CS342 Machine Learning – Assignment 1 Student ID: 1605217

Nature of Inputs and Target Variables

The given information are 4 chemical compounds for each lifeform (cryptonine, mermaidine, posidine, neraidine) each at 3 different chemical resolutions from probe A and probe B, plus a genetic attribute TNA only from probeA. In both tasks i.e. predicting the missing classifications (plant v.s. animal) and TNA given probe B data, all 3 chemical resolutions of the 4 chemical compounds are used as the features in the models. Whereas the probability of a sample being an animal and TNA are the targets.

One of the biological explanations would be plants and animals are different living organisms, and therefore in theory should possess different chemical compositions and genetic attribute TNA. The abovementioned models made use of this fact to perform classification.

Procedure

In both tasks, I first rearrange the probe A and probe B data since some of the chemical resolutions are mixed up within the same chemical compound.

In Task 1, each of the chemical resolutions of the chemical compounds in probe A and probe B are standardised by subtracting the mean and dividing by the standard deviation respectively. A k-NN model is learnt by fitting the scaled attributes of probe A and their respective class, provided in classA.csv. Then, this model is used to predict the probability of the samples being class 1 from probe B. Here, I picked k = 34 after doing 10-fold cross-validation to choose the optimal parameter. I picked the k-NN model because all the features are numeric and a distance metric is more suitable for clustering in this task. I then construct the ROC curve and compute the area under the curve (AUC) which gives 0.707.

In Task2, after standardising as above, I wrote a function that takes the choice of models, probe A attributes, probe A TNA and probe B attributes as input and returns its R^2, a metric measuring the goodness of fit of the input model. I fitted all the models (OLS, RidgeCV, LassoCV, kNNReg, DecisionTreeReg and RandomForest) and did 10-fold cross-validations to learn the optimal parameters. Parameters and R^2 are as follows.

OLS: $R^2 = 0.0692$

RidgeCV: alpha = 10. $R^2 = 0.0692$

LassoCV: alpha = 0.0212. $R^2 = 0.0617$

kNNReg: k = 6. $R^2 = 0.653$

DecisionTreeReg: max_features=0.8, max_depth=8, min_samples_leaf=0.06. $R^2 = 0.339$

RandomForest: max_features=log2, n_estimators=600, max_depth=6, min_samples_leaf=0.1. $R^2 = 0.207$

To conclude, kNNReg has the highest R^2 among the above models, therefore I chose it to predict the TNA of probe B.

Predictivity of Attributes

Chemical Compounds	AUC
Cryptonine	0.666
Mermaidine	0.678
Neraidine	0.640
Posidine	0.611

Table 1: AUC of the models fitting each chemical compound individually with class A.csv

Chemical Compounds	AUC
w/ Cryptonine	0.711
w/ Mermaidine	0.658
w/ Neraidine	0.700
w/ Posidine	0.733

Table 2: AUC of the models fitting all chemical compounds but leave one out at a time with class A.csv (e.g. w/Cryptonine means a model fitting Mermaidine, Posidine and Neraidine with class A.csv)

Correlations	Class A
C1	0.0380
C2	0.0270
C3	0.0242
M1	0.0609
M2	-0.249
M3	-0.156
N1	-0.00547
N2	0.156
N3	0.0334
P1	-0.127
P2	-0.114
P3	-0.0860

Table 3: Correlations of each chemical resolutions with class A.csv

Looking at the correlations (Table 3) gave me an initial thought that m2, m3, n2, p1, p2 (in descending magnitude) have a relatively large relationship with the prediction of class. However, correlation does not imply causation.

So, I produced models by fitting each chemical compound individually with class A data (Table 1) to discover further relationship with the area under ROC curve. I could see that Cryptonine and Mermaidine have slightly higher AUC score. Together with the correlations, I would deduce Mermaidine is the most predictive attribute.

To further confirm my deduction, I produced models by fitting all chemical compounds but leave one out at a time with class A data (Table 2). I could clearly see that missing out Mermaidine reduce the AUC of ROC curve significantly. Since the area under the ROC curve is a measure of how well a parameter can distinguish between two diagnostic groups, the above results imply that missing out the attribute Mermaidine reduces the predictability of the model.

Therefore, Mermaidine is the most predictive attribute for this classification problem.