Machine Learning

SCI321

Midterm exam

5 November 2015, 11:00
Duration of the exam is 1.5 hours

The exam was made by Maarten van Someren and peer-reviewed by Praveen Dakwale
Course examiner: Maarten van Someren

Student name

Teacher name

On the following page, you will find important information about the examination. Before starting with the examination you should read this information.
Instructions

1. You are allowed to use pen, pocket calculator.

2. Please write your name as well as the name of the teacher on the front page of the exam before you answer the questions. Write your name on every sheet of paper you hand in.

3. This exam consists of four questions on six pages.

4. Answering instruction

   (a) Write your answers on this exam, in the empty space after the question.

   (b) If you are asked to present an argument for your answer or to explain some issue, write your answers in correct sentences.

5. The number of points per question is specified in the table below.

6. Your mobile phone should be switched off and should be put in your bag. Your bag needs to be closed.

7. Warning against cheating: Do not cheat! In case of cheating, the maximum punishment will be exclusion to all examinations for one year (AS&P appendix 2, Regulations governing fraud and plagiarism).

<table>
<thead>
<tr>
<th>Question</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>3</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>2.</td>
<td>1</td>
<td>1.5</td>
<td>1.5</td>
<td>4</td>
</tr>
<tr>
<td>3.</td>
<td>2</td>
<td>2</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>4.</td>
<td>2</td>
<td>2</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
<td></td>
<td>15</td>
</tr>
</tbody>
</table>
**Question 1**

We use the multivariate Gaussian distribution to do "anomaly detection". Can this method detect anomalies with an XOR structure? For example, we have two variables, $X_1$ and $X_2$ ranging between 1 and 50. The standard deviation is 15 for both. Although values around 6 or 44 are themselves not unusual for $X_1$ and for $X_2$, but the combination of a **low** value for $X_1$ and a **high** value of $X_2$ and vice versa are are very unusual and should be detected. For example $(X_1 = 6, X_2 = 43)$ and $(X_1 = 44, X_2 = 7)$ should be detected as anomalous because this is extremely rare. However, $(X_1 = 46, X_2 = 47)$ and $(X_1 = 5, X_2 = 6)$ should not because this quite common. Is our approach able to do this? Explain your answer.

*If the data have the structure described above then this means that they have a (strong) covariance (positive or negative). This will be reflected in the multivariate Gaussian distribution that is estimated from the data. This means that points as described above will have low probability even if the values on individual variables have "normal" probabilities. So the answer is YES.*

**Question 2**

Consider the following training data for a univariate classification problem (for $Y$ 0 and 1 are class labels):

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

1. Intuitively it seems that the best classifier is to predict 1 if $X > 3.5$ and else 0. Which of the following learning algorithm(s) will learn this from the data given above: logistic regression, a neural network with a hidden layer of 2 nodes, decision tree learner. Explain your answer. The decision learner will create two interval with a boundary between 3 and 4. The threshold is then set at 3.5. Logistic regression will minimize the cost. The result will be that he threshold will be shifted to a lower value to reduce the cost. The same is true for the neural network. The neural network is expected to give the same result as logistic regression.

2. The cost function for Logistic Regression that Andrew Ng gives is: $-y . log(h_\theta(x)) - (1 - y). log(1 - h_\theta(x))$. What would be the cost for the model with a threshold at $X = 3.5$? You may calculate this or give a (simplified) expression.
To answer this we must first calculate the hypothesis that corresponds to a threshold at 3.5. This we find using:
\[
\frac{1}{1+e^{\theta_0+\theta_1\cdot3.5}} = 0.5 \quad \text{or} \quad \theta_0 + \theta_1 \cdot 3.5 = 0
\]

There are many possible models that all have this effect. We take \(\theta_0 = 0\) and \(\theta_1 = \frac{1}{3.5} = 0.28\). With this hypothesis \(h\) we can calculate the cost.

3. We could use linear regression as classifier, instead of Logistic Regression. Will logistic regression and linear regression find the same class boundary? Explain your answer.

No. In particular the value of 20 has a different effect on the cost. See lecture on introduction of Logistic Regression.

Question 3

We have a dataset that we use for supervised learning and we consider learning algorithm A. We first split the dataset into a train set and a test set. We take 10 random subsamples from the train set, apply A to each and use all 10 hypotheses to make predictions on the testset. We analyze the predictions for "y" on the testset and observe that they vary substantially between hypotheses based on different samples.

1. Is this an indication of a large bias component in the error, a large variance component or something else?

This suggests strong sensitivity to the data, large variance and thus overfitting.

2. Suppose that our algorithm A is in fact Logistic Regression. Does our analysis suggest that the value of the \(\lambda\) parameter should be higher or lower? Or is there no such implication? Explain your answer.

To prevent overfitting \(\lambda\) should be high. This penalizes large values of the parameters, thereby "flattening" the shape of the hypothesis.

3. Suppose that our algorithm A is in fact the decision tree learning algorithm. Does this analysis suggest that the size of the tree is too large or too small, or is there no such implication? Explain your answer.

Large decision trees are likely to overfit so the tree should be small.
Question 4

We have a neural network with only a single input node X, a hidden layer node H and an output node Y. There are no bias nodes. The parameter of the link (X, H) is 0.2 and of link (H, Y) is 0.1. The input value of a training example is -5. We do not use regularization.

1. What will be the activation of Y? (Calculate or give a simplified formula).
   
   The activation of node H, say $a_H$ is
   
   $\text{sigmoid}(0.2 \cdot -5) = \text{sigmoid}(-1)$. In the same way the activation of output node Y is then equal to $\text{sigmoid}(0.1 \cdot \text{sigmoid}(-1))$.

2. Suppose that the activation of Y that was found by forward propagation is equal to 0.8 and the actual value is 1 is. Will the $\theta$ parameters of (X, H) en (H, Y) become larger or smaller? Explain your answer.

   The activation of output node Y should increase. We should increase $\theta_{H,Y}$ to increase the value of the sigmoid function for the second step.