18s1: COMP9417 Machine Learning and Data Mining

Lectures: Introduction to Machine Learning and Data Mining
Topic: Questions from lecture topics
Version: with answers

Last revision: Fri Mar 2 16:17:18 AEDT 2018

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 ?

Answer

a) Linear regression aims to minimize the sum of squared errors[a]. This is defined as $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ where y_i is the actual value of the target (dependent) variable and $\hat{y}_i$ is the value predicted by the learned regression function for example i in the training set.

b) If the "line" (hyper-plane in general) passes through all of the values of the output in the training data, i.e., $\forall i \; y_i = \hat{y}_i$.

c) It is non-negative (because the error term is squared) and has a unique minimum (because the derivatives of the squared error term are linear).

[a] Slides 19 and 20 of the lecture muddy the waters a bit here; the slides talk about Mean Squared Error (MSE), which is what is minimised in the typical setup, but what is shown on the slides in the sum of squared errors, also known as the residual sum of squares (RSS).

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 ?

Answer

a) Essentially, features are the “interface” between the raw data and the model which is to be learned. For example, you might be given raw data on income which has been split into a number of age ranges and you want to agregate this into a single number, giving you a new feature, say “expected income”.

b) A task defines a mapping from input features to output, e.g., mapping from demographic features to income, whereas a model is a specific form of that mapping that can be learned by an algorithm, such as a linear regression equation, thereby defining a learning problem.

c) Yes, for example, you could use a linear classifier learning algorithm or a nearest neighbour learning algorithm for the same classification task using the same features.

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	−

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 ?

Answer

We are told on slide 26 that a linear classifier is defined by $\mathbf{w} \cdot \mathbf{x}_i = t$ for some example x_i to be classified, and in the lecture we mentioned that if $\mathbf{w} \cdot \mathbf{x}_i > t$ then x_i is classified *positive*. So we need to find (i.e., "learn") $\mathbf{w}$ and t from the training set.

First, we find the centroids $\mathbf{p}$ of the positives and $\mathbf{n}$ of the negatives (as was mentioned in the lecture, we can do this by taking the means of the respective subsets of labelled examples). We get $\mathbf{p} = (5, 8)$ and $\mathbf{n} = (3, 2)$. On slide 40 we are told that $\mathbf{w} = \mathbf{p} - \mathbf{n}$, so we obtain $\mathbf{w} = (2, 6)$.

Next, on slide 41 we are told that $t = (||\mathbf{p}||^2 - ||\mathbf{n}||^2)/2$, where $||\mathbf{x}||$ denotes the length of vector $\mathbf{x}$, i.e., $\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$ assuming $\mathbf{x}$ has n components. If you are not sure where this expression for t comes from, try deriving it from the definition of a linear classifier, and what we are told about the location of any point on the separating hyperplane of the basic linear classifier.

We can now obtain $t = ((5^2 + 8^2) - (3^2 + 2^2))/2 = 38$. Plugging in the two examples to be classified we find the following. $(2, 6) \cdot (2, 7) = 46 > 38$, so the first example is classified *positive*. $(2, 6) \cdot (3, 4) = 30 < 38$, so the second example is classified *negative*.

You could argue that the classification is what you would expect if you consider, for each example, whether it is closer to the centroid $\mathbf{p}$ of the positives, or to the centroid $\mathbf{n}$ of the negatives.

It is interesting to apply the method using the extended representation of "Homogeneous coordinates" described on slide 28. The following shows this for the same two examples. $(-38, 2, 6) \cdot (1, 2, 7) = 8 > 0$, so the first example is classified *positive*. $(-38, 2, 6) \cdot (1, 3, 4) = -8 < 0$, so the second example is classified *negative*.

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.

Answer

a) if we see at least one occurrence each of 'valuation' and 'manufacturing', or just at least one occurrence of 'valuation', we should predict 'business'. Note: these scenarios will give the same classification, but with different probabilities !

b) both classes are equally probable, so without any further information it is irrational to make a prediction; however, if we are told that one class is more probable *a priori* then we could use that fact to make a prediction by default.

c) we use *marginalisation* to average over the two possibilities, i.e., that 'valuation' did or did not occur. We compute the conditional probabilities for each class given this average evidence. We obtain the probabilities from Table 3. First, the formula to use is: $P(Y|\mathsf{valuation} = 0, \mathsf{manufacturing})P(\mathsf{valuation} = 0) + P(Y|\mathsf{valuation} = 1, \mathsf{manufacturing})P(\mathsf{valuation} = 1)$ and we evaluate this for **each** of the classes $Y = \mathsf{business}$ and $Y = \mathsf{general}$. For $Y = \mathsf{business}$ this evaluates to $(0.5 * 0.95) + (0.9 * 0.05) = 0.52$ and for $Y = \mathit{general}$ this evaluates to $(0.5 * 0.95) + (0.1 * 0.05) = .48$. Since 'valuation' mostly does NOT occur, we see that this is pretty close to the posterior probabilities of each class (0.5) in the second row of Table 3 when 'valuation' is KNOWN not to occur.

d) we need to multiply together the (independent) marginal likehood ratios to obtain the overall likelihood ratio, for each instantiation of the two keywords denoting whether the word appears in the document, or not. Letting X_1, X_2 stand for the occurrence of the keywords, the formula is $\frac{P(X_1|Y=\mathsf{business})}{P(X_1|Y=\mathsf{general})} \times \frac{P(X_2|Y=\mathsf{business})}{P(X_2|Y=\mathsf{general})}$. Expanding this out for each of the combinations of keyword occurrences, this gives:

$$\frac{P(\mathsf{valuation} = 0|Y = \mathsf{business})}{P(\mathsf{valuation} = 0|Y = \mathsf{general})} \times \frac{P(\mathsf{manufacturing} = 0|Y = \mathsf{business})}{P(\mathsf{manufacturing} = 0|Y = \mathsf{general})} = \frac{0.7}{0.9}\frac{0.6}{0.8} = 0.58 \qquad (0.3)$$

$$\frac{P(\mathsf{valuation} = 0|Y = \mathsf{business})}{P(\mathsf{valuation} = 0|Y = \mathsf{general})} \times \frac{P(\mathsf{manufacturing} = 1|Y = \mathsf{business})}{P(\mathsf{manufacturing} = 1|Y = \mathsf{general})} = \frac{0.7}{0.9}\frac{0.4}{0.2} = 1.55 \qquad (0.5)$$

$$\frac{P(\mathsf{valuation} = 1|Y = \mathsf{business})}{P(\mathsf{valuation} = 1|Y = \mathsf{general})} \times \frac{P(\mathsf{manufacturing} = 0|Y = \mathsf{business})}{P(\mathsf{manufacturing} = 0|Y = \mathsf{general})} = \frac{0.3}{0.1}\frac{0.6}{0.8} = 2.25 \qquad (0.6)$$

$$\frac{P(\mathsf{valuation} = 1|Y = \mathsf{business})}{P(\mathsf{valuation} = 1|Y = \mathsf{general})} \times \frac{P(\mathsf{manufacturing} = 1|Y = \mathsf{business})}{P(\mathsf{manufacturing} = 1|Y = \mathsf{general})} = \frac{0.3}{0.1}\frac{0.4}{0.2} = 6.00 \qquad (0.9)$$

The decision in all but the first row is: predict '$\mathsf{business}$'. This agrees with the decision rule using the posteriors in all but the second row (where both predictions are equally probable).