Mock Exam 2021 - Fall

January 2022

1 Multiple Choice

1.1 kNN

1.1.1 Question 1

Which of the following statement(s) is(are) true for k-NN?

- ☐ k represents the number of classes.
- ☒ The final outcome of the algorithm may change with the distance measure.
- ☒ CrossValidation can be used to find the optimal k.
- ☐ As k increases, we overfit the data eventually.

1.1.2 Question 2

(MCQ) Select true statements for the k-NN algorithm:

- ☐ It only works for two and three dimensions.
- ☐ It only works on data that is linearly-separable.
- ☐ It works best if k is close to the number of dimensions.
- ☐ It only works with Euclidean and Manhattan distance metrics.
- ☒ The decision boundary will be smoother if you increase k.

1.1.3 Question 3

(SCQ) Which class labels would 1-NN, 3-NN and 5-NN classifiers respectively predict for the data point $(3, 4)$, given the training data in the table below. Use the Euclidean distance to compute distances between data points.

X_1	X_2	Y
1	3	+
3	2	+
4	4	−
6	3	−
7	4	−
4	-1	+

☒ $(-, +, -)$

☐ $(+, +, -)$

☐ $(-, -, -)$

☐ $(-, -, +)$

☐ $(+, -, +)$

1.2 K-Means

1.2.1 Question 1

(MCQ) For which of the following tasks might K-means clustering be a suitable algorithm?

☒ From the users' usage patterns on a website, find out what different groups of users exists.

☐ Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

☐ Given historical weather records predict the mean speed of the wind for the following day.

☒ Given many emails, with a few annotated as `spam` and `non-spam`, you would like to classify every email as either of these two classes.

☒ Given a Spotify database of songs, where each song is represented by D features, find K distinct genres.

1.2.2 Question 2

(MCQ) K-Means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which of the following are the repeating steps?

☐ It defines the number of clusters K.

☒ It assigns each point to the nearest center.

☒ It updates each cluster centers given the points assigned to it.

☐ It defines the distance function.

☐ It tests on the cross-validation set.

1.2.3 Question 3

(MCQ) Alex plans to use the K-means algorithm to cluster the data for 200 people into males and females. The attributes include height (meters), weight (kilograms), 100 meter sprint test record (seconds), high jump test record (meters) and long jump test record (meters). Which of the following processes are well suited to this dataset?

- ☒ Scale each attribute appropriately
- ☐ Use the Euclidean distance with the raw data
- ☐ Set number of clusters to 100 in order to reduce the average within-cluster sum of squares
- ☐ Use the L_1 distance with the raw data

1.3 Logistic Regression

1.3.1 Question 1

(SCQ) Consider the following data point $\mathbf{x} = [1.0, 0.3, 0.5]^T$ and the model weights $\mathbf{w} = [0.2, 1.0, -3]^T$. Using logistic regression, which class would this data point be assigned to? ($e \approx 2.72$)

- ☒ 0
- ☐ 1

1.3.2 Question 2

(MCQ) Which of the following problems is appropriate to be solved using logistic regression?

- ☐ You are given a database of the salaries of employees in the company you work for. The data has the following features: the number of hours the employee works, the number of projects they have completed, and their age. Predict your salary for next year.
- ☒ You are given access to a medical database, containing the data of 10000 patients. This data includes what disease the patient has and their symptoms. You want to predict the disease of a new patient given their symptoms.
- ☒ You are given a database of aerial images of roads taken using an aircraft. You are also given the annotation images, where the pixels corresponding to roads are marked with 1 and the rest of the pixels are marked with 0. In a new aerial image, you want to detect where the road is.
- ☐ You are given a database of the history of a university course. The data has the following features: the year (Y), the number of students that passed the midterm (M), the number of students that attended the course regularly (C), the number of students that attended the exercise sessions regularly (E), and the number of students that passed the course (P). For this year, knowing M, C and E, predict P.

1.3.3 Question 3

(MCQ) Which of the following statements are true for logistic regression?

- ☐ Logistic regression is a regression method
- ☐ In logistic regression, we pass the result of the linear model $\mathbf{w}^T\mathbf{x}$ through a step function, which gives us a discrete output
- ☒ The prediction found as the output of the sigmoid function or the softmax function corresponds to the probability of the class assignment
- ☒ Logistic regression is robust to outliers, as opposed to classification performed using linear regression
- ☐ In logistic regression, it is impossible to use regularization

1.4 SVM

1.4.1 Question 1

Let $\mathbf{x} \in \mathbb{R}^2$ be a 2D data sample such that $\mathbf{x} = [x_1, x_2]^\top$. Consider the following three feature expansion functions:

(a) $\phi_1(\mathbf{x}) = [1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2]$,
(b) $\phi_2(\mathbf{x}) = [x_1, x_2]$,
(c) $\phi_3(\mathbf{x}) = [1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2, x_1^2x_2^2$.

For each ϕ decide whether it corresponds to a linear kernel, polynomial kernel or RBF kernel.

- ☐ (a) linear, (b) polynomial, (c) linear
- ☒ (a) polynomial, (b) linear, (c) polynomial
- ☐ (a) polynomial, (b) linear, (c) RBF
- ☐ (a) RBF, (b) linear, (c) RBF
- ☐ (a) linear, (b) polynomial, (c) RBF
- ☐ (a) RBF, (b) linear, (c) polynomial

1.4.2 Question 2

Recall that the SVM problem can be written as

$$\min_{\mathbf{w},\{\xi_i\}} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{subject to} \quad y_i \cdot (\mathbf{w}^T\mathbf{x_i}) \geq 1 - \xi_i, \forall i \text{ and } \xi_i \geq 0,$$

where $\mathbf{w}$ are trainable parameters (including the bias term) and $\mathbf{x_i}$ is the i-th data sample. Which of the following statements regarding this formulation are true?

- ☒ If $0 < \xi_i \leq 1$, sample i lies inside the margin.
- ☒ If $0 < \xi_i \leq 1$, sample i is correctly classified.
- ☐ If $\xi_i \geq 1$, we cannot tell whether the sample i was correctly classified.
- ☐ Setting C to 0 would prevent the SVM from misclassifying any sample.
- ☒ The bigger the C, the less misclassifications will the SVM make.
- ☒ The formulation above results in a strictly linear decision boundary in the original space.

1.4.3 Question 3

An SVM is trained with a linear kernel to do hard margin classification, i.e. the slack variables are not used, no misclassifications are allowed and the formulation is simply

$$\min_{\mathbf{w},\{\xi_i\}} \quad \frac{1}{2}||\mathbf{w}||^2$$
$$\text{subject to} \quad y_i \cdot (\mathbf{w}^T\mathbf{x_i}) \geq 1, \forall i.$$

We know the support vectors and the margins after the training. Now, we introduce a new data point and retrain. Select the correct statements:

- ☐ The margin will never change if the new data point is on the correct side of the original decision boundary and inside the margin.
- ☒ The margin will not change if the new data point is on the correct side of the original decision boundary and outside of the margin.
- ☒ The margin will change if the new data point is in between the original margin regardless of its positioning with respect to the decision boundary.
- ☐ The margins will not change if the new data point is within the original margin and on the correct side of the original decision boundary.
- ☐ The number of support vectors will not change regardless of the position of the new data point.

1.5 Neural Networks

1.5.1 Question 1

You want to estimate, given the day of the week (expressed as a one-hot vector) and the time of the day (expressed as a scalar), the chance of being checked by controllers (expressed as a scalar) on your way to EPFL on M1. To do so, you decide to use a MLP with two hidden layers, each of size 10, and no bias. How many trainable parameters are there in your network?

- ☒ 190
- ☐ 221
- ☐ 130
- ☐ 90

1.5.2 Question 2

You have trained a very deep MLP on your favorite classification task, and you observe severe overfitting. What can you do to solve this issue?

- ☐ Reduce the number of hidden layers
- ☐ Reduce the size of each hidden layer
- ☐ Regularize training through weight decay
- ☒ All of the above

1.5.3 Question 3

When training a neural network, we typically use mini-batching, i.e. for each iteration, we randomly select a subset of training samples to compute the gradients of the loss function with respect to parameters. Why do we do so?

- ☐ To reduce the number of iterations required to reach convergence.
- ☐ To improve robustness to outliers.
- ☒ To reduce the chance of falling into a local minima.
- ☐ All of the above

1.5.4 Question 4

We train a CNN architecture to regress a 3D human pose from an input RGB image. Our current architecture consists of 6 convolutional layers, each followed by a max pooling layer, and 3 fully connected layers. After our network converges, we inspect the training and validation errors and realize that the network severely overfits on the training dataset. What are our options to make the network generalize better?

- ☐ Stack more convolutional layers, the architecture is not deep enough.
- ☐ Stack more fully connected layers, the architecture is not deep enough.
- ☐ Remove the max pooling layers.
- ☒ Reduce the number of neurons in the first two fully connected layers.
- ☒ Augment our current loss function L by an L2 weights regularizer to get a new loss function $L' = L + \frac{1}{M}\sum_{i=1}^{M} w_i^2$, where $\{w_i | 1 \leq i \leq M\}$ is the set of all model's weights.
- ☒ Increase the size of our training dataset.
- ☐ Increase the size of our validation dataset.

2 Open Ended Questions

2.1 K-Means

During your exotic vacation, you took a beautiful photograph of a jungle and you would like to proudly send it to your mom. Her old phone can only read GIFs and upon exporting the photograph to this format you notice that the image quality rapidly degraded. The problem is that the GIF format uses an internal fixed palette of 256 colors. The default color palette is preset so as to cover the full color spectrum but your photograph mostly consists of shades of green. Therefore, you decide to compute your own palette of 256 colors, which would better suite your photograph. Let us represent the photograph as $I \in [0, 255]^{512\times512\times3}$. Being an aspiring data scientist, you decide to use k-means clustering to tackle this problem. In the questions below you will describe how you would use k-means clustering to retrieve 256 colors from your own photograph.

1. What is the dimension of a data sample and how many data samples do we have? How many clusters are we searching for?
2. Give a high-level overview of the k-means clustering algorithm with respect to the problem at hand.
3. One option to initialize the clusters is to use exactly the colors from the default palette. Why is this not a good idea and what is likely to happen if you do that?

4. A better option is to initialize the clusters by randomly picking colors from your own photograph. While this approach works, you notice that the small red bird, which contains the only red-ish pixels in your image, now appears green as well. Can you think of a better cluster initialization strategy which would be more likely to recover even the under-represented colors in your photograph?

5. Which of the standard objective functions would you use to measure the discrepancy between the original photograph and the produced GIF? Write down the formula and describe the variables you use.

6. You would like to send the photograph to your grandma as well, but she has an even older phone with limited memory and thus you would like to reduce the number of colors in the palette as much as possible so as to decrease the GIF file size while still maintaining a reasonable image quality. What method would you use to find a good trade-off between the number of colors and the image quality? Explain how this method works.

2.2 K-Means - Solutions

1. Each sample is three dimensional, i.e., $p_i \in [0, 255]^3$, and there are $P = 512 * 512$ samples in total. We search for 256 clusters.

2. 1 - Initialize the 256 cluster centers, e.g., by randomly choosing $\{\mu_i \in I : 1 < i < 256\}$. 2 - Iterate between the following two steps. Step 1 - assign each p_i to the closest μ_i. Step 2 - Recompute each μ_i as the centroid of its assigned samples.

3. Our photograph mostly contains shades of green whereas the color palette contains very different colors. Therefore, only a few of the palette colors would be close to our shades of green and upon convergence of the algorithm, many clusters will be empty. The image quality will thus be bad as the selected colors will not represent our image well.

4. Set μ_1 to a randomly-selected color from the photograph. Set μ_2 to the color from the photograph which is the most distant from μ_1 in the L2 sense. Set μ_3 to the color from the photograph which is the most distant from every color in $\{\mu_1, \mu_2\}$. Continue until all 256 cluster centers are initialized.

5. Mean squared error. $E = \frac{1}{512^2} \sum_{i=1}^{512^2} \|\mathbf{p_i} - \hat{\mathbf{p}}_\mathbf{i}\|^2$, where $\mathbf{p_i}, \hat{\mathbf{p}}_\mathbf{i}$ are the pixels of the original photograph and of the GIF, respectively.

6. We can use the elbow method. We plot either the MSE or the average within-cluster sum of squares as a function of number of clusters. We then select the number of clusters corresponding to the "elbow" in the generated plot.

3 MLPs

Given an input image $I \in [0, 255]^{128\times128\times3}$, you would like to find a corresponding segmentation mask $S = S(I) \in [0, 1]^{128\times128\times1}$, where each pixel $S_{i,j}$ of the output segmentation mask is one if the image pixel belongs to a cat, and zero otherwise. Since you are a novice at Machine Learning, you decide to use a Multi Layer Perceptron (MLP) with **one hidden layer of size 100** and **no bias** as machine learning model for this task.

1. How many parameters are there in your neural network architecture?
2. You are given a rather small set of image-mask pairs as training data and, consequently, your model is overfitting. How can you address this without gathering new data?
3. A friend of yours has collected a large-scale dataset to deal with the above-mentioned overfitting issue. Unfortunately, however, the simple architecture you have designed is now severely underfitting. Since discarding data is not a viable solution, how would you address this issue?

3.1 MLPs-Solution

1. The first layer maps an image of size $128 \times 128 \times 3$ to a feature vector of size 100, while the output one maps a feature vector of size 100 to a binary image of size 128×128. Consequently, we have $128 \times 128 \times 3 \times 100 + 100 \times 128 \times 128 = 6'553'600$ parameters.
2. To avoid overfitting, one might consider simply reducing the number of parameters in the network, or use one of the techniques presented during the course (early stopping, weight decay, dropout).
3. Increase the number of parameters, or consider changing the architecture altogether to a more powerful one (e.g., CNN).

4 Dimensionality Reduction

You have a large non-annotated dataset, consisting of $N = 100000$ samples with $D = 512$ features, on which you would like to run a clustering algorithm. Since you fear the curse of dimensionality, you decide to first reduce the number of features to $D' = 32$. Your favourite machine learning library contains the implementation of PCA, LDA, kernel PCA and (deep) autoencoders (AE), and you are pondering which one to choose.

1. Your first candidates are PCA and LDA. Which one would you use for your dataset and why?
2. If you use PCA to reduce the dimension from D to D', what is the number of values you have to store to be able to later reconstruct your D dimensional data? Assuming that you have no prior information about your dataset, is it guaranteed that you will be able to reconstruct your original data without any loss of information? If not, how many values would you have to store to guarantee a lossless reconstruction?

3. Now you are choosing between PCA and a very simple AE with just a single hidden layer of 32 neurons and with linear activation functions on both the hidden and output layer. If you want to store these models, would you have to use the same amount of memory?

4. What hyperparameters does each of the four aforementioned methods (PCA, LDA, kernel-PCA, deep AE) have? If a method has many, list three of them. Do not consider the target dimension as a hyperparameter.

4.1 Dimensionality Reduction - solution

- Since the dataset does not contain any labels, we cannot try to separate classes and must use PCA.

- We have to store the matrix $W \in \mathbb{R}^{512\times 32}$ and the dataset mean $\mathbf{x} \in \mathbb{R}^{512}$, so we have to store $512 \cdot 32 + 512$ values. In general, we cannot expect to be able to reconstruct the original data without loss of information. For that, we would have to store $W \in \mathbb{R}^{512\times 512}$, and thus we would have $512^2 + 512$ values in total.

- The simple AE will contain two distinct weight matrices $W_1 \in \mathbb{R}^{32\times 512}, W_2 \in \mathbb{R}^{512\times 32}$, where in general $W_1 \neq W_2^\top$, and thus it needs more memory for storing both matrices (and biases).

- PCA and LDA have no hyperparameters. Kernel-PCA involves choosing the kernel function and setting the hyperparameters related to the kernel function, e.g., the maximum degree in case of a polynomial kernel. A deep AE has multiple hyperparameters, such as the number of layers, the number of units in each layer, the choice of the activation function, the choice of regularizer, the regularization weight (if any), the learning rate, the number of iterations, etc.