Machine Learning, Main paper 2018 Solutions.

December 19, 2018

Learning outcome	Questions
demonstrate a knowledge and understanding of the main approaches to machine learning	All questions.
demonstrate the ability to apply the main approaches to unseen examples	1c, 2b, 2c, 3a, 3b
demonstrate an understanding of the differences, advantages and problems of the main approaches in machine learning	2b, 3c
demonstrate an understanding of the main limitations of current approaches to machine learning, and be able to discuss possible extensions to overcome these limitations	1b, 2a, 2b
demonstrate a practical understanding of the use of machine learning algorithms	Coursework

1. (a) Supervised: there are matched pairs of dependent and independent variables
   Unsupervised: independent variables are know but there are no matching dependent variables
   [2 marks]
   Supervised learning type: classification/regression. In both cases one can have multiple independent variables which can be continuous or categorical. In regression, the dependent variables are continuous; in classification, the dependent variables are categorical.
   [3 marks + 1 quality mark]

   (b) Term 1: the variance (noise) in the dataset Term 2: the variance of the function, which measures how sensitive the optimal fit is to the particular realisation of the training set.
   Term 3: the bias, which measures the ability of f to model the true underlying function h.

[3 marks + 1 quality mark]

Taken together, these terms imply that there is a tradeoff between the bias (capacity) of a model and its variance. Low bias models have a high variance and can overfit the training data. Low variance models have a high bias and cannot model the trend.

[2 marks + 1 quality mark]

(c) The model has low bias and high variance and is overfitting the data, causing the large fluctuations between the data points.

[2 marks + 1 quality mark]

It could be improved through use of a more biased model to reduce the sensitivity to noise, through regularisation to control the model weights, or through the acquisition of more data, which also effectively reduces the variance of the model.

[3 marks + 1 quality mark]

2. (a) J-L is a statement that there exists a mapping $f : \mathbb{R}^M \mapsto \mathbb{R}^K$ from a high-dimensional space to a random low-dimensional subspace that preserves relative distances between points.
This is important because it provides a way to overcome the curse of dimensionality by providing a way to map a problem from high-d to low-d.

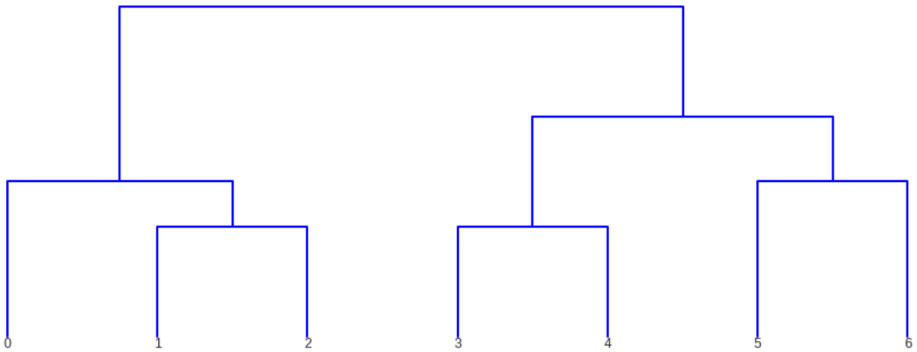
[4 marks for basic explanation + 2 quality marks for accurate explanation]

(b) Separability by hyperplanes implies that k-means will work here.

[1 mark for k-means + 1 quality mark for why]

Likely issues include poor initialisation (overcome by repeats); high dimensionality (overcome by random projection)

[4 marks for sound reasons and basic solutions + 1 quality mark for accuracy]



[5 marks for correct connectivity + 2 quality mark for approx accuracy of vertical heights]

3. (a) $\text{logit}(p_0) = 1/5$ leads to $5p_0 = p_1 + p_2$ (A)
$\text{logit}(p_1) = 0$ leads to $p_1 = p_0 + p_2$ (B)
We also know that $p_0 + p_1 + p_2 = 1$ (C)
Combining B with C gives $p_1 = 1/2$.
Sub B with $p_1 = 1/2$ into A gives $p_0 = 1/6$
Sub into C gives $p_2 = 1/3$.

[3 marks for correct answers + 1 quality mark for a clear and direct solution]

(b) First, check the assumptions. If assumptions of any method are clearly met then OK to use that method.
In either case, need to validate, either by cross-validation, or by train-validate-test split. You would choose the method by considering the best validation result and then finally testing on the held-out test set.

[5 marks for making basic points + 2 quality marks for a fuller explanation]

(c) Both methods based on ensembles of classifers (decision trees). Both use the average performance of a large number of learners.
Adaboost: Multiple learners trained sequentially. Importance of data points reweighted after each learner is used so that later learners are encouraged to correctly classify hard samples. Exponential loss used to heavily penalise incorrect classifications.
Random Forest: A bagging method in which multiple decision trees are trained on randomly sampled subsets of the features, which helps to reduce model variance without increasing bias.

[6 marks for making basic points + 3 quality marks for a fuller explanation]