Exam Machine Learning

Tuesday 10 June 2014, 18.30 - 21.15

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

QUESTION 1 (1 point)

We use ML to train a classifier to classify houses that are on sale as "within our budget" or not, on the basis of the surface, the number of bedrooms, the number of bath rooms and the value of the house next door. Unfortunately, of all houses that are on sale only 3% is within our budget.

1. On the basis of your intuition of what the underlying pattern, would a decision tree or logistic regression be a better method from a promising model for this domain? Explain your answer.

2. For a particular house we can make a prediction on the basis of the likelihood or of the a posteriori probability. Would these predictions be different for these data? If yes, which prediction would be better? Explain your answer.

QUESTION 2 (2 points)

The K-means clustering algorithm needs the number of clusters as input parameter. Andrew Ng recommends to use the "elbow method" to find the optimal number of clusters.

1. What is the objective or cost function of K-means clustering?

2. Explain how the "elbow method" works.

3. Is the "elbow method" minimzing the cost function?

4. The problem of finding the optimal number of clusters can be seen as a form of avoiding underfitting and overfitting. In supervised learning we use cross validation as a measure against overfitting. Would it be possible to apply a form of cross validation to optimize the number of clusters? Explain your answer.

QUESTION 3 (2 points)

Suppose that we have neural network with 1 hidden layer in which the activation of a node on the hidden layer is a linear combination of its inputs, so a function of the following form: $a(h) = \Sigma w_i.x_i$ where i is an index over input nodes. Similarly, the activation of the output nodes is $a(o) = \Sigma w_i.h_i$.

1. Can this type of neural network learn non-linear patterns? (Tip: write the output activation as a function of the input activations)

2. Suppose that we use the mean squared error as cost function. What are the gradient and the update rule for this type of network?

3. Now write the output activation of a *s*tandard neural network with 2 input nodes and two nodes on the hidden layer as a function of the input activations. Compare this with the network.

QUESTION 4 (1.5 points)

Consider the following dataset:

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. Suppose that the prior probability of p is in fact 0.4. What is the MAP classification of example 8 using Naive Bayes?

2. Naive Bayes assumes conditional independence between the variables. Give the formal definition of this assumption.

3. Suppose that we suspect that this assumption is false and that some pairs of variables may not be conditionally independent. We decide to construct a new variable as a function of $X1$ and $X2$. This new variables has a new value for each combination of values of $X1$ and $X2$, so $v1 = (X1 = aANDX2 = p)$, $v2 = (X1 = aANDX2 = q)$, $v3 = (X1 = bANDX2 = p)$ etc. Does this solve the problem? Explain your answer.

4. Will the classifier that is found when we use all variables (X1, X2 and X3) be better that the onse that uses only X1 and X2? Explain your answer.

QUESTION 5 (2 points)

Linear regression minimizes the mean squared error. Suppose that we have the following data:

ID	X	Y
1	1	7
2	3	5
3	5	7
4	6	2
new	5	??

1. Initialize w_0 and w_1 as 1 and $\alpha = 0.01$ and compute one iteration of gradient descent.

2. Plot the data in a graph, draw the regression function and estimate w_0 and w_1 from the plot. What would be $h(5)$ according to this regression function?

3. Note that we in fact have a datapoint for $x = 5$. The prediction by the regression function is probably different. What would, in your view, be the best prediction, the one based on the example or on the regression function? Explain your answer.

QUESTION 6 (1.5 points)

Consider the problem of insurance fraud detection. The data are credit card transactions that are labeled as correct (C) or (attempted) fraud (AF). Features are extracted from information about the transaction like time, location, amount of money, time between pervious and next transaction, subject of payment, etc. Some entries in the form are numbers, some are a choice from a list of possibilities and others are text. The proportion AF is 0.5%.

1. What is a good measure for evaluating the performance of learning algorithms on this problem? Explain why this is preferred over others.

2. One approach is to treat this as a supervised classification task and train a classifier. Are there other possibilities? If so, which seems best? Why?

3. Suppose that we change the labelling and label the AF claim forms with the type of fraud: "card copied", "limit exceeded", "card stolen". Would this change this change the preferred approach? Why (not)?