KTH Computer Science and Communication

Exam in DD2421 Machine Learning
2021-10-25, kl 14.00 – 2021-10-26, kl 14.00

Aids allowed: *calculator*, *language dictionary*. This is also an open book exam.

To take this exam you must be registered to this specific exam as well as to the course. Then you have an access to the Canvas page, "Tentamen för DD2421 TEN1/FDD3431 EXA1: 2021-10-25", https://canvas.kth.se/courses/33117. Read the instructions given there. For your submission go in "Assignments" on the top-left of the page where you also find "Code of conduct" – you can click submission buttons during the exam hours. Email is not accepted as a tool for submissions.

In order to pass this exam, your score x first needs to be 16 or more (out of 42, full point). In addition, given your points y from the Programming Challenge (out of 18, full point), the requirements on the total points, $p = x + y$, are preliminarily set for different grades as:

$$
\begin{aligned}
53 < p \leq 60 &\quad \rightarrow \quad A \\
46 < p \leq 53 &\quad \rightarrow \quad B \\
39 < p \leq 46 &\quad \rightarrow \quad C \\
32 < p \leq 39 &\quad \rightarrow \quad D \\
24 < p \leq 32 &\quad \rightarrow \quad E \quad \text{(A pass is guaranteed with the required points for 'E'.)} \\
0 \leq p \leq 24 &\quad \rightarrow \quad F
\end{aligned}
$$

This exam consists of sections **A**, **B**, and **C**, to which different zoom links are assigned in case of inquiries. **NB. Use different papers (answer sheets) for different sections.**

A Graded problems

Potential inquiries to be addressed to zoom link (A).

A-1 Terminology (5p)

For each term (a–e) in the left list, choose the explanation among the right list which *best* describes the term in the context of machine learning.

a) Posterior probability

b) RANSAC

c) k-means

d) Fisher's criterion

e) Dropout

1) An example of ensemble learning

2) Sudden drop of performance

3) Conditional probability taking into account the evidence

4) Probability at a later time

5) Probability before observation

6) Clustering method based on centroids

7) A strategey to generate k different models

8) Random strategy for amplitude compensation

9) Robust method to fit a model to data with outliers

10) An approach to find useful dimension for classification

A-2 Nearest Neighbor, Classification (4p)

Suppose that we take a data set, divide it into two parts of equal size, Part I and Part II. We try out two different classification procedures, by using Part I and Part II as our training set and test set, respectively, That is, we use half of the data for training, and the remaining half for testing.

a) First we use 1-Nearest Neighbor rule (1-NN) and get an average error rate (averaged over both test and training data sets) of 8%. What was the error rate with 1-nearest neighbor on the test set? Briefly reason the answer. (1p)

b) Next we use the Adaboost Algorithm and get an error rate of 10% on the training data. We also get the average error rate (averaged over both test and training data sets) of 12%. (1p) What was the error rate with the Adaboost Algorithm on the test set? Just answer the error rate. (1p)

c) Now, we swap the roles of Part I and Part II, and repeat the same experiments. On the test set (Part I), we get an error rate of 12% with both 1-NN and the Adaboost Algorithm. Based on all these results, by the cross-validation, indicate the method which we should prefer to use for classification of new observations, with a simple reasoning. (2p)

A-3 Regression with regularization (4p)

For a set of N training samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, each consisting of input vector $\mathbf{x}$ and output y, suppose we estimate the regression coefficients $\mathbf{w}^\top (\in \mathbf{R}^d) = \{w_1, \ldots, w_d\}$ in a linear regression model by minimizing

$$\sum_{n=1}^{N} (y_n - \mathbf{w}^\top \mathbf{x}_n)^2 + \lambda \sum_{i=1}^{d} |w_i|$$

with a value of λ for the regularization term.

a) What is this method called? (1p)

b) Now, let us consider a model with an extremely large λ and then models with gradually smaller values of λ (towards 0). The trainig error and test errors are then believed to change with respect to the bias-variance trade-off. Indicate which of {i,ii,iii,iv,v} is correct about the variance, and motivate your answer. The variance of the model will:
i. Remain constant.
ii. Steadily increase.
iii. Steadily decrease.
iv. Increase initially, and then eventually start decreasing in an inverted U shape.
v. Decrease initially, and then eventually start increasing in a U shape. (2p)

c) The method described is known to yield *sparse models*. Briefly explain what property of it enables the sparcity, in a short sentence. (1p)

A-4 PCA, Subspace Methods (3p)

Given a set of feature vectors which all belong to a specific class C (i.e. with an identical class label), we performed PCA on them and generated an orthonormal basis $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$ which spans a p-dimensional subsapce, $\mathcal{L}$, as the outcome. Provide an answer to the following questions.

a) We have a new input vector $\mathbf{x}$ whose class is unknown, and consider its projectiton length on $\mathcal{L}$. Describe how the projection length is represented, using a simple formula. (2p)

b) Now, we consider to solve a K-class classification problem with the Subspace Method and assume that a subspace $\mathcal{L}^{(j)}$ $(j = 1, ..., K)$ has been computed with training data for each class, respectively. Given a new input vector $\mathbf{x}$, we computed its projectiton length on each subspace as $S^{(j)}$ $(j = 1, ..., K)$. Among those $S^{(\alpha)}$, $S^{(\beta)}$, and $S^{(\gamma)}$ were the maximum, the minimum, and the closest to the average of all $S^{(j)\prime}s$, respectively. Based on this observation which class should $\mathbf{x}$ belong to? (1p)

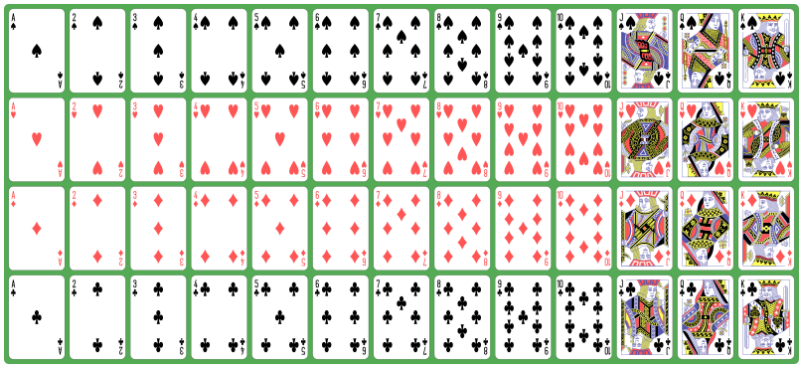


Figure 1. Playing cards consisting of 52 patterns.

A-5 Information Contents (6p)

Imagine that you are playing with Cards and randomly sample *four cards* out of the pile of 52 cards (see Figure 1) *with replacement*, i.e. you sequentially draw a card but return it to the pile each time you have seen what it is.

a) At each instance of drawing a card, what is the Shannon information content of the outcome with respect to the suit, one of {*Spades, Hearts, Diamonds, and Clubs*}, measured in bits? (2p)

b) You play a game with a rule that you win if the suits of all the *four* cards are of *the same colour*, either *black or red*. Otherwise you lose. How unpredictable is the outcome of this game (win or lose)? Answer in terms of entropy, measured in bits. (2p)

c) With respect to the outcome of the game in **b**), what is the expected information gain by drawing the first two card, i.e. by seeing (the suit colours of) the first card and the second card? (2p)

Note: if you do not have a calculator, answer with an expression but simplify it as much as possible.

walk	poopoos	doggo
1	0	Shoogee Nulnul
2		Shoogee Nulnul
3	1	Shoogee Nulnul
4	1	Shoogee Nulnul
5	0	Shoogee Nulnul
1	2	Max the Tax
2		Max the Tax
3	2	Max the Tax
4	3	Max the Tax
5	3	Max the Tax

Table 1. The number of poopoos emitted by two doggos on five walks together.

B Graded problems

Potential inquiries to be addressed to zoom link (B).

B-1 Warming up With Bayes' (3p)

Before I took a COVID test, the doctor said $x\%$ of the people in my area do not have COVID, but 10% of them are testing positive. A few days later the doctor called and said my test was positive, and that the probability I have COVID given this positive test is $y\%$. Find and plot the relationship between x and y such that the above analysis is correct. Show your work.

B-2 Maximum likelihood estimation (3p)

Consider the data in Table 1. Assume all observations are independent, but there are two missing. Assume the number of poopoos for each doggo is distributed Poisson. For each doggo, fit the number of its poopoos with a Poisson distribution using maximum likelihood estimation of the parameters, and compute those parameters. Then estimate the probability the two missing numbers are the same, assuming that the doggos make poopoos independent of each other. (In this part it is ok to write a small computer program.) Show your work.

B-3 Maximum a posteriori estimation (3p)

Consider the data in Table 1. Assume all observations are independent. For each doggo, fit its number of poopoos with a Poisson distribution using maximum a posteriori estimation of the parameters, and compute those parameters. Assume the prior distribution of the parameters $f_\Theta(\theta_k)$ is exponential with parameter γ. Find the value of γ such that for Shoogee Nulnul $\lambda_0 = 0.4$, and for Max the Tax $\lambda_1 = 2$. Show your work.

B-4 Probabilistic Linear Regression (3p)

Say we believe the conditional probability of the number of poopoos emitted by Max the Tax y given Shoogee Nulnul emitted x poopoos on the same walk is Poisson with parameter $wx + b$, i.e.,

$$P(y|x, w) = \text{Poisson}(wx + b). \tag{1}$$

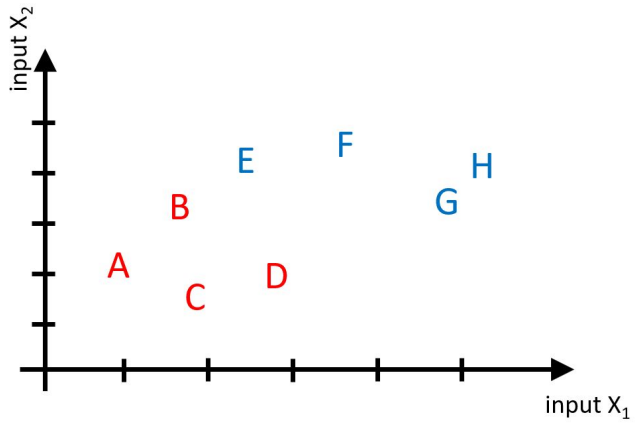
Assuming each walk is independent, find the maximum likelihood estimates of w and b. Show your work.

C Graded problems

Potential inquiries to be addressed to zoom link (C).

C-1 Support Vector Classification (4p)

The following diagram shows a small data set consisting of four RED samples (A, B, C, D) and four BLUE samples (E, F, G, H). This data set can be linearly separated.

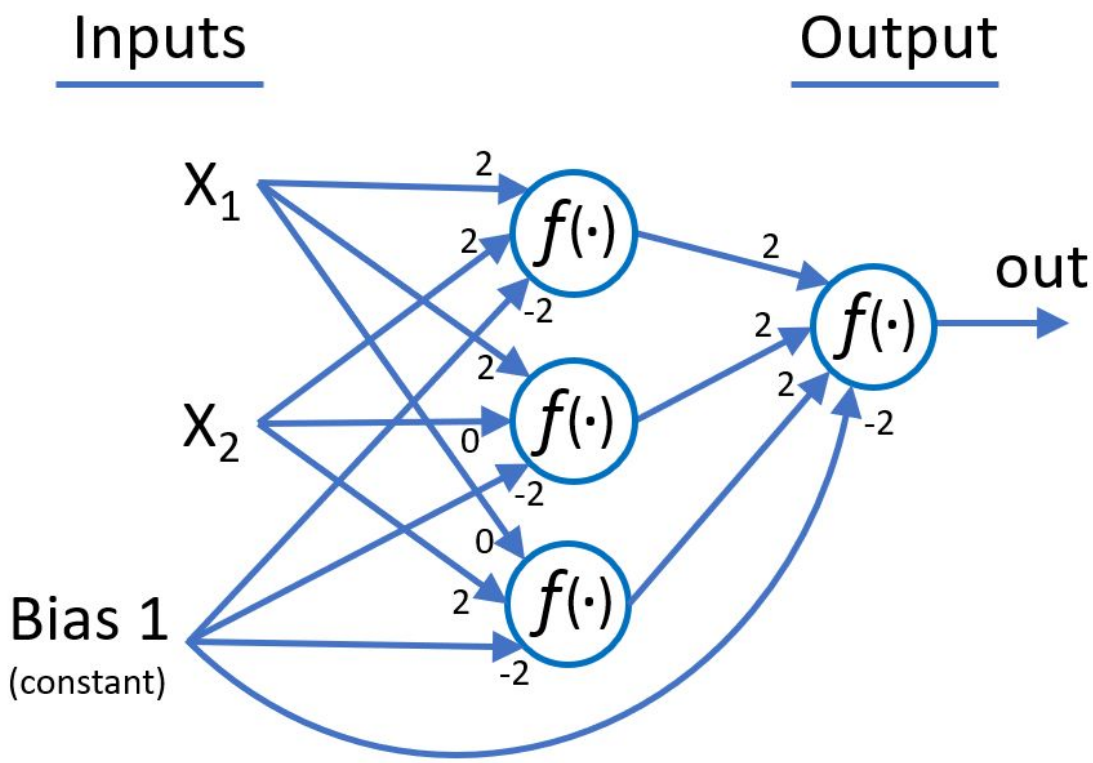


a) We use a linear support vector machine (SVM) without kernel function to correctly separate the BLUE and the RED class. Which of the data points (A-H) can be removed for training without changing the resulting SVM decision boundary? No explanation needed; name the point(s) the can be removed. (2p)

b) Assume someone suggests using a non-linear kernel for the SVM classification of the above data set (A-H). Give one argument in favor and one argument against using non-linear SVM classification for such a data set. **USE KEYWORDS!** (2p)

C-2 Neuronal Networks (4p)

The following diagram shows a simple neuronal network with step activation functions in all neurons.



$f(\cdot)$ = step neuronal activation function

$$\text{step}\left(\sum_i (inp_i * wi)\right) = \begin{cases} 1 \text{ if } (\cdot) \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

a) Draw a 2D diagram of the input space (two dimensional plot of X1 and X2) and show for which area of the input space the network produces a positive output. (2p)

b) Can this network be implemented in a single neuron with linear activation function (yes/no)? Explain **in KEYWORDS**. (1p)

c) Assume all multiplicative weights in the network double their value (the small numbers right next to the neuron). What happens with the output? (1p)