2024 AIML427 Big Data: Assignment 2

This assignment has 100 marks and is at on 11:59 pm, Monday, 6^{th} May 2024. Please submit your answers as a single *.pdf* file. Make sure you read the Assessment section before writing the report. This assignment contributes 25% to your overall course grade.

Any questions about Parts 1 or 2 should be directed to Bach; any questions about Parts 3 should go to Qi.

1 Manifold Learning [40 marks]

In class, we discussed a variety of different manifold learning methods, which we broadly categorised as “classic” statistical methods or “modern” ML-based methods. In this question, you are expected to further explore the differences between these classes of methods, and compared to PCA (as a linear dimensionality reduction method). You should use *no more than 3 pages* to answer the following questions.

1. Find a reasonably high-dimensional (at least 100 dimensions) dataset that is interesting to you. It should also have at least 100 instances, but preferably more. Describe the dataset (name, related task/what it is used for, number of features and instances, reference) and justify your choice.
2. Using your choice of library, apply PCA to the dataset, and present your results. You should show visualisation(s) of the PCs found and also comment on the explained variance.
3. Pick one of the “classic” manifold learning methods and apply it to the same data. Show visualisation(s) and compare the results to that of PCA (e.g. for an embedding with two dimensions). Highlight any differences between the two methods, and hypothesise why they may have occurred.
4. Pick one of the “modern” manifold learning methods and apply it to the same data. Show visualisation(s), compare and contrast the results to the two previous methods, with analysis of any differences seen.
5. Finally, pick one of the two manifold learning methods for further analyse. Your method will have “tunable parameters” — parameters that you can change to get different results. Pick one such parameter, and explore how sensitive the embedding is to changes in this parameter. You should explain the role of this parameter in the manifold learning algorithm, how you tested its effect, and show the results found.

2 Clustering [30 marks]

The NCI60 dataset (from the Stanford NC160 Cancer Microarray Project) consists of p = 6,830 gene expression measurements for each of n = 64 cancer cell lines. (Sourced from *An Introduction to Statistical Learning*).

In this question, you will be clustering the genes, rather than individual cancer cell lines. This can be seen as a form of feature clustering — i.e. what genes are most related?

For each clustering method, you will need to visualise the clustering results (partition) for that method. Given that there are 64 dimensions for each gene, for visualising the clustering results, you should use PCA to reduce the dimensionality to 2D so that you can plot the found clusters.

It is recommended you use either R or Python for this question, as they both have libraries to interact with this data, as shown below. You should use *no more than 3 pages* to answer the following questions.

2.1 R:

```r
library(ISLR)
nci.data = NCI60$data
X = scale(t(nci.data))
P = X %*% prcomp(X)$rotation
```

X will be the numpy array of interest — note that we have transposed our data so our rows are the genes. P is the principal components of the data. You will use the the first 2 PCs, i.e. the first 2 columns of P, to visualise the clusters.

2.2 Python:

For Python (or any other language), you will need to first download `nci60_data.csv` from the course homepage.

```python
import pandas as pd
from sklearn.preprocessing import scale
from sklearn.decomposition import PCA

nci_data = pd.read_csv("nci60_data.csv", index_col=0)
X = scale(nci_data.T)
P = PCA().fit_transform(X)
```

X will be the data matrix of interest — note that we have transposed our data so our rows are the genes. P is the principal components of the data. You will use the the first 2 PCs, i.e. the first 2 columns of P, to visualise the clusters.

2.3 Tasks:

1. Carry out hierarchical clustering with Euclidean distance and complete linkage.
   (a) Describe the resulting clustering for *3 to 6* clusters.
   (b) Plot the first 2 principal components against each other with the colour argument set equal to the cluster labels. What can you deduce/observe about the clustering?

2. Repeat the cluster analysis using correlation-based distance and complete linkage. NB: you will need to precompute the correlations and pass them into your clustering method. Compare the clusters with those found above.

3. Finally, carry out K-means clustering for *3 to 6* clusters. Compare the clusters of K-means with that of the above two approaches. Which of the hierarchical clustering results is more similar to that of K-means? Why?

3 Regression [30 marks]

In the lecture, we considered the case in which the features/predictors appeared only linearly in the regression model. The simplest type of nonlinearity we could add to the model is pairwise interactions of the features. If x_j and x_k are distinct features, this means we also consider $x_j x_k$ as a feature. Pairwise interactions are rather straightforward to implement in R:

$$X = model.matrix(balance \sim . * ., Credit)[, -1] \tag{1}$$

becomes the new design matrix. The construction $. * .$ means consider all pairwise multiplications of distinct features.

Repeat the analysis for the **Credit** dataset (we have done it in the lecture on Week 7) with pairwise interactions of the features. You will find it convenient to set $grid = 10^\wedge seq(3, -1, 100)$ and $thresh = 1e-10$.

Answer the following questions:

1. How many predictors are there, i.e. what is p?

2. How did you generate your training and test sets?

3. Select the tuning parameter for the ridge regression model using cross-validation, and show the process.

4. Select the tuning parameter for the lasso regression model using cross-validation, and show the process. How many features have been selected by the lasso?

5. Compare and discuss the final form of the model from the linear regression, ridge regression, and lasso regression.

6. Compare the test errors for the linear model, ridge regression model, and lasso model.

7. Plot a comparison of the test predictions for the three approaches.

NB Please show how you generated your test and training sets – in particular the RNG seed you used – so we can replicate your results. Remember to report this in the following questions when applicable.

Assessment

Format: You can use any font to write the report, with a minimum of single spacing and 11 point size (hand writing is not permitted unless with approval from the lecturers). Reports exceeding the maximum page limit will be penalised. Any additional material such as code or figures/tables can be included in an appendix, which will not count towards the page limit.

Communication: a key skill required of a scientist is the ability to communicate effectively. No matter the scientific merit of a report, if it is illegible, grammatically incorrect, mispunctuated, ambiguous, or contains misspellings, it is less effective and marks will be deducted.

Marking Criteria: The final report will be submitted to Turnitin for a plagiarism check. Late submissions without a pre-arranged extension will be penalised as per the course outline. The usual mark checking procedures in place for all assessment apply to this report. The assessment of the reports will account of the understanding of big data, clarity and accuracy of answer, presentation, organisation, layout and referencing.

Submission: You are required to submit a single *.pdf* report through the web submission system from the AIML427 course website *by the due time*.