Machine Learning

Solutions

Resit Examinations 2020

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 67, which will be rescaled to a mark out of 100.

Question 1

(a) In the notation used in the lectures, the quantities needed to solve a univariate unregularised least squares regression problem are:

- The vector of *independent* variables $\mathbf{x}$ with components $\{x_i\}_{i=1}^{N}$.
- The vector of *dependent* variables $\mathbf{y}$ with components $\{y_i\}_{i=1}^{N}$.
- The vector of model parameters $\mathbf{w}$ with components $\{w_i\}_{i=1}^{M}$.
- The basis states $\{\phi_i(x)\}_{i=1}^{M}$.

Explain how to construct the *normal equations* for unregularised regression from these quantities. You do *not* need to derive the normal equations from first principles. **[5 marks]**

(b) Explain the meaning of *bias* and *variance* in the context of a regression problem, illustrating your answer with appropriate diagrams. **[7 marks]**

(c) Explain the principle of regularisation and write down the general expression for the regularised least-squares loss function. Give two examples of regularisation functions and explain their effect. **[8 marks]**

Model answer / LOs / Creativity:

Learning outcomes: demonstrate a knowledge and understanding of the main approaches to machine learning (1); demonstrate an understanding of the differences, advantages and problems of the main approaches in machine learning (3).

(a) The normal equations are $\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^{\mathsf{T}}\mathbf{y} = 0$, where the components of $\mathbf{\Phi}$ are $\Phi_{ij} = \phi_j(x_i)$.

(b)
- Bias: ability of model to represent the data (low bias is good)
- Variance: sensitivity of model to noise in 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.

[7]

(c)

- Add term to loss function to penalise solutions with a certain structure or characteristic and therefore encourage solutions that do not have this characteristic. [2]

  $$\mathcal{L}(\mathbf{w}) = \|\mathbf{y} - \mathbf{f}(\mathbf{x}, \mathbf{w})\|^2 + \lambda R(\mathbf{w})$$

  [2]

- L_2 regularisation penalises large values of the model parameters and therefore encourages "shrinkage". [2]

- L_1 regularisation also penalises large values of the model parameters and encourages "shrinkage", but also tends to encourage sparsity in the parameter vector. [2]

Question 2

(a) A *decision stump* is a decision tree containing only one split on the most informative variable. Using the principle of maximising information gain, determine which variable should be used to form a decision stump for the data shown in the table below. **[5 marks]**

x_0	x_1	x_2	y
0	1	1	0
0	0	1	1
1	1	0	0
1	0	1	1

(b) Explain the random forest algorithm for classification. **[7 marks]**

(c) A *labelled* dataset contains 500 samples, each of which is from a 5-dimensional space. It is known that there are three (3) classes of data (A, B, C) in this dataset and each sample is drawn from one of those classes. The number of training points in each of the classes is A: 50; B: 250: C:200 . The classes are known to not be fully separable by three hyperplanes.

Explain how you would choose an algorithm to classify this dataset, what difficulties may be encountered, and how you would overcome them. **[8 marks]**

Model answer / LOs / Creativity:

Learning outcomes: demonstrate the ability to apply the main approaches to unseen examples (2); demonstrate a practical understanding of the use of machine learning algorithms (5). Parts a and c are creative.

(a) A sharp-eyed student will notice that the dataset can be split on x_1 to give homogeneous subgroups. A more pedestrian but just as acceptable approach is to compute the information gain for each variable. Both are acceptable. [5]

(b) The key points are:

- Multiple decision trees
- Each tree trained on random subspace
- Bagging: each tree trained on random sample from data (with replacement)
- Decisions are taken by the majority vote of the trees

[7 marks]

(c) The main points here are:

- Low-d space so dimensionality reduction not necessary.

- Not separable by hyperplanes therefore no point considering LDA.
- Classes are unbalanced so need to take care with a majority voting technique like *k*-nearest neighbours.
- It will be necessary to cross-validate a range of techniques on this data as a *a priori* selection looks difficult.

[8]

Question 3

(a) The Johnson-Lindenstrauss lemma can be stated as:

$$1 - \varepsilon \leq \frac{\|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|^2}{\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \leq 1 + \varepsilon$$

where $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^M$; $f : \mathbb{R}^M \mapsto \mathbb{R}^K$; $0 < \varepsilon < 1$ and $K < M$.

Explain the implications of this lemma and their relevance to machine learning. **[6 marks]**

(b) The table below contains a set of data with two variables. Each column contains one datapoint. *Sketch* the dendrogram for agglomerative hierarchical clustering using single-linkage on this dataset.

x_0	2.0	2.0	3.0	2.0	1.0	5.0	6.0
x_1	1.0	2.0	3.0	5.0	6.0	6.0	6.0

[7 marks]

(c) A common modification to the k-nearest neighbours algorithms is the *weighted* k-nearest neighbours algorithm.

(i) Describe how the *weighted* k-nearest neighbours algorithm works.

(ii) Sketch one example of a situation in which this method will give rise to an incorrect decision. Explain your reasoning.

[7 marks]

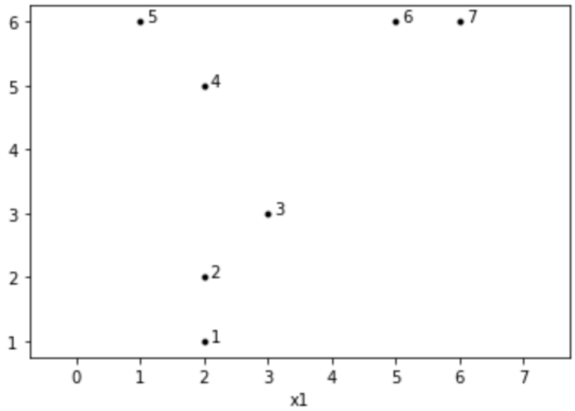
Model answer / LOs / Creativity:

Learning outcomes: demonstrate an understanding of the differences, advantages and problems of the main approaches in machine learning (3); demonstrate an understanding of the main limitations of current approaches to machine learning, and be able to discuss possible extensions to overcome these limitations (4); demonstrate a practical understanding of the use of machine learning algorithms (5). Parts (b) and (c)i are creative.

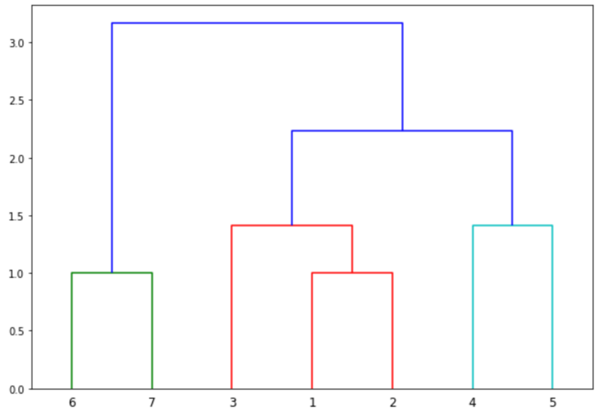
(a) J-L is a statement that there exists a mapping $f : \mathbb{R}^M \mapsto \mathbb{R}^K$ from a high-dimensional space to a random low-dimensional subspace that preserves relative distances between points.

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.

(b) The data is plotted below:



from which the dendrogram can be easily derived as



(c) (i) The pseudocode is:

```
for each training point p
  for each class C
    compute the mean distance between p and the points in C
  assign p to the class with the minimum mean distance
```

[4]

(ii) This can happen when the two classes are close and one of them is very compact. Any feasible example will suffice here as long as the reasoning is sound. [3]