17s1: COMP9417 Machine Learning and Data Mining

Lectures: Introduction to Machine Learning and Data Mining
Topic: Questions from lecture topics
Last revision: Thu Mar 9 09:37:08 AEDT 2017

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 ?

Question 3 Suppose you run a learning algorithm that returns a basic linear classifier using the *homogeneous coordinates* representation on a set of training data and obtain the follow weight vector $\mathrm{w} = (-0.4, 0.3, 0.2)$. For each of the following examples, what would the classification be using the weight vector w ?

a) $\mathrm{x}_1 = (0.9, 1.1)$?

b) $\mathrm{x}_2 = (0.3, 1.2)$?

c) $\mathrm{x}_3 = (0.0, 2.0)$?

Question 4 You want to use a probabilistic model to learn to classify text files as containing either 'business' or 'general' news articles. To illustrate, we will only consider the presence or absence of two *keywords*, 'valuation' and 'manufacturing' in the text files. To simplify we assume the two classes are mutually exclusive, i.e., text files can only have one class, either 'business' or 'general'.

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

Table 2 shows the marginal likelihoods of independently observing each of the keywords given each class.

Table 1: 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 2: 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

a) using the data from Table 1, 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 2 compute the likelihood ratios for each of the four possible patterns of occurrence the keywords