CSCI 567 Machine Learning

Assignment 04

Team Members:
Ira Deshmukh (3648692173)
Prem Tibadiya (6351018440)

We will be using our final late day for this submission and it will be registered as Prem Tibadiya's late day.

Question 01 :

1.1 :

Classification Error :

$$\varepsilon_{1,L} = 50/(150+50) = 0.25$$

$$\varepsilon_{1,R} = 50/(150+50) = 0.25$$

$$\varepsilon_{2,L} = 0/(0+100) = 0$$

$$\varepsilon_{2,R} = 100/(200+100) \approx 0.33$$

Entropy :

$$H_{1,L} = -(0.25) - \left(\frac{150}{150+50}\right)\ln\left(\frac{150}{150+50}\right) - \frac{50}{150+50}\ln\left(\frac{50}{150+50}\right)$$

$$\approx 0.56$$

$$H_{1,R} = -\left(\frac{50}{150+50}\right)\ln\left(\frac{50}{150+50}\right) - \left(\frac{150}{150+50}\right)\ln\left(\frac{150}{150+50}\right)$$

$$\approx 0.56$$

$$H_{2,L} = -\left(\frac{0}{0+100}\right)\ln\left(\frac{0}{0+100}\right) - \left(\frac{100}{0+100}\right)\ln\left(\frac{100}{0+100}\right)$$

$$= 0$$

$$H_{2,R} = -\left(\frac{200}{200+100}\right)\ln\left(\frac{200}{200+100}\right) - \left(\frac{100}{100+200}\right)\ln\left(\frac{100}{100+200}\right)$$

$$\approx 0.64$$

Gini Impurity:

$$G_{1,L} = 1 - \left(\frac{150}{150+50}\right)^2 - \left(\frac{50}{150+50}\right)^2 = 0.375$$

$$G_{1,R} = 1 - \left(\frac{50}{150+50}\right)^2 - \left(\frac{150}{150+50}\right)^2 = 0.375$$

$$G_{2,L} = 1 - \left(\frac{0}{0+100}\right)^2 - \left(\frac{100}{0+100}\right)^2 = 0$$

$$G_{2,R} = 1 - \left(\frac{200}{200+100}\right)^2 - \left(\frac{100}{200+100}\right)^2 \approx 0.44$$

1.2:

Let $p_1 = \frac{150+50}{400} = 0.5$ be the fraction of samples that belong to left leaf of T_1, and $p_2 = \frac{0+100}{400} = 0.25$ be the fraction of samples that belong to left leaf T_2.

∴ Classification Error:

$$E_1 = p_1 E_{1,L} + (1-p_1) E_{1,R}$$
$$= (0.5)(0.25) + (0.5)(0.25)$$
$$= 0.25$$

$$E_2 = p_2 E_{2,L} + (1-p_2) E_{2,R}$$
$$= (0.25)(0) + (0.75)(0.33)$$
$$= 0.2475$$

So, they are almost equally good in terms of classification error.

Conditional Entropy:

$$E_1 = p_1 H_{1,L} + (1-p_1) H_{1,R}$$
$$= (0.5)(0.56) + (0.5)(0.56)$$
$$= 0.56$$

$$E_2 = p_2 H_{2,L} + (1-p_2) H_{2,R}$$

$$= (0.25)(0) + (0.75)(0.64)$$

$$= 0.48$$

∴ In terms of conditional entropy, T_2 performs better.

Gini Impurity:

$$E_1 = p_1 G_{1,L} + (1-p_1) G_{1,R}$$

$$= (0.5)(0.375) + (0.5)(0.375)$$

$$= 0.375$$

$$E_2 = p_2 G_{2,L} + (1-p_2) G_{2,R}$$

$$= (0.25)(0) + (0.75)(0.44)$$

$$= 0.33$$

∴ T_2 performs better in terms of weighted gini impurity.

∴ T_2 appears to be a better split.

Based on the above calculations, conditional entropy seems to be more suitable.

Question 02 :

2.1 :

To find $\pi_1, \ldots, \pi_{jK}$, we solve

$$\underset{\pi}{\operatorname{argmax}} \sum_i \sum_j \pi_{ij} \ln \pi_j$$

$$\text{s.t. } \pi_j \geq 0$$

$$\sum_j^K \pi_j = 1$$

with $a_j = \sum_i \pi_{ij}$

$$\therefore w_j^* = \frac{\sum_i \pi_{ij}}{\sum_j \sum_i \pi_{ij}} = \frac{\sum_i \pi_{ij}}{\sum_i (1)} = \frac{\sum_i \pi_{ij}}{N}$$

To find μ_j and σ_j, we solve for each j.

$$\underset{\mu_j, \Sigma_j}{\operatorname{argmax}} \sum_i \pi_{ij} \ln \mathcal{N}(x_i \mid \mu_j, \Sigma_j)$$

$$= \underset{\mu_j, \Sigma_j}{\operatorname{argmax}} \sum_i \pi_{ij} \ln \left[\frac{1}{(\sqrt{2\pi}\sigma_j)^D} \exp\left(-\frac{1}{2\sigma_j^2} \| x_i - \mu_j \|^2 \right) \right]$$

$$= \underset{\mu_j, \Sigma_j}{\operatorname{argmax}} \sum_i \pi_{ij} \left(-D \ln \sigma_j - \frac{\| x_i - \mu_j \|^2}{2\sigma_j^2} \right)$$

where D is the length of x_i.

Let derivative w.r.t $\mu_j = 0$: $\frac{1}{\sigma_j^2} \sum_i \pi_{ij} (x_i - \mu_j) = 0$

$$\therefore \mu_j^* = \frac{\sum_i \pi_{ij} x_i}{\sum_i \pi_{ij}}$$

$\therefore$ Derivative w.r.t $\sigma_j = 0$: $\sum_i \pi_{ij} \left(-\frac{D}{\sigma_j} + \frac{\| x_i - \mu_j \|^2}{\sigma_j^3} \right) = 0$

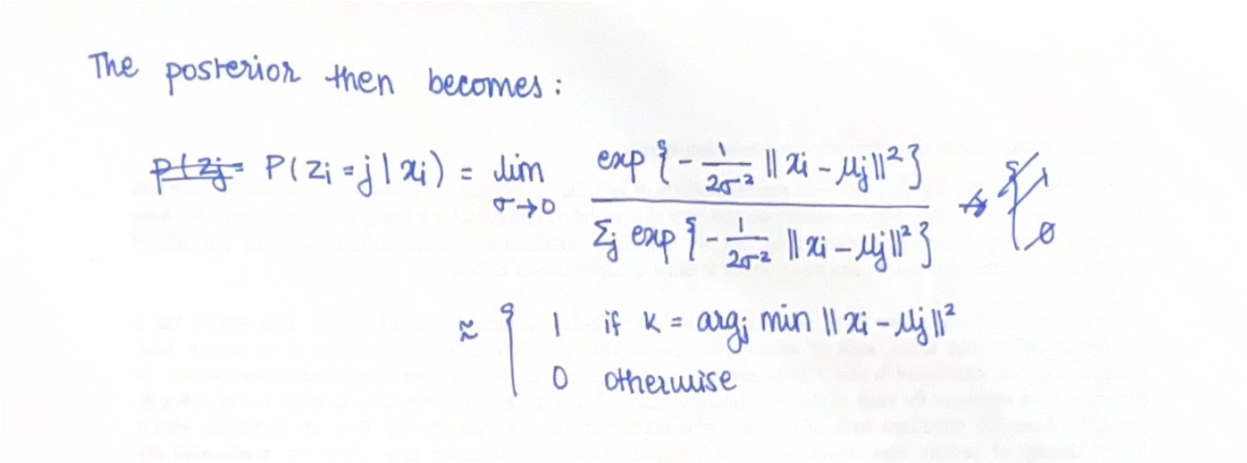
$$\therefore (\sigma_j^*)^2 = \frac{\sum_i \pi_{ij} \| x_i - \mu_j^* \|^2}{D \sum_i \pi_{ij}}$$

2.2 :

Set all $\sigma_j = \sigma \to 0$ and $\pi_j = \frac{1}{k}$, we have

$$p(x_i, z_i = j) \propto \exp - \frac{1}{2\sigma_j^2} \| x_i - \mu_j \|^2$$

where constant terms are ignored.



Question 03:

3.1

In this graph, we can clearly see that as the max depth of the decision tree increases, the performance of the model over the training set increases continuously. On the other hand as the max_depth value increases, the performance over the test set increases initially but after a certain point, it starts to decrease rapidly. The tree starts to overfit the training set and therefore is not able to generalize over the unseen points in the test set.

Random forests can combat increase in variance by averaging over multiple trees, but are not immune to overfitting. Thus, random forests perform comparatively better as compared to decision trees.

3.2

In this graph, we can clearly see that the performance of the model sharply increases and then stagnates at a certain level.
This means that choosing a large number of estimators in a random forest model is not the best idea. Although it will not degrade the model.
n_estimators should range Between 10 and 50 so as to not add on redundant decision trees.

3.3

We can clearly see that the Random Forest model is overfitting when the parameter value is very low, but the model performance quickly rises up and rectifies the issue of overfitting. But when we keep on increasing the value of the parameter the model slowly drifts towards underfitting. This is because we have controlled the growth of the tree by setting a minimum sample criterion for terminal nodes.

3.4

We can see that the performance of the model initially increases as the number of max_samples increases. But, after a certain point, the train_score keeps on increasing. But

the test_score saturates and even starts decreasing towards the end, which clearly means that the model starts to overfit.

3.5

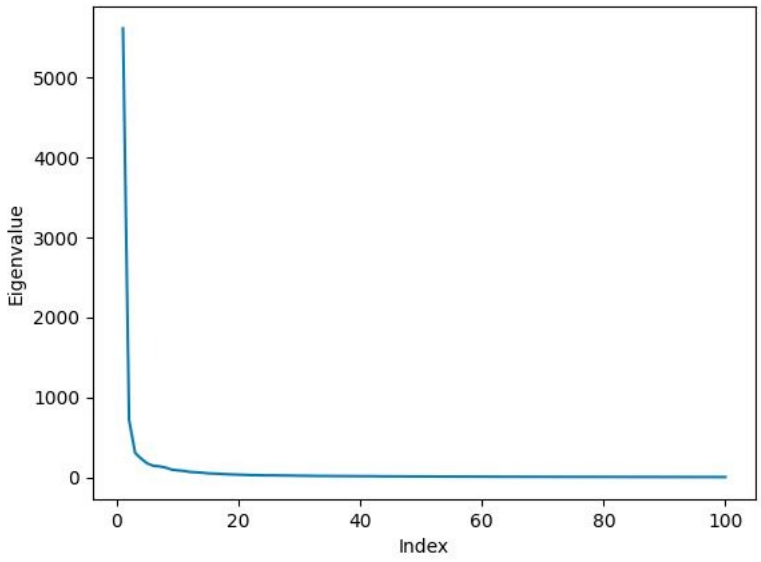
We can see that the performance of the model rises sharply and then saturates fairly quickly. Since it is not necessary to give the full data to all the decision trees in the random forest after a certain point given extra data does not change the accuracy score of the tree and the data becomes redundant.

3.6

The pixels are focusing on the main subject of the image rather than the background. Pixels with high importance are required for prediction task. It helps in increasing the confidence of the prediction. Giving importance to certain pixels will help in reducing the computation of the algorithm as well as it will narrow down the focus area from the entire image to a certain part.

Question 04:

Q 4.1)



- Do the eigenvalues seem to decay?
- Yes. The eigenvalues seems to decay as you can see in the graph.

- What percent of the variance in the data is explained by the first 100 eigenvalues we calculated (note that there are 10,000 eigenvalues in total)?
- By finding the cumulative sum of percentage of variance explained by each of the pca we get this result.

```
variance = pca.explained_variance_ratio_.cumsum()
print(variance[99])
```

- output : 0.7566978267579538

- This means that 75.669% of the variance in the data is explained by the first 100 eigenvalues we calculated.

Q 4.2)

- Try finding the closest words to some common words, such as “learning”, “university”, “california”, and comment on your observations.
- 'mother' : most similar word 'father'
- 'learning' : most similar word we got 'knowledge'
- 'university' : most similar word we got 'college'
- 'california' : most similar word we got 'texas'

As you can see some of the similar words are either synonyms or has close meaning to each other. And for "california" which a place gives us "texas" which is also a place.

So these words have somehow same semantics.

Q 4.3)

- Can you find 5 interesting eigenvectors, and point out what semantic or syntactic structures they capture?
- i) ['broadcast',
  'broadcasting',
  'chart',
  'championship',
  'fm',
  'baseball',
  'racing',
  'abc',
  'stadium',
  'tournament',
  'arena',
  'basketball',
  'wrestling',
  'comics',
  'concert',
  'rugby',
  'bus',
  'cbs',
  'cricket',
  'studios']
  - This eigenvectors has captured the semantics part of sports.
- iii) ['specific',
  'actress',
  'starring',
  'jr',
  'kevin',
  'anthony',
  'born',
  'processes',
  'chris',

'require',
'jimmy',
'gary',
'steve',
'types',
'mike',
'billy',
'brian',
'bobby',
'politician',
'johnny']

- This eigenvectors has captured the semantics part of movies.

- ii) ['district', 'county', 'university', 'national', 'council', 'you', 'album', 'love', 'me', 'regional', 'government', 'south', 'northern', 'east', 'north', 'department', 'united', 'my', 'college', 'union']

- This eigenvectors has captured the semantics part of some organisation.

- iv) ['km',
'jpg',
'located',
'adjacent',
'ft',
'mile',
'creek',
'moral',
'junction',
'bay',
'operated',
'northeast',
'southwest',
'route',
'railway',
'highway',
'nearby',
'rail',
'lake',
'mm']

- This eigenvectors has captured the semantics part of some kind of places or direction.

- v) ['born', 'john', 'james', 'david', 'robert', 'william', 'george', 'jr', 'thomas', 'michael', 'richard', 'paul', 'mike', 'charles', 'chris', 'steve', 'frank', 'peter', 'jim', 'scott']

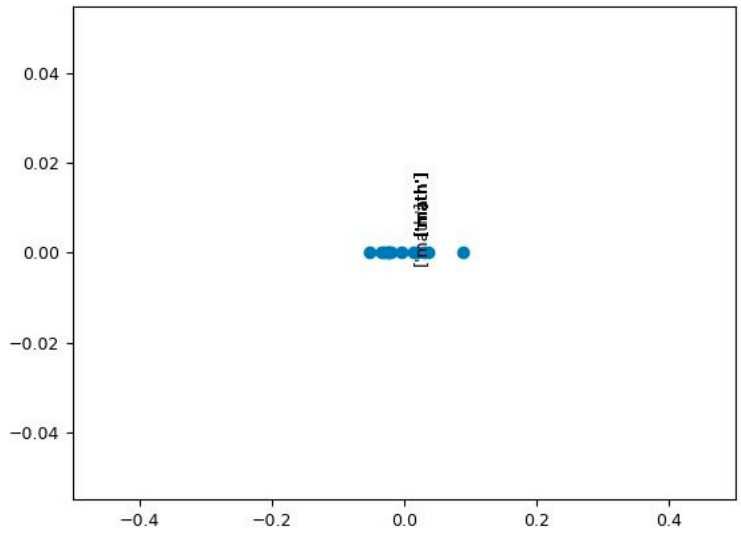
- This eigenvectors has captured the semantic part name of a person.

- Can you do this for all 100 eigenvectors?

- No. As we go down the order the eigenvalues of a vector decreases so the variance captured by these vectors is less thereby capturing little semantics or syntactic structure.

Q 4.4)

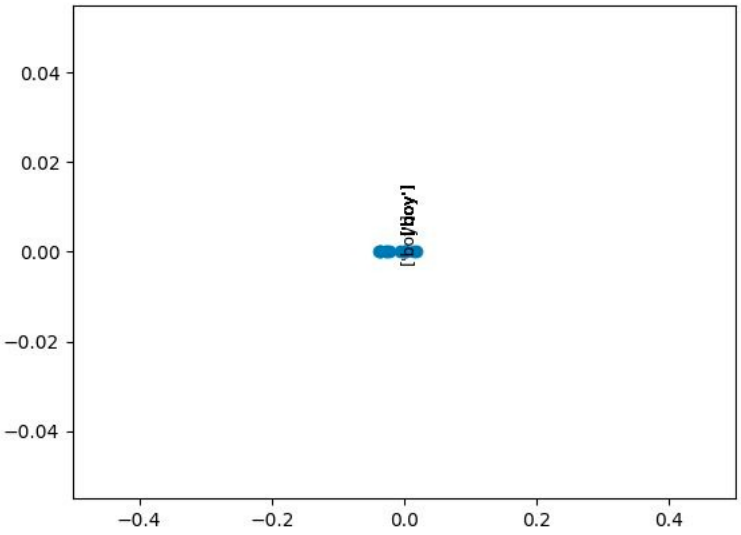
- Q 4.4.1)



What do you observe?

- All are clustered at th origin as all are strong gendered biased words these words wont cause any problem while using word embedding.

- Q 4.4.2)



What do you observe? Why do you think this is the case? Do you see a potential problem with this? Remember that word embeddings are extensively used across NLP.

Suppose LinkedIn used such word embeddings to find suitable candidates for a job or to find candidates who best match a search term or job description. What might

be the result of this?

- There is bias in the plot. Because word embedding is gendered bias. Yes this will relate biased result. So in LinkedIn case the particular jobs would only be searces by a particular gender which would cause a problem. It wont be able to reach out to less biased gender.

Q 4.5) Look at the incorrect/correct answers of the approach and comment on the results. For example, what types of analogy questions seem to be harder to answer correctly for this approach?

- Analogy questions which has noun in it are seem harder to solve.