Statistical Foundations of Data Analytics

LAB 8: **Unsupervised Learning**

YOUR TEAM

Last name: Yuan
First name: Tian

Last name: Sam
First name: Greenes

Last name: Nicholas
First name: Toronga

Part I: Analytics in Medicine

Using principal component analysis to detect cancer.

In our journey through this course, we have applied analytics across many areas: Restaurants (Yelp), E-commerce (the App Store), Real Estate, Board Games, Movies, Politics, Dating, Weather prediction, and Sports (basketball). In the first part of this lab we will study medical analytics. Analytics is now helping doctors when diagnosing patients.

In the western world, cancer is one of the most destructive and nefarious affections of the 21st century. Most of us, if not all, know someone who has been diagnosed with cancer. And, among women, breast cancer is one of the most common types of cancer. But thanks to the advancements in medicine, breast cancer fatality has decreased astronomically.

As you know detecting cancer is quite difficult in the early stages because the human body is extremely complex, and many factors interact jointly when there is a disease, but machine learning and AI were born to tackle complexity. And thanks to Principal Component Analysis, doctors are now able to study all the human complexity and be more effective at detecting cancer.

In this lab, we will learn how doctors and hospitals are using PCA to diagnose cancer. I have gathered a dataset with information about approximately 1000 real breast cancer tissue diagnostics. Each observation includes the information about the characteristics of breast-tissue samples, and the outcome of the diagnostic.

Factors

- **Diagnosis:** The diagnosis of breast tissue tumor (M = malignant/cancerogenous, B = benign).
- **radius:** mean of distances from the center to points on the perimeter of the tissue.
- **texture:** higher values indicate lumpy texture. Lower values represent plain texture.
- **perimeter:** mean size of the core tumor.
- **area:** area of tumor.
- **smoothness:** mean of local variation in radius lengths.
- **compactness:** how compact is the tissue.
- **concavity:** Higher values indicate that the tissue is highly concave (as opposed to convex).
- **concave points:** Number of concave points in the tissue.
- **symmetry:** symmetry of breast tissue sample.

1. Unsupervised learning (5 points)

A) What is unsupervised leaning? How does it differ from supervised learning?

Your answer:

Unsupervised learning is when we only have access to a set of characteristics, and we try to get interesting findings about each characteristic. It differs from supervised learning because in unsupervised learning there is no response variable, and we are not trying to predict how each characteristic affects the response variable.

B) What are the challenges of unsupervised learning?

Your answer:

Unsupervised learning is fairly new, only having been around for a few years, and so it is not very deeply understood. Also, in unsupervised learning there is no clear task at hand, and therefore it is difficult to test the accuracy of our results.

C) What is unsupervised learning used for? List three potential applications.

Your answer:

Image processing

Face recognition

Human Genome

2. Principal Component Analysis (20 points)

We want to understand how to efficiently distinguish benign and malign (i.e., cancerogenous) tumors, in order to help doctors become better at diagnosing breast cancer.

To this end, we need to understand how the characteristics of a breast tissue correlate with each other. In our dataset, we have **9** factors describing the characteristics of each breast tissue sample.

A) If we wanted to use scatter plots to study how the "tissue variables" (9 of them) correlate with each other, how many scatter plots would we need to analyze these correlations?

Your answer:

(9*8)/2 = 36

36 scatterplots

B) What is the role of principal component analysis?

Your answer:

We use Principal component analysis to see how the data is structured, and to see the relationships between the characteristics. Principal component analysis finds a low dimensional representation of the data that captures variability as much as possible.

C) What are the principal components of these nine characteristics? How are these components calculated?

Your answer:

PCA looks for a small number of dimensions that capture variability of observations as much as possible. These are known as principal components. Components should be uncorrelated with previous components.

D) What are the loadings in a principal component? What would they describe?

Your answer:

The loadings are the weight attributed to each characteristic used to find the first principal component. The squares of the loadings must add up to 1. They describe the relative importance of each characteristic.

Suppose we have a dataset with n=8 observations and p=5 characteristics:

- {(5, 2, 3,4, 3), { 2, 7, 4,1,5}, {5,6,2,3,3}, {5,6,2,2,2}, {4,3,6,8,9},{1,9,2,7,8}, {3,1,2,4,3},{2,6,2,3,2}

E) How many principal components could be calculated in this dataset?

Your answer:

Min{n-1,p}

Min{7,5}

5 principal components

F) How do we find the first component in the above dataset with n=8 observations? Show the optimization equation that we are trying to solve (i.e., with the actual numbers):

Your answer:

To find the first principal component Z_1, we need to find the loadings for each variable. To do this, we maximize the weighted average of Z_1^2 over all 8 observations, subject to the square of each loading's weight having the sum of 1.

$Z_{11} = \phi_{11}*5$

$Z_{21} = \phi_{21}*2$

$Z_{31} = \phi_{31}*5$

$Z_{41} = \phi_{41}*5$

$Z_{51} = \phi_{51}*4$

$Z_{61} = \phi_{61}*1$

$Z_{71} = \phi_{71}*3$

$Z_{81} = \phi_{81}*2$

maximize $_{\phi 11,.., \phi 81}\{(1/8)*(\Sigma_{j=1...8} Z_{all}^2)\}$ subject to $\Sigma_{j=1...8}\ \phi^2_{j=1} = 1$

G) Find the first three principal components in the cancer_data dataset. Paste the output below:

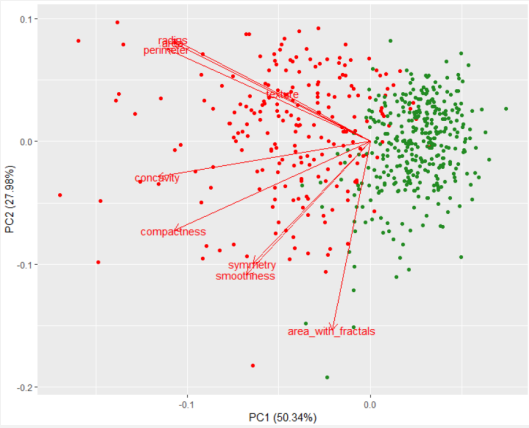
	PC1	PC2	PC3
radius	-0.40145560	0.3110165	-0.14068523
texture	-0.17743680	0.1455855	0.94723823
perimeter	-0.41462241	0.2816479	-0.13068215
area	-0.40155309	0.3019291	-0.13955781
smoothness	-0.25200939	-0.4037446	-0.17567296
compactness	-0.40011072	-0.2689709	0.04363335
concavity	-0.43292085	-0.1073958	0.02751865
symmetry	-0.23912193	-0.3702213	0.02043331
area_with_fractals	-0.07743304	-0.5723934	0.11159547

(H) Interpret the results of the first component (limit yourself to discussing the key things this component tells you).

Your answer:

The loadings of the first principal component tell us that most of the variability found across diagnoses can be explained by radius, perimeter, area, compactness, and concavity. In general, it seems as if “size” variables are what distinguish tumors from each other.

I) Plot the first two components using autoplot(). Make sure that, in the same plot, malign observations are in red, and benign observations are in green (please print in colour).



J) Interpret the PCA plot. Discuss the results, and what you've learned about benign and malign cancer tissues. Imagine you are informing a doctor who has little expertise about PCA (no jargon).

Commented [MOU1]: Why does it have to be two paragraphs in particular? Just seems very specific to me

Your answer:

From the PCA, tumors with larger sizes, including larger perimeter, area, radius, tend to have lumpy texture, therefore have a higher chance to be malign than those with smaller sizes and plain texture. Tissues which are smoother, symmetric are less likely to be malign tumors.

K) Are there any variables that are highly reliable indicators that a breast tissue is benign or malign? Based on the plot, is it easier to detect a benign or a malign tissue?

Your answer:

Texture, compactness and concavity are highly reliable as there are many dots, which indicates the observations, surround the arrows. Based on the plot, it is easier to detect a malign tissue because nearly all the arrows and variables are indicating the prediction of malign tumors (red part). The plot is mostly telling us which characteristics a malign tissue is likely to have. By contrast, there are very few arrows help us conclude the characteristics of a benign tissue.

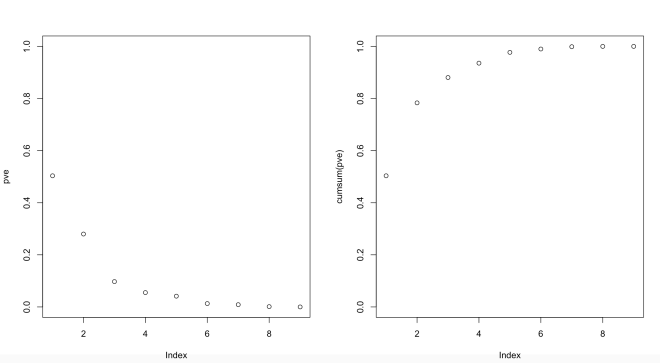
L) If you were going to use a classification technique (e.g., logistic regression or LDA), which variables would you include as main predictors of a breast tissue diagnosis?

Your answer:

Radius, concavity, smoothness and compactness

M) What percentage of the variance is explained by each component? Create the percent-of-variance-explained plots and paste them below:

Your answer:




```
> pve
[1] 5.033882e-01 2.798363e-01 9.730049e-02 5.510968e-02 4.138936e-02
[6] 1.295599e-02 8.749572e-03 1.238987e-03 3.143879e-05
```

N) Use your insights from PCA to create a logistic regression, picking only four predictors. Paste your output below (note, this is an open-ended exercise--- answers might differ across groups):

Your answer:

```
Call:
glm(formula = diagnosis ~ compactness + concavity + smoothness +
    radius, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.34871  -0.25874  -0.08319   0.06205   2.85563

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -24.282      2.856  -8.502  < 2e-16 ***
compactness   -7.316      8.184  -0.894 0.371358
concavity     23.742      5.476   4.335 1.45e-05 ***
smoothness    73.948     19.036   3.885 0.000102 ***
radius         1.075      0.123   8.740  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 212.32  on 564  degrees of freedom
AIC: 222.32

Number of Fisher Scoring iterations: 7
```

O) What is the error rate of your classification model? Would you trust this model to detect cancer? Why or Why not?

Commented [MOU2]: Would add “Would you trust this model to detect cancer, why or why not?”

Your answer:

```
> forest_model

Call:
 randomForest(formula = diagnosis ~ compactness + concavity +      smoothness + radius, ntree = 500, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 5.98%
Confusion matrix:
    B   M class.error
B 342  15  0.04201681
M  19 193  0.08962264
```

I would not trust this model because an error rate of close to 6% is too high when it comes to something as serious and life-changing as detecting cancer. If 6% of people are told they have cancer when they don’t or even worse, told they don’t have it when they do, there will be serious issues. I would trust the model if it had an error rate of at most 1%.

Part II: Analytics in Music

Using clustering to understand music.

Music companies are taking advantage of analytics to cater better products to their users. In fact, Spotify is a highly data-driven company, that creates recommendations for its users, by using clustering analysis. In this second part of the lab, we will learn how Spotify clusters music to create playlists based on genres or moods. We will also use analytics to study how music has changed across decades.

To this end, I have gathered a dataset with the characteristics about 5200 of songs. I have collected the following information for each song:

Factors

Labels

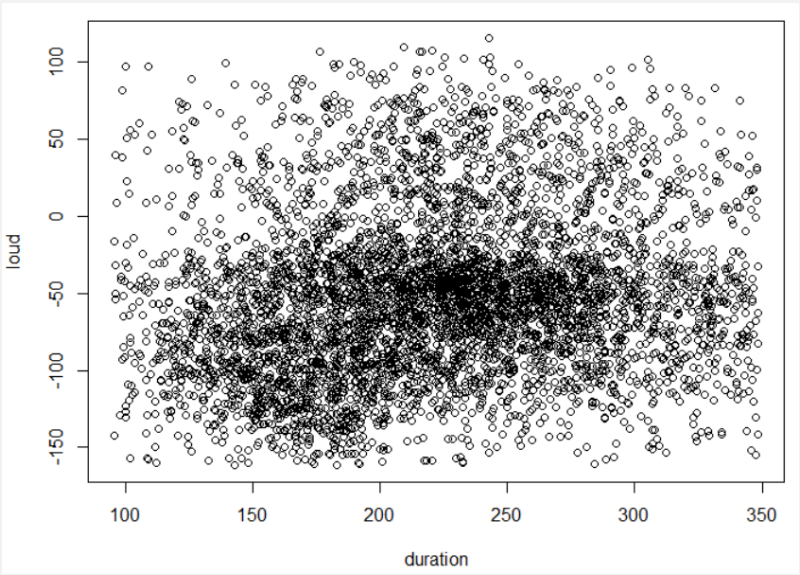
- **Artist_name:** Name of the artist/band.
- **Song_title:** title of the track
- **Track_id:** identifier code for the track

Factors:

- **duration:** song's duration, in seconds
- **loudness:** The overall loudness of a track in decibels (dB).
- **tempo:** The overall estimated tempo of a track in beats per minute (BPM).
- **Key:** The musical key the track is in. Integers map the pitch class.
- **decade:** the decade the song was released in (1950s, 1960s, 1970s, and so on).

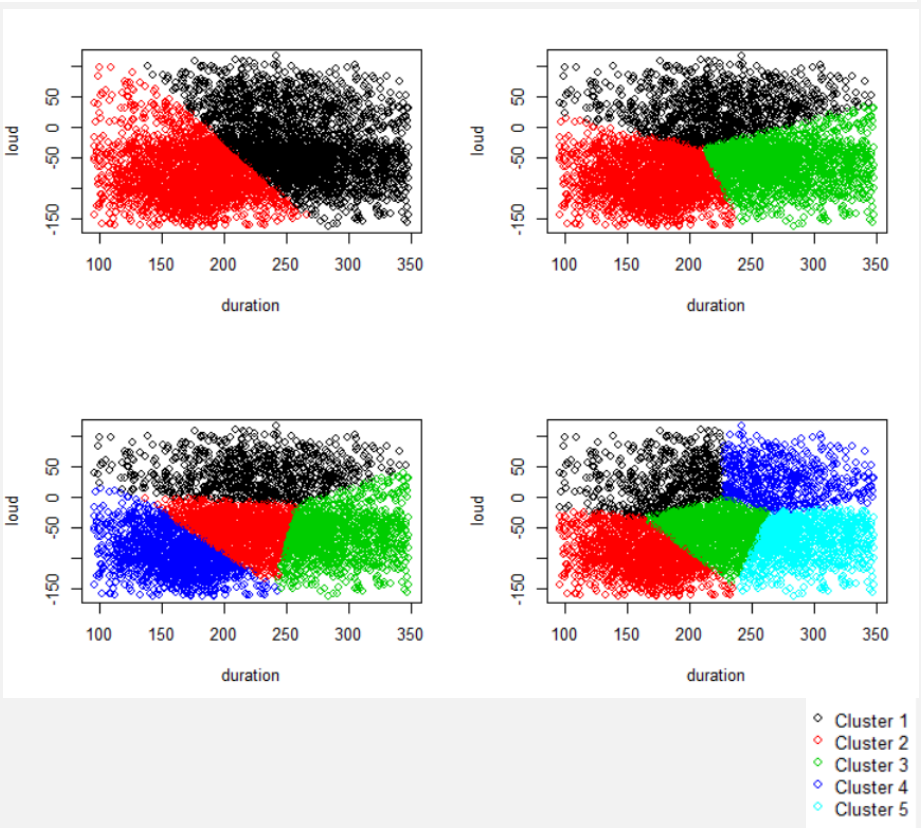
1. Clustering (15 points)

A) Create a scatterplot of the dataset. As the x-axis, I want the duration of the song. As the y-axis, I want the loudness of the song.



B) Let's perform a K-means clustering analysis of the data. Let's run four distinct analyses, dividing the data into: (i) two clusters; (ii) three clusters; (ii) four clusters; and (iv) five clusters. Paste a "clustered scatterplot" for each of the four distinct analyses (make it a 2x2 scatterplot matri):

Commented [MOU3]: Would add the 2x2 requirement in the question description as well


loud
duration
100
150
200
250
300
350
-150
-50
0
50
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

C) Let's now focus on the analysis with five clusters. Paste the means of each cluster for each variable (i.e., the numerical output from R) **for the 5-cluster analysis.**

```
> km.5
K-means clustering with 5 clusters of sizes 573, 1416, 1595, 575, 1048

Cluster means:
  duration      loud
1 181.3973  26.84230
2 160.1424 -95.43118
3 221.8441 -54.94376
4 273.6070  28.04067
5 291.8432 -72.29707
```

D) How would you describe to a music executive each of the five clusters (i.e., in terms of low loudness/medium/high loudness; short/medium/long duration):

- Cluster 1: medium-high loudness, short duration
- Cluster 2: low loudness, short duration
- Cluster 3: medium loudness, medium/long duration
- Cluster 4: high loudness, long duration
- Cluster 5: low loudness, high duration

E) Based on your 5-cluster analysis, how would you describe the following bands/musicians, in terms of the loudness and duration of their songs? Justify your answer:

Aerosmith

- Description: medium loudness, short duration
- Explanation: Aerosmith's music is generally loud and short but also has two songs in the 4th cluster, characterizing them as very loud and long. They are a popular rock band, which justifies the loudness and relatively short songs.

Bob Wills

- Description: very quiet, long duration
- Explanation: Bob Wills' songs are all in the 5th cluster. He was an American western swing musician who was mainly active in the 1960s to 80s, when louder electronic music was not popularized.

Rise Against

- Description: long, medium loudness
- Explanation: Rise Against's songs place mostly in the 3rd cluster (one song in 2 and another in 4). They are a punk rock band, which may contribute to the loudness and the length.

Kings of Leon

- Description: very loud and very long
- Explanation: All Kings of Leon songs are in the 4th cluster. They are an American rock band with long guitar solos which might explain this phenomenon.

Sick of it all

- Description: medium loudness, medium/long duration
- Explanation: All their songs are in the third cluster. They are a punk band that have songs of average length and loudness

Bob Dylan

- Description: Medium loudness, relatively short duration
- Explanation: The majority of Bob Dylan songs are in cluster 3. This is because he is sort of a folk/rock singer, with pretty average length songs and loudness.

F) How has the loudness and duration of music changed across decades? To make this analysis, you should first look at the proportion of songs that are in each cluster, for each decade. Then discuss how the loudness and duration of each song has changed. Include tables or plots to justify your answer. Hint: You should look at the proportion of observations (in each decade) that are found in each cluster. (Note: not all songs have information about the decade. Feel free to ignore these observations for your analysis).

Commented [MOU4]: This is quite open ended as a question, maybe provide them with a little more guideline / example of what you're looking for?

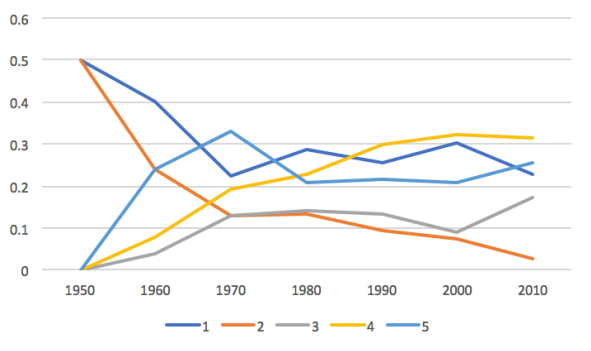
Your answer:

Count of songs in each cluster over each decade:

decade	km.5.cluster 1	2	3	4	5
1950	1	1	0	0	0
1960	10	6	1	2	6
1970	21	12	12	18	31
1980	43	20	21	34	31
1990	114	42	60	132	96
2000	345	83	105	369	237
2010	8	1	6	11	9

Proportions of songs in each cluster in each decade:

	1	2	3	4	5
1950	50%	50%	0%	0%	0%
1960	40%	24%	4%	8%	24%
1970	22%	13%	13%	19%	33%
1980	29%	13%	14%	23%	21%
1990	26%	9%	14%	30%	22%
2000	30%	7%	9%	32%	21%
2010	23%	3%	17%	31%	26%



The proportion of songs in cluster 4 constantly increases over the decades. Cluster 4 is the longest and loudest cluster showing that songs got longer and louder over time. The proportion of songs in cluster two severely decreases over time, and cluster 2 is the cluster with the shortest and quietest songs, which further proves the point that songs got louder and longer from 1950 until 2010

2. Bonus (1 point)

This is your last lab question! To earn your bonus point, search for the loudest song in the dataset, and play it on Youtube. While you are pumping up the volume of this song, answer the following:

You gotta feeling? Check one:

Yes:___X____

No: _____

Now, tell me… you gotta feeling about what?:

That tonight's gonna be a good night.

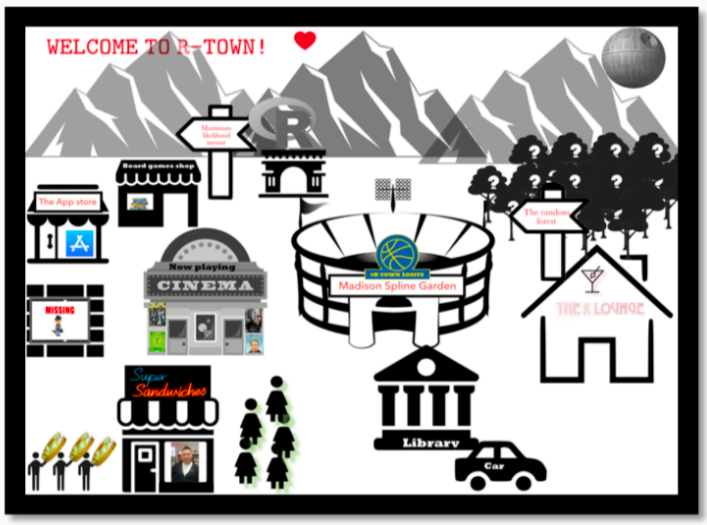
Due date: Beginning of lecture, last week of Class.

Finally, I'd like to take this opportunity to thank our class TAs: Colin Donahoe, Thais Lewko, Carly Matz, Fiona McCarten, and Sean Mitro. They spent a lot of time working on these labs, proofreading through typos, making the questions were easy to understand, and even baking delicious cookies! No matter how busy (or how sleep-deprived) they were, they made sure to have ridiculously good labs and solution keys, improving your slides, and having your grades back within one week. They really went beyond their duty to make this course a more pleasant one, and to help us build this nice analytics community. You could make their day happier if you could surprise them with a short thank-you note for their effort. ☺

Thank you for all the hard work. The course was very enjoyable.

CONGRATULATIONS!

YOU'VE MADE IT THROUGH ALL EIGHT LABS! YOU'VE RIGHTFULLY EARNED YOUR PASS TO BECOME A CITIZEN OF R-TOWN. WELCOME TO YOUR NEW HOME!




```
>INSTALL.PACKAGES("WINTER BREAK")
>REQUIRE(LOTS_OF_REST)
```