KTH Computer Science
and Communication

Exam in DD2421 Machine Learning
2020-10-23, kl 10.00 – 2020-10-24, kl 10.00

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

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, find the explanation from the right list which *best* describes how the term is used in machine learning.

a) k-means

b) Dropout

c) k-fold cross validation

d) Expectation Maximization

e) k-nearest neighbour

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

2) Algorithm to learn with latent variables

3) An approach to train artificial neural networks

4) A strategey to generate k different models

5) Method for estimating the mean of k observations

6) Clustering method based on centroids

7) Class prediction by a majority vote

8) Estimating expected value

9) A technique for assessing a model while exploiting available data for training and testing

10) An approach to find useful dimension for classification

Solution: a-6, b-3, c-9, d-2, e-7

A-2 Entropy and Decision Trees/Forests (5p)

a) Indicate the standard strategy for selecting a question (attribute) at each node in decision trees.

i. To minimize the expected reduction of the entropy.
ii. To maximize the expected reduction of the entropy.
iii. To split the samples at the node into two groups of equal size.

Simply indicate your choice. (1p)

b) Consider a game where you throw two dice, one red and one green. You win if the green die has a smaller number, and you lose if the red die has a smaller number. If they are equal, it is a draw. How unpredictable is the outcome of this game (win, lose, or a draw)? Answer in terms of entropy, measured in bits. (2p)

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

c) Suppose we have trained a Decision Forest using four bootstrapped samples from a data set containing three classes, {Red, Yellow, Green}. We then applied the model to a specific test input, x, and observed four estimates of class distributions from the four trees, respectively:

$\{P_1(\text{Red}|x), P_1(\text{Yellow}|x), P_1(\text{Green}|x)\}$ = {0.1, 0.5, 0.4}
$\{P_2(\text{Red}|x), P_2(\text{Yellow}|x), P_2(\text{Green}|x)\}$ = {0.2, 0.4, 0.4}
$\{P_3(\text{Red}|x), P_3(\text{Yellow}|x), P_3(\text{Green}|x)\}$ = {0.3, 0.4, 0.3}
$\{P_4(\text{Red}|x), P_4(\text{Yellow}|x), P_4(\text{Green}|x)\}$ = {0.2, 0.2, 0.6}

Now, let us consider using the average probabilities for combining these results into a single class prediction. What will the final class prediction be? *Motivate your answer by short phrases*. (2p)

Solution: a)-ii. **b**)-Let $f(p_1, p_2, p_3) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - p_3 \log_2 p_3$
Entropy: $f(15/36, 6/36, 15/36) \approx 1.483$
c)-Green. (0.2,0.375,0.425)

A-3 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 half of the data for training, and the remaining half for testing.

First we use 1-nearest neighbor and get an average error rate (averaged over both test and training data sets) of 3%. Next we use the subspace method and get an error rate of 3% on the training data. We also get the average error rate (averaged over both test and training data sets) of 4%.

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 the subspace method 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)

Solution:

a) 0%. Training error for 1-NN is always zero.
b) 6%. Given the answer in **a**), the testing error is 6%.
a) 5%.
c) The subspace method because it achieves lower error rate on the test data (5% < 6%).

A-4 The bias-variance decomposition (4p)

a) Let us consider a classifier, function $f(\mathbf{x})$ of input vector $\mathbf{x}$, and the following concepts:

$$\begin{aligned} \hat{f}(\mathbf{x}) &: \text{ prediction function (= model) estimated with a set of data samples, } \mathcal{D} \\ E_{\mathcal{D}}[\hat{f}(\mathbf{x})] &: \text{ the average of models due to different sample set} \end{aligned}$$

Show the bias and variance of the classifier *in formulae* referring these terms. (2p)

b) Derive that the mean square error (MSE) for estimating $f(\mathbf{x})$ can be decomposed into a two-fold representation consisting of the terms of bias and variance. (2p)

Solution:

a) Bias: $E_{\mathcal{D}}[\hat{f}(\mathbf{x})] - f(\mathbf{x})$
Variance: $E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - E_{\mathcal{D}}[\hat{f}(\mathbf{x})])^2]$

b)

$$\begin{aligned} & E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2] \\ = \; & E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - E_{\mathcal{D}}[\hat{f}(\mathbf{x})] + E_{\mathcal{D}}[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2] \\ = \; & E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - E_{\mathcal{D}}[\hat{f}(\mathbf{x})])^2 + (E_{\mathcal{D}}[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2 + 2(\hat{f}(\mathbf{x}) - E_{\mathcal{D}}[\hat{f}(\mathbf{x})])(E_{\mathcal{D}}[\hat{f}(\mathbf{x})] - f(\mathbf{x}))] \\ = \; & E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - E_{\mathcal{D}}[\hat{f}(\mathbf{x})])^2] + (E_{\mathcal{D}}[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2 \\ = \; & Variance + (Bias)^2 \end{aligned}$$

A-5 Regression with regularization (4p)

a) In regression, one way for performing regularization is to introduce an additional term, so-called shrinkage penalty. Which one of the three methods includes the additional term. (1p)

i. Logistic regression.
ii. Ridge regression.
iii. k-NN regression.

b) 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

$$\sum_{n=1}^{N} (y_n - \mathbf{w}^\top \mathbf{x}_n)^2 \quad \text{subject to} \quad \sum_{i=1}^{d} |w_i| \le s$$

for a particular value of s. As we *increase* s from 0, the trainig error and test errors (residual sum of squares, RSS) are believed to change. Indicate which of {i,ii,iii,iv,v} is correct about the *test* error. (1p)

i. Remain constant.
ii. Steadily decrease.
iii. Steadily increase.
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.

Justify your answer. (1p)

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

Solution:

a) ii.

b) iv.
When $s = 0$, all w_i are zero; the model is extremely simple, predicting a constant and has no variance with the prediction being far from actual value (thus with high bias). As we increase s, all w_i increase from zero toward their least square estimate values while steadily decreasing the *training* error to the ordinary least square RSS (and also decreasing bias as the model continues to better fit training data). The values of w_i then become more dependent on training data, thus increasing the variance. The *test* error initially decreases as well, but eventually start increasing due to overfitting to the training data.

c) The variable selection property.

B Graded problems

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

B-1 Parameter Estimation (12p)

Consider a dataset of N independent observations $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, where each $\mathbf{x}_n$ is a p-dimensional real vector. Assume the random variable Y_n is distributed Gaussian with a mean $\boldsymbol{\beta}^T\mathbf{x}_n$ and known variance σ^2. In other words, $Y_n|\mathbf{x}_n, \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}^T\mathbf{x}_n, \sigma^2)$. Assume each dimension of the column vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is independent and identically distributed Bernoulli in $\{0, 1\}$ with known parameter τ, i.e., $P[\beta_i = 1] = \tau$.

a) Show that the maximum likelihood estimate of the parameters $\boldsymbol{\beta}$ is given by solving

$$\boldsymbol{\beta}_{\text{ML}} = \arg \min_{\boldsymbol{\beta} \in \{0,1\}^p} \|\mathbf{X}\boldsymbol{\beta}\|_2^2 - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} \qquad (1)$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, $\mathbf{X}^T = [\mathbf{x}_1|\mathbf{x}_2|\ldots|\mathbf{x}_N]$ and $\|\cdot\|_2$ is the Euclidean norm. Show your work. (4p)

b) Which of the following is the correct interpretation of equation (1). (1p)

Find the $\boldsymbol{\beta}$ that makes $\mathbf{X}\boldsymbol{\beta}$

1. short in a Euclidean sense and point in the same direction as $\mathbf{y}$.
2. short in a Euclidean sense and point orthogonal to $\mathbf{y}$.
3. short in a Euclidean sense and point in the opposite direction as $\mathbf{y}$.

c) Show that the maximum a posterior estimate of the parameters $\boldsymbol{\beta}$ is given by solving

$$\boldsymbol{\beta}_{\text{MAP}} = \arg \max_{\boldsymbol{\beta} \in \{0,1\}^p} \lambda\|\boldsymbol{\beta}\|_1 + 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \|\mathbf{X}\boldsymbol{\beta}\|_2^2 \qquad (2)$$

where $\|\cdot\|_1$ is the Manhattan distance, and $\lambda = 2\sigma^2 \log[\tau/(1-\tau)]$. Show your work. (4p)

d) Which of the following is the correct interpretation of equation (2) for $\tau < 0.5$. (1p)

Find the $\boldsymbol{\beta}$ that is

1. sparse, and makes $\mathbf{X}\boldsymbol{\beta}$ short in a Euclidean sense and point in the same direction as $\mathbf{y}$.
2. not sparse, and makes $\mathbf{X}\boldsymbol{\beta}$ short in a Euclidean sense and point in the same direction as $\mathbf{y}$.
3. not sparse, and makes $\mathbf{X}\boldsymbol{\beta}$ short in a Euclidean sense and point in the opposite direction as $\mathbf{y}$.

e) What happens to $\boldsymbol{\beta}_{\text{ML}}$ as $\tau \to 0$? (1p)

f) What happens to $\boldsymbol{\beta}_{\text{MAP}}$ as $\tau \to 0$? (1p)

Solution:

a) The log-likelihood of the dataset is given by

$$\log Pr(\mathcal{D}|\boldsymbol{\beta}) = \sum_{n=1}^{N} \log Pr(y_n|\mathbf{x}_n, \boldsymbol{\beta}) \tag{3}$$

The maximum likelihood estimate of $\boldsymbol{\beta}$ is defined

$$\begin{aligned}
\boldsymbol{\beta}_{\text{ML}} &= \arg \max_{\boldsymbol{\beta}\in\{0,1\}^p} \log Pr(\mathcal{D}|\boldsymbol{\beta}) && (4)\\
&= \arg \max_{\boldsymbol{\beta}\in\{0,1\}^p} \sum_{n=1}^{N} \log Pr(y_n|\mathbf{x}_n, \boldsymbol{\beta}) && (5)\\
&= \arg \max_{\boldsymbol{\beta}\in\{0,1\}^p} \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_n - \mathbf{x}_n^T\boldsymbol{\beta})^2 && (6)\\
&= \arg \min_{\boldsymbol{\beta}\in\{0,1\}^p} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^T\boldsymbol{\beta})^2. && (7)
\end{aligned}$$

where the restriction $\boldsymbol{\beta} \in \{0,1\}^p$ comes from the prior knowledge that each of its p elements is in $\{0,1\}$. Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ and $\mathbf{X}^T = [\mathbf{x}_1|\mathbf{x}_2|\ldots|\mathbf{x}_N]$. Then

$$\begin{aligned}
\boldsymbol{\beta}_{\text{ML}} &= \arg \min_{\boldsymbol{\beta}\in\{0,1\}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \arg \min_{\boldsymbol{\beta}\in\{0,1\}^p} \|\mathbf{y}\|^2 - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \|\mathbf{X}\boldsymbol{\beta}\|^2 \\
&= \arg \min_{\boldsymbol{\beta}\in\{0,1\}^p} \|\mathbf{X}\boldsymbol{\beta}\|^2 - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta}. && (8)
\end{aligned}$$

b) 1. Find the $\boldsymbol{\beta}$ that makes $\mathbf{X}\boldsymbol{\beta}$ short in a Euclidean sense and point in the same direction as $\mathbf{y}$.

c) The maximum a posteriori estimate of $\boldsymbol{\beta}$ is defined by

$$\begin{aligned}
\boldsymbol{\beta}_{\text{MAP}} &= \arg\max_{\boldsymbol{\beta}} Pr(\boldsymbol{\beta}'|\mathcal{D}) = \arg\max_{\boldsymbol{\beta}} Pr(\mathcal{D}|\boldsymbol{\beta})Pr(\boldsymbol{\beta}) && (9)\\
&= \arg\max_{\boldsymbol{\beta}} \log Pr(\boldsymbol{\beta}) + \log Pr(\mathcal{D}|\boldsymbol{\beta}). && (10)
\end{aligned}$$

From the first part, we know

$$\begin{aligned}
\log Pr(\mathcal{D}|\boldsymbol{\beta}) &= \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_n - \mathbf{x}_n^T\boldsymbol{\beta})^2 && (11)\\
&= C - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 && (12)
\end{aligned}$$

With the assumption that elements of $\boldsymbol{\beta}$ are independent, then

$$\begin{aligned}
\log Pr(\boldsymbol{\beta}) &= \log \prod_{i=1}^{p} \tau^{\beta_i}(1-\tau)^{1-\beta_i} = \sum_{i=1}^{p} \beta_i \log\tau + (1-\beta_i)\log(1-\tau) && (13)\\
&= \|\boldsymbol{\beta}\|_1 \log\tau + (p - \|\boldsymbol{\beta}\|_1)\log(1-\tau) && (14)\\
&= \|\boldsymbol{\beta}\|_1 \log\frac{\tau}{1-\tau} + p\log(1-\tau) && (15)
\end{aligned}$$

Hence,

$$\boldsymbol{\beta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\beta}\in\{0,1\}^p} \|\boldsymbol{\beta}\|_1 \log\frac{\tau}{1-\tau} - \frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 \tag{16}$$

$$= \arg\max_{\boldsymbol{\beta}\in\{0,1\}^p} 2\sigma^2 \log\frac{\tau}{1-\tau}\|\boldsymbol{\beta}\|_1 - \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 \tag{17}$$

$$= \arg\max_{\boldsymbol{\beta}\in\{0,1\}^p} \lambda\|\boldsymbol{\beta}\|_1 + 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \|\mathbf{X}\boldsymbol{\beta}\|_2^2 \tag{18}$$

where $\lambda = 2\sigma^2\log[\tau/(1-\tau)]$.

d) 1. Find the $\boldsymbol{\beta}$ that is sparse, and makes $\mathbf{X}\boldsymbol{\beta}$ short in a Euclidean sense and point in the same direction as $\mathbf{y}$.

e) What happens to $\boldsymbol{\beta}_{\mathrm{ML}}$ as $\tau \to 0$? Nothing.

f) What happens to $\boldsymbol{\beta}_{\mathrm{MAP}}$ as $\tau \to 0$? It becomes more and more sparse, until it is all zeros.

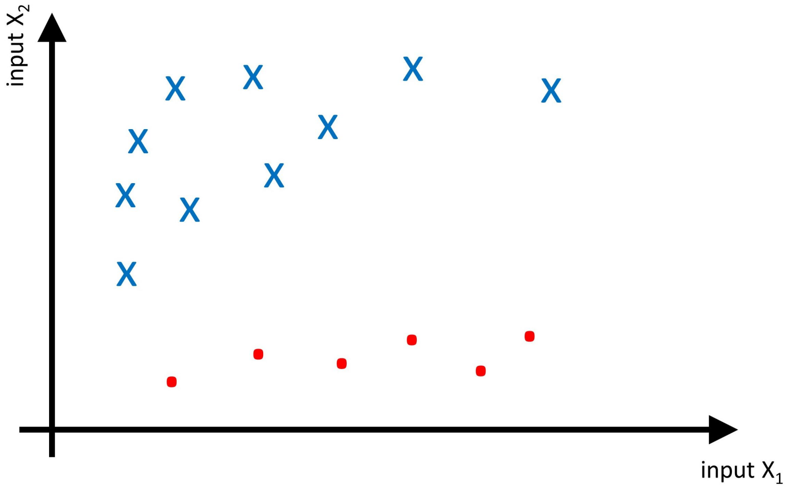
C Graded problems

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

C-1 Linear Separation Methods (4p)

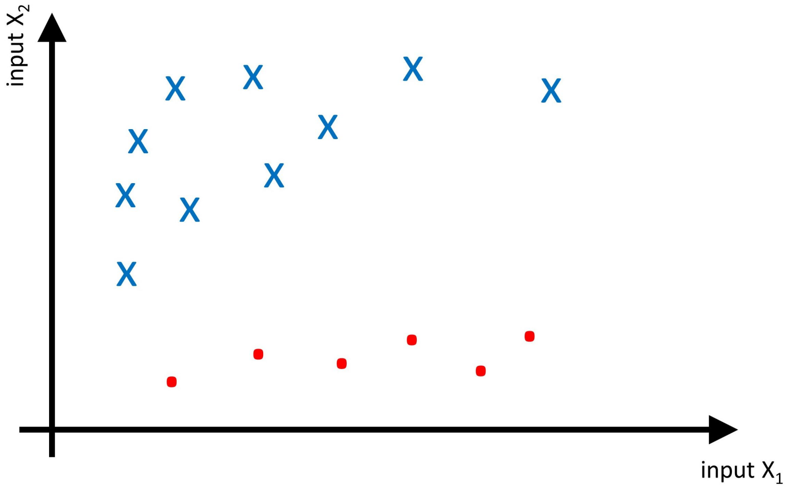
The following four diagrams (a)-(d) all show the same two-dimensional data points, which shall be separated by machine learning into the two classes blue X and red *.
For each of the mentioned **LINEAR CLASSIFICATION METHODS** draw a possible valid separating hyperplane in the diagrams.
Add a **BRIEF JUSTIFICATION (AT MOST 3 KEYWORDS!)** per diagram that explains your reasoning for the drawn separating hyperplane.

a) **linear neuron without bias** (bias=0).



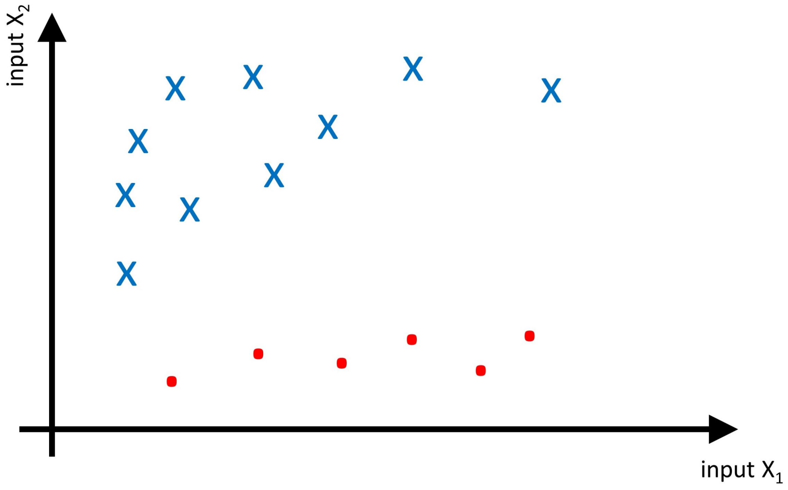
(1p)

b) **linear neuron with bias**, trained by gradient descent learning rule. Training stopped as soon as all data is classified correctly.



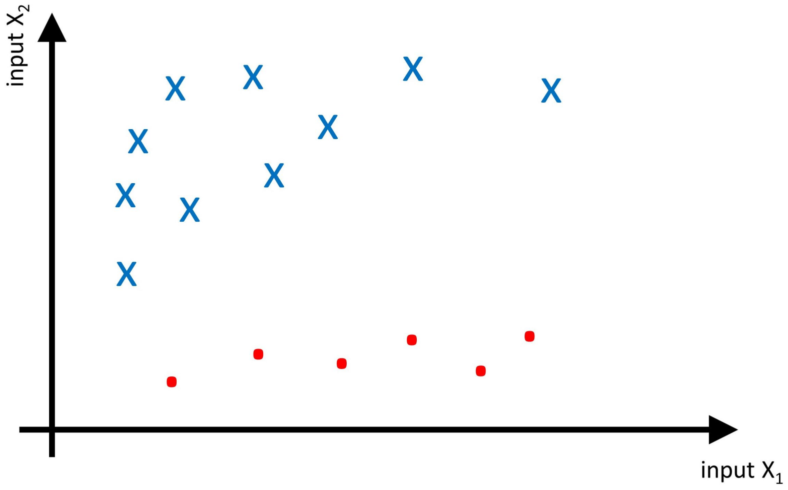
(1p)

c) **support vector machine** (SVM) WITHOUT kernel function.



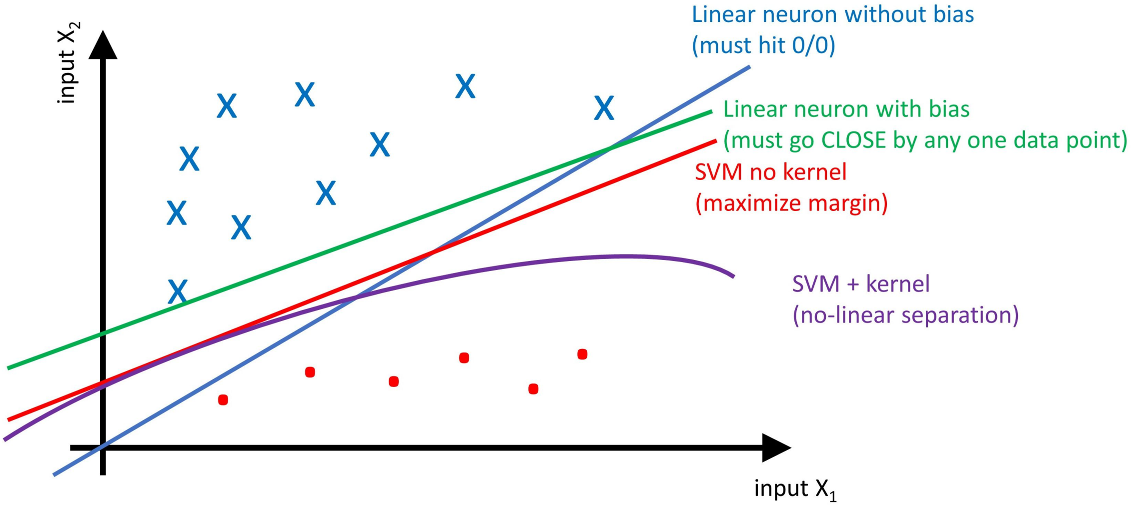
(1p)

d) **support vector machine WITH kernel function**.



(1p)

Solution:

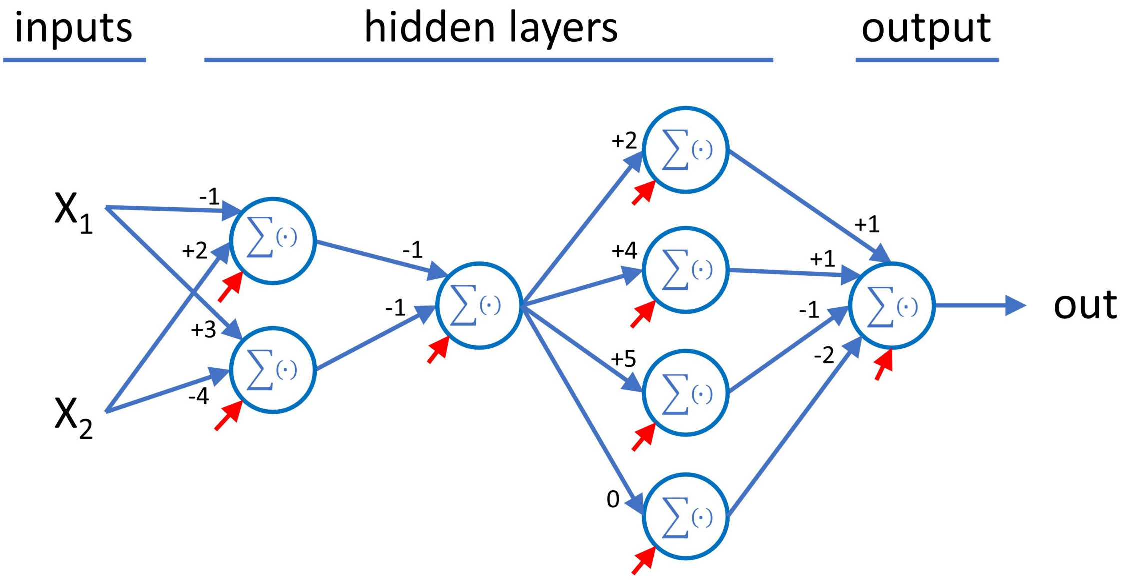


C-2 Neuronal Networks (2p)

Simplify the following **Neuronal Network** (NN) as much as possible.
The NN consists of only linear transfer units in all layers.

a) Sketch the smallest NN (with linear units) that performs the same function at its output. Label all weights and biases of your simplified NN. (1p)

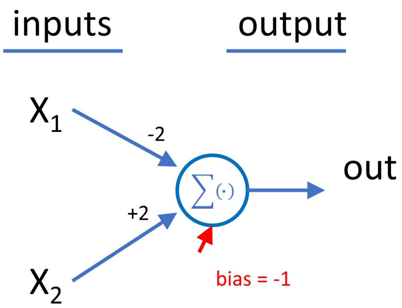
b) Explain **IN KEYWORDS** how you justify your solution. (1p)



↗ = bias input (constant **-1 bias for all neurons**)

$\sum(\cdot)$ = linear neuronal activation function: output = $\sum_i (inp_i * w_i)$

Solution:

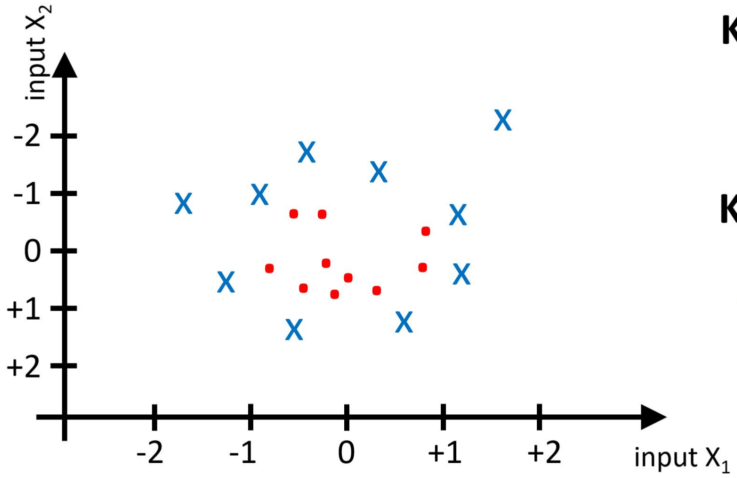


The "network's" output is out = -2*x_1 + 2*x_2 -1

A (large) network of linear units can be collapsed into a single linear unit, as any linear function of a linear function (of a linear function ...) is still a linear function. Hence, a single linear unit is sufficient.
How to find the weights? Write out the output function of the original network, simplify, and find out = -2*x1 + 2*x2 -1.

C-3 Support Vector Machines (2p)

The following diagram shows two-dimensional data samples of two NON-LINEARLY separable classes, displayed as blue X and red *. A project engineer has implemented **a support vector machine (SVM) with two different kernel functions** to separate the data points.



Kernel Function 1

$$\mathcal{K}(\vec{x},\vec{y}) = (\vec{x}^T\vec{y}+1)^p$$

Kernel Function 2

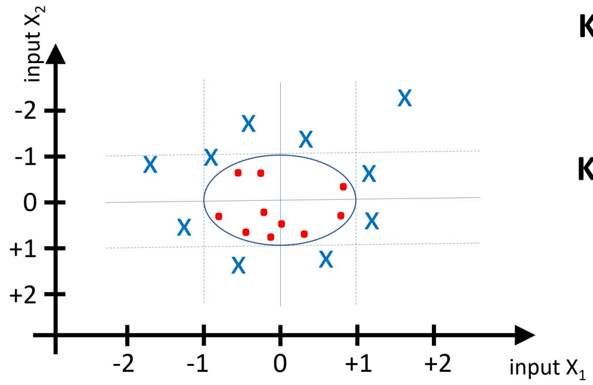
$$\mathcal{K}(\vec{x},\vec{y}) = e^{-\frac{1}{2\rho^2}||\vec{x}-\vec{y}||^2}$$

Unfortunately, the engineer left the project, and now it is your task to adjust the parameters that are contained in the already existing kernel functions. Pick **one of the two possible kernel functions** (state which you picked),

a) find an example value of parameter(s) that allow the SVM to correctly separate data points, if such is possible. Justify **with keywords**. (1p)

b) find an example value of parameter(s) that does NOT enable the SVM to correctly separate data points, if such is possible. Justify **with keywords**. (1p)

Solution:



Kernel Function 1

$$\mathcal{K}(\vec{x},\vec{y}) = (\vec{x}^T\vec{y}+1)^p$$

Kernel Function 2

$$\mathcal{K}(\vec{x},\vec{y}) = e^{-\frac{1}{2\rho^2}||\vec{x}-\vec{y}||^2}$$

Function 1
any integer value p >= 2 will create an additional dimension that allows classification whereas p=0 will not allow correct classification.
Function 2
any value p != 0 will allow classification.