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.

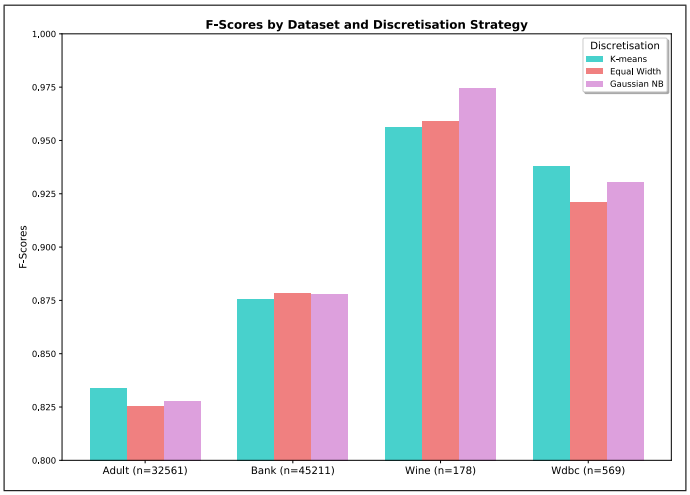

Figure 1: F-Scores ($\beta = 1$) for various discretisation methods.

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 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.

need to talk aout differences between baselines

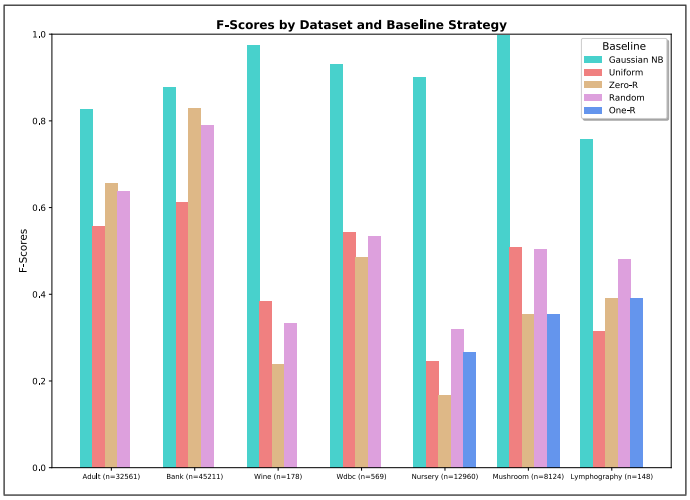

Figure 2: F-Scores ($\beta = 1$) for various baselines.

Question 4

		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
WBDC	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	0.977	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.	0.763	0.751	0.763	0.757

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

Question 5

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