PAPER C395

INTRODUCTION TO MACHINE LEARNING

Tuesday 17 March 2020, 10:00
Duration: 90 minutes
Post-processing time: 30 minutes
Answer TWO questions

Paper contains 3 questions

1 a You are given three different problems below. For each problem, classify it as either *supervised learning*, *unsupervised learning*, or *reinforcement learning*. Also justify (**in one sentence**) why you classified each problem as such.

i) A book distributor has a collection of books, which it has classified into different categories, e.g. "Young adults", "Science fiction", "Biography", and "Horror". It wants to use this information to build a system to classify its new products automatically.

ii) A supermarket has a database of its customers, and wants to automatically discover and group its customers into different market segments to target them separately.

iii) A group of aviation companies wants to develop a machine learning algorithm which can predict the (x, y) coordinates pinpointing the location of a plane that has crashed. To develop this, the companies have collected historical plane crash data, which include the coordinates of the planes' crash sites.

b You are given the dataset below. Each row is a sample email annotated as spam (or not), given whether or not a word appears in the email (0 indicates that the word does not appear in the email, 1 indicates that it does).

#	cash	win	debt	home	Spam?
1	0	1	0	0	no
2	1	1	1	0	yes
3	1	0	0	0	no
4	0	0	1	0	no
5	1	1	0	1	yes
6	1	1	1	1	yes
7	0	1	1	0	yes
8	1	0	1	1	no
9	0	0	0	0	no
10	1	1	0	0	yes

i) Using the *Information Gain* metric, which attribute will be selected as the *root* node of a decision tree classifier? Please show all calculations (including the Information Gain of all candidate nodes) to justify your answer.

ii) Give one reason why one would need to prune a decision tree. Also describe (in one or two sentences) how a validation set can be useful in performing pruning.

c Consider the training dataset below for a single-variable *regression* problem.

i	$x^{(i)}$	$y^{(i)}$	$d(x^{(q)}, x^{(i)})$	$w_q^{(i)}$
1	1.5	3.16	???	???
2	2.3	1.45	???	???
3	3.0	1.07	???	???
4	3.8	2.01	???	???
5	4.9	4.51	???	???

i) At test time, you are given a query $x^{(q)} = 4.2$.
Given $d(x^{(q)}, x^{(i)}) = |x^{(q)} - x^{(i)}|$ and $w_q^{(i)} = \frac{1}{d(x^{(q)}, x^{(i)})}$, where $|x|$ indicates the absolute value of x, please complete the table above.

ii) Predict the output $y^{(q)}$ for $x^{(q)} = 4.2$ using the k-nearest neighbours regression algorithm with $d(x^{(q)}, x^{(i)})$ as its distance measure, and assuming $k = 3$. Show your calculation.

iii) Now predict the output $y^{(q)}$ for $x^{(q)} = 4.2$ using the *locally weighted* k-nearest neighbours regression algorithm. Use $k = 3$, the distance measure $d(x^{(q)}, x^{(i)})$, and the weights $w_q^{(i)}$. Show your calculation.

The three parts carry, respectively, 30%, 40%, and 30% of the marks.

2 a Consider a fully-connected feedforward neural network for regression with 1 hidden layer and a single output neuron. Both the hidden and the output layers use sigmoid activation. Mean squared error is used as the loss function for optimisation.

Write out the necessary calculations for updating both the connection weights and bias weights in the last layer using gradient descent. You are given the matrix of input features X with N datapoints, output values $\hat{Y}$ from the network, the desired targets (labels) Y, and the current network weights.

b Explain the concept of overfitting. Name 3 methods you can use to deal with overfitting and explain how each of them helps.

c A bank has developed a machine learning model for automatically identifying fraudulent card transactions, which will then be manually reviewed. You run some examples through the model and get the following results:

Example nr	**True label**	**Predicted label**
1	fraud	real
2	real	real
3	real	fraud
4	real	real
5	fraud	real
6	real	real
7	real	real
8	fraud	fraud
9	real	real
10	real	real
11	real	real
12	fraud	real
13	fraud	fraud
14	real	real

i) Construct the confusion matrix for this output.

ii) Calculate the classification accuracy, along with precision, recall and F1 for both classes.

iii) Analyse the results. Are there any issues? If so, which metrics identify them?

The three parts carry, respectively, 40%, 30%, and 30% of the marks.

3 a Suppose at an update step, the K-means algorithm computes 3 cluster centroids: $\mu_1 = \langle -3, -1 \rangle$, $\mu_2 = \langle 1, 2 \rangle$, and $\mu_3 = \langle -4, 1 \rangle$. It then executes a cluster assignment step. Assume that the algorithm uses a Euclidean distance measure. To which cluster will the training example $x^{(i)} = \langle -2, 0 \rangle$ be assigned after the cluster assignment step? Justify your answer by showing your calculations.

b Consider below the parameters $\theta = \{\pi_k, \mu_k, \sigma_k^2 : k = 1, 2, 3\}$ of a univariate Gaussian Mixture Model with 3 components that has been fitted to a set of training examples, where π_k is the mixing proportion of component k, μ_k the mean of component k, and σ_k^2 the variance for component k:

k	π_k	μ_k	σ_k^2
1	0.5	-2	1
2	0.3	1	4
3	0.2	4	0.25

Suppose you are given an example $x^{(i)} = 0$ at test time. Compute the probability density for $p(x^{(i)}|\theta)$ given the parameters of the Gaussian Mixture Model above. Show your calculations.

Hint: the Gaussian distribution is defined as: $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$

c When developing a neural network, which activation function and loss function would you use in the output layer for the following applications? Justify your decisions.

i) Predicting the temperature for tomorrow, based on the weather today.

ii) Generating text by predicting the next word in the sequence based on the previous words.

iii) Detecting whether the camera image from a self-driving car contains a stop sign.

d Hyperparameters are something that we need to deal with when designing machine learning models.

i) Explain what hyperparameters are and give 2 examples.

ii) Given a dataset of 10,000 datapoints, how would you use it to find good hyperparameters?

The four parts carry, respectively, 20%, 30%, 30%, and 20% of the marks.