OkCupid Assignment

Son Do
Princeton University
sqdo@princeton.edu

Max Land
Princeton University
mland@princeton.edu

Abstract

Predicting if two people are a match is not easy. Yet, that has not discouraged researches from trying. With the increasing popularity of dating apps such as OKCupid, there is plenty of data to create models to predict matches. However, it should be noted so much more can be done with the data. Dating profiles contain a plethora information– demographics, personal preferences, etc. Thus, in this assignment we take on the OKCupid data set consisting a collection of 59,946 adult dating profiles and challenge ourselves to identify underlying patterns and connections among features. With a focus on user response to essay questions, we perform latent variable analysis using incremental PCA and K-means clustering. We find there are certain clusters of words that are mostly spoken by a single age group. We also discover there are words that are prevalent among different topics in LDA.

1 Introduction

Online dating is becoming more and more prevalent in our day and age. 15% of American adults have used online or mobile dating apps, including 27% of adults aged 18-24. Most apps require users to create dating profiles. Dating profiles often contain demographic information as well response to essay questions. Then, users can view other peoples' profiles and "match" with other users they prefer. As a result, there is a wealth of data available to explore and identify interesting patterns as well as make predictions based of a users dating profile.

In this report, we investigate OKCupid data. The data consist of past public profiles of 59,946 OKCupid users. These users lived within 25 miles of San Francisco, had active profiles on June 26, 2012, were online at some point after June 26, 2011, and had at least one picture in their profile. Data were scraped from users' public profiles on June 30, 2012. Each user had 31 variables or features makeup their dating profile. We focus on the user response to essay questions as features to discover latent variables (such as words and age). This allows us to analyze the correlation between age and certain user preferences to determine the best focus for researchers interested predicting match making moving forward. To do so, we used K-Means Clustering and LDA.

2 Related Work

Before starting this assignment, we recognized that the section of the OKCupid data set we were working with was both high dimensional and sparse. We wanted to focus on the essay questions, because we thought there would be the most variance among user outputs and we wanted to analyze people's responses. Thus, we generated a bag- of-words representation of user responses to essay questions. In a paper on PCA [3], the author mentioned how a bag-of-words representation of a text will typically be a very sparsely populated vector living in a very high-dimensional space. Furthermore, the paper showed PCA was effective in reducing the dimension of the feature space while retaining as much information as possible. The paper also compared the performance of PCA vs LSA and found that PCA performed much better. This motivated us to use PCA to perform feature engineering.

However, after attempting to apply PCA to our bag-of-words, we quickly found that PCA was extremely slow and inefficient– our computers ran out of memory. This is due to the sheer size of our bag-of-words which resulted in large, high dimensional vectors. So, we looked for alternative solutions. In a paper on Incremental PCA (IPCA) [2], it found that IPCA had the following advantages: low memory demand, low computational complexity, and had better recognition accuracy and less execution time then normal PCA. Thus, we used IPCA in this assignment.

3 Methods

3.1 Data processing

We are given the dating profiles of 59,946 OKCupid users. For each user, 31 variables are available, including typical user information (such as sex, sexual orientation, age, and ethnicity), lifestyle information (such as diet, drinking habits, smoking habits), and text responses to the 10 essay questions posed to all OkCupid users. For this assignment, we generate a bag-of-words representation of all 10 essay questions using our own script. We ignore empty essay responses. This resulted in a bag-of-words of around 29,000. We randomly split the data 80/20 into train/test set (80% train, 20% test). When we were doing LDA we created a bag of words that simply counted words that appeared more than 10 times in an essay question. For the K Means clustering, we used a tf-idf bag of words representation in order to better represent text as tf-idf lowers the weight of words that appear often such as "a" or "and" as they tell little about the text. The formula for the tf-idf vector weights are:

$$\text{tf-idf}(x_i) \triangleq [\text{tf}(x_{ij}) \times \text{idf}(j)]_{j=1}^{V}]$$

3.2 Latent Variable Models

We used two different models from the SciKitLearn Python libraries. The two models are as follows:

1. **K Means Clustering** (KMeans): K Means is an algorithm that is broken into 3 steps. The first is setting each sample to a random cluster (n different clusters). The second is finding the mean of each of the clusters and then (3rd step) assigning each sample to a new cluster that is closest to that cluster. Step 2 and 3 is repeated until there is no change in cluster assignments.
2. **Latent Dirichlet Allocation** (LDA): Described below in our spotlight Method.

3.3 Evaluation

In order to evaluate our methods you use three different evaluation techniques. For Kmeans Clustering we used the Inertia to see which was the best means. For Incremental PCA, we used the reconstruction error. And finally for LDA, we used the log-likelihood to evaluate the strength my findings. We go in more detail for each method below.

Log-likelihood in a nut shell is a way to estimate if the parameters for a given model is the best. We use this to ensure that the parameters of our LDA will give us the best outcomes. Log-likelihood is caluculated by maximizing the "likelihood" of seeing a specific data given some parameters[1] . Log-Likelihood can be written as:

$$max_{\beta}\left\{\sum_{i} ln\{p(y_i|\beta)\}\right\} \tag{1}$$

KMeans Inertia:

KMeans Inertia is also known as the objective function. It is simply the sum of all the distances of each sample from its cluster mean. For this project we use the euclidean distance to calculate the distances between each user's answers and their cluster mean. u_j is the mean of the cluster j that x_i is apart of. We used this to find the best number of clusters to set the kmeans algorithm with. We found that 10 clusters was the best in describing the data set as it provided the smallest inertia. The objection function is:

$$\sum_{i-0}^{n} min\left\{||x_i - u_j||^2\right\} \tag{2}$$

4 Spotlight Classification Method: Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a type of a Probabilistic Topic Model. Topic models take an unorganized collection of documents as input and outputs a set of topics across the documents; for each document, the proportion of that document related to each topic. For reference, topic is defined to be a distribution over a fixed vocabulary.

In generative probabilistic modeling, we treat our data as arising from a generative process that includes hidden variables. This generative process defines a joint probability distribution over both the observed and hidden random variables. We perform data analysis by using that joint distribution to compute the posterior distribution – the conditional distribution of the hidden variables given the observed variables. For a LDA, the observed variables are the words of the documents; the hidden variables are the topic structure. We can see the generative model for LDA below:

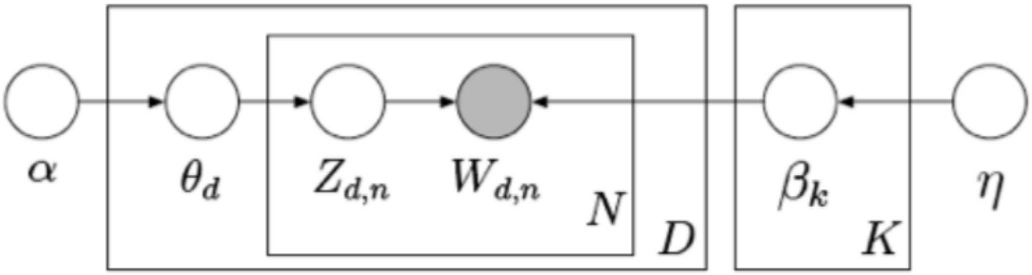


Figure 1: For n = 1: N words, d = 1 : D documents, k = 1: K topics
$\theta_d \sim Dir(\alpha)$ $\beta_k \sim Dir(\eta)$ $z_{d,n} \sim Mult(\theta_d)$ $w_{d,n} \sim Mult(\beta_{z_{d,n}})$

Given a collection of documents, we can infer:

- Per-word topic assignment $z_{d,n}$
- Per-document topic proportion θ_d
- Per-corpus topic distribution β_k

Then, we use posterior expectations to perform topic-based searches, document comparisons, document assignments, etc.

Written out, the per-document posterior distribution for LDA is as follows:

$$\begin{aligned} p(w_{d,1:N}, z_{d,1:N}, \theta_d|\alpha, \beta) &= p(\theta_d|\alpha)p(z_{d,1:N}, |\theta_d)p(w_{d,1:N}|\beta, z_{d,1:N}) \\ &= p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_{d,n}, |\theta_d)p(w_{d,n}|\beta, z_{d,n}) \end{aligned}$$

Overall, LDA assigns each word in a document to a topic according to the document-specific topic proportion, and draws from the word distribution of that topic. It is also a method to reduce a set of samples to a low dimensional latent space: a set of "topics". It trades off two goals:

- In each document, assign words to a few high probability topics.
- In each topic, assign high probability to a few terms.

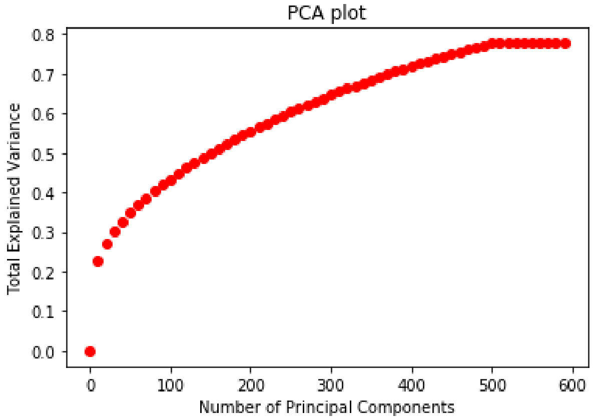
These goals are at odds with each other. Thus, trading off these goals results in recovering a few groups of commonly co-occurring words.

On separate note, LDA assumes that each word is independent, each topic is independent, and each document is independent. Because of the independence assumption implicit in the Dirichlet distribution, LDA is unable to capture the correlation between different topics.

5 Results

5.1 PCA

Prior to running PCA, we perform an analysis on the total explained variance over the number of principle components.



From the graph, we can see it levels off around 500 components. However, we determined that 10 components was the best by finding the best reconstruction errors to inertia ratio of the K-Means Clustering Algorithm. This is because from 0 to 10 components, there is the largest jump in total explained variance. The first 10 components explain a little over 20% of the total variance, whereas the next 490 explain 55% of the total variance. So, each individual component in the 10-component group explains a higher % of total explained variance than each individual component in the 500-component group. Furthermore, the a lower number of components was preferred for K-Means Clustering due the curse of high dimensionality. This is because K-Means Clustering implements Euclidean distance, which is skewed in high dimensions. This is substantiated by the fact that 10 components had the smallest K-Means inertia of: 1457.107.

After performing PCA, we report the PCA reconstruction error:

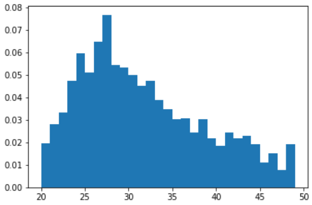
- PCA Train Diff: 0.70245

- PCA Test Diff: 0.70389

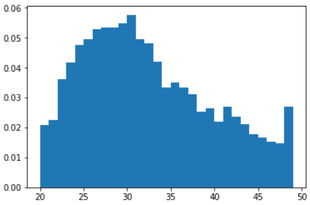
These reconstruction error are from doing PCA to find 10 Principal Components. We only put the lowest value of the Reconstruction error and inertia to conserve space.

5.2 K-Means Cluster

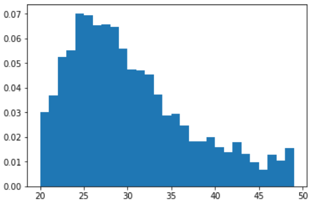
We calculate the K-Means clusters after running PCA, and then report the histogram of the age distribution of each cluster. The rest of the clusters can be found in the Appendix.



(a) K1: There are spikes on certain ages with the largest spike at age 28



(b) K3: There is a large spike at the 48 and greater age



(c) K9: Lower age group (age 20 and younger) is higher than all other clusters

K-Means Cluster	('WORD', TFIDF)
K0	('lacto', 25.17), ('pockets', 24.53), ('addicts', 24.22), ('junction', 23.50), ('kinks', 23.40)
K1	('007', 1.00), ('06', 1.00), ('14', 1.00), ('17', 1.00), ('18th', 1.00)
K2	('smilemusic', 21.18), ('witty', 21.01), ('roots', 20.77), ('grrrl', 20.44), ('magic', 20.22)
K3	('immune', 19.20), ('frank', 19.05), ('comes', 18.55), ('paleontology', 17.87), ('centric', 17.56)
K4	('name', 18.94), ('surgery', 16.83), ('discussed', 16.81), ('persistent', 16.81), ('yorke', 16.61)
K5	('lolnan', 23.67), ('kayaks', 23.63), ('eerily', 23.34), ('beekeeper', 22.98), ('promotion', 22.93)
K6	('mingling', 24.14), ('nanlooking', 23.54), ('soprano', 23.52), ('summarizes', 22.87), ('evocative', 21.29)
K7	('doings', 17.07), ('alli', 16.25), ('stilettos', 16.23), ('liberty', 15.73), ('pick', 15.66)
K8	('decompressing', 23.06), ('poets', 19.53), ('advancing', 18.70), ('godfrey', 18.36), ('csueb', 18.31)
K9	('htm', 14.79), ('guidelines', 14.74), ('exclusion', 14.20), ('comedies', 13.91), ('quantities', 13.90)

5.3 LDA

Next, we performed LDA analysis on the data with varying n-components and learning rates. We report the result with the lowest log-likelihood of -37003748.165. The following topics were achieved with 10-components and a 0.7 learning rate. The order of the words corresponds to the number of times it was assigned to each topic.

Topic	List of words
0	music movies like food books love favorite shows tv rock
1	new san francisco bay years friends city work area ve
2	strong em br com http www youtube target watch blank
3	fi nbsp sci psychology nannannannannannannannannannan fantasy action philosophy comedy computers
4	love life people amp world br art open things music
5	br music like food amp making things people love books
6	love like friends just im good music want family know
7	love friends good enjoy food family life music like new
8	br like don just really people things good think know
9	interests class href ilink br games music science rock star

6 Discussion and Conclusion

In this paper, we used three latent variable models to analysis the essay answers of each user in the OkCupid data set to see if we can learn more about the different users and their preferences. We choose to focus mostly on the differences between age.

When we did our K-means clustering we found that there were some clear trends in words with different age group and clusters. All the data was centered heavily around the age 25-30 group as the data set was centered around that certain age group. Thus, the majority of the essay answers would have been from that age group as well, swaying the clusters. The K1 cluster, were often for users that used words that had low tf-idf scores and didn't really belong in any of the other clusters. As a result, the corresponding histogram had a lot of sporadic distribution. Thus, we looked for spikes in the two outliers (under 25 and over 35). This best represented in the K9 and K3 clusters. These clusters had a higher percentage of their samples being in the outlier age groups. This means that older people (according to the K3 cluster) tend to use words such as "frank", "immune", or "paleontology" which I find pretty funny since we consider "frank" to be used much more in the older community. For the K9 cluster we see that the younger demographic often based there essays more on 'comedies' which we think is a much younger trait.

Looking at our LDA results, we see that many of topics share words in common i.e. "love", "friends", "food" to a pretty extreme extent. For example, "love" appears as the forefront of topic 4, 5, and 6 which means it was the most common in each of those topics. This suggests the individual topics are not that distinct and share a lot of overlap. This is consistent with the flaw in LDA highlighted in the spotlight classifier: LDA is unable to capture the correlation between different topics. This is because LDA assumes that each word is independent, each topic is independent, and each document is independent. From this, we can conclude the appearance of these words ("love", "friends", "food") across most clusters suggest that user responses are generic to some extent. In addition, each topic contains words that many would consider "common". For example, topic 7 contains: ("love friends good enjoy food family life music like new"). It is hard to intuitively label these words under a specific category. This further substantiates the fact user responses all share an underlying generic pool of vocabulary.

Overall, we found there are latent variables present in the essay responses, especially a correlation between age and vocabulary. In addition, we found users' responses aren't as unique as they think. Thus, in order to stand out and get a match, users might want to consider refraining from including such popular words.

6.1 Next Steps

We found that this project was gave us some great insight on different demographics of people, especially in relation to age. The data seems to defend our assumptions that there is an underlying correlation between age and vocabulary. Some further extension of this paper for those who are looking to extend our paper are:

1. Use different data sets that have information on people in other cities
2. See if other data from other dating online tools have similar results.

We recognize that San Francisco's population is skewed towards young adults and is not representative of the US as a whole, which is why we recommend the two extensions above.

References

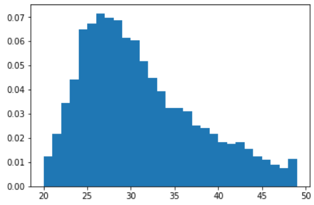
[1] Balaban, Jonathan. "A Gentle Introduction to Maximum Likelihood Estimation." Towards Data Science, Towards Data Science, 20 Feb. 2018, towardsdatascience.com/a-gentle-introduction-to-maximum-likelihood-estimation-9fbff27ea12f.

[2] I. Dagher, "Incremental PCA-LDA algorithm," 2010 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Taranto, 2010, pp. 97-101. doi: 10.1109/CIMSA.2010.5611752

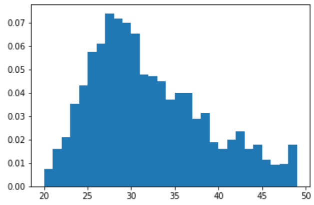
[3] Ljungberg, Benjamin Fayyazuddin. "Dimensionality reduction for bag-of-words models : PCA vs LSA." (2017). 2012, Pages 21-31, ISSN 0169-7439,

7 Appendix

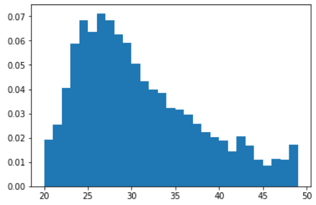
7.1 K-Means Cluster



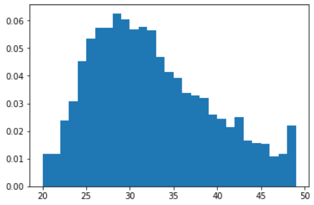
(a) K0: There is big cluster around the age 26, but it is much evenly distributed than other clusters



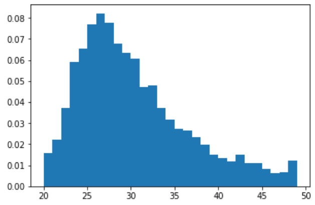
(b) K2: Most of the distribution is on the 25-30 age



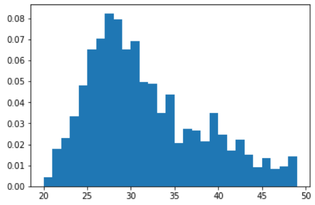
(c) K4: Most of the distribution is centered more heavily on the 20 to 25 age group



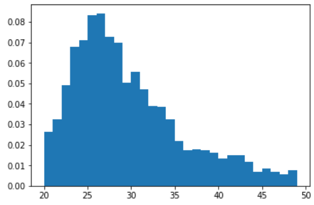
(a) K5: There a very even distribution compared to the other clusters



(b) K6: Most of the distribution is centered on the 25 to 30 age group again with a very small distribution at the older age (35 and up)



(c) K7: A very spiking distribution similar to cluster K1



(a) K8: This distribution is heavily swayed towards the 20 to 25 age group with a very low percentage in the over 35 age group