sijia.chen1@student.unsw.edu.au

18s1: COMP9417 Machine Learning and Data Mining

Lectures: Introduction to Machine Learning and Data Mining
Topic: Questions from lecture topics
Last revision: Fri Mar 2 16:17:18 AEDT 2018

week1

Introduction

Some questions and exercises from the first course lecture, an "Introduction to Machine Learning and Data Mining", focusing on reviewing some basic concepts and terminology, and building some intuitions about machine learning.

Question 1

a) What is the function that Linear Regression is trying to minimize ?

b) Under what conditions would the value of this function be zero ?

c) Can you suggest any other properties of this function ?

Question 2 Machine learning has a fair amount of terminology which it is important to get to know.

a) Why do we need features ?

b) What is a the difference between a “task”, a ”model” and a “learning problem” ?

c) Can different learning algorithms be applied to the same tasks and features ?

p34

Table 1: Training set for basic linear classifier.

positive examples			negative examples		
x_1	x_2	**class**	x_1	x_2	**class**
4	7	+	2	2	−
5	9	+	3	1	−
6	8	+	4	3	−

p41,28

Table 2: Instances to be classified ('?' indicates unknown class to be predicted).

x_1	x_2	**class**	x_1	x_2	**class**
2	7	?	3	4	?

Question 3 Suppose you are given the following training set of 6 examples where each example is described using 2 numeric features, "x_1" and "x_2". Your training set, shown in Table 1, contains 3 positive examples of the target class, labelled "+", and 3 negative examples of the target class, labelled "−".

Use the information on linear classification on slides 26–27 and the *basic linear classifier* on slides 40–41 to derive the weight vector $\mathbf{w}$ and threshold t for a basic linear classifier from the training set above.

Now "run" your classifier to classify the 2 instances shown in Table 2. Is the classification what you would have expected ? Why ?

Table 3: Posterior probability distribution of classes given word occurrence (bold font indicates more probable class).

valuation	manufacturing	$P(Y = \text{business} \mid \text{valuation}, \text{manufacturing})$	$P(Y = \text{general} \mid \text{valuation}, \text{manufacturing})$
0	0	0.3	**0.7**
0	1	0.5	0.5
1	0	**0.6**	0.4
1	1	0.9	**0.1**

Table 4: Marginal likelihoods: think of these as probabilities of observing the data items (words) independently of any others, given the repective classes.

Y	$P(\text{valuation} = 1 \mid Y)$	$P(\text{valuation} = 0 \mid Y)$
business	0.3	0.7
general	0.1	0.9

Y	$P(\text{manufacturing} = 1 \mid Y)$	$P(\text{manufacturing} = 0 \mid Y)$
business	0.4	0.6
general	0.2	0.8

Question 4 To answer this question you will need to refer to the probabilistic approach described on slides 47–55 from the lecture. Imagine you are asked to use a probabilistic model to learn to classify text files containing news articles as either 'business' or 'general'. To illustrate, we will only consider the presence or absence of two *keywords*, 'valuation' and 'manufacturing' in the text files. For simplicity we will further assume the two classes are mutually exclusive, i.e., text files can only have one class, either 'business' or 'general'.

Shown in Table 3 are the probabilities of the classes given the presence (1) or absence (0) of the keywords in the text.

Table 4 shows the marginal likelihoods of independently observing each of the keywords given each class.

a) using the data from Table 3, what two patterns of occurrence of keywords in a text file lead to a prediction of 'business' ?

b) what prediction should be made if we have an occurrence of 'manufacturing' but NOT 'valuation' in a text file ?

c) suppose we are given a text file to classify, and we know that 'manufacturing' occurs in the text file, but we know some words are missing from the file for some reason, and we are uncertain if 'valuation' occurred or not. However, we do know that the probability of 'valuation' occuring in any text file is 0.05. Compute the probability of each class for the given text file.

d) using the values from Table 4 compute the likelihood ratios for each of the four possible patterns of occurrence of the keywords.