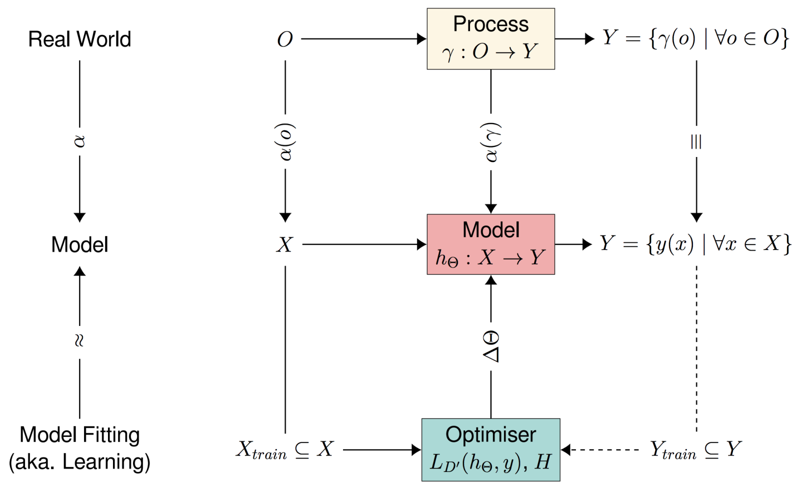
Exercises

Machine Learning - Basics

1. Recall from the lecture notes the following figure. It depicts abstractly a typical machine learning system and it illustrates how the system is linked to the real world. Answer the following questions:



(a) A machine learning system aims to find the best parameters Θ of the model function h_Θ by minimizing the loss $L_{D'}(h_\Theta, y)$ on a dataset D'.
☐ Yes ☐ No ☐ Don't know

(b) One key difference between the ideal target function γ and the model function h_Θ is that γ acts on extracted features whereas h_Θ acts on real-world objects.
☐ Yes ☐ No ☐ Don't know

(c) A distinction between (X_{train}, Y_{train}) and (X, Y) wouldn't be necessary if optimization methods were more efficient on larger datasets.
☐ Yes ☐ No ☐ Don't know

(d) α can introduce modeling errors into the system.
☐ Yes ☐ No ☐ Don't know

(e) Learning requires a trade-off between fitting the seen data perfectly and sticking to prior-assumptions about the unseen data distribution.
☐ Yes ☐ No ☐ Don't know

(f) The "No-free-lunch-theorem" states that the best parameters Θ of the model function can only be found by exhaustive search.
☐ Yes ☐ No ☐ Don't know

2. You are given a dataset of different kinds of beer from a beer tasting jury.
Each row represents one data object.

ID	REGION	PRICE	AWARD
0	Lower Bavaria	18.70	bronze
1	Upper Bavaria	19.90	silver
2	Upper Franconia	7.20	silver
3	Lower Bavaria	16.50	gold
4	Lower Franconia	11.80	bronze
5	Lower Bavaria	17.40	gold
6	Upper Bavaria	24.50	silver
7	Lower Bavaria	13.90	bronze

(a) Specify the type of the attributes *ID*, *PRICE* and *AWARD* according to the *four* attribute types introduced in the lecture.

ID :

PRICE :

AWARD :

(b) Calculate the mode of the attribute *REGION*.

Mode :

(c) Calculate the sample mean $\bar{x}_{price}$ of attribute *PRICE*. Show your workings and round the result to two decimal places.

$\bar{x}_{price} \approx$

(d) Name and describe briefly **three** preprocessing methods.

3. We use Simple Linear Regression to model the relationship between a dog's weight x and its running speed y. Below, you are given the definition of a Simple Linear Regression model with the sum of squares (SSE) as error measure and $n = 4$ samples of dogs' running speeds. The **parameters** of the model **are given** as: $\beta_0 = 3.00$ and $\beta_1 = 1.82$.

$$f(x) = \beta_0 + \beta_1 \cdot x$$

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Dog	1	2	3	4
Weight (kg)	16	2	10	18
Running Speed (km/h)	37	5	21	30

(a) Use the given parameters to calculate the four predicted values $f(x_i)$ and the SSE of this model. Round the results to two decimal places.

$f(x_1) =$

$f(x_2) =$

$f(x_3) =$

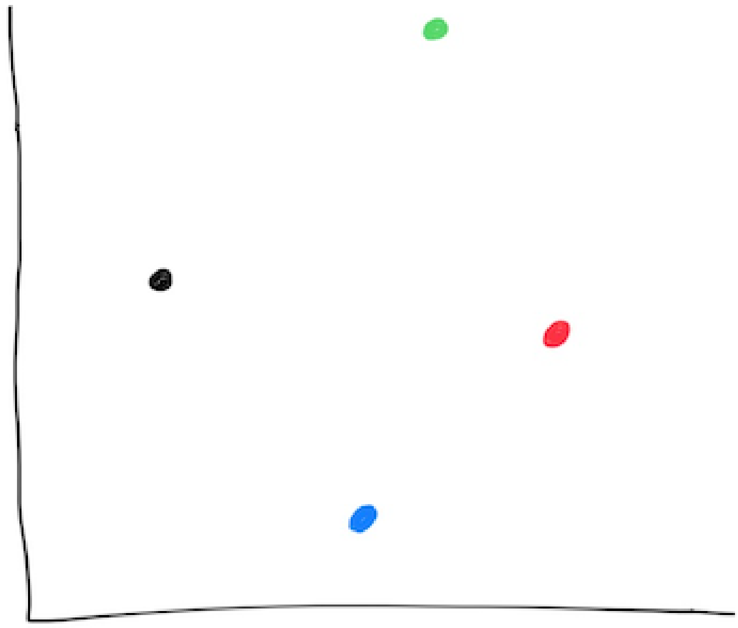
$f(x_4) =$

$SSE(3.00, 1.82) =$

(b) The parameter β_1 is correct. The parameter β_0 is **incorrect**. Derive the correct value of β_0 from the SSE error measure by using the method of least squares. Show your workings and round the result to two decimal places.

$\beta_0 \approx$ $\boxed{}$

4. Given the following illustration of a feature space with four data points - each from a different class. Draw the decision boundary of a 1-nearest neighbour classifier that uses euclidean distance.



5. Suppose you are given a training data set $D = \{(\mathbf{x}, y)_i\}_{i=1}^{N}$ consisting of N training examples $(\mathbf{x}, y)$ and you want to implement the k-NN classifier in its original (naive) form.

   (a) Describe the training process and the memory requirement of such a classifier.

   (b) How many distance evaluations are necessary during prediction for a single test example?

   (c) Describe how the hyper-parameter k affects the generalization capability of the k-NN classifier.

Machine Learning - Evaluation

6. In a multiclass classification problem with three classes $Y = \{1, 2, 3\}$, the figure below shows a confusion matrix for a classifier $\hat{y} = h(\mathbf{x})$ evaluated on some dataset $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.

		Predicted $\hat{y}$		
		1	2	3
True y	1	4	3	3
	2	1	9	0
	3	80	10	10

(a) How many examples are in each of the three classes (according to D)?

$N_1 =$

$N_2 =$

$N_3 =$

(b) What is the probability that a *random classifier* would predict class 1, 2 or 3?

P(C=1) =

P(C=2) =

P(C=3) =

(c) For class 2, determine the number of *True Positives* TP_2, *False Positives* FP_2, *False Negatives* FN_2 and *True Negatives* TN_2.

$TP_2 =$

$FP_2 =$

$FN_2 =$

$TN_2 =$

(d) Calculate the *macro*-averaged precision over all classes. Show your workings and round the result to two decimal places.

$\pi_{macro} \approx$ $\boxed{}$

(e) Among the three predicted classes, for which predicted class can we be most certain that its prediction is in fact correct? Briefly explain your answer.

Class :

7. Cross-validation is a general model evaluation principle with the purpose of estimating the generalization capability of a predictor. Given a dataset D, check whether the following statements are true or false.

(a) Holdout validation estimates the error on all samples in D.
☐ Yes ☐ No ☐ Don't know

(b) In .632-Bootstrapping, in each iteration the training set is formed by drawing samples from D with replacement.
☐ Yes ☐ No ☐ Don't know

(c) In .632-Bootstrapping, in each iteration there is a small chance that a sample is both in the training set and the test set.
☐ Yes ☐ No ☐ Don't know

(d) For k-fold cross-validation, we have to specify also the size of the training set and the size of the test set a-priori.
☐ Yes ☐ No ☐ Don't know

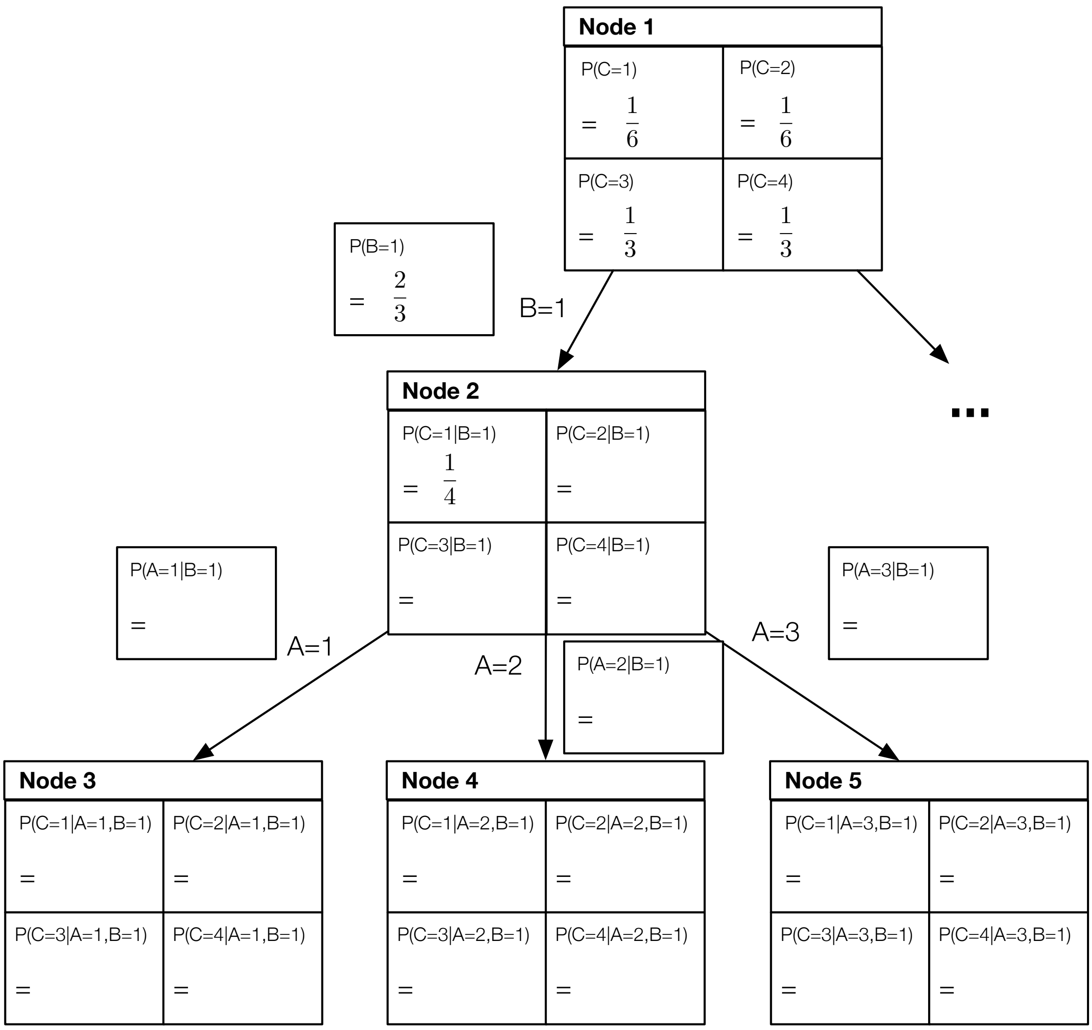
(e) The above-mentioned validation principles are applicable to both classification and regression problems.
☐ Yes ☐ No ☐ Don't know

Decision Trees

8. Given the following dataset with categorical attributes A, B and the class attribute C. Someone started running the **ID3-algorithm** with *misclassification-impurity* on this dataset. Unfortunately, the algorithm is not finished yet. We have to calculate the impurity reduction for one split manually.

A	B	Class
2	1	1
3	1	1
1	1	2
2	1	2
2	1	3
1	1	3
1	1	3
1	1	3
2	2	4
2	2	4
1	2	4
3	2	4

(a) The figure below illustrates the current state of the tree after attribute B was selected in the root node (Node 1). Complete the figure by filling in the corresponding statistics (relative frequencies) in the empty boxes.



(b) Calculate the misclassification-impurities of nodes $2, 3, 4$ and 5.
NOTE: $D_{A=a,B=b}$ denotes the set of samples which have values $A = a$ and $B = b$.

$\iota_{mis}(D_{B=1}) =$ $\iota_{mis}(D_{A=1,B=1}) =$

$\iota_{mis}(D_{A=2,B=1}) =$ $\iota_{mis}(D_{A=3,B=1}) =$

(c) Calculate the impurity reduction $\Delta\iota_{mis}(D_{B=1}, \{D_{A=1,B=1}, D_{A=2,B=1}, D_{A=3,B=1}\})$, when **node 2** splits on attribute A. Show your workings.

$\Delta\iota_{mis}(D_{B=1}, \{D_{A=1,B=1}, D_{A=2,B=1}, D_{A=3,B=1}\}) =$

(d) Determine the class labels c_i associated with nodes $2, 3, 4$ and 5.
NOTE: c_i is the class label associated with node i.

c_2 : c_3 : c_4 : c_5 :

(e) Is it strictly necessary to calculate the impurity reduction for **node 2** at this stage? Answer with YES or NO and briefly explain your answer.

Answer :

(f) Which class would the tree predict for a sample with values $A = 2$ and $B = 2$? Answer by stating the class label and briefly explain your answer.

Class Label :

Neural Networks

9. Given the following dataset with attributes x_1, x_2 and associated class labels $y(\mathbf{x})$, we want to learn the weights of a perceptron such that the perceptron classifies the four samples correctly.

   - threshold function is the heaviside step function $\varphi(x) = \max(sign(x), 0)$
   - learning rate is $\eta = 0.4$
   - weights are initialized as $\mathbf{w} = (0.5, -1, 1)$

x_1	x_2	$y(\mathbf{x})$
1	0	0
0	0	0
1	1	0
0	1	1

Apply the *perceptron training algorithm* on the four samples in the given order. Iterate over the samples only once and **fill in the following table** with intermediate results of the algorithm.

NOTE: For notational convenience, we added a dummy attribute $x_0 = 1$ to the dataset. The vector $\mathbf{x} = (x_0, x_1, x_2)$ denotes a sample of the dataset and the vector $\mathbf{w} = (w_0, w_1, w_2)$ denotes the weights of the perceptron. i is the iteration counter and the last column contains the new weights after applying a weight update.

i	x_0	x_1	x_2	$y(\mathbf{x})$	$\mathbf{w}^T\mathbf{x}$	$\varphi(\mathbf{w}^T\mathbf{x})$	err	Δw_0	Δw_1	Δw_2	w_0	w_1	w_2
0		–		–	–	–	–		–		0.5	-1.0	1.0
1	1	1	0	0									
2	1	0	0	0									
3	1	1	1	0									
4	1	0	1	1									

10. Answer the following questions

(a) In the perceptron training algorithm, the error induced by a sample is proportional to the distance of that sample to the hyperplane defined by parameters $\mathbf{w}$.
☐ Yes ☐ No ☐ Don't know

(b) If two classes are linearly separable, the perceptron training algorithm will converge.
☐ Yes ☐ No ☐ Don't know

(c) If two classes are not linearly separable, the perceptron training algorithm will alternate between exactly two states of the weight vector.
☐ Yes ☐ No ☐ Don't know

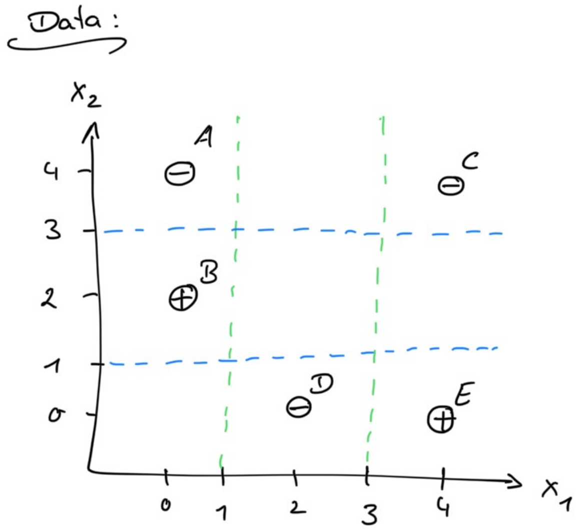
(d) Gradient descent-based optimization algorithms are used for training neural networks because they find the global optimum independent of the network's architecture.
☐ Yes ☐ No ☐ Don't know

(e) The *mini-batch* gradient descent algorithm sums the weight adaptations over a small subset of samples before applying an update.
☐ Yes ☐ No ☐ Don't know

11. Describe the steps involved in training a multi-layer perceptron for *binary classification*. Assume that the network is trained for a *single epoch* using *stochastic gradient descent* with mini-batches. Give a conceptual description of the process, name the relevant quantities to be computed and use adequate terminology.

Boosting - Adaboost

12. You are given the following data set together with a set of decision tree stumps. Apply the Adaboost algorithm for two rounds to determine the ensemble classifier $H(\mathbf{x})$. Write down the weights of all training examples and the error rates of all decision stumps in each round.



Decision Stumps:

$x_1 > 1$

$x_1 > 3$

$x_2 < 1$

$x_2 > 3$

Remarks:

- The data set consists of five points (A, B, C, D, E). Data points B and E belong to the positive class (+). Data points A, C, D belong to the negative class (-).
- Decision tree stumps are to be interpreted in the following way: A stump, such as $x_2 < 1$, means that this stump classifies all data points with an x_2-value smaller than 1 as (+) and all other points as (-). Another stump, such as $x_1 > 3$, classifies all data points with an x_1-value larger than 3 as (+) and all other points as (-).
- To determine the best classifier, choose the one whose error rate ϵ^t is furthest from 0.5. I.e. the one that maximizes $|\epsilon^t - 0.5|$.
- As a tie breaker when determining the best classifier, use the topmost classifier (ordered as below).
- The voting power of the chosen classifier is calculated as: $\alpha^t = \frac{1}{2}\ln(\frac{1-\epsilon^t}{\epsilon^t})$