**Abstract**

Linear classification models are simple yet powerful methods of machine learning that can result in very high accuracy rates with little computing power. In this project we investigated the performance of linear classification models on two benchmark datasets. The two models we tested were logistic regression and linear discriminant analysis (LDA). We tested the accuracy of each model using k-fold cross validation with 5 folds. We found that both models were able to perform significantly better on the cancer data set, with logistic regression achieving 100% accuracy on one of our runs of 5-fold cross validation on the cancer data set (compared to a maximum accuracy of less than 70% on the wine data set). We found that \_\_\_\_\_\_\_ achieved better accuracy than \_\_\_\_\_\_\_\_ and was significantly \_\_\_\_\_\_\_\_ to train.

**Introduction**

The first of the two tasks presented were to determine whether a tumour was malignant or benign based on features such as radius, area, smoothness, and symmetry. This data was first used in 1993 in the International Symposium on Electronic Imaging to try and improve medical standards globally [1]. From this dataset 35% of tumours were classified as malignant and the rest were benign.

The second task was to determine whether a wine tasted good or bad based on features such as density, pH, and residual sugar. This data was initially used by researcher Paulo Cortez and his team in hopes of being able to reverse engineer the process of making tasty wine [2]. From this dataset 53.5% of wine was classified as good tasting and the rest was bad tasting.

The higher ratio of positive to negative classes in the tumour dataset compared to the wine dataset made it significantly easier to achieve a high accuracy on the model, since it could simply predict “Good tasting” every single time and record at 65% accuracy rate.

**Datasets**

The wine dataset behaved very well and didn’t have any malformed features. For preprocessing we binarized all outputs, with outputs of 6-10 being converted to 1, and outputs 1-5 being converted to 0.

The tumour dataset was not as cooperative and had many malformed features. The first step was to remove the IDs from each training example since it would skew the results. We had to loop through the data and remove all training examples with missing data to prevent the program from breaking. Next we transformed the output of all malignant tumours to 1 and the output of all benign tumours to 0.

For both datasets we standardized the data by setting the mean of each feature to 0 and the standard deviation of each feature to 1.

An ethical concern arises when dealing with data from people, such as cancer data; as you don’t want any patients to be identified. Which is why these data sets are anonymized, so patients are only identified by an id, which is the first value in the breast cancer data csv.

**Results:**

**Running Times:**

**Learning Rate:**

|  |  |
| --- | --- |
| Learning rate | K Fold Results |
| 1 / 1\_000\_000\_000\_000\_000 | CANCER: 0.9708029197080292  CANCER: 0.9562043795620438  CANCER: 1.0  CANCER: 0.9852941176470589  CANCER: 0.9632352941176471  WINE: 0.696875  WINE: 0.65625  WINE: 0.61875  WINE: 0.603125  WINE: 0.5893416927899686 |
| 1 / 1\_000\_000\_000\_000 | CANCER: 0.9708029197080292  CANCER: 0.9562043795620438  CANCER: 1.0  CANCER: 0.9852941176470589  CANCER: 0.9632352941176471  WINE: 0.696875  WINE: 0.65625  WINE: 0.61875  WINE: 0.603125  WINE: 0.5893416927899686 |
| 1 / 1\_000\_000\_000 | CANCER: 0.9708029197080292  CANCER: 0.9562043795620438  CANCER: 1.0  CANCER: 0.9852941176470589  CANCER: 0.9632352941176471  WINE: 0.696875  WINE: 0.65625  WINE: 0.61875  WINE: 0.603125  WINE: 0.5893416927899686 |
| 1\_000\_000\_000\_000 | CANCER: 0.9708029197080292  CANCER: 0.9562043795620438  CANCER: 1.0  CANCER: 0.9852941176470589  CANCER: 0.9632352941176471  WINE: 0.696875  WINE: 0.584375  WINE: 0.5125  WINE: 0.4625  WINE: 0.48589341692789967 |
| 1\_000\_000 | CANCER: 0.9562043795620438  CANCER: 1.0  CANCER: 0.9852941176470589  CANCER: 0.9632352941176471  WINE: 0.696875  WINE: 0.584375  WINE: 0.5125  WINE: 0.4625  WINE: 0.48589341692789967 |
| 1 | CANCER: 0.9708029197080292  CANCER: 0.9562043795620438  CANCER: 1.0  CANCER: 0.9852941176470589  CANCER: 0.9632352941176471  WINE: 0.696875  WINE: 0.584375  WINE: 0.509375  WINE: 0.4625  WINE: 0.48589341692789967 |
| 1 / 100\_000 | CANCER: 0.9708029197080292  CANCER: 0.9562043795620438  CANCER: 1.0  CANCER: 0.9852941176470589  CANCER: 0.9632352941176471  WINE: 0.696875  WINE: 0.65625  WINE: 0.615625  WINE: 0.603125  WINE: 0.5924764890282131 |

**Discussion and Conclusion**

When looking at different learning rates, all of the evaluation accuracy results of the cancer data set are identical, and appear to be independent of the learning rate. However the wine model accuracy was effected by the learning rate. This could be because the model that represents the cancer data has only one local minimum.

An important takeaway from this project is that preprocessing of data needs to be taken very seriously, as many problems can arise if not treated properly. If a certain feature has a very high mean compared to the rest, it could skew the algorithm to over adjust for that specific feature.

The most important finding was realizing the true power of these simple linear classification algorithms. Operating times were similar on a small scale with only a thousand or so datapoints but choosing the correct algorithm would make a large difference if they were implemented on massive datasets. Tinkering with learning rates and testing new interaction terms can clearly have a big impact on the final performance of the algorithms. In the future a good idea would be to have the computer loop through many possible interaction terms and learning rates and then choose the best ones.

**Statement of Contributions**

Petar preprocessed the datasets and implemented k-fold cross validation. Ethan implemented LDA. Alex implemented logistic regression and tested learning rates and interactions. All members helped each on their respective duties and contributed to the final report.

**Sources**

[1] W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.

In Decision Support Systems, Elsevier, 47(4):547-553, 2009.