Machine Learning 2018 Supplementary paper Solutions

December 18, 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	1b, 1c, 2a
demonstrate an understanding of the differences, advantages and problems of the main approaches in machine learning	2b
demonstrate an understanding of the main limitations of current approaches to machine learning, and be able to discuss possible extensions to overcome these limitations	2a, 2b, 3b
demonstrate a practical understanding of the use of machine learning algorithms	Coursework

1. (a) They are classification methods.
   [1 mark]
   Possible answers include:
   LDA: A: generative; D: strong statistical assumptions
   kNN: A: zero training time; D: very costly inference
   LR: A: can apply regression principles; D: same risks as in regression
   [6 marks + 2 quality marks for avoiding "obvious" statements.]

   (b) The boundary is approx at $x = 0$ and is straight.
   [2 marks]
   The boundary would move to the right ($x > 0$) whilst remaining straight.
   [2 marks]

Different covariance means QDA should be used instead, yielding a quadratic decision boundary.

[2 marks]

(c) This is a classification problem.

[1 marks]

Variables are on different scales so should be normalised.
Independent categorical variables need to be encoded (eg as one-hot vectors)
Check for correlated variables (eg height and weight) which could
[3 marks + 1 quality mark for insight.]

2. (a) Any example that contains two obvious clusters that cannot be separated by a straight line will receive credit. This reason should be the basis of the explanation as to why k-means won't work.

[2 marks + 1 quality mark for insight.]

The answer to the second part depends on the diagram. Almost all diagrams that I can think of will lead them agglomerative hierarchical clustering as the solution because GMMs will certainly fail if the classes are not linearly separable. A few may suggest an alternative distance metric and this should be credited.

[3 marks + 2 quality mark for creativity/insight.]

(b) Curse of dimensionality implies convergence of distances, concentration of mass at the edges of the space etc. Means that distances are not a good measure of similarity in high dimensions, and that probability distributions need to be sampled much for densely near the edges where most of their mass is.

[4 marks + 2 quality marks for linking concepts together]

(c) 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.

[4 marks + 2 quality marks for clarity]

3. (a) Bias: ability of model to represent the data (low bias is good)
Variance: sensitivity of model to the training data. Low bias typically requires complex model which is prone to being sensitive to the data (high variance).

A suitable diagram that illustrates this should be appropriately credited.

[4 marks + 2 quality marks for coherence]

(b) $R(\mathbf{w})$ encourages the model weights to take certain properties by increasing the loss for weights that do not have those properties.

[2 marks + 1 quality marks for coherence]

One option is $R(\mathbf{w}) = (\sum_i w_i^2 - 1)^2$, which is zero when the sum of the weights is 1, and we square it to prevent it taking negative values. Could also take the absolute value which isn't differentiable but that's not required.

[3 marks + 2 quality marks for explaining the reasoning]

(c) Cross validation or similar is what is needed here. The general principle is that this can only be done empirically through such a scheme. The key point is that the value that gives the best combined result on the training/validation sets is the one that should be chosen. A sketch of the test/validation loss would be appropriate

[4 marks + 2 quality mark for coherence]