Machine Learning 1 June 29, 2021

Name: .....

Matr.Nr.: | | | | | | | |

Group: A B C D

There is exactly one correct answer per multiple choice question.
Each correctly answered question gives 2 points (in total there are 60 points).
*Selecting multiple answers or a wrong answer is considered as a mistake. Each mistake gives **1 negative point**.*

	a	b	c	d
1	○	○	○	○
2	○	○	○	○
3	○	○	○	○
4	○	○	○	○
5	○	○	○	○
6	○	○	○	○
7	○	○	○	○
8	○	○	○	○
9	○	○	○	○
10	○	○	○	○
11	○	○	○	○
12	○	○	○	○
13	○	○	○	○
14	○	○	○	○
15	○	○	○	○
16	○	○	○	○
17	○	○	○	○
18	○	○	○	○
19	○	○	○	○
20	○	○	○	○
21	○	○	○	○
22	○	○	○	○
23	○	○	○	○
24	○	○	○	○
25	○	○	○	○
26	○	○	○	○
27	○	○	○	○
28	○	○	○	○
29	○	○	○	○
30	○	○	○	○

Machine Learning 1 June 29, 2021

IGI Exam - Questions Sheet

Group A

Name: ..... Matr.Nr.: .....

There is exactly one correct answer per multiple choice question.
Each correctly answered question gives 2 points (in total there are 60 points).
Selecting multiple answers or a wrong answer is considered as a mistake. Each mistake gives ***1 negative point****.*

1) K-means

   (a) uses Euclidean distance to assign data points to the closest cluster.

   (b) can be used not only for continuous, but also for categorical variables.

   (c) always converges to the global minimum of the cost function J.

   (d) uses Euclidean distance to measure the distance between cluster means.

2) Gradient Descent algorithm

   (a) can lead to convergence to a local, instead of global minima, depending on the cost function.

   (b) requires a learning rate, and partial derivatives of the cost function w.r.t. each component of the optimal vector Θ to be specified.

   (c) can be used with many different cost functions.

   (d) all answers are correct.

3) Logistic regression:

   (a) the cost function can be minimized by Gradient Descent or variants of Gradient Descent, such as SGD (Stochastic Gradient Descent), or second order optimization methods.

   (b) can be used for regression problems, where the labels are continuous values.

   (c) uses a non-convex cost function.

   (d) as the final solution finds a hypothesis for which the decision boundary is always linear.

4) You apply 5-fold cross-validation for model selection using a training set with 1000 samples. How many samples will you use in the validation set at each round of cross-validation?

   (a) 5000

   (b) 200

   (c) 800

   (d) 1000

5) Consider support vector machines. If we apply the kernel trick to transform the original feature space into a higher-dimensional space that makes the data linearly separable, how many support vectors will we end up with?

   (a) It depends on the training data.

   (b) At most four.

   (c) It depends on the dimensionality of the transformed higher-dimensional space.

   (d) It depends on the dimensionality of the input.

6) Which one of the following models is likely to be subject to *underfitting*?

   (a) Handwritten digit recognition with deep convolutional neural networks.

   (b) Approximating the y=sin(x) function from (x,y) pairs using a multilayer perceptron.

   (c) Image-based medical diagnosis using a logistic regression model.

   (d) A single hidden layer feedforward neural network solving the XOR task.

7) Consider the problem of estimating the continuous valued price of a house in $ from several explanatory measurements (e.g., house size, distance to the railroad, etc.) to be solved using a feedforward neural network and sufficient training data. When designing your network architecture for this problem, the output layer should have a ...

   (a) Linear activation

   (b) Softmax activation

   (c) Sigmoid activation

   (d) Any of these

8) How many of the following statements about learning in neural networks is correct?

   - Backpropagation algorithm begins by initializing all network weights with zero values.
   - Backpropagation algorithm helps to compute the error gradients for all parameters through the use of Bayes' rule.
   - In mini-batch learning, weights are updated after all training set samples are seen.
   - Parameters of a neural network are optimized by minimizing the error through gradient descent on the test set.

   (a) 3

   (b) 0

   (c) 1

   (d) 2

9) If you perform *whitening* on data samples, the transformed samples have ...

   (a) zero mean

   (b) all of these

   (c) same variance

   (d) decorrelated features

10) In linear regression, the Mean-Squared Error function (MSE)

   (a) is used only to derive the closed-form (analytical) solution.

   (b) MSE is not used for linear regression, but for logistic regression.

   (c) is used for both optimization of the optimal parameters Θ by Gradient Descent, and the derivation of the closed-form (analytical) solution.

   (d) is used only when the optimal parameters Θ are optimized by Gradient Descent.

11) Consider a supervised classification task on a given data set. Target values, or labels, are included in:

   (a) training dataset.

   (b) labels are not included in data for supervised classification.

   (c) training and test datasets.

   (d) test dataset.

12) How can we keep the size and depth of a decision tree small while maintaining good performance?

(a) We monitor the error as we grow the tree and keep the tree where the error rate does not change significantly.

(b) While we grow our tree, we can compute the optimal partitioning of the tree and keep the nodes that are most important, i.e. have the lowest split cost.

(c) We grow the tree until there are no more worthwhile nodes that we can split. Then we keep removing the leaves and nodes that do not change the error rate significantly.

(d) We grow a large number of trees randomly and choose the tree with the lowest complexity and cost.

13) Precision and recall are important classification metrics to measure the performance of a classifier. Precision (P) and recall (R) are defined using true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

(a) $P = \frac{TP}{TP+FP}, R = \frac{TN}{TN+FN}$

(b) $P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}$

(c) $P = \frac{TP}{TN+FP}, R = \frac{TP}{TP+FP}$

(d) $P = \frac{TN}{TP+FP}, R = \frac{TP}{TP+FN}$

14) Consider modelling dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ with probability distribution $P(\mathbf{x})$. Which of the following statements is true?

(a) Kernel density estimation is a parametric technique.

(b) None of the other statements are true.

(c) When the distribution is deduced directly from the data, without a-priori assumptions about data distribution, we call the approach parametric modelling.

(d) When we make a-priori assumptions about data distribution, the approach is called nonparametric modelling.

15) Consider a regression problem on a given data set. How many of the following statements makes sense as a criterion for model selection?

- Selecting the model with the lowest error on the validation set.
- Selecting the model with the highest classification accuracy on the validation set.
- Selecting the model with the highest classification accuracy on the training set.

(a) 3

(b) 2

(c) 1

(d) 0

16) Which of the following statements about the Gaussian distribution is **incorrect**?

(a) Many sets of data show signs of normality or can be appropriately translated into a version of Gaussian distribution.

(b) Under certain conditions, the sum of random variables has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases.

(c) Variance and mean of the one-dimensional Gaussian distribution are always positive.

(d) Many natural phenomena in real life can be approximated by the Gaussian distribution.

17) You are given a data set with 90 observations where each observation is represented by a 10-dimensional feature vector. If you want to perform PCA to represent your data using only two dimensions through the principal components, you should select the eigenvectors of the data covariance matrix which corresponds to the ... eigenvalues out of a total of ... eigenvalues obtained after eigendecomposition.

(a) two smallest / 90

(b) two largest / 10

(c) two largest / 90

(d) two smallest / 10

18) For a linear regression problem with nonlinear features:

(a) the design matrix is extended by adding new features that are independent of the original feature x.

(b) there is no closed-form (analytical) solution.

(c) we are able to include a prior knowledge about the regression problem.

(d) all answers are correct.

19) Consider the AdaBoost algorithm.

(a) At iteration m, we compute $f_m(x) = f_{m-1}(x) + \nu\beta_m\phi(x;\gamma_m)$ and update all parameters $\beta_1, \ldots, \beta_m$ and $\gamma_1, \ldots, \gamma_m$.

(b) AdaBoost minimizes the margin on the training set.

(c) We want to minimize the loss $\sum_{i=1}^{N} \exp\left(-\tilde{y}_i f(x_i)\right)$, where $\tilde{y}_i \in \{-1, 1\}$.

(d) The analytical solution is given by $w^* = \frac{1}{2}\log\frac{\hat{\pi}}{1-\hat{\pi}}, \hat{\pi} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i = 1)$

20) Logistic (sigmoid) function, used in logistic regression is:

(a) bounded (that is, output values are always in a certain range) and differentiable.

(b) all answers are correct.

(c) thresholded at value 0.5 to predict the class 0 or 1.

(d) is interpretable, and it tells us how certain the prediction is. Its values can be interpreted as probabilities.

21) Which of the following statements about the Bayes rule is **incorrect**?

(a) Bayes rule can be derived from the product rule and the definition of conditional probability.

(b) Bayes rule can only be used on continuous random variables.

(c) Bayes rule is used in the derivation of Maximum a-Posteriori (MAP) estimation.

(d) Bayes rule describes the relationship between the posterior, the prior and the likelihood.

22) Consider a maximum likelihood estimation of the data X, coming from a three-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. What is true about the estimated parameter $\hat{\boldsymbol{\mu}}$?

(a) The parameter $\hat{\boldsymbol{\mu}}$ is one-dimensional (a scalar).

(b) The parameter $\hat{\boldsymbol{\mu}}$ is the value of $\boldsymbol{\mu}$ that minimizes the likelihood $p(X|\boldsymbol{\mu})$.

(c) The parameter $\hat{\boldsymbol{\mu}}$ is the value of $\boldsymbol{\mu}$ for which likelihood $p(X|\boldsymbol{\mu})$ and log likelihood $Log(p(X|\boldsymbol{\mu}))$ are equal.

(d) The parameter $\hat{\boldsymbol{\mu}}$ is the value of $\boldsymbol{\mu}$ that maximizes the likelihood $p(X|\boldsymbol{\mu})$.

23) Which of the following statements is *incorrect*?

(a) Decision trees implicitly choose the most important variables.

(b) Decision trees scale well to large datasets.

(c) Random forests have good predictive performance.

(d) Small changes to input data has little effects on the structure of the tree.

24) It can be shown, that ...

(a) ... we can use boosting to increase the performance on the training set arbitrarily high.

(b) ... we can use bagging to increase the performance on the training set arbitrarily high.

(c) ... adaptive basis function models can come within a factor of 2 of the best possible performance.

(d) none of the other statements are true.

25) Which of the following statements about reinforcement learning is *incorrect?*

(a) Reinforcement learning methods typically take a long time to reach expert-level performance.

(b) The agent's policy uniquely determines the agents behaviour and translates observed states into actions to be taken.

(c) Since our data is i.i.d., we can predict future rewards using past states.

(d) Reinforcement learning can implicitly find hidden structures in collections of unlabeled data.

26) Which of the following statements is **incorrect**?

(a) If the value of the cost function during optimization oscillates or increases, one should decrease the learning rate.

(b) MSE (Mean-Squared-Error) cost function is a convex, quadratic function and it has a unique minimum.

(c) Gradient Descent may get stuck in local minima, but if we run it for enough iterations, it will always find the global minima.

(d) Logistic (sigmoid) function, used in logistic regression, is differentiable, hence we can apply Gradient Descent.

27) Which of the following is correct for transformations with Fisher LDA?

(a) Within-class covariances are minimized.

(b) None of them are correct.

(c) Increases class separability by decorrelating the features.

(d) Distance between class means is minimized.

28) Which of the following statements refers to a property of support vector machines?

(a) The distance between a training example and the margin is $\frac{2}{\|w_0\|}$.

(b) Since the support vectors have to be stored to classify new inputs after training, one can argue that SVMs have properties of a parametric and a non-parametric machine learning model.

(c) The separation hyperplane is influenced by the scaling of $\|w_0\|$ or b_0.

(d) The parameters $\alpha_i, \ldots, \alpha_N$ in our support vector machine cost function $J(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i \left(y^{(i)}(w^T x^{(i)} + b) - 1\right)$ control the amount of regularization.

29) Which of the statements contains only unsupervised learning methods?

(a) Gaussian mixture model, PCA and k-means clustering.

(b) Neural networks, k-nearest neighbour and Gaussian mixture model.

(c) Neural networks, PCA and SVM.

(d) k-nearest neighbour, k-means clustering and SVM.

30) Expectation-Maximization (EM) algorithm

(a) has covariance matrices of mixture components that are shared by all components. This means, if used for clustering, the clusters are of equal size and of same shape.

(b) is an iterative algorithm, but a closed-form solution for assigning points to clusters can be derived analytically.

(c) is a probabilistic approach to finding clusters in data.

(d) all answers are correct.

Name: ..... Matr.Nr.: .....

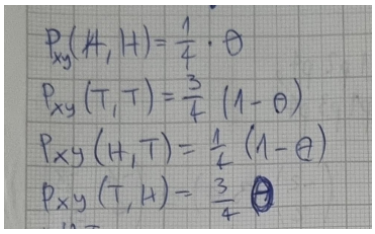
Note: Do not write outside of the boxes!

1. You are simultaneously tossing two (unfair) coins. Each trial therefore consists of two tosses. You know the marginal probabilities of tossing heads (H) or tail (T) for each of the two coins X and Y:

   $P_X(H) = \frac{1}{4}$
   $P_X(T) = \frac{3}{4}$
   $P_Y(H) = \theta$
   $P_Y(T) = 1 - \theta$,

   where θ is the parameter of the marginal distribution $P_Y(y)$.

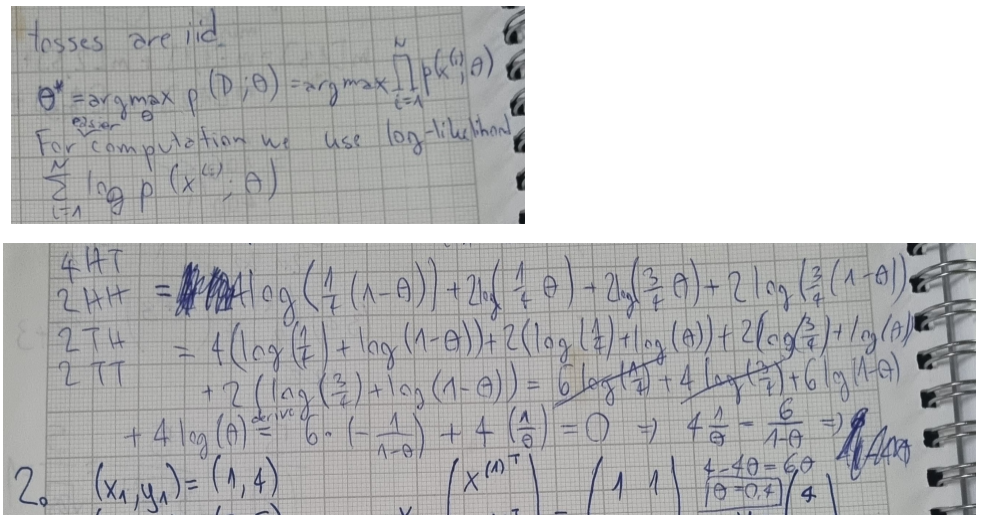
   Calculate the joint probability $P_{XY}(x, y)$ for each possible outcome, assuming that the tosses of each of the coins are independent. [3 points]



You perform 10 trials, and observe these samples from such a joint distribution:

$XY = \{HT, HH, HT, TH, HT, TT, TT, HH, HT, TH\}$.

Given these data, calculate the maximum likelihood estimate of parameter θ, assuming independence of the joint tosses. [7 points]



Name: ..... Matr.Nr.: .....

Note: Do not write outside of the boxes!

2. Consider a linear regression problem. For simplicity, given are only two points, and we want to fit a line through them. The given points are: $(x_1, y_1) = (1, 4)$, and $(x_2, y_2) = (2, 5)$. We want to find $y = f(x)$.

   First, specify vectors x and y for these two points. [1 point]

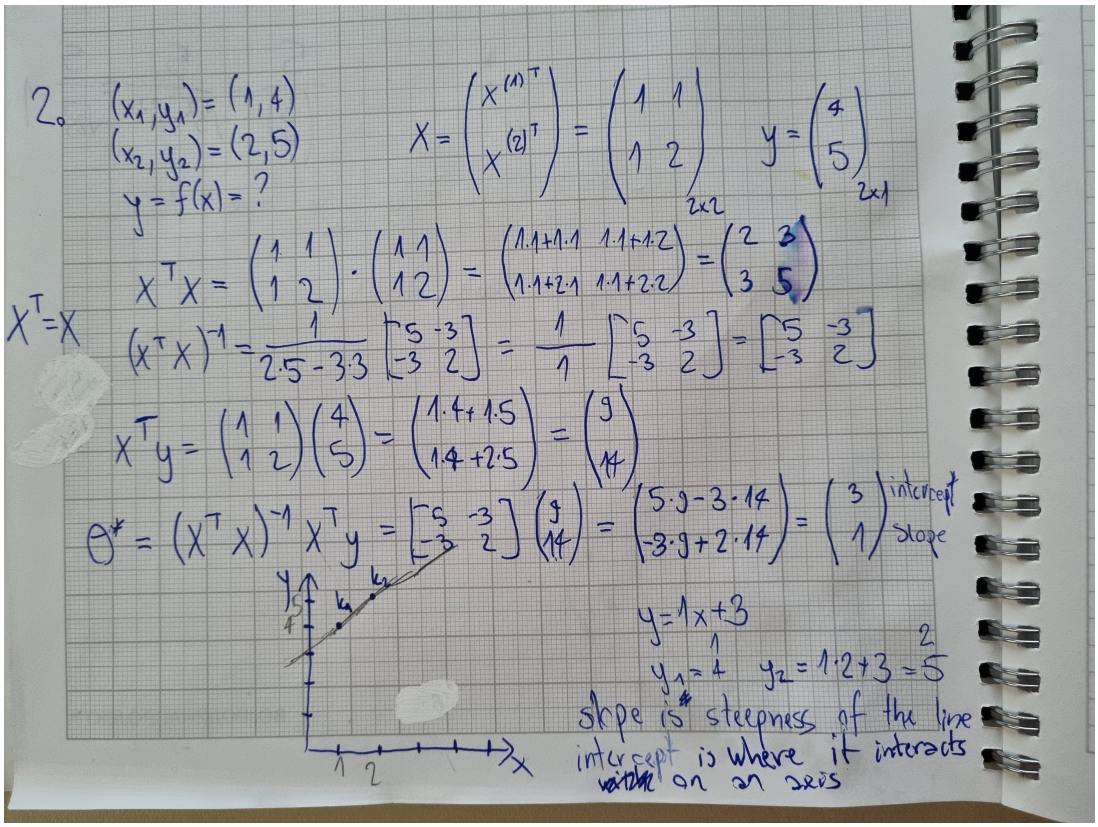
   Form an appropriate design matrix X that contains zero feature, and the feature x. (The dimension of this matrix must be 2x2.) [1 point]

   Write down the expression that represents the analytical solution for linear regression problem to find optimal parameters Θ^*, and calculate Θ^* using that expression. [6 points]

   For calculating the inverse of a 2x2 matrix use the following formula:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

   Finally, sketch two points in a coordinate system, and draw a line with the coefficients that you calculated (parameters Θ^*). What is the slope of this line? What is the intercept? [2 points]



Name: ..... Matr.Nr.: .....

Note: Do not write outside of the boxes!

3. Consider that you are given data $\{x_1, x_2, \ldots, x_N\}$ where $x_n \in \mathbb{R}^D$. Derive the PCA solution to transform your D-dimensional data into one dimension (i.e., using one principal component) using the method of Lagrange multipliers on the constrained optimization objective [6 points].

   Clearly define your variables, state the optimization objective and note which statistical properties of linear transformations are used (if any). Mark your final answers for the following: (a) How can we calculate the projection vector? [2 points] (b) What is the variance of the transformed data? [2 points]

Using the eigen decomposition we can write the PCA problem as

$$\max_{\boldsymbol{b}^T\boldsymbol{b}=1} \boldsymbol{b}^T C\, \boldsymbol{b} = \underbrace{\boldsymbol{b}^T V}_{\boldsymbol{a}^T}\; E\; \underbrace{V^T \boldsymbol{b}}_{\boldsymbol{a}} = \sum_{i=1}^{D} \lambda_i a_i^2$$

- $a_i = (\boldsymbol{v}_i^T \boldsymbol{b})$
- Evidently, $a_i^2 \geq 0$
- Since $\boldsymbol{b}$ is a unit vector and V is orthonormal, also $\boldsymbol{a}$ is a unit vector, i.e. $\sum_{i=1}^{D} a_i^2 = 1$
- Thus, the values a_i^2 are non-negative and sum to one!
- The variance is

$$\lambda_1 a_1^2 + \lambda_2 a_2^2 + \cdots + \lambda_D a_D^2$$

- The variance becomes maximal when $a_1^2 = 1$:

$$\lambda_1 1 + \lambda_2 0 + \cdots + \lambda_D 0$$

- Since $a_i = (\boldsymbol{v}_i^T \boldsymbol{b})$, this exactly happens if $\boldsymbol{b} = \boldsymbol{v}_1$!
- **The first eigen vector $\boldsymbol{v}_1$ of the covariance matrix is the first principal direction $\boldsymbol{b}_1$**
- The corresponding eigen value λ_1 is the variance which is "captured" in this direction (**explained variance**)

Name: ..... Matr.Nr.: .....

Note: Do not write outside of the boxes!

4. Consider adaptive basis function models. What's the central idea of adaptive basis function models and how does this relate to decision trees? [2 points]

Describe the process of growing a decision tree in detail. What do the nodes in the tree represent and how do we compute their successors / children? [4 points]

At first we split until there are no worthwhile nodes to be split and afterwards we compute the costs and could rearrange the tree if necessary.
Nodes represent the possible outcome of each action.
Based on the certain probability represented we make a decision of the outcomes for their children.

How do we know when we have to stop growing the tree? [2 points]

If there are no more worthwhile nodes to be split we can stop the algorithm.

```
split(L):
if stopping criterion is true for L then
    Learn p(Y) from class proportions of samples assigned to L
    return
```

Briefly describe how we can deal with the disadvantages of decision trees. [2 points]

Since they are very interpretable but don't perform too well in practice we can build random forests by bundling/ensambling decision trees together.