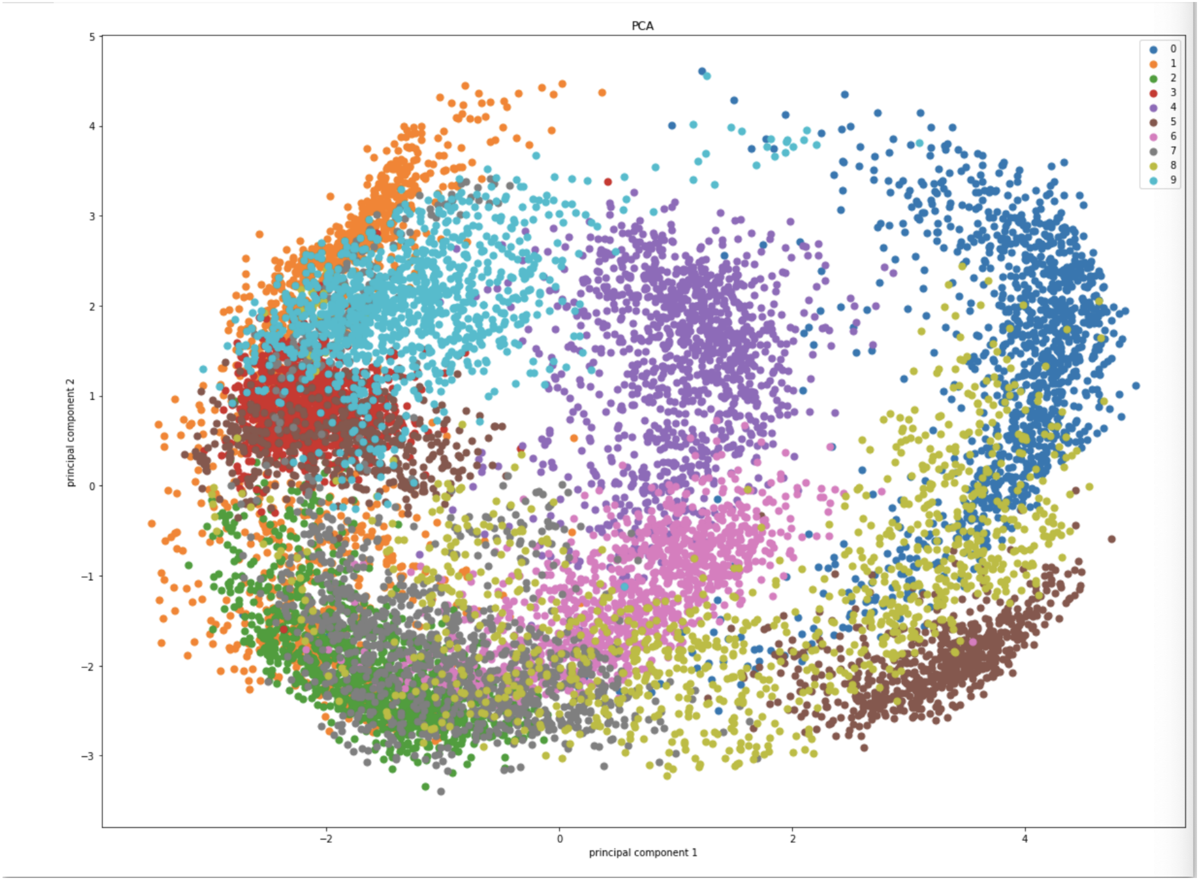
Terry Shipman
tshipm5

Homework 3
CSC 3730 Machine Learning

11-05-2020

Problem 1: [6 pt] Programming: Dimension reduction

Show the figure in your report. How do you interpret the results? Are digits ‘0’, ‘2’ well separated from each other? Are digits ‘4’ and ‘6’ well separated from each other?



i. Interpretation: By doing the PCA dimensional reduction a data set of 16 features has been reduced to the 9 key features for hand writing analysis. While these features have not been labeled, reducing the number of dimensions in the data set has allowed for some observations to be made. By knowing that this is handwriting samples and each target being a number, it it can be inferred that each class is that number. When looking at the variance in the principal components, numbers 0, 5, and 8 have the highest values on principal component 1’s axis. Numbers 1, 2, 3, 7 and 9 have the lowest values on principal component 1’s axis. Numbers 4 and 6 are toward the center of the principal component 1 axis.

Comparing the principal component 2 axis has numbers 0, 1, 3, 4, and 9 are in the positive area of principal component 2 and numbers 2, 5, 6, 7, and 8 are in the negative area of the principal component 2’s axis.

My interpretation is that this plot maps the shape of the figure (round, has straight lines, curves, etc) and the distance to the edge of the sample area.

ii. The digits ’0’ and ’2’ are separated from each other. There is some sparse intermingling of 0 with 2, but it is minimal.

iii. The digits ’4’ and ’6’ are not well separated and there is intermingling of the data points at the edges of these two classes.

Problem 2: [4 pt] Association rules

What is the confidence of the rule B → J?
Answer: {B}: 3 baskets, {BJ}: 1 basket, B → J *Confidence:* 1 ÷ 3 = 1/3

What is the confidence of the rule M → P?
Answer: {M}: 4 baskets, {MP}: 3, M → P *Confidence:* 3 ÷ 4 = 3/4

Problem 3: [10 pt] Programming: Association rule mining.

a. [5pt] Using the default Apriori parameters, what are the top 10 rules?

Answer:
Best rules found:
1. biscuits=t frozen foods=t fruit=t total=high 788 → bread and cake=t 723 <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)
2. baking needs=t biscuits=t fruit=t total=high 760 → bread and cake=t 696 <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)
3. baking needs=t frozen foods=t fruit=t total=high 770 → bread and cake=t 705 <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)
4. biscuits=t fruit=t vegetables=t total=high 815 → bread and cake=t 746 <conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)
5. party snack foods=t fruit=t total=high 854 → bread and cake=t 779 <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)
6. biscuits=t frozen foods=t vegetables=t total=high 797 → bread and cake=t 725 <conf:(0.91)> lift:(1.26) lev:(0.03) [151] conv:(3.06)
7. baking needs=t biscuits=t vegetables=t total=high 772 → bread and cake=t 701 <conf:(0.91)> lift:(1.26) lev:(0.03) [145] conv:(3.01)
8. biscuits=t fruit=t total=high 954 → bread and cake=t 866 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(3)
9. frozen foods=t fruit=t vegetables=t total=high 834 → bread and cake=t 757 <conf:(0.91)> lift:(1.26) lev:(0.03) [156] conv:(3)
10. frozen foods=t fruit=t total=high 969 → bread and cake=t 877 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92)

b. [5pt] Now change the lowerBoundMinSupport to 0.5, and you will have the following parameters: Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.5 -S -1.0 –c -1, re-run the algorithm, what are the top rules? Why?

Answer: *There were no Best Rules found for the lowerBoundMinSupport = 0.5.*
By increasing the lower bound minimum support from 0.1 to 0.5, the starting occurrences to match sets of items has increased. Instead of matching at a lower threshold where 10% would allow more items to be included, the set of items have to match 50% of the time to be considered. This limits the number of items that will be considered to pair associations. By setting this data set with the lower bound minimum support at 0.5, this eliminated matching of items with fewer associations.

Problem 4: [10 pt] Recommendation systems

a) [2 pt] Can you briefly describe what is content-based recommendation and what is collaborative filtering?

Content-based recommendation: A recommendation system where the user will be recommended items similar to the ones the user preferred in the past. This is used in information retrieval and information filtering.
Collaborative filtering A recommendation system where the user will be recommended items that people with similar tastes and preferences liked in the past.

If we know a user likes movie ”50 first dates”, which one to recommend ”The wedding planner” or ”Kill Bill”? Is this content-based or collaborative?

Answer: If it is known that the user likes the movie ”50 first dates”, it would be a fair assumption that the user likes romantic comedies. The recommendation would be the movie ”The wedding planner”. This would be an example of *content-based recommendation*.

b) [8 pt] Predict user John’s rating on item b using user-user collaborative filtering. Key steps: subtract mean ratings for John, Alice and Bob respectively. Compute cosine similarities between (John, Alice) and (John, Bob), and the weighted average of (Alice,b) and (Bob,b).

Answer:

Mean Ratings:
John: $\frac{12}{4} = \textit{3}$ Alice: $\frac{9}{4} = \textit{2.25}$ Bob: $\frac{6}{4} = \textit{1.5}$ b: $\frac{4}{4} = \textit{1}$

Cosine Similarities:
Sim(John,Alice) =

$$\frac{(2\times\ 2.75)\ +\ ((-3)\times\ 0.75)\ +\ (2\times\ -2.25)\ +\ ((-1)\times\ (-1.25))}{\sqrt{2^2\ +\ (-3)^2\ +\ 2^2\ (-1)^2}\ +\ \sqrt{(2.75)^2\ +\ (0.75)^2\ +\ (-2.25)^2\ +\ (-1.25)^2}} = \frac{0}{16.25} = \textit{0}$$

Sim(John,Bob) =

$$\frac{(2\times\ (-1.5))\ +\ ((-3)\times\ (-0.5))\ +\ (2\times\ (-1.5))\ +\ ((-1)\times\ (-3.5))}{\sqrt{2^2\ +\ (-3)^2\ +\ 2^2\ (-1)^2}\ +\ \sqrt{(-1.5)^2\ +\ (-0.5)^2\ +\ (-1.5)^2\ +\ (-3.5)^2}} = \frac{-1}{\sqrt{17.49}} \approx \textit{-0.06}$$

Sim(Alice,b) =

$$\frac{(2.75)\times(-1)\ +\ (0.75)\times(2)\ +\ (-2.25)\times\ 0\ +\ (-1.25)\times(-1)\ +}{\sqrt{(2.75)^2\ +\ (0.75)^2\ +\ (-2.25)^2\ +\ (-1.25)^2}\ +\ \sqrt{(-1)^2\ +\ 2^2\ +\ 0^2\ +\ (-1)^2}} = \frac{0}{9.41} = \textit{0}$$

Sim(Bob,b) =

$$\frac{(-1.5)\times(-1)\ +\ ((-0.5)\times\ 2\ +\ (-1.5)\times\ 0\ +\ (-3.5)\times\ (-1)\ +}{\sqrt{(-1.5)^2\ +\ (-0.5)^2\ +\ (-1.5)^2\ +\ (-3.5)^2}\ +\ \sqrt{(-1)^2\ +\ 2^2\ +\ 0^2\ +\ (-1)^2}} = \frac{4}{10.10} \approx \textit{0.40}$$

Weighted Averages:

weight(Alice,b) = $\frac{0}{0\ +\ |0.40|} = \textit{0}$

weight(Bob,b) = $\frac{0.40}{0\ +\ |0.40|} = \textit{1}$

Prediction of John’s rating of item b:

$Ratings\_Prediction(John, b)\ =\ 3\ +\ (0\ \times\ (3-2.25)\ +\ (1\ \times(1\ -\ 1.5)$
Ratings_Prediction(John,b) = *2.5*

Problem 5: [20 pt] Programming: Implement a simple recommender system for video games. Load the training data (‘rating.txt’), where each row contains a triple (user_id, item_id, rating).

a) [5] What is the average rating of user U845167387?

Answer: The average rating for U845167387 is 4.1.

b) [10] For each user, a set is defined as his/her liked items (ratings>3.0). Jaccard similarity between two users is the similarity between two sets of liked items. E.g., each user has rated several items with ratings. (‘U1’,’I43’,4.0), (‘U1’,‘I24’, 3.0),(‘U1’,‘I12’,1.0), (‘U2‘,’I43’,5.0), (‘U2‘, ‘I65’, 4.0), (‘U2‘, ‘I87’,2.0). The liked item set for U1 is ‘I43’, the like item set for U2 is ‘I43’, ‘I65’, the Jaccard similarity between U1 and U2 is thus 0.5.
Find users/neighbors whose Jaccard similarity to U845167387 is larger than (>0.2). (If there are multiple users satisfying the condition, list all of their ids).

Answer: These are the neighbors that the Jaccard similarity is larger than 0.2 compared to ’U845167387’:

U645280931, U332042974, U746401474, U597403119,
U065309156, U386262628, U000268981, U992842409,
U317796054, U634705862, U201272639

c) [5] Recommend games to the user U845167387. Given the neighbors computed in b), what are the game ids that are liked by at least one of the neighbors?

Answer: The following are the game ids that are liked by at least one of the neighbors.

I981874545 I822939121 I792989869 I232939121