0.1091,which is close to 0 ,therefore we can say that quality of the class is more or less is same according to all the questions.

1.1.2 how we used the data?

We considered only 28 questions as attributes because they have same (0-5) values and if we consider all attributes then it will be a complex task
Eg- we will have to consider student entry only if they have attended the class ,If not then the values for the questions are misleading clustering. Therefore assume only 28 ques as attributes .
there are 28 dimensions which is hard to visualize therefore, for visualizing data we used PCA ,but PCA reduces the quality of data so we got the result without using PCA also.
When we perform PCA to reduce the 28 dimensions to two dimesnsions we get values from -10 to 10.
pca components -

[-0.17872911 -0.18696044 -0.18218529 -0.18417011 -0.19021407 -0.18708119 -0.1878324 -0.18678649 -0.18239155 -0.19236264 -0.18669482 -0.1862382 -0.19227288 -0.19118139 -0.19023804 -0.19628846 -0.18088329 -0.19357879 -0.19273593 -0.19319851 -0.19110602 -0.19085911 -0.19483935 -0.19313335 -0.18889574 -0.19086936 -0.18975553 -0.18866989

0.23223503 0.11551155 0.24533527 0.20717759 0.20075314 0.24290761 0.24901578 0.12919618 0.1891172 0.1105148 0.21203229 -0.1061603 -0.15629705 -0.15533847 -0.04865705 -0.26259518 -0.1290584 -0.15363393 -0.19670072 -0.22007368 -0.22347198 -0.10278122 -0.06210583 -0.20787213 -0.12045026 -0.07204025 -0.21401477
]

Chapter 2

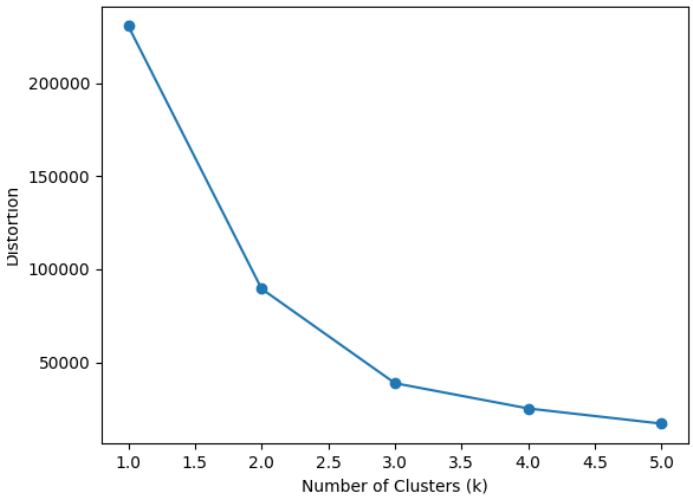
K-means

2.1 How Algorithm works?

Pick the number of cluster. Lets call this number k.
Randomly pick k observations as initial centroids.
Assign each observation based on the nearest centroid (used Euclidean distance).
Move the centroid to the center of the observations that were assigned to it.
Repeat the steps until the centroids position doesn't change or up to some iterative value.We took k ranges from 1 to 5.

2.2 Our work

We chose to use Kmeans++ version of Kmeans which initializes the centroids far from each other therefore, making convergence faster.
We chose k from 1 to 5 range and then found mean square error for each k value. The value of k which gives the least squared error value is used to make a model.



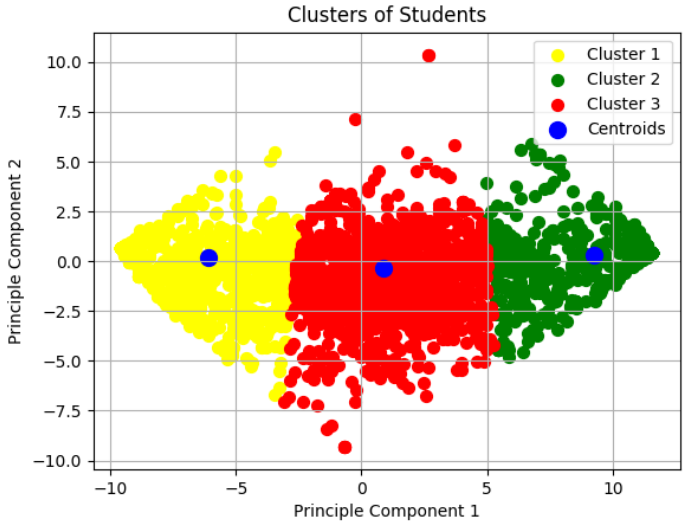
The graph for different k values is shown Here.

code:

```
sqerror= []
Krange = range(1, 6)

for i in Krange:
model = KMeans(
nclusters=i,
init='k-means++',
njobs=-1,
randomstate=1)
model.fit(Xpca)
sqerror.append(model.inertia)
```

Again the steps are repeated ,but with value of k as 3.
Model is fitted with xpca values(2d array which we get after using pca for reduction of 28 to 2 dimensions). 3 clusters represents -Satisfied ,Dissatified or neutral students. Here is the graph for 3 cluster where x and y are PCA dimensions.



Counter(cluster 3: 2358, cluster 2: 2222, cluster 1: 1240)
For NO PCA we will use X which has all 28 attributes with all 5082 rows as a parameter in model.fit(X)
this gives us Counter(3: 2358, 2: 2222, 1: 1239).

2.3 Conclusions

We can infer that 1239 students are not satisfied.If we dig in more, then we can say that students which has instructor 3 and for which cluster is "1" are not staisfied ,the number of

such students are 868 ,which is more than half of the 1239 ,that means instr 3 is not doing his job properly.

Chapter 3

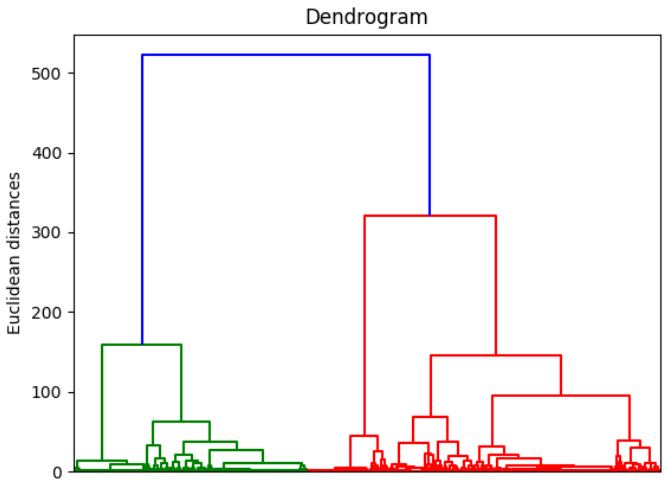
Agglomerative

3.1 How Algorithm works?

The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity.
Its also known as AGNES (Agglomerative Nesting).
The algorithm starts by treating each object as a singleton cluster.
Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.
The result is a tree-based representation of the objects, named dendrogram.

3.2 Our work

We focused on the Ward's linkage algorithm.
Wards linkage will merge clusters that lead to minimum increase of Sum of Square Error (SSE). The code below will generate the following dendrogram



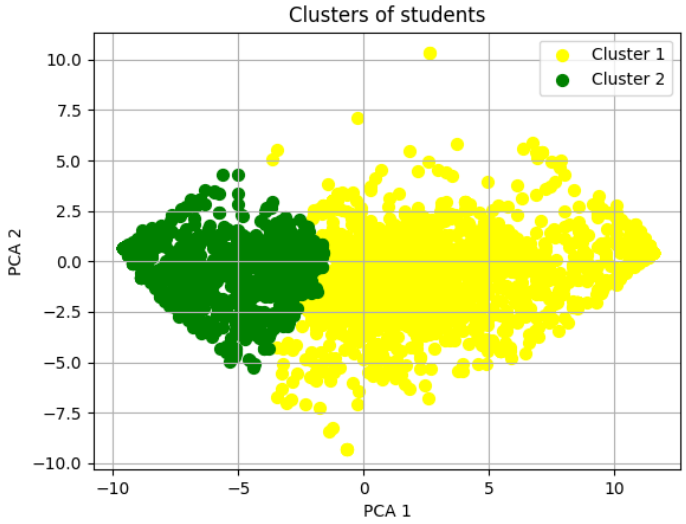
The graph for different k values is shown Here.

code:
dendrogram = hier.dendrogram(hier.linkage(X_pca,method = 'ward'))
plt.title('Dendrogram')
plt.xlabel('questions')
plt.ylabel('Euclidean distances')
plt.show()

Notice that there are two colours : green and red (exclude blue, which is the root).This mean the best k number of clusters according to Wards linkage (method = ward) is two.

The best k number of clusters will be different if we change the linkage algorithm. On fitting the model with the best k we found earlier :

code:
model = AgglomerativeClustering(n_clusters = 2, affinity ='euclidean', linkage ='ward')
y = model.fit_predict(X_pca)
Around 60% students belong to the first cluster. Similar to K-Means implementation, we are going to check the difference between the two clusters by checking their mean. Here is thescatter plot of the clusters.



Mean of cluster 1 : 2.44248184711,STD :0.829040130493 Mean of cluster 2 : 4.30968815481,STD :0.480492591784.

3.3 Conclusions

We prefer the result of K-Means with k=3 from Elbow method because having three different groups seems to be more meaningful (Dissatisfied, Neutral, and Satisfied). Smaller number of clusters (Dissatisfied/Neutral Satisfied) from the Wards linkage lead to a larger gap between the mean of each cluster.

Chapter 4

Meanshift Algorithm

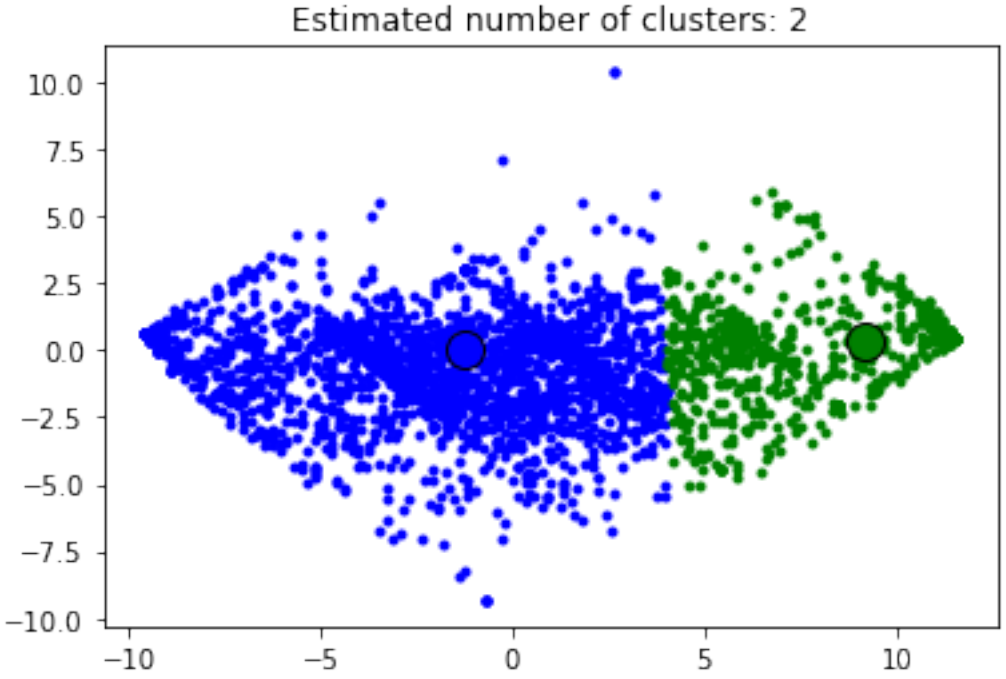
4.1 how algorithm works?

Mean shift clustering is a sliding-window-based algorithm that attempts to find dense areas of data points. It is a centroid-based algorithm meaning that the goal is to locate the center points of each class, which works by updating candidates for center points to be the mean of the points within the sliding-window. These candidate windows are then filtered in a post-processing stage to eliminate near-duplicates, forming the final set of center points and their corresponding groups.
Step-1 - We begin with a circular sliding window centered at a point C (randomly selected) and having radius r as the kernel. Mean shift is a hill climbing algorithm which involves shifting this kernel iteratively to a higher density region on each step until convergence.
step-2 - At every iteration the sliding window is shifted towards regions of higher density by shifting the center point to the mean of the points within the window. The density within the sliding window is proportional to the number of points inside it. Naturally, by shifting to the mean of the points in the window it will gradually move towards areas of higher point density.
step-3 - We continue shifting the sliding window according to the mean until there is no direction at which a shift can accommodate more points inside the kernel. Check out the graphic above; we keep moving the circle until we no longer are increasing the density.

4.2 Analysis

We chose gaussian kernel because it gives more smooth and accurate results.In my report bandwidth refers to the width of the kernel.
First to use the sklearn means shift ,we have to calculate the bandwidth for our dataset.Accuracy of the dataset depends on the bandwidth.So sklearn provides estimate bandwidth method which we used. After pca dimensional reduction bandwidth estimated was 4.7626742338772905 and with X (i.e all 28 ques and 5820 records) we got bandwidth as 6.185285800070237.Then,this bandwidth was feeded to train the model .
for PCA we got 2 clusters with counts- Counter(0: 4468, 1: 1352).
for X we got 3 clusters with count-Counter(0: 4456, 1: 1363, 2: 1).

This algorithm gives us set of dissatisfied and satistisfied students.
output for mean and std deviation were
Mean cluster 1 : 3.693582938994754,STD :0.7472712892223174
Mean cluster 2 : 1.5092455621301768,STD :0.5142820734247068
Graph shown below shows the clusters:



4.3 Conclusion

kmeans gives more deeper view than mean shift in this case but here assumption of number of clusters is not done like how we have to do in kmeans.
More than 70 percent of data points lies in one cluster.
This merges satisfied and neutral students together,which makes it look like more students are satisfied which in actual is not correct.

Chapter 5

Conclusions

We used 3 algorithms to analyze the data and it depends ,if the user want the clustering to be in more detailed one then ,he should go for k -means which gives the cluster of neutral students as well.
I chose mean shift algorithm because its a simple and flexible clustering algorithm and the main advantage of mean shift over k means is we dont have to know the number of clusters beforehand.There is a drawback of this algorithm i.e. to choose a proper bandwidth.
But for our dataset mean shift algorithm is not able to show its magic ,but personally I like this algorithm and its working so we used it.
Link for the git hub repository- paridhi-gupta1998/student_review_analysis

Bibliography

Appendices

Appendix A

An Appendix of Some Kind

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Appendix B

Another Appendix

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.