Linnæus University
Department of Computer Science
Jonas Nordqvist
Jonas Lundberg
Rafael Messias Martins

Exam in Introduction to Machine Learning, 2DV516, 7.5 hp
June2, 2020, 8.00–13.00

The exam consists of 7 questions. The maximum number of points for the exam is 40.

1. You have a dataset $\mathcal{D}$ consisting of n =100,000 datapoints and $p = 30$ features. To each datapoint there is a corresponding binary label. Your objective is to investigate which of the three hypothetical machine learning models A, B and C will have the best performance. Describe your method of finding the best model, and how you will estimate the performance of the best model. (4p)

2. In Figure 1 a neural network with three binary inputs (x_1, x_2, x_3), one hidden layer with two nodes and one output is given. All activation functions $f_h(x)$ in the hidden units and the output unit are step functions defined as

$$f_h(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

The network should output the following: if *exactly* two inputs are active then the output should be 1 otherwise 0. (6p)

(a) Find weights and biases solving this problem.

(b) Produce the matrices $W^{(2)}$ and $W^{(3)}$ which carries all the weights for your answer in a).

(c) Is this a linear problem in the sense that a classifier producing a linear decision boundary could solve the problem? Motivate your answer!

Hint: one hidden unit should activate if all inputs are active, and one should activate if two or more are active.

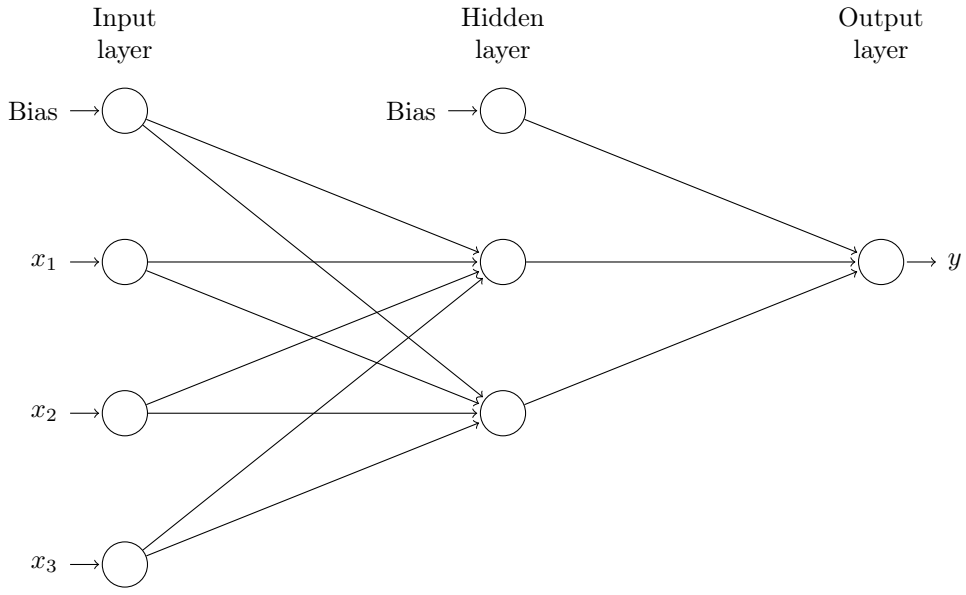

Figure 1: A neural network with one hidden layer consisting of two nodes.

3. Recall that the purity gain of a binary split in a decision tree is computed as follows

$$\Delta = I(\text{parent}) - \sum_{j=1}^{2} \frac{N(v_j)}{N} I(v_j),$$

where I is an impurity measure, and $N(v_j)$ is the number of elements in the node v_j (*i.e.* the jth subregion after the split), and N is the number of elements in the parent node (*i.e.* the region in which we perform the split). Suppose that our impurity measure is the Gini index which is defined, for the mth region, by

$$G_m = 1 - \sum_{k=1}^{K} \widehat{p}_{mk}^2,$$

where $\widehat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class.

x_1	191	271	201	162	188	169	213	241	191	237
x_2	101	122	115	99	116	98	110	101	91	122
Class	B	A	A	C	B	C	A	A	B	A

Table 1: A data table containing the features x_1 and x_2

Given the three-class data in Table 1 compute the purity gain from both the splits S_1, where the split is taken as $x_1 \geq 170$ and S_2, where the split is taken as $x_2 \geq 102$. Based on your computations, which is the preferred split? (6p)

4. Consider a three-class classification task with only three training examples as seen in Figure 2. Suppose that a maximal margin classifier is used and trained on the mentioned training examples, together with either (i) one-vs-one binarization or (ii) one-vs-all binarization. (6p)

1. In each of these cases (i) and (ii). How many classifiers needs to be trained?
2. If this was a four-class problem, how many classifiers would be trained in each of the cases (i) and (ii)?
3. Are the same decision boundaries produced by (i) and (ii)? Motivate your answer, and if necessary provide illustrations to aid your explanations.
4. In this particular example is either (i) or (ii) preferred over the other? Motivate your answer. (Note that one is not preferred over the other in a objective sense, so it is the argument itself which is important)

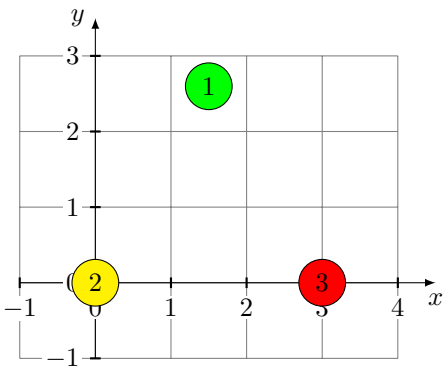

Figure 2: Three training examples from the classes with labels $1, 2$ and 3. Note that the distance between each pair of training examples is the same.

5. A kNN regression using $k = 3$, and the Manhattan distance is used to predict the price of houses. The data for the model is presented in Table 2. Recall that the Manhattan distance between two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is given by

$$|\mathbf{x} - \mathbf{y}| = \sum_{i=1}^{n} |x_i - y_i|.$$ (6p)

price (MSEK)	4.3	7.3	2.2	1.9	10.1	3	3	6.6	8	4.2	4.1	8.3
size (m^2)	150	170	130	100	208	140	145	160	124	135	141	98
# **of rooms**	5	6	4	4	7	5	5	5	4	4	4	5
distance to centrum (km)	2.1	1.2	4.4	5.4	12.1	5	6.3	0.3	0.7	5.5	3.2	11.1

Table 2: Housing price data

(a) What is the predicted cost of a house of 150 m^2 in size, 5 rooms and which lies 4 km from the city centre?

(b) Looking at the table it seems as the most expensive houses are those which is either located near the city centre or far away. Let x_i denote the distance to the city centre for the ith training example. Find a transformation of this feature such that examples which are either far from the city centre or near have are close (for this feature). Change x_i to your transformed feature and repeat (a). How is the prediction changed?

(c) A potential source of unwanted effects is that the features are not normalized (set at a similar scale). What effect does the fact that the data is non-normalized have in this example?

6. Different Dimensionality Reduction (DR) methods may differ in many ways, such as (a) the required format of the input data, (b) the goal of the reduction, i.e., what characteristics of the original data are chosen to be kept in the final output, and (c) the computational methods used to reach their goal. Describe and compare two of the DR methods we discussed in the lecture, PCA and MDS, regarding these three aspects. (6p)

Good Luck!