Academic year 2018-2019, January 22nd, 8h30

Machine learning

Clearly write your name on every answer sheet, as well as on all question sheets. Answer each question separately, and clearly mention the number of every question next to your answer. If you don't provide an answer for a question, clearly mention the question number and write "No answer". You can answer either in Dutch or English.
No laptops, calculators, PDAs, phones or Internet access is allowed. Hand in all question sheets.

---

1. **Love thy nearest neighbour as thyself (3 pt)**

   (a) Give two advantages of decision trees over KNN and two advantages of KNN over decision trees **(1 pt)**

   (b) Usually we perform a Z-score normalization to make features more comparable. Give a situation where this would be beneficial for a KNN classifier and a situation where this would not be beneficial for a KNN classifier. **(1 pt)**

   (c) Suppose you have a training set with N data points in it, and you are using k-Nearest Neighbours. For which of the following values of k are you most likely to observe overfitting: K=1, K=N/2 or K=N ? Explain your answer. **(1 pt)**

   (d) Provide a 2D dataset where 1-Nearest Neighbor (1-NN) has a lower leave-one-out cross-validation error (LOO error) than an SVM with linear kernel, and an example where 1-NN has a higher LOO error than SVM with linear kernel (draw the 2D dataset). **(1 pt)**

2. **There are no certainties, at most probabilities (5 pt)**

   (a) You are playing a game with two coins. Coin 1 has a probability θ of landing heads, while coin 2 has a probability 2θ of landing heads. You flip these coins several times and record your results:

Coin	Result
1	Head
2	Tail
2	Tail
2	Tail
2	Head

   Formulate the log-likelihood of the data given θ and calculate the maximum likelihood estimate for θ. **(2 pt)**

(b) Consider a Hidden Markov Model with states $Y_t \in \{S_1, S_2, S_3\}$, observations $X_t \in \{A, B, C\}$, and parameters

$\pi_1 = 1$	$a_{11} = 1/2$	$a_{12} = 1/4$	$a_{13} = 1/4$	$b_1(A) = 1/2$	$b_1(B) = 1/2$	$b_1(C) = 0$
$\pi_2 = 0$	$a_{21} = 0$	$a_{22} = 1/2$	$a_{23} = 1/2$	$b_2(A) = 1/2$	$b_2(B) = 0$	$b_2(C) = 1/2$
$\pi_3 = 0$	$a_{31} = 0$	$a_{32} = 0$	$a_{33} = 1$	$b_3(A) = 0$	$b_3(B) = 1/2$	$b_3(C) = 1/2$

i. What is $P(Y_5 = S_3)$? **(1 pt)**

ii. Suppose we observe AABCABC, what is $P(Y_5 = S_3 | X_{1:7} =$AABCABC) ? **(1 pt)**

iii. Write down the sequence of $Y_{1:7}$ with the maximal posterior probability assuming the observation AABCABC. What is that posterior probability ? **(1 pt)**

3. **The best angle from which to approach a problem is the Try-angle (4 pt)**
Answer the following questions short but to the point in at most 5 sentences.

(a) State how to apply early stopping in the context of learning neural networks using Gradient Descent. Why is it necessary to use a validation set (instead of simply using the test set) when using early stopping? **(1 pt)**

(b) Describe how to train a standard autoencoder network. **(1 pt)**

(c) Briefly describe three different ways of making the learning process go faster when using back-propagation to learn a feedforward multilayer neural network. **(1 pt)**

(d) Describe two ways of avoiding overfitting. **(1 pt)**

4. **Design is where science and art break even (3 pt)**
Your friend from the Ghent University Hospital has heard about machine learning, and wonders if it would be applicable to his problem where he is studying a rare type of cancer. He calls you in for help and asks you to design a rigorous machine learning experiment. The dataset consists of 300 healthy patients and 15 patients with the rare cancer type, for which the doctor has collected gene expression data measuring the values of 10,000 continuously-valued features. The doctor has heard about Deep Learning and wonders if that would be applicable to his data, but he also wants you to compare to some basic machine learning techniques, such as a KNN classifier, a Naive Bayesian classifier, a Support Vector Machine (with linear kernel) and a Random Forest model.

Describe in detail how you would design an experiment for this problem, including comparison of the models, hyperparameter tuning, the performance measure(s) you will use and any other techniques you think are necessary to approach this problem adequately.

5. **True or False (4 pt)**
Are the following statements True or False ? If True, explain in at most two sentences. If False, explain why or give a counterexample in at most two sentences.

(a) We would expect the support vectors to remain the same in general as we move from a linear kernel to higher order polynomial kernels.

(b) For a neural network, the number of hidden nodes affects more the trade-off between underfitting and overfitting than the learning rate.

(c) The log-likelihood of the data will always increase through successive iterations of the expectation maximization (EM) algorithm.

(d) The α coefficients assigned to the classifiers assembled by boosting are always non-negative.

(e) Overfitting is more likely when the set of training data is small.

(f) PCA and Spectral Clustering perform eigen-decomposition on two different matrices. However, the size of these two matrices are the same.

(g) As the number of data points grows to infinity, the MAP estimate approaches the MLE estimate for all possible priors. In other words, given enough data, the choice of prior is irrelevant.

(h) Consider the parameter k in k-fold cross-validation. With a higher value of k, on average we will have a lower value of the cross-validation error.