Exam Machine Learning

7 Aug 2014, 10:00 - 13:00

Answers in bold italics.

You may use Peter Flach's book (the book itself or a printout) and a calculator.

QUESTION 1 (10 points)

Suppose that we apply logistic regression with two predicting variables x_1 and x_2 to a dataset.

1. Suppose that we find the following values for the parameter vector: $\theta_0 = 1$, $\theta_1 = 1$ and $\theta_2 = 1$. Draw the decision boundary in a graph (with x_1 and x_2 as axes).

   From function; the line for which $1 + x_1 + x_2 = 0$

2. Now, assume that we find some values for the parameters θ_0, θ_1 and θ_2 for which the cost function $J(\theta_0, \theta_1, \theta_2) = 0$. Which of the following statements must be true? Explain your answer:

   (a) This is not possible. Because of the definition of $J(\theta_0, \theta_1, \theta_2)$ there can be no θ_0, θ_1, and θ_2 such that $J(\theta_0, \theta_1, \theta_2 = 0)$

   (b) For this $y(i) = 0$ for every $i = 1, 2, ..., m$.

   (c) Probably the gradient reached a local minimum rather than a global minimum.

   (d) For this we must have $h_\theta(x(i)) = y(i)$ for each training example $(x(i), y(i))$

   if the loss is 0 then the sum of the errors of all data points must be 0; this means that only **d** ***is correct. Loss cannot be less than 0, so it must be a global maximum***

QUESTION 2 (10 points)

Consider three "architectures" for neural networks. Each is based on an input vector X, a weight vector θ, has a "bias" variable with corresponding weight and a single output Y The input values (in X) are all either 1 or

-1. Each architecture multiplies the input values with the corresponding parameters (setting the bias "input" variable to 1).

The architectures vary in how they combine the weighted inputs. Architecture **A** combines the values that result from multiplying input with parameters inputs by adding them and passes the result to the output Y. Architecture **B** also adds the resulting values but outputs 1 if the result is positive, otherwise it outputs 0. Architecture **C** instead applies the logistic function to the result.

1. Give a reasonable cost function (that we would minimize when training the network) for each architecture **A**, **B** and **C**.

   A: mean squared error, B: equal $\rightarrow 0$***, different*** $\rightarrow 1$***, C: mean squared error (of die van logistic regression)***

2. Describe the type of prediction model that each architecture **A**, **B** and **C** can learn.

   A: lineaire (regressie) functie, geeft continue waarde, B & C: lineaire discriminant functie

3. Now suppose that we add an extra (hidden) layer of the same type to each architecture between inputs and output. Does this change your answer to the previous two questions? Explain your answer.

   voor de loss maakt dit niet uit; voor het type functie wel: A: blijft hetzelfde; B: ??; C: nu ook niet-lineaire functies

QUESTION 3 (10 points)

We apply a neural network to a dataset. The resulting accuracy is very high and we are concerned that it overfits.

1. How can we find out if the network overfits the data?

   met kruisvalidatie of althans train/testset

2. If this is the case then which of the following actions is/are likely to reduce this problem? For each action indicate if this is a possible solution and indicate why:

   (a) reduce the number of nodes on the hidden layer

   (b) increase the learning rate

   (c) increase the number of iterations

(d) removing the "bias" input

(e) including a pre-processing step in which "outliers" are removed and then use the "cleaned" dataset

helpt: reduce hidden layer, increase regularizer, remove outliers (al kan dat laatste ook andere effecten hebben; een beetje subtiel)

QUESTION 4 (10 points)

This question delves into the issues of bias and variance.

1. What is meant by "bias variance decomposition" of the error? Give definitions of the bias and variance components in words.

   see book (p.93-94)/ bias-variance sheets

2. Suppose that we use the mean squared error. Give a formal definition of mean squared error, bias and variance.

   see book (eq. 3.2)/ bias-variance sheets

3. Suppose that a learning algorithm makes error E on a dataset with a bias component B and a variance component V. Is it possible to find a learning algorithm that reduces both B and V?

   ja, ondanks het "bias-variance dilemma" kan dat best. Stel dat het eerste algorithme een foute "learning bias" had en ook nog overfit. Als we een ander algorithme kiezen met een betere en sterkere learning bias dan worden zowel B als V kleiner.

4. Can we measure or estimate B and V?

   Yes. See bias/variance slides

QUESTION 5 (10 points)

Consider the following data set:

ID	X1	X2	Y
1	a	g	p
2	a	h	q
3	b	h	p
4	b	g	q
5	c	g	p
6	c	h	p
7	a	g	p
8	b	g	?

1. What is the 'prior' probability of p, $P(p)$?

   Aantal p/totaal $= \frac{5}{7}$***; opmerking: dit is niet een echte prior zoals Bayes die bedoelt***

2. What is the definition of 'likelihood' in Maximum Likelihood Estimation?

   $P(data|klasse)$

3. What is the Maximum Likelihood Estimate for **Y** in example 8 given the data in the table and applying the Naive Bayes Assumption? Compute this using the data in the table.

   ML: de waarde die de grootste likelihood heeft; dus hier P(X1, X2 — Y); voor example 8 de grootste van P(X1=b en X2 = g — Y=p) en P(X1=b en X2 = g — Y=q)

4. The dataset above is actually quite small. Suppose that we know (from another source than our data) that in the domain the frequency of p is four times that of q. What should we do? Will this change our estimate for example 8?

   in dat geval gebruiken we ook de "prior" en krijgen we de Naive Bayes Classifier; dit betekent dat we de likelihood vermenigvuldigen met de prior en de grootste gebruiken als schatting