Assignment 1

Rohan Hitchcock and Patrick Randell

Question 1

We applied the both K-means and equal-width discretisation techniques to all numeric attributes in the four datasets with numeric attributed. We then assessed the naïve Bayes model generated from these modified data sets by computing the average F-score over a cross validation split with ten partitions (Figure 1). We see that in all datasets the non-discretised model (using the usual Gaussian naïve Bayes approach for numeric attributes) performs the same or better than models trained on data discretised using the equal width approach. We suggest that this reflects the fact that some information is necessarily lost when discretising numeric data.

On the other hand the models trained on data discretised using K-means performed better on the WDBC and Adult datasets than the models trained using numeric data. The models trained on numeric data assume that the data has a Gaussian distribution, so we suggest that the K-means approach can perform better when this assumption is violated. K-means is well-suited to finding natural groupings within data, which may mean the model is more adaptable to different attribute distributions.

To investigate this idea we plotted histograms of each numeric attribute each of the datasets. Qualitatively, the proportion of numeric attributes judged to be highly non-Gaussian for WDBC (0.60) and Adult (0.67) was higher than the proportion for Wine (0.15). This supports the hypothesis that the K-means approach performs better than the standard Gaussian approach when the data is non-Gaussian, although further investigation is warranted.

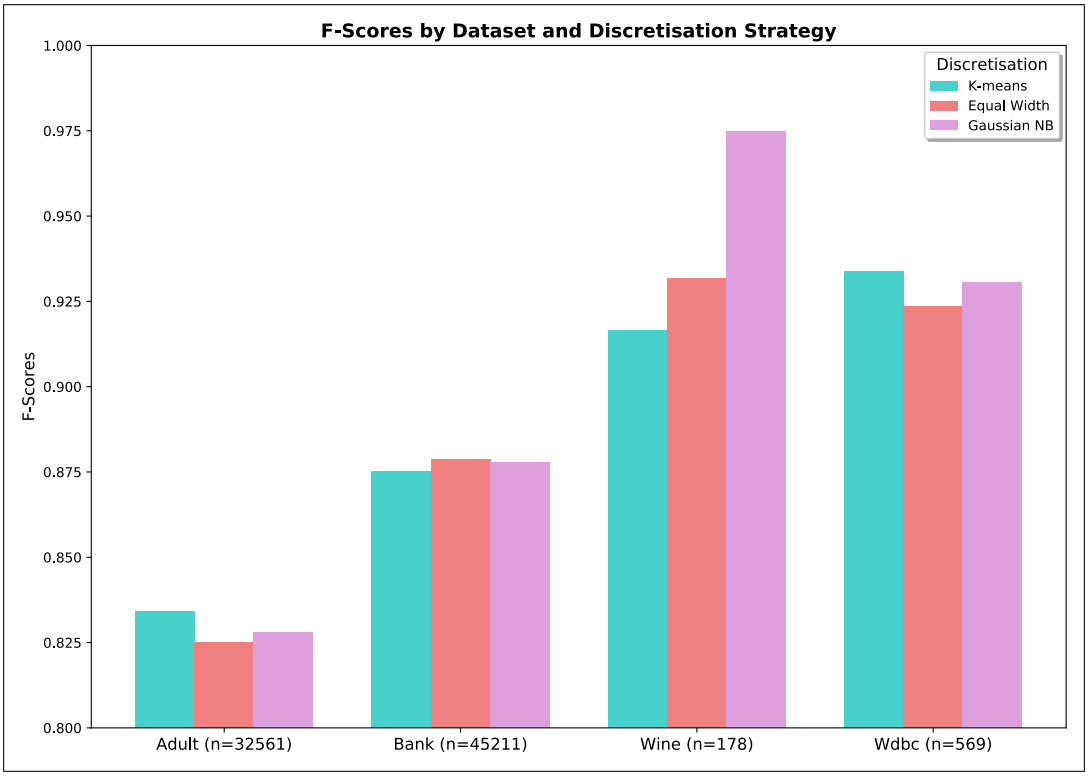

Figure 1: F-Scores ($\beta = 1$) for various discretisation methods.

Question 2

We looked at four baseline classifiers: Uniform (class is chosen uniformly at random), Random (class is chosen randomly according to frequencies of data in the training set), Zero-R (most frequent class is chosen) and One-R (class is predicted using one discrete attribute). These baselines were compared to the naïve Bayes classifier by computing their average F-score over a cross validation split with ten partitions.

The naïve Bayes classifier performed better than all four baselines in all datasets, although not to the same degree. In particular, in the Adult and Bank datasets the performance of the Random and Zero-R classifiers was much closer to the naïve Bayes classifier. This could be for two reasons: either data is not suited to naïve Bayes classification because it violates some assumption (e.g. numeric attributes are highly non-Gaussian, or attributes are highly correlated), or the class is inherently less predictable from the attributes. If the first case were true we would expect the naïve Bayes classifier to have a high error rate, which would be reflected as a lower F-score as compared to other datasets. Therefore we suggest that it is more likely that the class is less predictable from the attributes in the Adult and Bank datasets. In this case we would expect all conditional probabilities $P(x|c)$ to be similar for all attribute values x and classes c, which would mean the naïve Bayes classifier would be dominated by the class priors. This would mean naïve Bayes would perform more like the Random and Zero-R baselines, which is exactly what we see in Figure 4. The better performance of the baselines in the Adult and Bank datasets can accounted for by the higher relative frequencies of a particular class in these datasets; these baselines are more likely to be correct if one class is significantly more common.

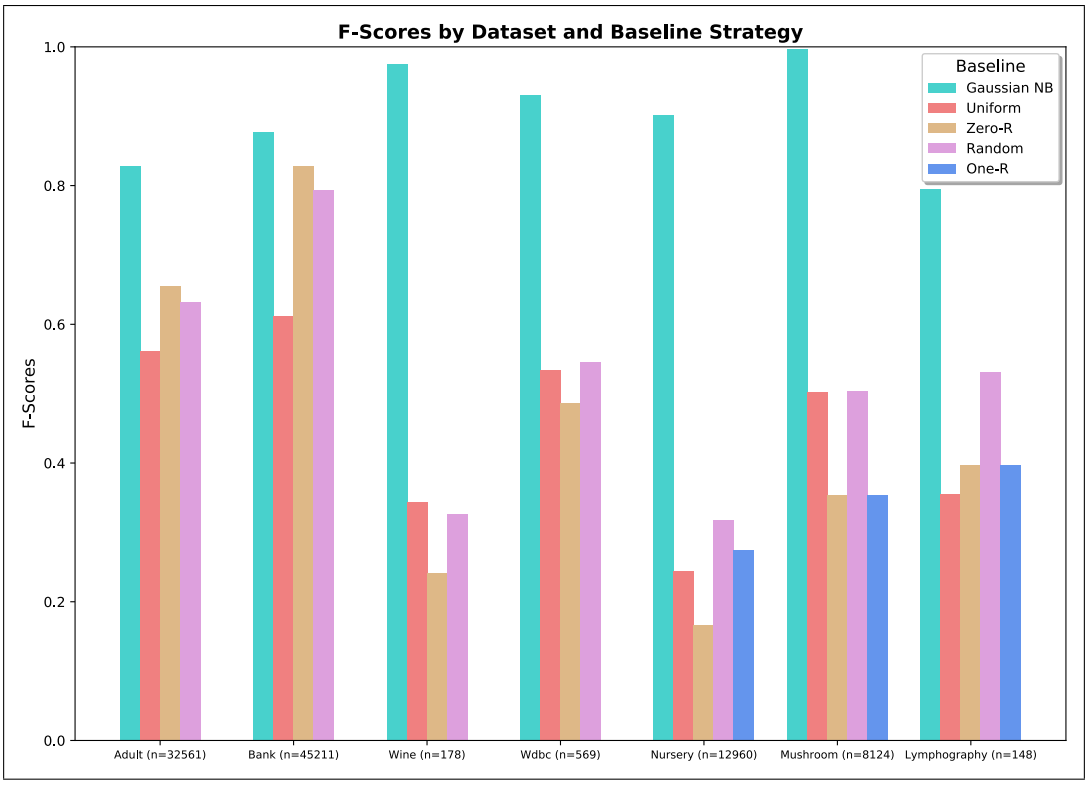

Figure 2: F-Scores ($\beta = 1$) for various baselines.

Question 4

We applied cross validation by splitting each dataset into 10 randomised partitions. We then trained a Naïve Bayers model on 9 of the partitions, and tested on the remainding 1. This was done 10 times, testing on each partition in turn.

The final Accuracy, Precision, Recall and F-scores were found by averaging the results from testing on each isolated partition. Figure 3 compares these averaged results to the results from training and testing on the entire dataset.

From the table we can see that the Cross-Validation results are never greater than the No Split results, however they are often very close.

Cross-Validation is important to ensure that your model is as good as you believe, but our results support the notion that a good model's performance is consistent when trained on enough data. Training on a smaller fraction of the original dataset (Setting k to a lower value) may have shown a greater difference in performance, however in our testing we found that this had little effect on larger datasets.

Lymphography perfomed significantly worse when using Cross-Validation when compared to the other datasets. This is likely due to its small size (148 instances), meaning that the 10% less data each model was trained on significantly impacted its predictive performance.

		Accuracy	Precision	Recall	F-Score ($\beta = 1$)
Mushroom	No Split	0.997	0.997	0.997	0.997
	Cross Val.	0.997	0.997	0.997	0.997
Adult	No Split	0.833	0.823	0.833	0.828
	Cross Val.	0.833	0.823	0.833	0.828
Nursery	No Split	0.903	0.906	0.904	0.904
	Cross Val.	0.903	0.901	0.903	0.902
WDBC	No Split	0.940	0.940	0.940	0.940
	Cross Val.	0.930	0.931	0.930	0.930
Wine	No Split	0.989	0.989	0.989	0.989
	Cross Val.	0.972	0.978	0.972	0.975
Bank	No Split	0.877	0.880	0.877	0.878
	Cross Val.	0.877	0.879	0.877	0.878
Lymphography	No Split	0.892	0.893	0.892	0.893
	Cross Val.	0.784	0.807	0.784	0.795

Figure 3: No train-test split vs. cross validation (10 partitions).

Question 5

Without smoothing, a Naïve Bayers model gives events that are unlikely (Events we have not seen in training for that particular class, but are not impossible) a probability of 0. Smoothing is a method of giving a non-zero probability to such events, rather than skipping them. With larger datasets, the likelihood of unseen events arising in testing data, that were not encountered in training, is lower. This means smoothing has less effect on overall performance.

As can be seen from Figures 4, 5 and 6, the two smallest datasets (Mushroom and Lymphography) are greater impacted by different k values for add-k (Laplace) smoothing.

For Mushroom, the class can take on 2 values, and it may be the case that these two classes are quite distinct in terms of the attribute values that they generally comprise of. Giving these 'unseen events' a probability may be over estimating the likelihood of these extremely rare events, edging out the wrong class over the other.

For Lymphography, it is likely that the size of the dataset is the major cause of this behaviour. Due to its 14 attributes, many having more than 2 values, it is very likely that unseen events will be encountered during cross-validation splits. Giving a probability to these events may simply allow the characteristics of each class to show better, as the likelihoods of one class may have been non-uniformly decreased due to a higher number of unseen events for that class.

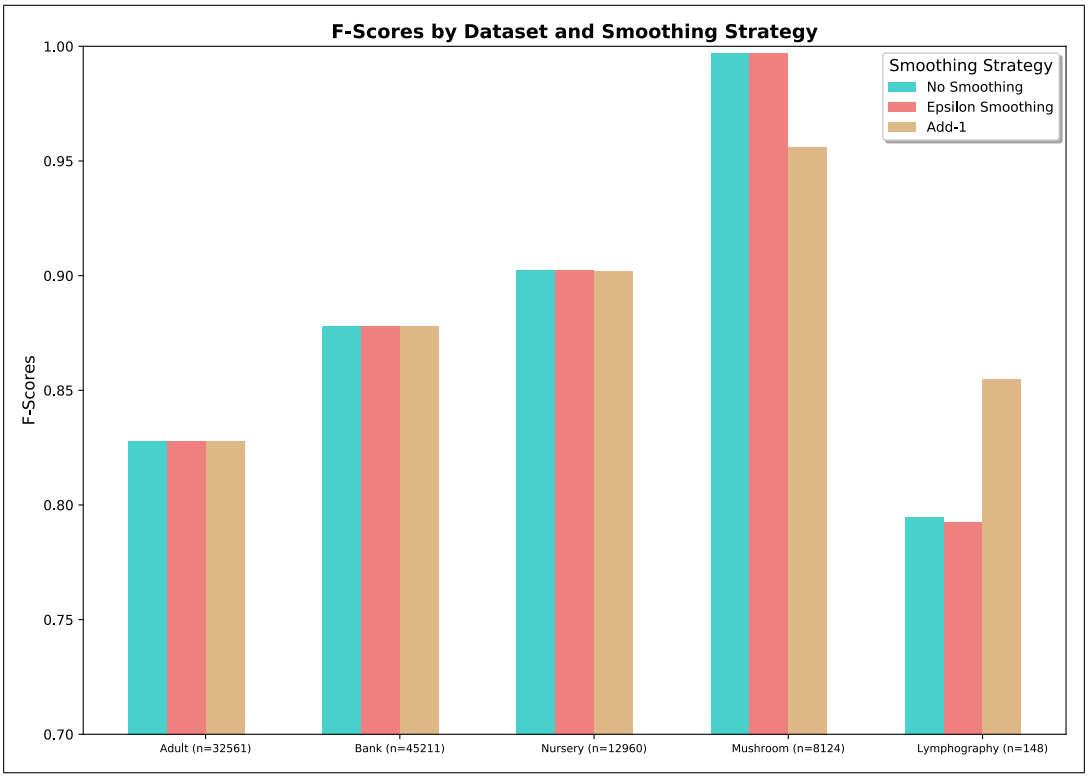

Figure 4: F-Scores ($\beta = 1$) for various baselines.