KTH Computer Science
and Communication

Exam in DD2421 Machine Learning
2023-03-16, kl 14.00 – 18.00

Aids allowed: *calculator*, *language dictionary*.

To take this exam you must be registered to this specific exam as well as to the course.

In order to pass this exam, your score x first needs to be 20 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:

$$\begin{aligned}
54 < p \leq 60 &\quad \rightarrow \quad A \\
48 < p \leq 54 &\quad \rightarrow \quad B \\
42 < p \leq 48 &\quad \rightarrow \quad C \\
36 < p \leq 42 &\quad \rightarrow \quad D \\
29 < p \leq 36 &\quad \rightarrow \quad E \quad \text{(A pass is guaranteed with the required points for 'E'.)} \\
0 \leq p \leq 29 &\quad \rightarrow \quad F
\end{aligned}$$

This exam consists of sections **A**, **B**, and **C**. **NB. Use different papers (answer sheets) for different sections.**

A Graded problems

Potential inquiries to be addressed to Atsuto Maki.

A-1 Terminology (4p)

For each term (a–d) in the left list, find the explanation from the right list which *best* describes how the term is used in machine learning.

a) Dropout

b) RANSAC

c) Occams's razor

d) k-fold cross validation

1) An approach to find useful dimension for classification

2) Random strategy for area compression

3) An approach to train artificial neural networks

4) Algorithm to learn with latent variables

5) A robust method to fit a model to data with outliers

6) A technique for margin maximization

7) A technique for assessing a model by exploiting available data for training and testing

8) A principle to choose the simplest explanation

A-2 Nearest Neighbor, Classification (4p)

Suppose that we take a data set, divide it into training and test sets, and then try out two different classification procedures. We use *three-quarters* of the data for training, and the remaining *one-quarter* for testing. First we use *Adaboost* and get an error rate of 10% on the training data. We also get the average error rate (weighted average over both test and training data sets) of 12%. Next we use k-nearest neighbor (where $k = 1$) and get an average error rate (weighted average over both test and training data sets) of 10%.

Answer the following questions while motivating them.

a) What was the error rate with 1-nearest neighbor on the training set? (1p)

b) What was the error rate with 1-nearest neighbor on the test set? (1p)

c) What was the error rate with AdaBoost on the test set? (1p)

d) Based on these results, indicate the method which we should prefer to use for classification of new observations, with a simple reasoning. (1p)

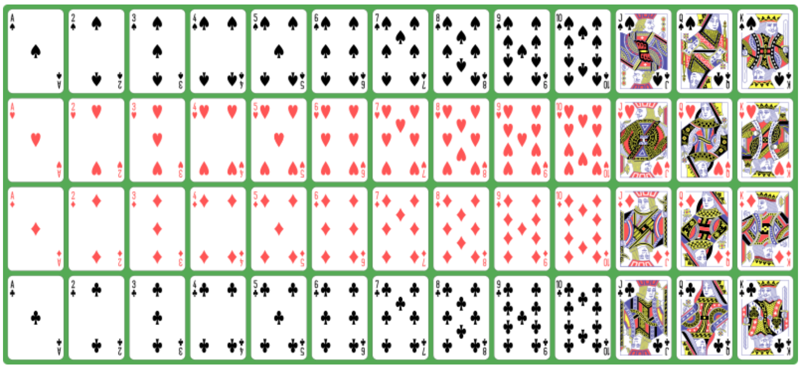


Figure 1. Playing cards consisting of 52 patterns.

A-3 Entropy and Decision Trees/Forests (6p)

a) Consider decision trees, and indicate a correct one as the basic strategy for selecting a question (attribute) at each node.

i. To minimize the expected reduction of the entropy.
ii. To maximize the expected reduction of the entropy.
iii. To minimize the expected reduction of gini impurity.

Simply indicate your choice. (1p)

b) Briefly explain the two kinds of randomness involved in the design of Decision Forests? (2p)

Now 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.

c) 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? Movivate your answer. (1p)

d) 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. With respect to the outcome of this game, 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? Movivate your answer. (2p)

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

A-4 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

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

for a particular value of λ.

Now, let us consider different models trained with different values of λ, starting from a very large value (infinity) and *decreasing* it down to 0. Then, for parts **a**) through **c**), indicate which of i. through v. is most likely to be the case.

Briefly justify each of your answers.

a) As we decrease λ, the variance of the model will:
i. Remain constant.
ii. Steadily increase.
iii. Steadily decrease.
iv. Decrease initially, and then eventually start increasing in a U shape.
v. Increase initially, and then eventually start decreasing in an inverted U shape. (1p)

b) Repeat **a**) for the training error (resisual sum of squares, RSS). (1p)

c) Repeat **a**) for test RSS. (2p)

A-5 PCA, Subspace Methods (4p)

We consider to solve a K-class classification problem with the Subspace Method and for that we compute a subspace $\mathcal{L}^{(j)}$ $(j = 1, ..., K)$ using training data for each class, respectively. That is, given a set of feature vectors (as training data) which belong to a specific class C (i.e. with an identical class label), we perform 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) Given that we compute $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$ as eigenvectors of the auto-correlation matrix Q based on the training data, how should we choose the eigenvectors in relation to the corresponding eigenvalues of Q? (1p)

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

c) Given $\mathbf{x}$, we computed its projectiton length on each subspace as $S^{(j)}$ $(j = 1, ..., K)$, respectively. For a few classes among those with labels $\{l,m,n\}$, we had the following observations:
$S^{(l)}$ was the minimum of all $S^{(j)}$'s,
$S^{(m)}$ was the maximum of all $S^{(j)}$'s, and
$S^{(n)}$ was the closest to the average of all $S^{(j)}$'s.
Which class should $\mathbf{x}$ belong to? Simply choose a class label. (1p)

B Graded problems

Potential inquiries to be addressed to Bob Sturm.

B-1 Identify Bayes' Theorem (1p)

For two events A and B, which of the following is Bayes' Theorem:

a) $P(A|B) = P(A,B)P(A)/P(B)$

b) $P(A|B) = P(B|A)P(A,B)/P(B)$

c) $P(A|B) = P(B|A)P(A)/P(B)$

d) $P(A|B) = P(A,B)/P(B)$

B-2 Conditioning (1p)

For three events A, B, and C, which of the following are not equivalent to $P(A,B,C)$:

a) $P(A,C|B)P(B)$

b) $P(A|B,C)P(B,C)$

c) $P(A,B|C)P(B)$

d) $P(A|C,B)P(B|C)P(C)$

B-3 Shoogee picks (2p)

Our dog Shoogee[1] (pictured) has a big box of N balls, all identical excerpt for color: one is red and the others are brown. But since Shoogee is colorblind she can't see a difference. She likes to take the balls out one by one at random and lay them beside the box. Find an expression for the expected number of balls Shoogee takes out of the box before taking out the one red ball. *Show your work.*

B-4 Inference by Minimizing Expected Loss (3p)

A good strategy for probabilistic inference given an observation x is to choose the value $y \in \mathcal{Y}$ that minimizes the expected loss:

$$y^*(x) = \arg\min_{y' \in \mathcal{Y}} E_{Y|X=x}[\mathcal{L}(y,y')] = \arg\min_{y' \in \mathcal{Y}} \int_{y \in \mathcal{Y}} \mathcal{L}(y,y') Pr(y|X=x) dy \tag{1}$$

where $\mathcal{L}(y,y')$ is a loss function. Define $\mathcal{L}(y,y') = (y-y')^2$ and assume $Y|X=x \sim \text{Unif}[0,2x]$, $x > 0$, that is a uniform distribution over the domain 0 and $2x$ for all positive x. Derive a formula for $y^*(x)$ in this case. *Show your work.*

B-5 Regression (1p)

[1] About the name. Being non-native Swedish speakers, we were amused hearing people on the television saying "tjugo-noll-noll". Except they don't say "shoogo", but "shoogee". "Shoogo" sounds like a cleaning product. "Shoogee" sounds like a dog's name. So we named her "Shoogee Nulnul".

Figure 2. Shoogee (dog) with one of her favorite things in the world.

Consider you have modelled $\mathcal{D} = ((x,y)_i : x, y \in \mathbb{R})$ with

$$y = w_0 + w_1 x^2 + \epsilon \quad (2)$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and $w_0 < 0$. Graph a solution in the x-y plane for $w_1 < 0$ and a solution for $w_1 > 0$. Label the intersection on the y-axis.

B-6 Probabilistic classification (1p)

I have a dataset of labeled feature vectors, $\mathcal{D} = ((\mathbf{x}, y)_i)$, where $y \in \{0,1\}$ and $\mathbf{x} = (x_1, x_2, x_3)$ is three dimensional. It looks like the first element of $\mathbf{x}$ is one of three characters $\{r, g, b\}$; the second element is one of two characters $\{y, n\}$, and the third element is a positive number. Which of the following is the Naive Bayes classifier for an observation $\mathbf{x}$?

a) $y(\mathbf{x}) = \arg\max_{y \in \{0,1\}} Pr(x_1, x_2, x_3|y)Pr(y)$

b) $y(\mathbf{x}) = \arg\max_{y \in \{0,1\}} Pr(x_1|y)Pr(x_2|y)Pr(x_3|y)Pr(y)$

c) $y(\mathbf{x}) = \arg\max_{y \in \{0,1\}} Pr(y|x_1)Pr(y|x_2)Pr(y|x_3)Pr(y)$

d) $y(\mathbf{x}) = \arg\max_{y \in \{0,1\}} Pr(y|x_1, x_2, x_3)Pr(y)$

B-7 Learning by Minimizing the Minimum Expected Loss (3p)

A good strategy for probabilistic learning is to choose the parameters $\theta \in \Theta$ of the conditional distribution $Pr(y|X = x, \theta)$ that minimize the minimum expected loss:

$$\begin{aligned}\theta^* &= \arg\min_{\theta\in\Theta} \left[\int_{x\in\mathcal{X}} \min_{y'\in\mathcal{Y}} E_{Y|X=x,\theta}[\mathcal{L}(y,y')] Pr(x) dx \right] \\ &= \arg\min_{\theta\in\Theta} \left(\int_{x\in\mathcal{X}} \left[\min_{y'\in\mathcal{Y}} \int_{y\in\mathcal{Y}} \mathcal{L}(y,y') Pr(y|X = x,\theta) dy \right] Pr(x) dx \right) \end{aligned} \tag{3}$$

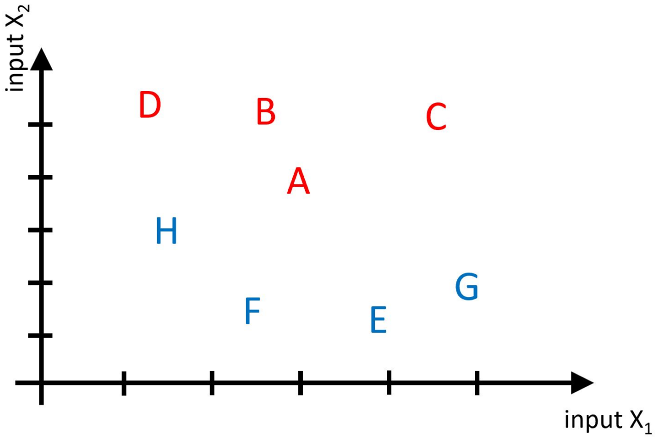
where $\mathcal{L}(y, y')$ is a loss function. Define $\mathcal{L}(y, y') = (y - y')^2$ and assume $Y|X = x \sim \text{Unif}[0, \theta x]$, $\theta > 0$, that is a uniform distribution over the domain 0 and θx, for all positive x. Show that $\theta^* \to 0$ results. *Show your work.*

C Graded problems

Potential inquiries to be addressed to Jörg Conradt.

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.

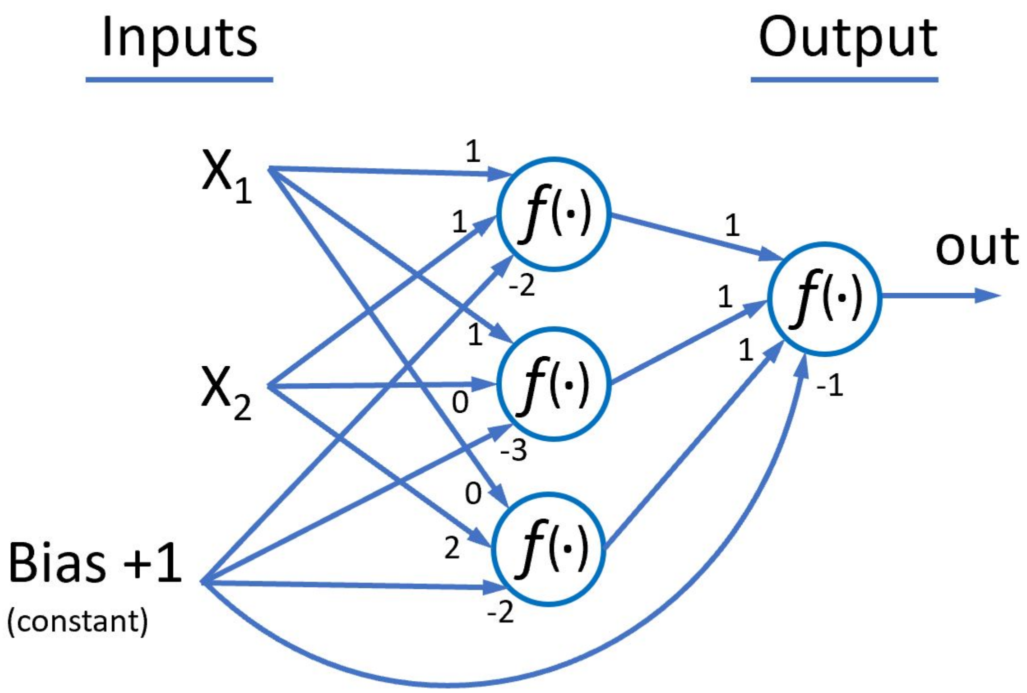


a) We use a linear support vector machine (SVM) without kernel function to correctly separate the BLUE and the RED class. Which of the data points (A-H) can be removed for training without changing the resulting SVM decision boundary? Name the point(s) that can be removed, and justify with **KEYWORDS** (or very short sentences). (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. The weight for incoming signals are shown directly at the neurons.



f (·) = 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 X_1 vs. X_2) 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 weights in the network double their value (the small numbers right next to the neuron input signal, including the weight for the constant bias input). What happens with the output? (1p)