KTH Computer Science and Communication

Exam in DD2421 Machine Learning
2022-10-21, 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.

	1) A principle to choose the simplest explanation
	2) The latent sampling solution
a) The LASSO	**3**) An approach to find useful dimension for classification
b) RANSAC	**4**) Random strategy for area compression
c) Occams's razor	**5**) Issues in data sparsity in space
d) Curse of dimensionality	**6**) An approach to regression that results in feature seletion
	7) Problems in computational costs
	8) Robust method to fit a model to data with outliers

Solution: a-6, b-8, c-1, d-5

A-2 Shannon Entropy (1p)

Consider a single toss of *skewed* coin (it is likely to show one side more than the other side). Regarding the uncertainty of the outcome {head, tail}, which one of the following is correct?

a) the entropy is smaller than one bit.

b) the entropy is equal to one bit.

c) the entropy is equal to two bits.

Solution: a

A-3 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 7%. What was the error rate with 1-nearest neighbor on the test set? Briefly reason the answer. (1p)

b) Next we use Bagging and get an error rate of 9% on the training data. We also get the average error rate (averaged over both test and training data sets) of 10%. (1p) What was the error rate with Bagging 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 10% with both 1-NN and Bagging. 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)

Solution:

a) 14%. (Training error for 1-NN is always zero, and therefore the testing error is 12%.)

b) 11%.

c) Bagging because it achieves lower error rate on the test data on average (10.5% < 12%).

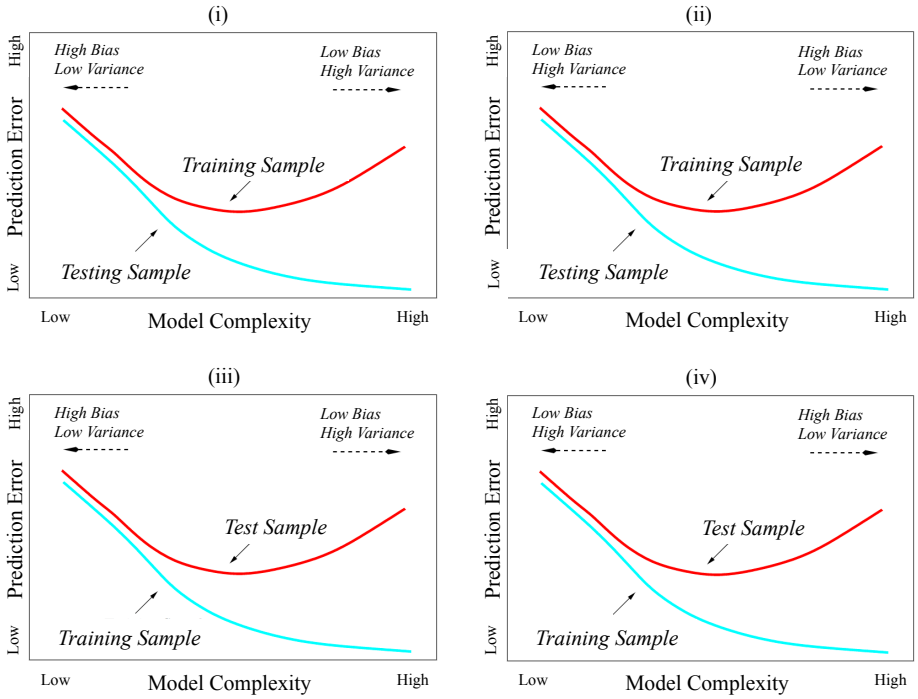


Figure 1. Typical behavior of prediction error plotted against model complexity.

A-4 Bias and Variance (5p)

a) One of the four subfigures (i)-(iv) in Figure 1 displays the typical trend of prediction error of a model for training and testing data with comments on its bias and variance, {high, low}. Which one of the four figures most well represents the general situation? (1p)

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

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

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

c) 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) (iii)

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

c)

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

A-5 Ensemble Methods (4p)

Briefly answer the following questions regarding ensemble methods of classification.

a) What are the two kinds of randomness involved in the design of Random Forests?

b) In Adaboost algorithm, each training sample is given a weight and it is updated according to some factors through an iteration of training weak classifiers. What are the two most dominant factors in updating the weights? How are they used?

Solution:

a) Randomness (i) in generating bootstrap replicas and (ii) in possible features to use at each node.

b) The update is according to (i) if the sample was misclassified (then the weight is increased, vice-versa), and (ii) the reliability of the weak classifier based on the training error used as the coefficient; the smaller the training error, the greater the reliability.

A-6 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, do we take eigenvectors corresponding to the largest, or smallest, eigenvalues of Q? (1p)

b) 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)

c) Given $\mathbf{x}$, 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 longest, the second longest, and the shortest, respectively. Based on this observation which class should $\mathbf{x}$ belong to? (1p)

Solution:

a) largest

b) $\sqrt{S}$ where $S = \Sigma_{i=1}^{p}(\mathbf{x}, \mathbf{u}_i)^2$

c) α

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

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

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

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

Solution: (a)

B-2 Conditioning (1p)

For three events A, B, and C, what is $P(A|B,C)P(B|C)P(C)$?

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

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

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

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

Solution: (c)

B-3 Shoogee gets a toy (2p)

Our dog Shoogee[1] (pictured) has two boxes of toys. One of them has 10 tennis balls. The other has one tennis ball and one small stuffed elephant. She brought me a tennis ball and then went to fetch another toy. What is the probability she brings back another tennis ball assuming she is equally likely to choose a toy from either box? *Show your work.*

Solution: In box 1 there are 10 tennis balls. In box 2 there is one tennis ball and one elephant. The probability she brings back another tennis ball given she brought back a ball the first time is equal to $1 - P(E|B)$, or one minus the probability she chooses the elephant conditioned on choosing the tennis ball. By Bayes' Theorem

$$1 - P(E|B) = 1 - \frac{P(B|E)P(E)}{P(B)} = 1 - P(E)/P(B) \tag{1}$$

since $P(B|E) = 1$. Now $P(E) = P(E|1)P(1) + P(E|2)P(2) = 0 + 0.25$, and $P(B) = P(B|1)P(1) + P(B|2)P(2) = 0.5 + 0.25 = 0.75$ – where the numbers denote the box from which Shoogee takes a toy. Finally,

$$1 - P(E|B) = 1 - P(E)/P(B) = 1 - 1/3 = 2/3. \tag{2}$$

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

B-4 Maximum likelihood estimation (3p)

You wish to model a labeled dataset $\mathcal{D} = ((x, y)_i : x, y \in \mathbb{R})$ of N observations with the following model:

$$y = w_0 + w_1 x^2 + \epsilon \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, i.e.,

$$Pr(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\epsilon^2/2\sigma^2}. \tag{4}$$

What is the maximum likelihood estimate of w_0 as a function of w_1? *Show your work.*

Solution: The random variable $Y|X = x$ is distributed

$$Y|X = x \sim \mathcal{N}(w_0 + w_1 x^2, \sigma^2). \tag{5}$$

[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".

The maximum likelihood optimality criterion says

$$
\begin{aligned}
(w_0, w_1)_{ML} &= \arg\max_{w_0,w_1} Pr(\mathcal{D}|w_0, w_1) && (6)\\
&= \arg\max_{w_0,w_1} \sum_i \log Pr(y_i|X = x_i)\\
&= \arg\max_{w_0,w_1} \sum_i \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-(w_0+w_1x_i^2))^2/2\sigma^2}\\
&= \arg\max_{w_0,w_1} \sum_i -\log\sqrt{2\pi\sigma^2} - (y_i - (w_0 + w_1x_i^2))^2/2\sigma^2\\
&= \arg\min_{w_0,w_1} \sum_i (y_i - (w_0 + w_1x_i^2))^2. && (7)
\end{aligned}
$$

Taking the partial derivative of this expression with respect to w_0 produces:

$$\frac{\partial}{\partial w_0} \sum_i (y_i - (w_0 + w_1x_i^2))^2 = \sum_i 2(y_i - (w_0 + w_1x_i^2))(-1). \quad (8)$$

Setting to zero and solving for w_0 gives:

$$w_0(w_1) = \frac{1}{N} \sum_i (y_i - w_1x_i^2). \quad (9)$$

B-5 Regression model (1p)

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

$$y = w_0 + w_1x^2 + \epsilon \quad (10)$$

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.

Solution: Both graphs will feature a parabola with a y-offset at w_0, but $w_1 < 0$ will be open downward and $w_1 > 0$ will be open upward.

B-6 Probabilistic classification, part 1 (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) $y(\mathbf{x}) = \arg\max_{y\in\{0,1\}} Pr(x_1|y)Pr(x_2|y)Pr(x_3|y)Pr(y)$

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

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

Solution: (a)

B-7 Probabilistic classification, part 2 (3p)

obs	x_1	x_2	x_3	y
1	r	y	0.2	0
2	r	y	0.5	0
3	g	y	0.1	1
4	b	y	0.3	0
5	g	n	0.5	0
6	b	y	0.9	1

Table 1. Dataset of six observations

Consider the dataset in table 1. Assume the first element of $\mathbf{x}$ is distributed Categorical, the second element is distributed Bernoulli, and the third element is distributed Uniform in $[0, 1]$. Using the Naive Bayes classifier, classify the observation $\mathbf{x} = (g, y, 0.5)$.

Solution: We will compute

$$Pr(x_1 = g|0)Pr(x_2 = y|0)Pr(0) \lessgtr Pr(x_1 = g|1)Pr(x_2 = y|1)Pr(1)$$

since we are assuming X_3 is distributed Uniform in $[0, 1]$. If the left hand side is greater, then $y = 0$. Otherwise, $y = 1$. From the table: we see $Pr(x_1 = g|0) = 1/4$ and $Pr(x_1 = g|1) = 1/2$; $Pr(x_2 = y|0) = 3/4$ and $Pr(x_2 = y|1) = 1/2$. Finally, $P(0) = 2/3$ and $P(1) = 1/3$. We see the left hand side is greater, and so we choose class 0.

C Graded problems

Potential inquiries to be addressed to Jörg Conradt.

C-1 Multiple-Choice: Artificial Neural Networks (1p)

What is the underlying principle in *backpropagation* to train an artificial neural network?
Do not justify your answer. Instead select **one** option of 1, 2, or 3.

1. The weights are modified to minimize the mismatch between the actual and the desired output.
2. A Gaussian distribution is used to approximate the training data.
3. The number of hyper-planes is maximized using a dual formulation.

Solution: 1

C-2 Multiple-Choice: Support Vector Machine (1p)

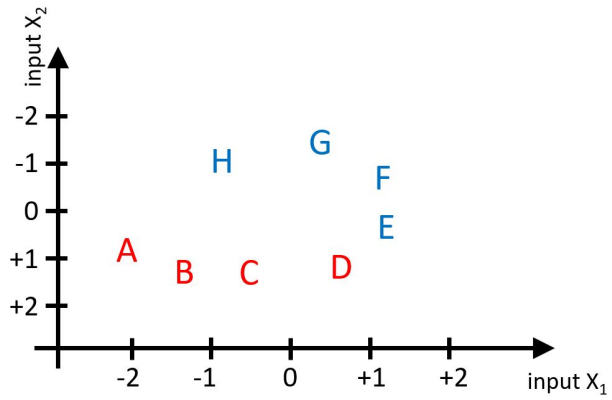
Assume a *kernel*-function in a support vector machine.
What does this function mathematically correspond to?
Do not justify your answer. Instead select **one** option of 1, 2, or 3.

1. The generalised distance between any data point and the decision boundary.
2. The scalar product between two data points transformed into a higher dimensional space.
3. The midpoint of the training data, computed separately for each class.

Solution: 2

C-3 Support Vector Classification (3p)

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.

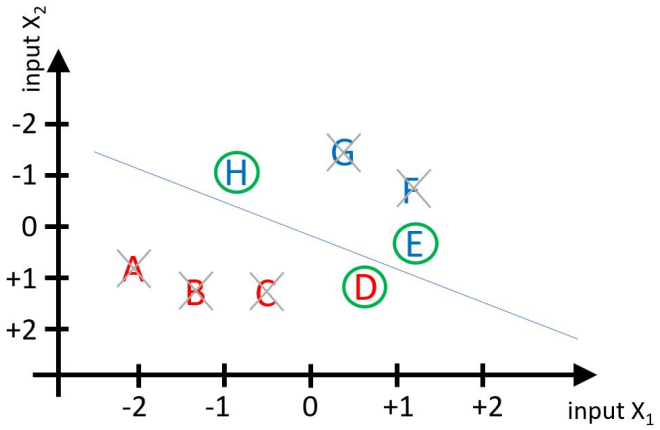


a) We use a linear support vector machine (SVM) without kernel function to correctly separate the BLUE and the RED class. Which of the data points (A-H) can be removed for training without changing the resulting SVM decision boundary? No explanation needed; name the point(s). (1p)

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)

Solution:

a) The blue line shows the linear decision boundary between the two classes.



The data points in green circles are support vectors for the linear decision boundary. If any of those data points change, the decision boundary changes. In contrast, we can remove all other data points: **A, B, C, F, and G**.

b) at least one of each; max 1 point for (+) and 1 point for (-)

+) The decision boundary margin might get wider with a non-linear kernel.

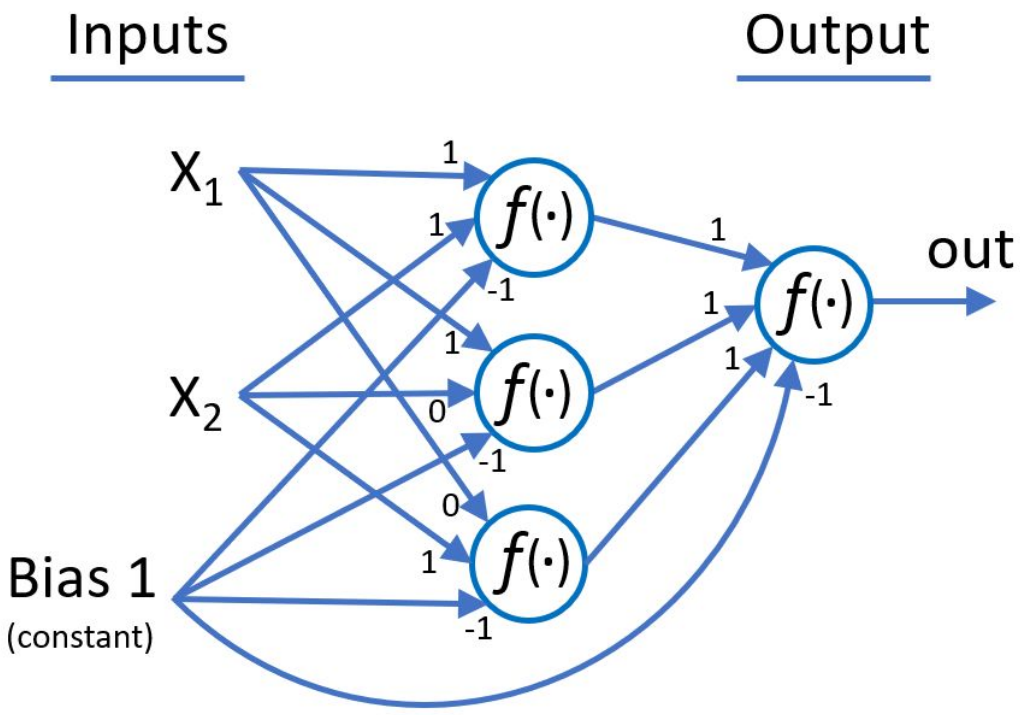
+) The same learning approach is likely to work for additional (possibly more complex) data.

-) More computing resources required.

-) The algorithm is more difficult to implement.

C-4 Neuronal Networks (3p)

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



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

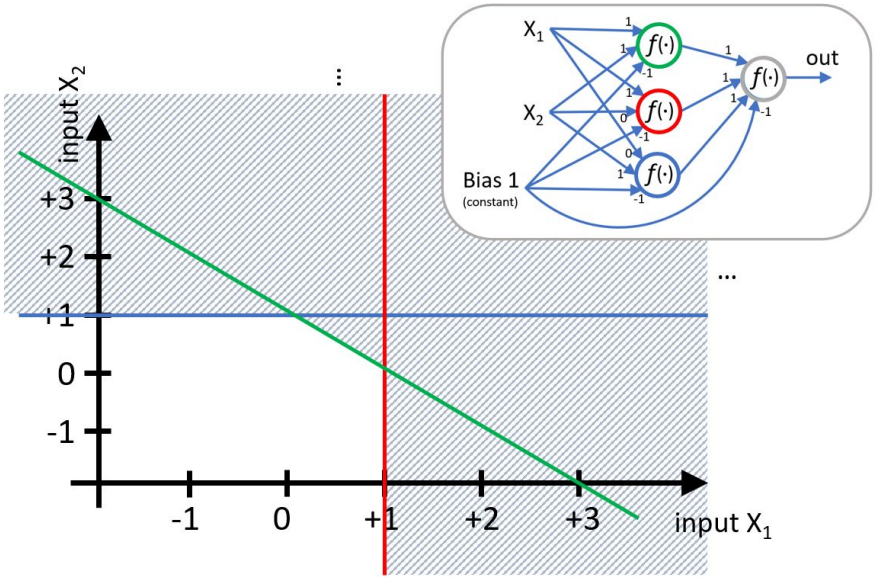
$$\text{step}\left(\sum_i (inp_i * wi)\right) = \begin{cases} 1 \text{ if } (\cdot) \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

a) Draw an input space diagram (two dimensional plot of X_1 and X_2) and show for which area of the input space the network produces a positive output. (1p)

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. What happens with the output? (1p)

Solution:



a) Diagram shown above. Each separation line (colored) is given by one of the three neurons in the first layer. The output neuron is active if at least one of those neurons is active (OR function). Hence, the shaded region corresponds to the area of the input space (x1,x2), where the network generates a positive output.

b) The output function of this network is highly non-linear (a step with a piece-wise linear decision boundary in input space); hence a single linear neuron **CANNOT** implement this function.

c) **No change at the output**, as all input (including bias) for all neurons doubles. Therefore, all total input that was below zero is still below zero; all total input that was above zero is still above zero. All neurons show the exact same output signals.