ECM3420 - Learning from Data

Coursework 1: Decision Trees Learning and K-Nearest Neighbours

Reflection

Exploratory data analysis

I first looked at the summary statistics of each input variable, to get an idea of how they are distributed. Immediately we see the data contains some anomalous records. For example, there exists a person who has been pregnant 17 times, and another who has a BMI of 0. The dataset appears to use zeros for missing data points, which could cause inaccurate correlations between attributes that are often missing together.

Next, I standardise the data so I can compare distributions of input attributes in a violin plot. It shows the minimum, median and maximum values of each attribute, and a kernel density plot. Unlike the mean, the median is unaffected by outliers.

Logically, I expected there to be a positive correlation between age and number of pregnancies (older people likely to have more children). However, after age ~50, I expect the number of pregnancies to remain constant (women losing their fertility). To get a better understanding of their correlation, I use sns.jointplot to plot two things on one chart: a scatter plot between age and number of pregnancies (coloured by target label), and a linear regression line-of-best-fit. Colouring by target label did not reveal much additional insight, but was still relevant enough to keep in the chart. Looking at the regression gradient, there is clearly a positive correlation. The positive Pearson correlation coefficient confirms this ($0 <= \sim 0.54 <= 1$).

Classification

I first perform 5-fold cross-validation to compare how Decision Tree accuracy changes with respect to the splitting criterion: Gini Index (purity) or Entropy (information gain). Data was not standardised in pre-processing, as Decision Trees are concerned only with ordering and standardisation preserves order. The bar chart implies Entropy is only ~2% more accurate than Gini on average. This small difference is expected: they measure the same thing so are expected to be almost interchangeable. However, the box plot shows an outlier in Entropy fold accuracies has skewed the results. In the folds plot, we see Gini outperform Entropy on the 5^{th} fold by ~4%, and this is expected, as it is the outlier.

Then I perform 5-fold cross-validation to compare how KNN accuracy changes with respect to $k = \{1,3,5\}$. This time I standardised the data (zero-score

normalisation), so that the distance metric (Euclidean) was unbiased to the different units of input variables. The plot implies k = 3 is the optimal, with k = 5 being ~1% less accurate, and k = 1 being ~4% less accurate. We expect this: underfitting likely at 1 neighbour, potential overfitting at 5 neighbours. We also expect k = 1 to be worst, because noise will have a much higher influence on the result. An improvement would be to test a wider range of values of k.

Both experiments could be improved by doing them on multiple larger data sets, with multiple cross-validations (different splits).

Classification parameters DT

I use a 70:30 split to evaluate how Decision Tree accuracy changes with respect to two of its parameters: min_samples_split (minimum number of samples required to split an internal node) and max_depth (maximum tree depth).

Accuracies at min_samples_split = {2, 3} are the same, so likely resulted in the same tree. Between minimum split of 3 and 5, accuracy appears to be decreasing, however only by ~1%. This parameter is used to counter-act overfitting, and the ideal min_samples_split values tend to be between 1 to 40 [1], so this experiment could be improved by testing min_samples_split = {1,2,..,39,40}.

When max_depth = 4, it is ~2% more accurate than at max_depth = {3, 5, 6}. There is a ~0.5% increase in accuracy between max_depth of 5 and 6, but it is small and likely insignificant. At max_depth = {5, 6} the tree can grow deeper so is likely overfitting, and at max_depth = 3 it is limited so is likely underfitting.

Both experiments could be improved by doing them on multiple larger data sets, and by using multiple cross-validation instead of a 70:30 split.

References

[1] Rafael Gomes Mantovani, Tomáš Horváth, Ricardo Cerri, Sylvio Barbon Junior, Joaquin Vanschoren, André Carlos Ponce de Leon Ferreira de Carvalho, “An empirical study on hyperparameter tuning of decision trees”

Student number: 670021141