EPFL

Profs. Nicolas Flammarion and Martin Jaggi
Machine Learning – CS-433 - IC
20.01.2022 from 08h15 to 11h15 in STCC
Duration : 180 minutes

1

Student One

SCIPER : **111111**

Do not turn the page before the start of the exam. This document is double-sided, has 20 pages, the last ones are possibly blank. Do not unstaple.

- This is a closed book exam. No electronic devices of any kind.
- Place on your desk: your student ID, writing utensils, one double-sided A4 page cheat sheet (hand-written or 11pt min font size) if you have one; place all other personal items below your desk.
- You each have a different exam.
- Only answers in this booklet count. No extra loose answer sheets. You can use the last two pages as scrap paper.
- For the **multiple choice** questions, we give :
  - $+2$ points if your answer is correct,
  - 0 points if you give no answer or more than one,
  - -0.5 points if your answer is incorrect.
- For the **true/false** questions, we give :
  - $+1$ points if your answer is correct,
  - 0 points if you give no answer or more than one,
  - -1 points if your answer is incorrect.
- Use a **black or dark blue ballpen** and clearly erase with **correction fluid** if necessary.
- If a question turns out to be wrong or ambiguous, we may decide to nullify it.

Respectez les consignes suivantes | Observe this guidelines | Beachten Sie bitte die unten stehenden Richtlinien

choisir une réponse | select an answer Antwort auswählen	ne PAS choisir une réponse | NOT select an answer NICHT Antwort auswählen	Corriger une réponse | Correct an answer Antwort korrigieren
☒ ☑ ■	☐	☐ ☐

ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte

First part: multiple choice questions

For each question, mark the box corresponding to the correct answer. Each question has **exactly one** correct answer.

Robustness to outliers

We consider a classification problem on linearly separable data. Our dataset had an outlier—a point that is very far from the other datapoints in distance (and also far from margins in SVM but still correctly classified by the SVM classifier).

We trained the SVM, logistic regression and 1-nearest-neighbour models on this dataset. We tested trained models on a test set that comes from the same distribution as training set, but doesn't have any outlier points. After that we removed the outlier and retrained our models.

Question 1 After retraining, which classifier will **change** its decision boundary around the test points.

- ☐ Logistic regression
- ☐ SVM
- ☐ 1-nearest-neighbors classifier
- ☐ All of them

SVM

Question 2 For any vector $\mathbf{v} \in \mathbb{R}^D$ let $\|\mathbf{v}\|_2 := \sqrt{v_1^2 + \cdots + v_D^2}$ denote the Euclidean norm. The hard-margin SVM problem for linearly separable points in $\mathbb{R}^D$ is to minimize the Euclidean norm $\|\mathbf{w}\|_2$ under some constraints. What are the additional constraints for this optimization problem?

- ☐ $\mathbf{w}^\top \mathbf{x}_n \geq 1 \ \forall n \in \{1, \cdots, N\}$
- ☐ $\frac{y_n}{\mathbf{w}^\top \mathbf{x}_n} \geq 1 \ \forall n \in \{1, \cdots, N\}$
- ☐ $y_n \mathbf{w}^\top \mathbf{x}_n \geq 1 \ \forall n \in \{1, \cdots, N\}$
- ☐ $y_n + \mathbf{w}^\top \mathbf{x}_n \geq 1 \ \forall n \in \{1, \cdots, N\}$

Cross Validation

Question 3 Consider the K-fold cross validation on a linear regression model with a sufficiently large amount of training data. When K is large, the computational complexity of the K-fold cross validation with respect to K is of order

- ☐ $\mathcal{O}(1/K)$.
- ☐ $\mathcal{O}(K)$.
- ☐ $\mathcal{O}(1)$.
- ☐ $\mathcal{O}(K(K-1))$.

Bias-Variance Decomposition

Question 4 Consider a regression model where data (x, y) is generated by input x uniformly randomly sampled from $[0, 1]$ and $y(x) = x^2 + \varepsilon$, where ε is random noise with mean 0 and variance 1. Two models are carried out for regression: model A is a trained quadratic function $g(x; \mathbf{w}) = w_2 x^2 + w_1 x + w_0$ where $\mathbf{w} = (w_0, w_1, w_2)^\top \in \mathbb{R}^3$, and model B is a constant function $h(x) = 1/2$. Then compared to model B, model A has

- ☐ higher bias, higher variance.
- ☐ lower bias, higher variance.
- ☐ higher bias, lower variance.
- ☐ lower bias, lower variance.

Optimization

Question 5 Let n be an integer such that $n \geq 2$ and let $A \in \mathbb{R}^{n \times n}$, and $\mathbf{x} \in \mathbb{R}^n$, consider the function $f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}$ defined over $\mathbb{R}^n$. Which of the following is the gradient of the function f?

- ☐ $A^\top \mathbf{x} + A\mathbf{x}$
- ☐ $2A^\top \mathbf{x}$
- ☐ $2\mathbf{x}^\top A$
- ☐ $2A\mathbf{x}$

Question 6 Which of the following functions reaches a global maximum on the set I? (Note that $[.,.]$ and $(.,.)$ denote closed and open intervals respectively)

- $f_1(x) = -x^4$, $I = [-5, 5]$
- $f_2(x) = \arccos(x)$, $I = (-1, 1)$
- $f_3(x) = x\exp(-x)$, $I = (-\infty, 0)$
- $f_4(x) = \sin(\cos(x))\sin(x)$, $I = \mathbb{R}_+$

- ☐ f_1, f_4
- ☐ f_1, f_3, f_4
- ☐ f_1, f_2, f_3, f_4
- ☐ f_1, f_2, f_4

Linear Models

Question 7 Consider a linear regression model on a dataset which we split into a training set and a test set. After training, our model gives a mean-squared error of 0.1 on the training set and a mean-squared error of 5.3 on the test set. Recall that the mean-squared error (MSE) is given by:

$$MSE_{\mathbf{w}}(\mathbf{y}, \mathbf{X}) = \frac{1}{2N} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\top \mathbf{w})^2$$

Which of the following statements is **correct** ?

- ☐ Ridge regression can help reduce the gap between the training MSE and the test MSE.
- ☐ Retraining the model with feature augmentation (e.g. adding polynomial features) will increase the training MSE.
- ☐ Retraining while discarding some training samples will likely reduce the gap between the train MSE and the test MSE.
- ☐ Using cross-validation can help decrease the training MSE of this very model.

Regularization

Question 8 Which of the following statements is **incorrect** ?

Training a model with L_1-regularization ...

- ☐ can reduce the storage cost of the final model.
- ☐ is used to help escaping local minima during training.
- ☐ can reduce overfitting.
- ☐ can be named Lasso regression when in combination with an MSE loss function and a linear model.

Neural networks

Let $f : \mathbb{R}^D \to \mathbb{R}$ be an L-hidden layer multi-layer perceptron (MLP) such that

$$f(\mathbf{x}) = \sigma_{L+1}\big(\mathbf{w}^\top \sigma_L(\mathbf{W}_L \sigma_{L-1}(\mathbf{W}_{L-1} \dots \sigma_1(\mathbf{W}_1 \mathbf{x})))\big),$$

with $\mathbf{w} \in \mathbb{R}^M$, $\mathbf{W}_1 \in \mathbb{R}^{M \times D}$ and $\mathbf{W}_\ell \in \mathbb{R}^{M \times M}$ for $\ell = 2, \dots, L$, and σ_i for $i = 1, \dots, L+1$ is an entry-wise activation function. For any MLP f and a classification threshold τ let $C_{f,\tau}$ be a binary classifier that outputs YES for a given input $\mathbf{x}$ if $f(\mathbf{x}) \leq \tau$ and NO otherwise.

Question 9 Assume σ_{L+1} is the element-wise **sigmoid** function and $C_{f,\frac{1}{2}}$ is able to obtain a high accuracy on a given binary classification task T. Let g be the MLP obtained by multiplying the parameters **in the last layer** of f, i.e. $\mathbf{w}$, by 2. Moreover, let h be the MLP obtained by replacing σ_{L+1} with element-wise **ReLU**. Finally, let q be the MLP obtained by doing both of these actions. Which of the following is true?

$$ReLU(x) = max\{x, 0\}$$

$$Sigmoid(x) = \frac{1}{1 + e^{-x}}$$

- ☐ $C_{h,0}$ may have an accuracy significantly lower than $C_{f,\frac{1}{2}}$ on T
- ☐ $C_{g,\frac{1}{2}}$, $C_{h,0}$, and $C_{q,0}$ have the same accuracy as $C_{f,\frac{1}{2}}$ on T
- ☐ $C_{g,\frac{1}{2}}$ may have an accuracy significantly lower than $C_{f,\frac{1}{2}}$ on T
- ☐ $C_{q,0}$ may have an accuracy significantly lower than $C_{f,\frac{1}{2}}$ on T

Question 10 Assume the weights in f were obtained by initializing **all parameters to zero** and running SGD. Assume there exists τ such that $C_{f,\tau}$ is a good classifier. Let g be the MLP obtained by randomly keeping one neuron per layer and removing the other $M-1$ neurons from each layer, and multiplying all weights except for the first layer by M (the number of neurons in each layer in the original network). What is the probability that $C_{g,\tau}$ is a good classifier?

- ☐ Less than $\frac{1}{M^L}$
- ☐ Between $\frac{1}{M^L}$ and $\frac{1}{M}$
- ☐ 1
- ☐ Not possible to determine with the given information.

Question 11 Which of the following techniques do *not* improve the generalization performance in deep learning?

- ☐ None. All techniques here improve generalization.
- ☐ Dropout
- ☐ Data augmentation
- ☐ Tuning the optimizer
- ☐ L2 regularization

Adversarial Examples

Consider a linear model $\hat{y} = \mathbf{x}^\top \mathbf{w}$ with the squared loss under an ℓ_∞-bounded adversarial perturbation. For a single point $(\mathbf{x}, y)$, it corresponds to the following objective:

$$\max_{\tilde{\mathbf{x}}:\ \|\mathbf{x}-\tilde{\mathbf{x}}\|_\infty \le \varepsilon} \left(y - \tilde{\mathbf{x}}^\top \mathbf{w}\right)^2, \tag{OP}$$

where $\|\mathbf{x}-\tilde{\mathbf{x}}\|_\infty \le \varepsilon$ denotes the ℓ_∞-norm, i.e. $|x_i - \tilde{x}_i| \le \varepsilon$ for every i.

Question 12 Assume that $\mathbf{w} = (3, -2)^\top$, $\mathbf{x} = (-1, 2)^\top$, $y = 2$. What is the maximum value of the optimization problem in Eq. (OP)?

- ☐ $(9 + 5\varepsilon)^2$
- ☐ $(3 + 10\varepsilon)^2$
- ☐ $(10 - \varepsilon)^2$
- ☐ Other
- ☐ $(5 + 9\varepsilon)^2$

Question 13 Assume that $\mathbf{w} = (3, -2)^\top$, $\mathbf{x} = (-1, 2)^\top$, $y = 2$. What is the optimal $\tilde{\mathbf{x}}^\star$ that maximizes the objective in Eq. (OP)?

- ☐ $(-1+\varepsilon, 2)^\top$
- ☐ $(-1-\varepsilon, 2-\varepsilon)^\top$
- ☐ $(-1-\varepsilon, 2)^\top$
- ☐ $(-1+\varepsilon, 2+\varepsilon)^\top$
- ☐ Other

KNN

Question 14 The KNN algorithm needs a notion of distance to assess which points are "nearest". Identify the distance measures that can be used in the KNN algorithm

(a) Euclidean Distance : distance associated to the L_2 norm $\|\mathbf{x}\|_2 := \sqrt{x_1^2 + \cdots + x_D^2}$
(b) Manhattan Distance : distance associated to the L_1 norm $\|\mathbf{x}\|_1 := |x_1| + \cdots + |x_D|$
(c) Distance associated to the L_4 norm $\|\mathbf{x}\|_4 := \left(|x_1|^4 + \cdots + |x_D|^4\right)^{1/4}$

- [ ] only a and c
- [ ] only b
- [ ] only c
- [ ] a, b and c
- [ ] only b and c
- [ ] only a and b
- [ ] only a

Linear Regression

Question 15 Assume we are doing linear regression with Mean-Squared Loss and L2-regularization on four one-dimensional data points. Our prediction model can be written as $f(x) = ax+b$ and the optimization problem can be written as

$$a^\star, b^\star = \underset{a,b}{\operatorname{argmin}} \sum_{n=1}^{4} [y_n - f(x_n)]^2 + \lambda a^2.$$

Assume that our data points are $Y = [1, 3, 2, 4]$ and $X = [-2, -1, 0, 3]$. For example $y_1 = 1$ and $x_1 = -2$. What is the optimal value for the bias, $b^\star$?

- [ ] 2
- [ ] None of the above answers.
- [ ] Depends on the value of λ
- [ ] 3
- [ ] 2.5

Subgradients

Question 16 Consider the Parametric ReLU function defined as

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ ax & \text{otherwise} \end{cases}$$

where $a \in \mathbb{R}$ is an arbitrary number. Which of the following statements is true regarding the subgradients of $f(x)$ at $x = 0$?

- [ ] If a subgradient exists, then it is not unique.
- [ ] A subgradient does not exist at $x = 0$.
- [ ] A subgradient exists even though $f(x)$ is not necessarily differentiable at $x = 0$.
- [ ] None of the mentioned answers.

Logistic regression

Consider a binary classification task as in Figure 1, which consists of 14 two-dimensional linearly separable samples (circles corresponds to label $y = 1$ and pluses corresponds to label $y = 0$). We would like to predict the label $y = 1$ of a sample (x_1, x_2) when the following holds true

$$\Pr(y = 1|x_1, x_2, w_1, w_2) = \frac{1}{1 + \exp(-w_1 x_1 - w_2 x_2)} > 0.5$$

where w_1 and w_2 are parameters of the model.

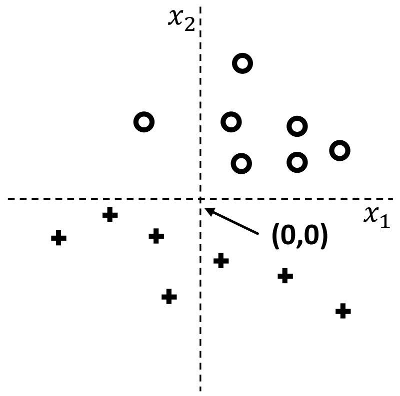

Figure 1: Two-dimensional dataset.

Question 17 If we obtain the (w_1, w_2) by optimizing the following objective

$$-\sum_{n=1}^{N} \log \Pr(y_n|x_{n1}, x_{n2}, w_1, w_2) + \frac{C}{2} w_2^2$$

where C is very large, then the decision boundary will be close to which of the following lines?

- ☐ $x_1 + x_2 = 0$
- ☐ $x_1 = 0$
- ☐ $x_2 = 0$
- ☐ $x_1 - x_2 = 0$

Recommender systems and word vectors

Question 18 What *alternates* in Alternating Least Squares for Matrix Factorization for a movie recommender system?

- ☐ recommendation steps and optimization steps
- ☐ updates based on different movie rating examples from the training set
- ☐ expectation steps and maximization steps
- ☐ updates to user embeddings and updates to movie embeddings

Question 19 Which NLP model architectures can differentiate between the sentences "I have to read this book." and "I have this book to read."?

(a) a convolutional model based on word2vec vectors
(b) a recurrent neural network based on GloVe word vectors
(c) a bag-of-words model based on GloVe word vectors

- ☐ only b and c
- ☐ only b
- ☐ only a and c
- ☐ a, b and c
- ☐ only c
- ☐ only a
- ☐ only a and b

Generative Networks

Question 20 Consider a Generative Adversarial Network (GAN) which successfully produces images of goats. Which of the following statements is **false**?

- ☐ After the training, the discriminator loss should ideally reach a constant value.
- ☐ The generator aims to learn the distribution of goat images.
- ☐ The generator can produce unseen images of goats.
- ☐ The discriminator can be used to classify images as goat vs non-goat.

Clustering

Question 21 For any vector $\mathbf{v} \in \mathbb{R}^D$ let $\|\mathbf{v}\|_2 := \sqrt{v_1^2 + \cdots + v_D^2}$ denote the Euclidean norm. Let $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ be a dataset of $N \geq 2$ distinct points. For any integer $K \geq 1$, consider the following value:

$$L_K^\star = \inf_{\substack{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K \in \mathbb{R}^D \\ z_{nk} \in \{0,1\} \text{ s.t. } \sum_{k=1}^K z_{nk} = 1}} \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K z_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2$$

The statements below are all true except one. Which of the following statement is *false*?

- ☐ For $2 \leq K < N$ and with the initial means randomly chosen as K data points, the K-means algorithm with K clusters is **not** guaranteed to reach the optimal $L_K^\star$ loss value.
- ☐ For $2 \leq K < N$, $L_K^\star$ is hard to compute.
- ☐ For $K \geq N$, $L_K^\star = 0$.
- ☐ The sequence $(L_K^\star)_{1 \leq K \leq N}$ is **not** necessarily strictly decreasing.
- ☐ $L_1^\star$ corresponds to the population variance of the dataset.

Question 22 You are given the data $(x_n, y_n)_{1 \leq n \leq N} \in \mathbb{R}^2$ illustrated in Figure 2 which you want to cluster into an inner ring and an outer ring (hence a number of clusters $K = 2$). Which of the following statement(s) is/are correct?

(a) There exists some initialization such that K-means clustering succeeds.

(b) There exists an appropriate feature expansion such that K-means (with standard initialization) succeeds.

(c) There exists an appropriate feature expansion such that the Expectation Maximization algorithm (with standard initialization) for a Gaussian Mixture Model succeeds.

- ☐ Only a and c
- ☐ Only a
- ☐ All of them
- ☐ Only b and c
- ☐ None of them
- ☐ Only a and b
- ☐ Only c
- ☐ Only b

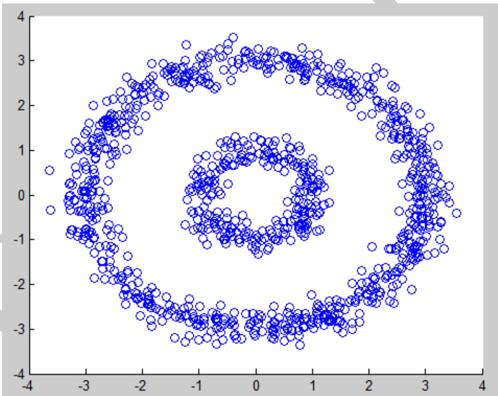

Figure 2: Two-dimensional dataset

Linear Algebra

Question 23 Given a matrix $\mathbf{X}$ of shape $D \times N$ with a singular value decomposition (SVD), $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, suppose $\mathbf{X}$ has rank K and $\mathbf{A} = \mathbf{X}\mathbf{X}^\top$

Which one of the following statements is **false**?

- ☐ The eigenvalues of $\mathbf{A}$ are the singular values of $\mathbf{X}$
- ☐ $\mathbf{A}$ is positive semi-definite, i.e all eigenvalues of $\mathbf{A}$ are non-negative
- ☐ The eigendecomposition of $\mathbf{A}$ is also a singular value decomposition (SVD) of $\mathbf{A}$
- ☐ A vector $\mathbf{x}$ that can be expressed as a linear combination of the last $D - K$ columns of $\mathbf{U}$, i.e $\mathbf{x} = \sum_{i=K+1}^{D} \mathbf{w}_i \mathbf{u}_i$ (where $\mathbf{u}_i$ is the i-th column of $\mathbf{U}$), lies in the null space of $\mathbf{X}^\top$

Second part: true/false questions

For each question, mark the box (without erasing) TRUE if the statement is **always true** and the box FALSE if it is **not always true** (i.e., it is sometimes false).

Question 24 (Neural Networks) Weight sharing allows CNNs to deal with image data without using too many parameters. However, weight sharing increases the variance of a model.

☐ TRUE ☐ FALSE

Question 25 (Kernels) For any vector $\mathbf{v} \in \mathbb{R}^D$ let $\|\mathbf{v}\|_2 := \sqrt{v_1^2 + \cdots + v_D^2}$ denote the Euclidean norm. Define the function $k(\mathbf{x}, \mathbf{x}') = \frac{1}{1-\mathbf{x}^\top \mathbf{x}'}$, on the set $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^D$ such that $\|\mathbf{x}\|_2 < 1$ and $\|\mathbf{x}'\|_2 < 1$. The function $k(\mathbf{x}, \mathbf{x}')$ is a valid kernel.

☐ TRUE ☐ FALSE

Question 26 (Bias/Variance Decomposition) Consider a linear regression model where the data is generated by input $\mathbf{x}$ and output $y = \mathbf{w}^\top \mathbf{x} + \varepsilon$, where $\mathbf{w}$ is a fixed vector and ε is a Gaussian noise with zero mean and σ^2 variance. Then there is no machine learning algorithm that can achieve a training error lower than σ^2.

☐ TRUE ☐ FALSE

Question 27 (Logistic regression) Consider a binary classification task $L(\mathbf{w}) = \sum_{n=1}^{N} -y_n \mathbf{x}_n^\top \mathbf{w} + \log(1 + e^{\mathbf{x}_n^\top \mathbf{w}})$ with linearly separable samples and $y \in \{0, 1\}$. Then there exists an optimal $\mathbf{w}$ with exact 0 loss and 100% training accuracy.

☐ TRUE ☐ FALSE

Question 28 (Linear Models) For any vector $\mathbf{v} \in \mathbb{R}^D$ let $\|\mathbf{v}\|_2 := \sqrt{v_1^2 + \cdots + v_D^2}$ denote the Euclidean norm. For $\mathbf{y} \in \mathbb{R}^D$, $\mathbf{X} \in \mathbb{R}^{D \times D}$, the solution of the least squares problem:

$$\mathbf{w}^\star = \underset{\mathbf{w} \in R^D}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

is always unique.

☐ TRUE ☐ FALSE

Question 29 (PCA) Your friend performed Principal Component Analysis on some data and claimed that he retained at least 95% of the variance using k principal components. This is equivalent to $\frac{\sum_{i \geq k+1} \lambda_i}{\sum_i \lambda_i} \leq 0.05$, where $\lambda_1, \ldots, \lambda_k, \ldots$ are the eigenvalues associated to each principal component, sorted in a *non-increasing* order.

☐ TRUE ☐ FALSE

Question 30 (Adversarial robustness) Let $\|\mathbf{v}\|_\infty := \max_i |v_i|$ denote the ℓ_∞-norm. Assume a binary classification problem with $y \in \{-1, 1\}$. Then the adversarial zero-one loss $\max_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_\infty \leq \varepsilon} \mathbb{1}_{yf(\mathbf{x}+\boldsymbol{\delta}) \leq 0}$ ($\mathbb{1}_C$ is equal to 1 if the condition C is true and to 0 if C is false) is always upper bounded by the adversarial hinge loss $\max_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_\infty \leq \varepsilon} \max\{0, 1 - yf(\mathbf{x} + \boldsymbol{\delta})\}$.

☐ TRUE ☐ FALSE

Question 31 (Expectation Maximization) If we run the Expectation Maximization (EM) algorithm on the log-likelihood for a Gaussian Mixture Model, then the sequence of log-likelihoods $\{\mathcal{L}(\boldsymbol{\theta}^{(t)})\}_{t\in\mathbb{N}}$ is guaranteed to be non-decreasing.

☐ TRUE ☐ FALSE

Question 32 (FastText Supervised Classifier) The FastText supervised classifier can be modelled as a two-layer linear (one hidden layer) neural network with a non-linear loss function at the output.

☐ TRUE ☐ FALSE

Question 33 (Matrix Factorization) Consider the matrix factorization objective function $\frac{1}{2}\sum_{(d,n)\in\Omega} \left[x_{dn} - (\mathbf{W}\mathbf{Z}^\top)_{dn}\right]^2$. When minimizing this objective with SGD over the embedding matrices $\mathbf{W}$ and $\mathbf{Z}$, you should initialize $\mathbf{W}$ and $\mathbf{Z}$ with zeros.

☐ TRUE ☐ FALSE

Question 34 (MSE and Neural Networks) The mean squared error (MSE) is convex w.r.t the parameters of a multi layer perceptron with more than one hidden layer and **linear** activation function.

☐ TRUE ☐ FALSE

Third part, open questions

Answer in the space provided! Your answer must be justified with all steps. Leave the check-boxes empty, they are used for the grading.

PCA

Let $\mathbf{x}_1, ..., \mathbf{x}_N$ be a dataset of N vectors in $\mathbb{R}^D$.

Question 35: (*1 point.*) What does it mean for the data vectors $\mathbf{x}_1, ..., \mathbf{x}_N$ to be centered, as for principle component analysis (PCA) to be meaningful? Use the notation x_{nd} for individual entries.

☐0 ☐1

Question 36: (*1 point.*) Write down the covariance matrix of the dataset $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N) \in \mathbb{R}^{D\times N}$, *and* state its dimensions. (Note that for PCA we assume data to be already centered, as in the previous question)

☐0 ☐1

Now let $\mathbf{x}$ be a random vector distributed according to the uniform distribution over the finite normalized dataset $\mathbf{x}_1, ..., \mathbf{x}_N$ from above. Consider the problem of finding a unit vector, $\mathbf{w} \in \mathbb{R}^D$, such that the random variable $\mathbf{w}^\top \mathbf{x}$ has *maximal* variance.

Question 37: (*1 point.*) What is the variance of the random variable $\mathbf{w}^\top \mathbf{x}$ over the randomness of $\mathbf{x}$?

☐0 ☐1

Question 38: (*2 points.*) Show that the solution of the problem of $\text{argmax}_{\mathbf{w}:\|\mathbf{w}\|=1} \text{Var}[\mathbf{w}^\top \mathbf{x}]$ is to set $\mathbf{w}$ to be the first principle vector of $\mathbf{x}_1, ..., \mathbf{x}_N$.

☐0 ☐1 ☐2

Question 39: (*1 point.*) Explain in words what the above result says about how PCA relates to the dataset.

$\square_0 \; \square_1$

The Perceptron

Setting

Let us consider a binary classification problem with a training set $S = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ such that:

$$\mathbf{x}_n \in \mathbb{R}^D, \text{ and } y_n \in \{-1, 1\}, \text{ for all } n = 1, \cdots, N,$$

where N, D are integers such that $N, D \geq 1$.
We consider the Perceptron classifier which classifies $\mathbf{x} \in \mathbb{R}^D$ following the rule:

$$f_{\mathbf{w},b}(\mathbf{x}) = \operatorname{sgn}(\mathbf{w}^\top \mathbf{x} + b),$$

where $\mathbf{w} \in \mathbb{R}^D$ is the weight vector, $b \in \mathbb{R}$ is the threshold, and the sign function is defined as

$$\operatorname{sgn}(z) = \begin{cases} +1 \text{ if } z \geq 0 \\ -1 \text{ if } z < 0 \end{cases}$$

Question 40: (*1 point.*) As seen in the course, explain how we can ignore the threshold b and only deal with classifiers passing through the origin, i.e., of the form $f_{\mathbf{w}}(\mathbf{x}) = \operatorname{sgn}(\mathbf{w}^\top \mathbf{x})$.

$\square_0 \; \square_1$

For the remainder of the exercise, we proceed without this additive threshold, as explained in the previous question, and only consider classifiers of the form $f_{\mathbf{w}}(\mathbf{x}) = \operatorname{sgn}(\mathbf{w}^\top \mathbf{x})$. We make the following two assumptions:

- **Bounded input:** There exists a real number $R \geq 0$ such that for all $n = 1, \cdots, N$ we have

$$\|\mathbf{x}_n\|_2 \leq R,$$

  where for any vector $\mathbf{z} \in \mathbb{R}^D$ we use $\|\mathbf{z}\|_2$ to refer to the Euclidean norm of $\mathbf{z}$, i.e, $\|\mathbf{z}\|_2 = \sqrt{\sum_{d=1}^D z_d^2}$.

- **Linearly separable data:** There exists $\mathbf{w}_\star \in \mathbb{R}^D$ and a real number $\gamma > 0$ such that for all $n = 1, \cdots, N$ we have

$$y_n \mathbf{w}_\star^\top \mathbf{x}_n \geq \gamma.$$

We will use the Perceptron algorithm to train the Perceptron classifier and find a separating hyperplane. Let t denote the number of parameter updates we have performed and $\mathbf{w}_t$ the weight vector after t updates. We initialize with $\mathbf{w}_0 = \mathbf{0}$. If this weight vector is already a separating hyperplane, we are done. If not, we pick an arbitrary point $\mathbf{x}_i$ that is currently misclassified. This point is used to update the weight vector $\mathbf{w}_t$ as

$$\mathbf{w}_{t+1} := \mathbf{w}_t + y_i \mathbf{x}_i \quad \text{where} \quad y_i \mathbf{x}_i^\top \mathbf{w}_t \leq 0.$$

The algorithm is formally written below.

Algorithm 1: Perceptron algorithm

$t \leftarrow 0$; $\mathbf{w}_t \leftarrow 0$;
while *there exists* $j \in \{1, \cdots, N\}$ *such that* $y_j \mathbf{x}_j^\top \mathbf{w}_t \leq 0$ **do**
  Pick an arbitrary $i \in \{1, \cdots, N\}$ such that $y_i \mathbf{x}_i^\top \mathbf{w}_t \leq 0$;
  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i \mathbf{x}_i$;
  $t \leftarrow t + 1$
end
return $\mathbf{w}_t$

We will study the convergence of this learning algorithm.

Convergence

Question 41: (*2 points.*) For $t > 0$, show that when making the t^{th} update according to the Perceptron update, we have:

$$\mathbf{w}_\star^\top \mathbf{w}_t \geq \mathbf{w}_\star^\top \mathbf{w}_{t-1} + \gamma$$

☐0 ☐1 ☐2

Question 42: (*1 point.*) Show that it implies that

$$\mathbf{w}_\star^\top \mathbf{w}_t \geq t\gamma$$

☐0 ☐1

Question 43: (*1 point.*) Using the previous question, derive a lower-bound on $\|\mathbf{w}_t\|_2$ depending on t, γ and $\|\mathbf{w}_\star\|_2$.
(We recall that a lower bound is a constant $C > 0$ such that $\|\mathbf{w}_t\|_2 \geq C$. Note that we will only accept non-trivial lower-bounds as correct answers.)

☐0 ☐1

Question 44: (*3 points.*) Derive the following upper bound on the squared norm $\|\mathbf{w}_t\|_2^2$:

$$\|\mathbf{w}_t\|_2^2 \leq \|\mathbf{w}_{t-1}\|_2^2 + R^2$$

☐0 ☐1 ☐2 ☐3

Question 45: (*1 point.*) Show that it implies that

$$\|\mathbf{w}_t\|_2^2 \leq tR^2$$

☐0 ☐1

Question 46: (*3 points.*) Combine the results obtained in the previous questions to obtain an upper bound on t depending on the quantities R, γ and $\mathbf{w}_\star$.

☐0 ☐1 ☐2 ☐3

Question 47: (*2 points.*) Qualitatively analyze the previous result. What have you proven? Interpret the dependency on R, γ and $\mathbf{w}_\star$.

☐0 ☐1 ☐2

Perceptron and margin

Let us remind that we define the max-margin $M_\star$ as

$$M_\star = \max_{\mathbf{w}\in\mathbb{R}^D, \|\mathbf{w}\|_2=1} M \text{ such that } y_n \mathbf{x}_n^\top \mathbf{w} \geq M \text{ for } n = 1, \cdots, N$$

and a max-margin separating hyperplane $\bar{\mathbf{w}}$ as a solution of this problem:

$$\bar{\mathbf{w}} \in \arg\max_{\mathbf{w}\in\mathbb{R}^D, \|\mathbf{w}\|_2=1} M \text{ such that } y_n \mathbf{x}_n^\top \mathbf{w} \geq M \text{ for } i = 1, \cdots, N$$

Question 48: (*2 points.*) Bound the number of perceptron updates t using the quantities R and $M_\star$. Prove your result.

☐0 ☐1 ☐2

Question 49: (*1 point.*) Does it imply that the output of the Perceptron algorithm is a max-margin separating hyperplane?

☐0 ☐1