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.

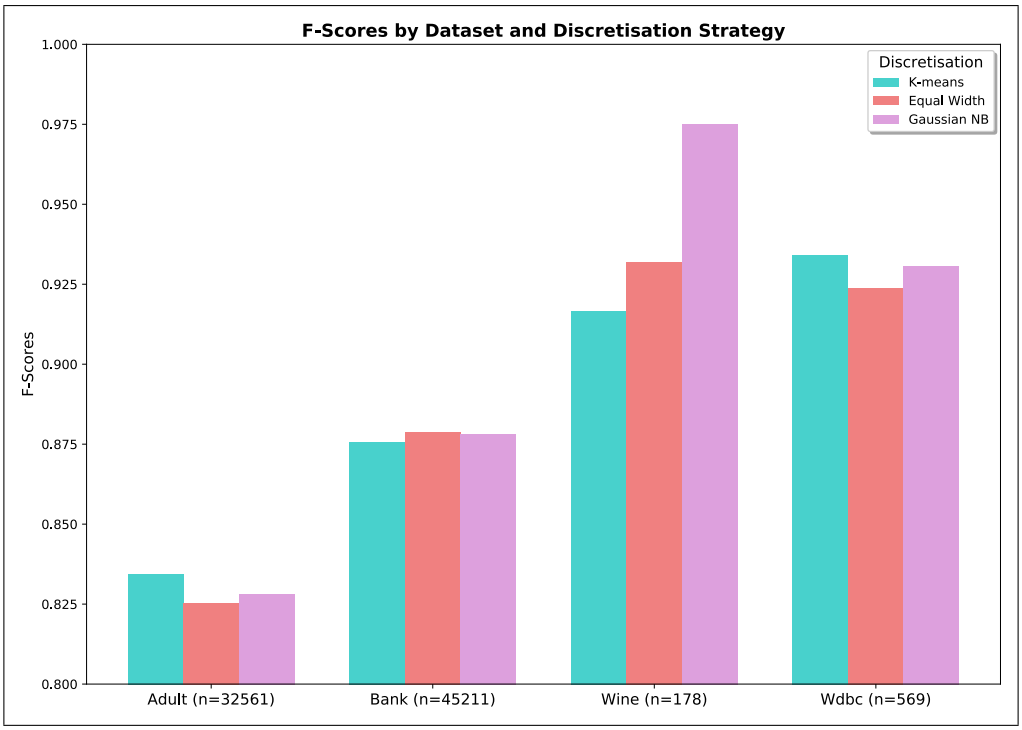

Figure 1: F-Scores ($\beta = 1$) for various discretisation methods.

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 2. 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.

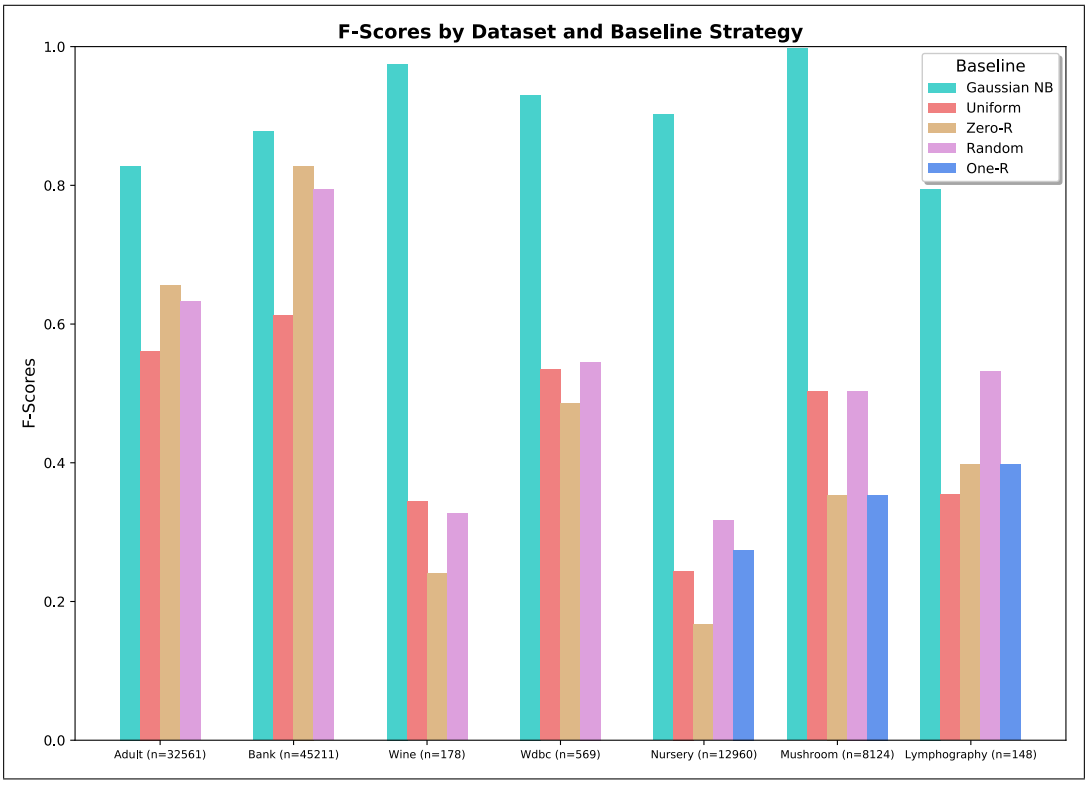

Figure 2: F-Scores ($\beta = 1$) for various baselines.

Question 4

We implemented a cross validation strategy to evaluate model performance. In our implementation the dataset is split into ten randomised partitions. We then trained a Naïve Bayes model on nine of the partitions, and used the remaining partition for testing. This process was repeated ten times, using each partition for testing once. The final accuracy, precision, recall and f-scores were calculated by averaging the results from each cross validation run. Figure 3 compares these averaged results to the results from training and testing on the entire dataset.

From Figure 3 we can see that the cross-validation results are never better than the results obtained when no train-test split is used (no-split), which is to be expected. In most datasets the cross-validation performance came very close to the no-split performance. This tells us that our model is generalising well, and that the good performance in the no-split runs is as a result of genuine learning rather than over-fitting.

The datasets with the largest discrepancy between no-split and cross-validation performance (wine and lymphography) are also the two smallest datasets. We therefore suggest that this difference in performance is because of a lack of training data, rather than evidence over-fitting in the no-split case. Lymphography also comprises entirely of discrete attributes and no smoothing techniques were used in this question, so less data results in more unobserved events (which have probability zero without smoothing) and hence even worse model performance. We will see in Question 5 that applying smoothing techniques had by far the biggest effect on lymphography.

		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 – that is, events we have not seen in training for that particular class, but are not impossible – a probability of 0. When these events arise in testing data this means the relative likelihood of the relevant class is assigned to be zero, which erases any information obtained from the other attributes of that test instance. Smoothing techniques attempt to get around this by estimating a non-zero probability for these rare events.

We implemented Laplace (a.k.a Add-k) smoothing, and compared the performance of a classifier using add-k smoothing for $k = 0.2, 1, 5$ to a classifier using no smoothing or simple epsilon smoothing (rare events are given a small fixed probability). In Figures 4, 5 and 6, we can see that the two smallest datasets (Mushroom and Lymphography) are most greatly impacted by smoothing, while the in the larger datasets virtually no difference was observed. We suggest that this is because in larger datasets the probability of rare events arising in the training data is much higher, and so there is much less chance for smoothing to take effect. Increasing the value of k exacerbated the difference in performance due to smoothing on the smaller datasets.

In general we would expect smoothing to improve the performance of a classifier (as it did in lymphography), but interestingly smoothing had an opposite effect on the mushroom dataset. In mushroom the class can take on 2 values, and it may be the case that these two classes have disjoint (or near-disjoint) sets of attribute values which can appear with each class. Giving unseen events a non-zero probability, when in fact the true probability in most cases should be zero, may lead to the worse performance, especially if only a few attributes are good predictors of the class (so the other attribute values do not provide information about which class to choose).

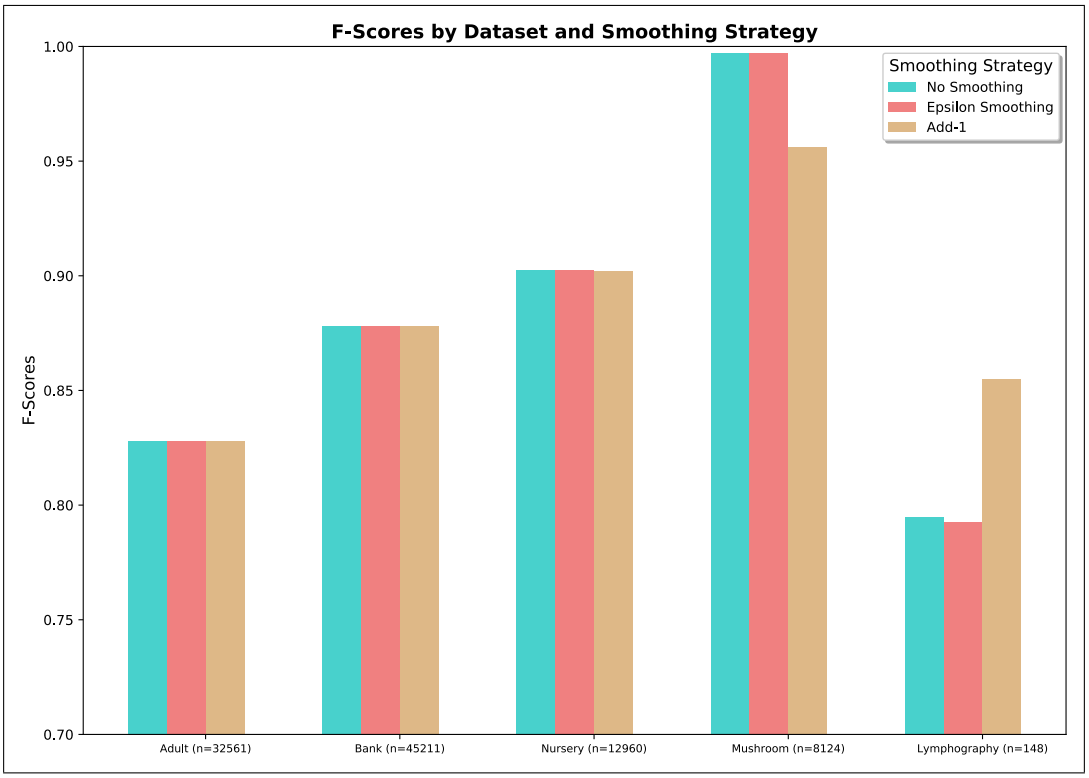

Figure 4: F-Scores ($\beta = 1$) for Add-1 smoothing.

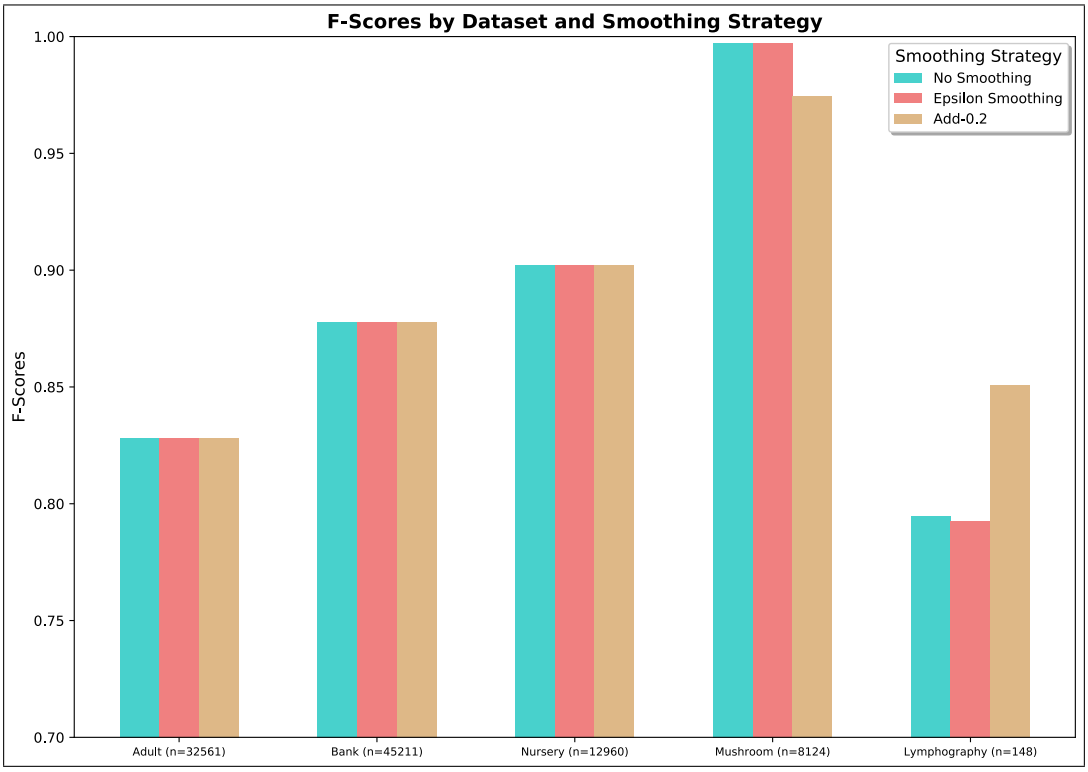

Figure 5: F-Scores ($\beta = 1$) for Add-0.2 smoothing.

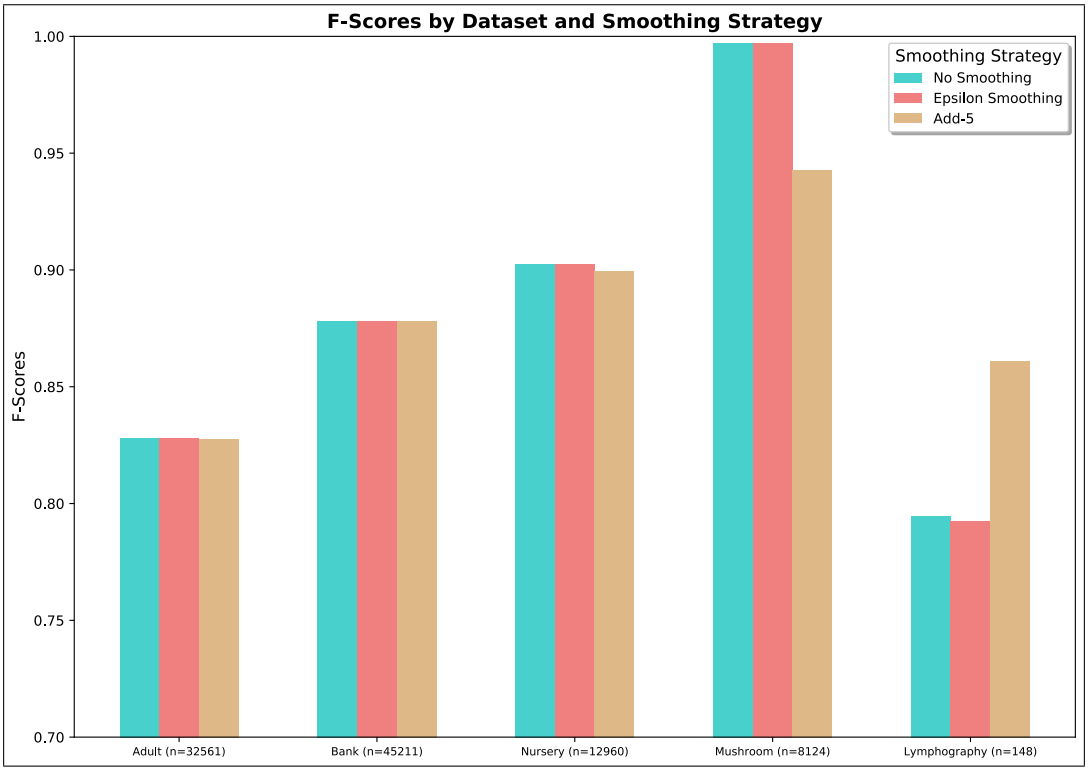

Figure 6: F-Scores ($\beta = 1$) for Add-5 smoothing.