Machine Learning

Solutions

Main Summer Examinations 2020

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60, which will be rescaled to a mark out of 100.

Question 1

(a) Describe two ways in which the tendency of decision trees to overfit their training data can be overcome **[5 marks]**

(b) Provide the mathematical definition of *information entropy* and sort the following binary strings from lowest to highest entropy. **[5 marks]**

(i) 1101101110

(ii) 1101010010

(iii) 1001000101

(c) The following table describes a binary classification dataset $\mathcal{D} = \{x_i, y_i, z_i, t_i\}_{i=0}^{7}$ with independent variables x, y, and z; and dependent variable t.

i	x_i	y_i	y_i	t_i
0	0	0	0	0
1	0	0	1	0
2	0	1	0	0
3	0	1	1	1
4	1	0	0	0
5	1	0	1	1
6	1	1	0	1
7	1	1	1	1

The data generating process returns $t = 1$ when two or more of the independent variables are 1.

Using the principle of maximum information gain to determine the order of the variable splits, construct the full decision tree for this dataset, using data points $i = \{0, 1, 2, 4, 6, 7\}$ as the training set. Test your tree on data points $i = \{3, 5\}$ and comment on your result. **[10 marks]**

You may find the following table of logarithms helpful.

x	1/4	1/3	1/2	2/3	3/4	1
$\ln(x)$	-1.386	-1.099	-0.693	-0.405	-0.288	0

Model answer / LOs / Creativity:

Learning outcomes: demonstrate the ability to apply the main approaches to unseen examples (2); 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). The creative part is part (c).

(a)
- Restrict the tree depth.
- Use in an ensemble, eg, through boosting, or in a random forest.

(b) Information entropy $S = \sum_i p(i) \ln p(i)$. The information entropy of the three strings is:

(i) $p(0) = 0.3, p(1) = 0.7 \quad \mapsto \quad S = -0.3 \ln 0.3 - 0.7 \ln 0.7 = -\ln 0.5 = 0.611$

(ii) $p(0) = 0.5, p(1) = 0.5 \quad \mapsto \quad S = -0.5 \ln 0.5 - 0.5 \ln 0.5 = -\ln 0.5 = 0.693$

(iii) $p(0) = 0.6, p(1) = 0.4 \quad \mapsto \quad S = -0.6 \ln 0.6 - 0.4 \ln 0.4 = 0.673$

So the order is i–iii–ii.

(c) The training set is:

i	x_i	y_i	z_i	t_i
0	0	0	0	0
1	0	0	1	0
2	0	1	0	0
4	1	0	0	0
6	1	1	0	1
7	1	1	1	1

First, we compute the parent entropy on the whole set, which has four zeros and two ones. The entropy of this is $S = -\frac{1}{3} \ln \frac{1}{3} - \frac{2}{3} \ln \frac{2}{3} = 0.637$.

Now we split on each of the variables

x_0: three zeros gives three zeros; three ones give one zero and two ones. The entropy of this is $S = \frac{1}{2} (-1 \ln 1 - 0 \ln 0) + \frac{1}{2} \left(-\frac{1}{3} \ln \frac{1}{3} - \frac{2}{3} \ln \frac{2}{3}\right) = 0.318$.

x_1: This is the same as x_0.

x_2: four zeros gives three zeros and one one; two ones give one zero and one ones. The entropy of this is $S = \frac{2}{3} \left(-\frac{3}{4} \ln \frac{3}{4} - \frac{1}{4} \ln \frac{1}{4}\right) + \frac{1}{3} \left(-\frac{1}{2} \ln \frac{1}{2} - \frac{1}{2} \ln \frac{1}{2}\right) = 0.51$.

So, the first split in the data should be made on either x or y.

- Choice 1: split on x. Selecting the samples where $x = 0$ we have:

i	x_i	y_i	z_i	t_i
0	0	0	0	0
1	0	0	1	0
2	0	1	0	0

This branch is homogeneous with entropy $S = 0$ and so it cannot be split further.

Selecting the samples where $x = 1$ we have:

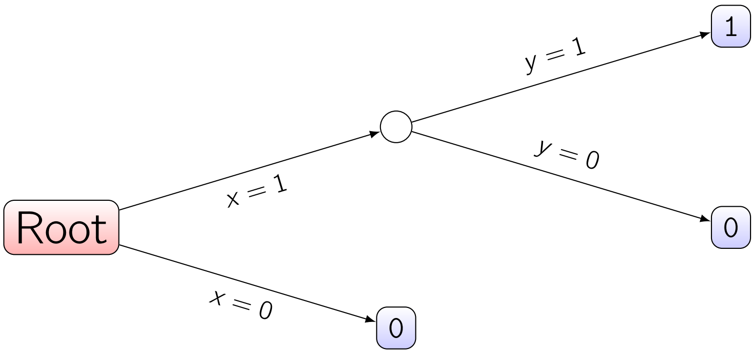
i	x_i	y_i	z_i	t_i
4	1	0	0	0
6	1	1	0	1
7	1	1	1	1

The parent state entropy is $S = -\frac{1}{3}\ln\frac{1}{3} - \frac{2}{3}\ln\frac{2}{3} = 0.637$.

Splitting on y: one zero gives one zero; two ones give two ones. The entropy $S = 0$.

Splitting on z: two zeros gives one one, one zero; one one gives one one. The entropy is $\frac{2}{3}\left(-\frac{1}{2}\ln\frac{1}{2} - \frac{1}{2}\ln\frac{1}{2}\right) + \frac{1}{3}(-\ln 1) = 0.462$.

Therefore, we split on y. This gets us to homogeneous leaf nodes and there is no need to split on z. The resulting tree is therefore:



- Choice 2: split on y. Selecting the samples where $y = 0$ we have:

i	x_i	y_i	z_i	t_i
0	0	0	0	0
1	0	0	1	0
4	1	0	0	0

This branch is homogeneous with entropy $S = 0$ and so it cannot be split further.

Selecting the samples where $x_0 = 1$ we have:

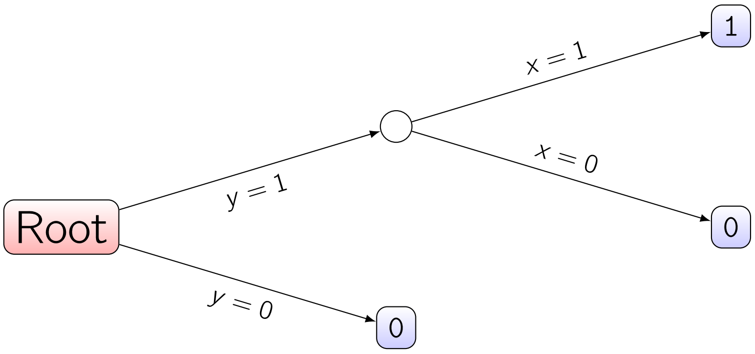
i	x_i	y_i	z_i	t_i
2	0	1	0	0
6	1	1	0	1
7	1	1	1	1

The parent state entropy is $S = -\frac{1}{3}\ln\frac{1}{3} - \frac{2}{3}\ln\frac{2}{3} = 0.637$.

Splitting on x: one zero gives one zero; two ones give two ones. The entropy $S = 0$.

Splitting on z: two zeros gives one one, one zero; one one gives one one. The entropy is $\frac{2}{3}\left(-\frac{1}{2}\ln\frac{1}{2} - \frac{1}{2}\ln\frac{1}{2}\right) + \frac{1}{3}(-\ln 1) = 0.462$.

Therefore, we split on x. This gets us to homogeneous leaf nodes and there is no need to split on z. The resulting tree is therefore:



Classifying data points 3 (011) and 5 (101) with either tree gives 0 incorrectly in both cases. The tree has failed to generalise.

Question 2

(a) L_2 regularisation is sometimes used to control the solution to regression problems. It is often referred to as a *shrinkage* method.

(i) What is meant by the term "shrinkage method"?

(ii) Explain why L_2 regularisation is classified as a shrinkage method, with reference to its probabilistic interpretation.

(iii) Give another example of a shrinkage method and explain how it influences the solutions of regression problem.

[5 marks]

(b) The expectation value of the least squares loss function can be written as

$$\mathbb{E}[\mathcal{L}] = \underbrace{\sigma^2}_{i} + \underbrace{\text{var}[f]}_{ii} + \underbrace{(h - \mathbb{E}[f])^2}_{iii}$$

Explain the meaning of the symbols σ, f, and h; and of the terms i, ii, and iii. **[6 marks]**

(c) Classification problems can be solved using a regression-type approach with the *logistic regression* algorithm. Explain the principles of binary logistic regression, how it can be extended to multi-class problems, and what advantages and disadvantages it has over other methods. **[9 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) (i) Shrinkage methods encourage model weights to be kept small.

(ii) It tries to minimise the sum of the squares of the model weights. This can be shown to be equivalent to imposing a Gaussian prior of mean zero on the weights and so large weights are unlikely.

(iii) L_1 is one examples, this tends to encourage sparse weights (mostly zeros)

(b)
- σ: Standard deviation of data generating process
- f: estimated function
- h: Mean of the data-generating process
- Term i: implicit loss due to measurement uncertainty
- Term ii: the variance in the estimated function as a consequence of the measurement uncertainty

- Term iii: the difference between the estimated function and the mean of the true data generating process (bias)

(c) The key points are:

- Assume one can construct a function that compute the probability of a data point being in one of two classes.
- Construct the log-odds of a point being in one of the classes.
- Fit the log-odds of the binary decision with a linear model.
- Form the joint probability density function of the data.
- Rewrite the joint PDF in terms of the fitted model.
- Maximise the likelihood to find optimal model parameters.
- Construct explicit decision rule.
- Handle multi-class case by pivoting against one class.
- Does not require the form of the PDF to be known.
- Very good out-of-the box.
- Can be easily controlled by regularisation.
- Tends to be better with more data.

A clear explanation of the method with be worth 5 marks. Extension to multiple classes is worth 2 marks. Advantages/disadvanatges worth 2 marks.

Question 3

(a) Sketch an example of a 2-dimensional dataset containing three classes of data point (unlabelled) that could not be separated by k-means clustering using Euclidean distance, indicating on your sketch how k-means would incorrectly partition the data. **[5 marks]**

(b) A researcher in the School of Chemistry has asked for your help. They have been collecting samples of water from different places around the world and have been trying to measure what is in the samples to understand the effects of environmental pollution. They have analysed all of the samples using a technique called mass spectrometry, which measures the number of molecules of a particular mass, for a range of different masses. For each sample, this produces a histogram that shows how many molecules of each mass were in the sample. This histogram can be represented as a vector, where each component corresponds to a mass value, and the value of the component is the number of molecules of that mass. The instrument used to obtain this data can measure the number of molecules at each of 1.2 million different mass values.

The researcher has samples from 10,000 different locations and wishes to separate them into groups of similar water composition. Suggest how you would do this, highlighting any potential problems you might encounter, and how you would solve them. **[7 marks]**

(c) The researcher uses another technique to identify 12 distinct types from 500 randomly chosen samples. Suggest how you might use this information to improve your results from part (b), again highlighting any potential problems you might encounter, and how you would solve them. **[8 marks]**

Model answer / LOs / Creativity:

Learning outcomes: demonstrate a knowledge and understanding of the main approaches to machine learning (1); demonstrate the ability to apply the main approaches to unseen examples (2); demonstrate an understanding of the main limitations of current approaches to machine learning, and be able to discuss possible extensions to overcome these limitations (3); demonstrate a practical understanding of the use of machine learning algorithms (4). The creative part is part (c).

(a)
- Any reasonable example that cannot be separated by three straight lines will suffice [3 marks]
- The incorrect partitions should be approximately midway between the cluster centroids [2 marks].

(b)
- This is a clustering problem so could be solved with k-means or hierarchical clustering. [3 marks]

- Issues will include high dimensionality causing slow clustering (solved by dimensionality reduction); lack of knowledge of number of clusters, which will need to be solved by cross-validation; the possibility that the sample might not be representative of the data. [4 marks]

(c) This is a very challenging open-ended question that is designed to test student's ability to think outside of the taught material. The key idea here is that the problem is now semi-supervised. Possible approaches could include:

- Train a classifier on the labelled points and use this to classify the unlabelled points.
- Use the labelled points to seed *k*-means clustering.

One potential problem is that there may be more than twelve clusters in the data.

Credit will be awarded for imaginative and self-consistent solutions that demonstrate insight into and understanding of the question. [8 marks]