CS57300: Homework 2

Due date: Wednesday February 18, midnight (submit via turnin)

Submit your homework in two files: one pdf containing your answers to the questions (with any associated plots), and one text file containing the code that you used for analysis (this can be R code or python code). Your homework must be typed. Use of Latex is recommended for the results, but not required.

In this assignment, you will use R (and optionally python) to explore, transform, and analyze the Yelp data you started to use in HW1. Based on your analysis you will formulate hypotheses about the data.

1 Principal Component Analysis (6 pts)

Consider the subset of the Yelp data comprised of the 35 numeric attributes.

(a) Run principal component analysis on the data.

(b) Plot the scree plot. Identify what number of components are needed to explain more than 95% of the variance in the data.

(c) Inspect the weights for the first principal component and identify how many of the 35 attributes have a significant weight in this component.

(d) Transform the data by removing the original column for *review_count* and replace it with a new column containing log-transformed values of *review_count*. Repeat the above analysis and discuss what if any changes you see in the results.

(e) Sample a random set of 100 examples from the original data. Repeat the above analysis and discuss what if any changes you see in the results.

2 Scoring and search (12 pts)

Consider the subset of the Yelp data with only the *review_count* and *tip_count* attributes.

(a) Run principal component analysis on the data. Report the eigenvector values (i.e., component weights) in the solution returned by R.

(b) Develop your own algorithm to search over possible basis vector solutions to choose the one with "best" score.

Recall that for p-dimensional data, a basis vector solution will be a set of p orthogonal p-dimensional vectors, each of norm 1. The eigenvector solution from applying PCA is one of the possible basis vector solutions—the one that maximizes the variance of the data along each dimension.

Since your data for this question is 2-dimensional, you will need to search for two 2-dimensional basis vectors: $b_1 = [v_1, v_2]$, $b_2 = [v_3, v_4]$. However, since the p^{th} basis vector is constrained by the solutions for the $[1, p-1]$ vectors, you will only need to search for the values in first basis vector (i.e., b_1). Moreover, since the basis vector must have a norm of 1, you will only need to search over the first value for the vector (i.e., v_1).

- Mean center your data.
- Using a step-size of 0.05, consider a search over the range [-0.95,-0.90, ... , 0.90, 0.95] for possible values of v_1.
- For each possible value of v_1, calculate a positive value for v_2 that constrains the basis vector $b_1 = [v_1, v_2]$ to have a norm of 1. (Note that searching over positive and negative values for v_1 and only positive values for v_2 will cover all directions.)
- For each choice of $b_1 = [v_1, v_2]$, project the mean-centered data onto the vector and calculate the PCA score function (i.e., the variance of the data when projected onto b_1).
- Plot the score as a function of v_1 and identify the solution ($b_1 = [v_1, v_2]$) with the best score. Compare it to the eigenvector solution returned by R and discuss any differences.

3 Transformations and associations (16 pts)

Consider the binary feature construction that you did in HW1 (e.g., Nightlife vs. not-Nightlife). In this question, you will construct binary features for values in the *category* and *city* attributes.

(a) Extract all the unique values in the *category* attribute by parsing the comma-separated lists (e.g., "`Mexican, Restaurants`" → two values, one for `Mexican` and one for `Restaurants`). Sort the list of values and choose the top 30. Construct binary features for each of these 30. (Note: you should figure out how to do this in a loop or a function, do not do it manually!)

(b) Repeat the same process of binary feature construction for the *city* attribute, but this time use the top 30 most frequent cities in the data (i.e., reverse sort by number of examples in the city). Note: you do not need to parse this attribute.

(c) For each pair of binary features (*category* vs. *city*; 30 × 30 pairs), determine whether there is any association by calculating χ^2 scores (using `chisq.test`) from a contingency table of counts, e.g.:

	City i	
Category j	0	1
0	N_{00}	N_{01}
1	N_{10}	N_{11}

Report the top five features combinations with the largest χ^2 scores, along with assessments of significance (i.e., p values), and discuss whether the correlations are interesting or expected, given your domain knowledge. *Note: If the values in the contingency table result in an undefined χ^2 score (e.g., when a column or row has all 0s), then just return a default value of 0 for the score.*

(d) Consider the feature pair with largest χ^2 score (let's call this pair A^{max}) and another feature pair with a score that is barely significant (i.e., A^{good} with p-value ≈ 0.05). Investigate the effect of sampling on the scores of these feature pairs.

- Repeat ten times:
  - Create ten random samples of the following sizes: $[16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192]$.
  - Calculate the χ^2 scores for A^{max} and A^{good} on each sample.
- Calculate the mean and standard deviation of the scores for each feature pair, for each sample size.
- Plot the χ^2 scores as a function of sample size. Your plot should include one curve for A^{max} and one curve for A^{good} and include error bars to show the standard deviation.
- Discuss the results. What effect does sample size have on significance? Does the effect vary across the two attributes?

4 Identifying hypotheses (6 pts)

The *stars* attribute corresponds to a rating for the business. The *review count* attribute records the number of reviews/ratings that the business received. Investigate how the binary features you created for the *city* and *categories* attributes, as well as the *latitude*, and *longitude* attributes relate to these two *stars* and *review count* attributes. Identify two hypotheses about the relationships between the features (one for *stars* and one for *review count*). For each of your hypotheses:

(a) Identify the type of hypothesis (descriptive vs. relational vs. causal; direction vs. non-directional).

(b) State the hypothesis and discuss how your analysis of the data led you to the conjecture.

(c) Include a plot to support your hypothesis.

Submission Instructions:

After logging into data.cs.purdue.edu, please follow these steps to submit your assignment:

1. Make a directory named $'yourName\_yourSurname'$ and copy all of your files there.
2. While in the upper level directory (if the files are in /homes/neville/jennifer_neville, go to /homes/neville), execute the following command:

   ```
   turnin -c cs57300 -p HW2 your_folder_name
   ```

   (e.g. your prof would use: turnin -c cs57300 -p HW2 jennifer_neville to submit her work)
   Keep in mind that old submissions are overwritten with new ones whenever you execute this command.

   You can verify the contents of your submission by executing the following command:

   ```
   turnin -v -c cs57300 -p HW2
   ```

   Do not forget the -v flag here, as otherwise your submission would be replaced with an empty one.

Your submission should include the following files:

1. Your answers & analysis in .pdf format.
2. Your source code in R and/or python.