



Profs. Nicolas Flammarion and Martin Jaggi
Machine Learning – CS-433 - IC
Wednesday 13.01.2021
from 16h15 to 19h15 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.

Linear Regression

Question 1 Under certain conditions, maximizing the log-likelihood is equivalent to minimizing mean-squared error for linear regression. The mean-squared error can be defined as $\mathcal{L}_{mse}(\mathbf{w}) := \frac{1}{2N} \sum_{n=1}^N (y_n - \tilde{\mathbf{x}}_n^\top \mathbf{w})^2$ and $y_n = \tilde{\mathbf{x}}_n^\top \mathbf{w} + \varepsilon_n$ is assumed for the probabilistic model. Which of following conditions is necessary for the equivalence?

- ☐ The noise parameter ε_n should have a normal distribution.
- ☐ The target variable y_n should have a normal distribution.
- ☐ The i.i.d. assumption on the variable \mathbf{w} .
- ☐ The conditional probability $p(y_n | \tilde{\mathbf{x}}_n, \mathbf{w})$ should follow a Laplacian distribution.
- ☐ The noise parameter ε_n should have non-zero mean.

SVMs versus Logistic Regression

Consider a classification problem on linearly separable data. We train an SVM model and a logistic regression model. For logistic regression (LR) we add a small regularization term (penalty on weights) in order to make the optimum well-defined. Each model gives us a margin. Consider a datapoint \mathbf{x}_0 that is correctly classified and strictly outside both margins.

Question 2 Which one of the following statements is **incorrect** ?

- ☐ There exists a direction in which we can slightly move \mathbf{x}_0 without changing the LR decision boundary after retraining.
- ☐ \mathbf{x}_0 isn't a support vector.
- ☐ There exists a direction in which we can arbitrarily move \mathbf{x}_0 without changing the SVM decision boundary after retraining.
- ☐ If we remove \mathbf{x}_0 from the dataset and retrain, this will change the LR decision boundary.
- ☐ If we remove \mathbf{x}_0 from the dataset and retrain, this will not change the SVM decision boundary.

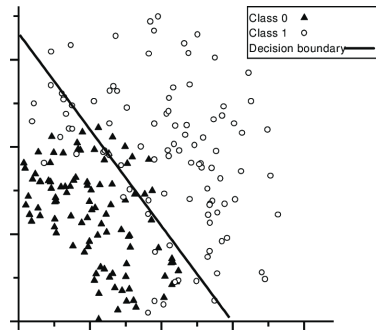
Logistic Regression Assumptions

Question 3 Binary logistic regression assumes a:

- ☐ Linear relationship between the input variables.
- ☐ Linear relationship between the observations.
- ☐ Linear relationship between the input variables and the logit (inverse of sigmoid) of the probability of the event that the outcome $Y = 1$.
- ☐ Linear relationship between the input variables and the probability of the event that the outcome $Y = 1$.



Decision boundary



Question 4 Which of these classifiers could have generated this decision boundary?

- ☐ 1-nearest-neighbor with L_2 distance & SVM.
- ☐ SVM & logistic regression.
- ☐ logistic regression & 1-nearest-neighbor with L_2 distance.
- ☐ None of the other options are correct.

Optimization Algorithm Complexity

Question 5 Consider a linear regression problem with N datapoints and D features. Finding the optimal parameters by running grid search while trying P values for each feature has approximately *the same* computational complexity as running:

- ☐ P iterations of Gradient Descent or NP iterations of Stochastic Gradient Descent.
- ☐ NP^D iterations of Gradient Descent or P^D iterations of Stochastic Gradient Descent.
- ☐ P^D iterations of Gradient Descent or NP^D iterations of Stochastic Gradient Descent.
- ☐ PD iterations of Gradient Descent or NPD iterations of Stochastic Gradient Descent.

Subgradients

Question 6 Consider the function $f(x) = -x^2$. Which of the following statements are true regarding subgradients of $f(x)$ at $x = 0$?

- ☐ A subgradient does not exist as $f(x)$ is differentiable at $x = 0$.
- ☐ A subgradient exists but is not unique.
- ☐ A subgradient exists and is unique.
- ☐ A subgradient does not exist even though $f(x)$ is differentiable at $x = 0$.



Train/Test Errors



Question 7 The above figure was produced by changing a hyperparameter in a classifier on a non-linearly separable training dataset. For which classifier could this picture be produced and how was its hyperparameter changed?

- ☐ Logistic regression, increasing regularization parameter λ .
- ☐ Logistic regression, decreasing regularization parameter λ .
- ☐ K-nearest neighbor classifier, decreasing number of neighbors k .
- ☐ K-nearest neighbor classifier, increasing number of neighbors k .

Exponential Families

Question 8 You are given two distributions over \mathbb{R} : Uniform on the interval $[a, b]$ and Gaussian with mean μ and variance σ^2 . Their respective probability density functions are

$$p_U(y|a, b) := \begin{cases} \frac{1}{b-a}, & \text{for } a \leq y \leq b, \\ 0 & \text{otherwise} \end{cases} \quad p_G(y|\mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Which one(s) belong to the exponential family?

- ☐ Only Uniform.
- ☐ Both of them.
- ☐ Only Gaussian.
- ☐ None of them.

Kernels

Let us assume that a kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is said to be *valid* if there exists $k \in \mathbb{N}$ and $\Phi : \mathcal{X} \rightarrow \mathbb{R}^k$ such that for all $(x, x') \in \mathcal{X} \times \mathcal{X}$, $K(x, x') = \Phi(x)^\top \Phi(x')$.

Question 9 Which one of the following kernels is **not** valid ?

- ☐ $\mathcal{X} = \mathbb{N}$, $K(x, x') = 2$.
- ☐ $\mathcal{X} = \mathbb{R}^d$, $K(x, x') = (x^\top x')^2$.
- ☐ $\mathcal{X} = \mathbb{R}$, $K(x, x') = \cos(x - x')$.
- ☐ All of the proposed kernels are in fact valid.
- ☐ $\mathcal{X} = \mathbb{Q}$, $K(x, x') = 2^{x+x'}$.
- ☐ $\mathcal{X} = \mathbb{R}^d$, $K(x, x') = x^\top A x'$, where A is a $d \times d$ symmetric positive semi-definite matrix.



Neural networks

Let $f_{\text{MLP}} : \mathbb{R}^d \rightarrow \mathbb{R}$ be an L -hidden layer multi-layer perceptron (MLP) such that

$$f_{\text{MLP}}(\mathbf{x}) = \mathbf{w}^\top \sigma(\mathbf{W}_L \sigma(\mathbf{W}_{L-1} \dots \sigma(\mathbf{W}_1 \mathbf{x}))),$$

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 is an entry-wise activation function.

Also, let $f_{\text{CNN}} : \mathbb{R}^d \rightarrow \mathbb{R}$ be an L' -hidden layer convolutional neural network (CNN) such that

$$f_{\text{CNN}}(\mathbf{x}) = \mathbf{w}^\top \sigma(\mathbf{w}_{L'} \star \sigma(\mathbf{w}_{L'-1} \star \dots \sigma(\mathbf{w}_1 \star \mathbf{x}))),$$

with $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{w}_\ell \in \mathbb{R}^K$ for $\ell = 1, \dots, L'$ and \star denoting the one-dimensional convolution operator with zero-padding, i.e., output of the convolution has the same dimensionality as the input.

Question 10 For each CNN neural network of the above form, there exists an MLP of the form f_{MLP} that can approximate f_{CNN} arbitrarily well, if

- ☐ $L = L'$ and $M = d$.
- ☐ $L > L'$ and $M > K$.
- ☐ another necessary condition different from these three.
- ☐ σ is a sigmoidal activation function.

Question 11 Let's assume σ is a tanh activation function. Thus, by flipping the signs of all of the weights leading in and out of a hidden neuron, the input-output mapping function represented by the network is unchanged. Besides, interchanging the values of all of the weights (i.e., by permuting the ordering of the hidden neurons within the layer) also leaves the network input-output mapping function unchanged. Suppose that, given the training data, SGD can find a solution with zero training loss, and the (absolute value) weights of such solution are non-zero and all unique. Choose the largest lower bound on the number of solutions (with zero training loss) achievable by f_{MLP} with $L=1$ and M hidden units on this dataset.

- ☐ $M!2^M$
- ☐ 1
- ☐ 2^M
- ☐ $M!$

Question 12 Regarding the weight updates in back-propagation,

- ☐ The output layer weights are not used for computing the error of the hidden layer.
- ☐ The weight changes are not proportional to the difference between the desired and actual outputs.
- ☐ A standard technique to initialize the weights is to set them exactly to 0.
- ☐ The weight change is also proportional to the input to the weight layer.



Minimum-Norm Adversarial Examples

Consider a binary classification problem with a linear classifier $f(\mathbf{x})$ given by

$$f(\mathbf{x}) = \begin{cases} 1, & \mathbf{w}^\top \mathbf{x} \geq 0, \\ -1, & \mathbf{w}^\top \mathbf{x} < 0, \end{cases}$$

where $\mathbf{x} \in \mathbb{R}^3$. Suppose that the weights of the linear model are equal to $\mathbf{w} = (4, 0, -3)$.

For the next two questions, we would like to find a *minimum-norm* adversarial example. Specifically, we are interested in solving the following optimization problem, for a given \mathbf{x} :

$$\min_{\boldsymbol{\delta} \in \mathbb{R}^3} \|\boldsymbol{\delta}\|_2 \quad \text{subject to} \quad \mathbf{w}^\top (\mathbf{x} + \boldsymbol{\delta}) = 0 \quad (\text{OP})$$

This leads to the point $\mathbf{x} + \boldsymbol{\delta}$ that lies exactly at the decision boundary and the perturbation $\boldsymbol{\delta}$ is the smallest in terms of the ℓ_2 -norm.

Question 13 What is the minimum value of the optimization problem Eq. (OP) for the point $\mathbf{x} = (-1, 3, 2)$?

- ☐ 2
- ☐ 3
- ☐ 1
- ☐ Other
- ☐ 1.5
- ☐ 0
- ☐ $\sqrt{2}$
- ☐ 4

Question 14 What is the optimum $\boldsymbol{\delta}^*$ that minimizes the objective in Eq. (OP) for the point $\mathbf{x} = (-1, 3, 2)$?

- ☐ $(1, -1, 0)$
- ☐ $(0, -1, 1)$
- ☐ $(-2, 0, 0)$
- ☐ $(1.2, 0, 1.6)$
- ☐ Other
- ☐ $(0, 2, 0)$
- ☐ $(-1.2, 0, 1.6)$



Question 15 Let $\mathcal{R}_p(f, \varepsilon)$ be the ℓ_p adversarial risk of a classifier $f : \mathbb{R}^d \rightarrow \{\pm 1\}$, i.e.,

$$\mathcal{R}_p(f, \varepsilon) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\max_{\tilde{\mathbf{x}} : \|\mathbf{x} - \tilde{\mathbf{x}}\|_p \leq \varepsilon} \mathbb{1}_{\{f(\tilde{\mathbf{x}}) \neq y\}} \right],$$

for $p = 1, 2, \infty$. Which of the following relationships between the adversarial risks is true?

- ☐ $\mathcal{R}_2(f, \varepsilon) \leq \mathcal{R}_1(f, 2\varepsilon)$
- ☐ $\mathcal{R}_\infty(f, \varepsilon) \leq \mathcal{R}_2(f, \sqrt{d}\varepsilon)$
- ☐ $\mathcal{R}_\infty(f, \varepsilon) \leq \mathcal{R}_1(f, \varepsilon)$
- ☐ $\mathcal{R}_\infty(f, \varepsilon) \leq \mathcal{R}_2(f, \varepsilon/d)$

Question 16 Identify the correct statement.

- ☐ None of the other options are correct.
- ☐ After training, and when the size of the vocabulary is large, a Skip-gram model would have higher space requirements than a GloVe model. We assume both models have the same number of dimensions (features), vocabulary, and are trained on the same corpus.
- ☐ Language models can be trained using either a multi-class (number of classes equal to the vocabulary size) classifier or a binary classifier to generate text.
- ☐ Language Models are useless for classification tasks in Natural Language Processing as they are only suited for text generation.

Question 17 Consider a matrix factorization problem of the form $\mathbf{X} = \mathbf{W}\mathbf{Z}^\top$ to obtain an item-user recommender system where x_{ij} denotes the rating given by j^{th} user to the i^{th} item. We use Root mean square error (RMSE) to gauge the quality of the factorization obtained. Select the correct option.

- ☐ Given a new item and a few ratings from existing users, we need to retrain the already trained recommender system from scratch to generate robust ratings for the user-item pairs containing this item.
- ☐ Regularization terms for \mathbf{W} and \mathbf{Z} in the form of their respective Frobenius norms are added to the RMSE so that the resulting objective function becomes convex.
- ☐ For obtaining a robust factorization of a matrix \mathbf{X} with D rows and N elements where $N \ll D$, the latent dimension K should lie somewhere between D and N .
- ☐ None of the other options are correct.

Cost functions

Question 18 Which statement is true for the Mean Squared Error (MSE) loss $\text{MSE}(\mathbf{x}, y) := (f_{\mathbf{w}}(\mathbf{x}) - y)^2$, with $f_{\mathbf{w}}$ a model parametrized by the weights \mathbf{w} ?

- ☐ MSE is not necessarily convex with respect to the weights of the model \mathbf{w} .
- ☐ MSE is more robust to outliers than Mean Absolute Error (MAE).
- ☐ For any ML task you are trying to solve, minimizing MSE will provably yield the best model.



Linear regression

Question 19 Which statement is true for linear regression?

- ☐ A linear regression model can be expressed as an inner product between feature vectors and a weight vector.
- ☐ Linear regression, when using 'usual' loss functions, works fine when the dataset contains many outliers.
- ☐ A good fit with linear regression implies a causal relationship between inputs and outputs.

Generative adversarial networks

Question 20 Every time you open the website `thispersondoesnotexist.com`, you see a fake picture of a person that was sampled from the distribution learned by a GAN. Which part of the GAN is deployed on this server?

- ☐ The discriminator.
- ☐ The generator.
- ☐ The discriminator and the generator.
- ☐ Only the last layer of the discriminator.

Principal component analysis

In principal component analysis, the left singular vectors \mathbf{U} of a data matrix \mathbf{X} of shape (d features, n datapoints) are used to create a new data matrix $\mathbf{X}' = \mathbf{U}^\top \mathbf{X}$.

Question 21 Which property always holds for the matrix \mathbf{X}' ?

- ☐ \mathbf{X}' is a square matrix.
- ☐ The mean of any row \mathbf{X}'_i is 0.
- ☐ \mathbf{X}' has only positive values.
- ☐ For any two rows i, j ($i \neq j$) from \mathbf{X}' , the dot product between the rows \mathbf{X}'_i and \mathbf{X}'_j is 0.

Question 22 To achieve dimensionality reduction, we keep only certain rows of the matrix \mathbf{X}' . We keep those rows that have:

- ☐ the lowest variance.
- ☐ the highest variance.
- ☐ smallest L2 norm.
- ☐ L2 norm closest to 1.



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 23 (Generalized Linear Models) Deep neural networks with logistic loss for binary classification are generalized linear models.

☐ TRUE ☐ FALSE

Question 24 (Weight initialization) The choice of weight initialization will not impact the optimization behavior of the neural network.

☐ TRUE ☐ FALSE

Question 25 (Support Vector Machines)

Reminder: The hard-margin problem for linearly separable points in \mathbb{R}^d is

$$\min_{\mathbf{w} \in \mathbb{R}^d \text{ s.t. } \forall i, y_i \mathbf{w}^\top \mathbf{x}_i \geq 1} \|\mathbf{w}\|.$$

Let $m > 1$ be a fixed number. Then there exists $\lambda > 0$ such that for every sample S of m examples which are linearly separable, the hard-SVM and the soft-SVM (with parameter λ) solutions will return exactly the same weight vector.

☐ TRUE ☐ FALSE

Question 26 (Linear Regression) You are given samples $\mathcal{S} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ where $\mathbf{x}_n \in \mathbb{R}^D$ and y_n are scalar values. You are solving linear regression using normal equations. You will always find the optimal weights with 0 training error in case of $N \leq D$.

☐ TRUE ☐ FALSE

Question 27 (Linear or Logistic Regression) Suppose you are given a dataset of tissue images from patients with and without a certain disease. You are supposed to train a model that predicts the probability that a patient has the disease. It is preferable to use logistic regression *over* linear regression.

☐ TRUE ☐ FALSE

Question 28 (Minima) Convex functions over a convex set have a unique global minimum.

☐ TRUE ☐ FALSE

Question 29 (Neural networks) Training only the first layer of a deep neural network using the logistic loss is equivalent to training a logistic regression over a transformed feature space.

☐ TRUE ☐ FALSE



Question 30 (Adversarial perturbations for linear models) Suppose you are given a linear classifier with the logistic loss. Is it true that generating the optimal adversarial perturbations by maximizing the loss under the ℓ_2 -norm constraint on the perturbation is an NP-hard optimization problem?

☐ TRUE ☐ FALSE

Question 31 (FastText supervised Classifier) The FastText supervised classifier can be modeled as a one-hidden-layer neural network.

☐ TRUE ☐ FALSE

Question 32 (SVD) The set of singular values of any rectangular matrix \mathbf{X} is equal to the set of eigenvalues for the square matrix $\mathbf{X}\mathbf{X}^\top$.

☐ TRUE ☐ FALSE

DRAFT



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.

K-Means

We will analyze the K -means algorithm and show that it always converge. Let us consider the K -means objective function:

$$\mathcal{L}(\mathbf{z}, \boldsymbol{\mu}) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2,$$

where $z_{nk} \in \{0, 1\}$ with $\sum_{k=1}^K z_{nk} = 1$ and $\boldsymbol{\mu}_k \in \mathbb{R}^D$ for $k = 1, \dots, K$ and $n = 1, \dots, N$.

Question 33: (3 points.) How would you choose the $\{z_{nk}\}_{n,k=1}^{N,K}$ to minimize $\mathcal{L}(\mathbf{z}, \boldsymbol{\mu})$ for given $\{\boldsymbol{\mu}_k\}_{k=1}^K$? Compute the closed-form formula for the z_{nk} . To which step of the K -means algorithm does it correspond?

☐ 0 ☐ 1 ☐ 2 ☐ 3

z_{nk} is an indicator that the point belongs to the z_k cluster
choosing the number of z correspond to the selection of number of clusters
 z will be a vector with one 1 entry and others 0

$$\sum_k z_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = \mathbf{0}.$$

$$z_{nk} = \begin{cases} 1 & \text{if } \mathbf{x}_n \in \mathbf{z}_k \quad K = \underset{1 \leq j \leq K}{\operatorname{argmin}} \|\mathbf{x}_n - \boldsymbol{\mu}_j\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$



Question 34: (3 points.) How would you choose $\{\mu_k\}_{k=1}^K$ to minimize $\mathcal{L}(\mathbf{z}, \boldsymbol{\mu})$ for given $\{z_{nk}\}_{n,k=1}^{N,K}$? Compute the closed-form formula for the μ_k . To which step of the K -means algorithm does it correspond?

☐ 0 ☐ 1 ☐ 2 ☐ 3

μ must be initialized

$$2) \sum_{k=1}^K z_{kn} (x_n - \mu_k) = 0$$

$$\mu_k = \frac{\sum_{n=1}^N z_{kn} x_n}{\sum_{n=1}^N z_{kn}}$$

Question 35: (2 points.) Using the two previous questions, show that the K -means algorithm always converges. Does it converge to a global minimum of the function \mathcal{L} ? Justify your answer.

☐ 0 ☐ 1 ☐ 2

* The quadratic form of the distance from the mean is a convex function, that has a global minimum.

- Each step is decreasing
- Lower bound of the objective is 0

It not necessarily converges to a global minimum, one example could be having three points $(0, 1, 2)$, with $K=2$ one solution could be $\mu(0, 1.5)$ or $\mu(0.5, 2)$

$$O(K^N)$$



Shift of the Eigenvalue Spectrum

Question 36: (1 points.) Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ be two symmetric matrices. Assume that $\mathbf{v} \in \mathbb{R}^n$ is an eigenvector for both matrices with associated eigenvalues λ_A and λ_B respectively. Show that \mathbf{v} is an eigenvector of the matrix $\mathbf{A} + \mathbf{B}$. What is the corresponding eigenvalue?

₀ ₁

$$\begin{aligned} & \mathbf{v} \lambda_A \mathbf{v}^T + \mathbf{v} \lambda_B \mathbf{v}^T \\ & \mathbf{v} (\lambda_A + \lambda_B) \mathbf{v}^T \\ & \mathbf{v} (\lambda_A + \lambda_B) \mathbf{v}^T \mathbf{v} = \mathbf{v} (\lambda_A + \lambda_B) \end{aligned}$$

Question 37: (2 points.) We consider now the ridge regression problem:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2N} \sum_{n=1}^N [y_n - \mathbf{x}_n^T \mathbf{w}]^2 + \lambda \|\mathbf{w}\|_2^2,$$

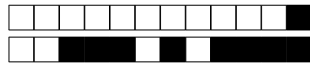
where the data $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ are such that the feature vector $\mathbf{x}_n \in \mathbb{R}^D$ and the response variable $y_n \in \mathbb{R}$.

Compute the closed-form solution $\mathbf{w}_{\text{ridge}}^*$ of this problem, providing the required justifications. State the final result using the data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$.

₀ ₁ ₂

I will take the derivative wrt \mathbf{w} as that's the parameter we want to minimize

$$\begin{aligned} & \frac{1}{2N} \sum_{n=1}^N (y_n - \mathbf{x}_n^T \mathbf{w}) + 2\lambda \mathbf{w} = 0 \quad \lambda^* = 2\lambda \\ & \mathbf{X}\mathbf{Y} - \mathbf{X}\mathbf{X}^T \mathbf{w} + \lambda^* \mathbf{w} = 0 \\ & \mathbf{X}\mathbf{Y} - (\mathbf{X}\mathbf{X}^T - \lambda^* \mathbf{I}_D) \mathbf{w} = 0 \\ & (\mathbf{X}\mathbf{X}^T - \lambda^* \mathbf{I}_D) \mathbf{w} = \mathbf{X}\mathbf{Y} \\ & \mathbf{w} = (\mathbf{X}\mathbf{X}^T - \lambda^* \mathbf{I}_D)^{-1} \mathbf{X}\mathbf{Y} \quad \lambda^* = 2\lambda \end{aligned}$$



Question 38: (2 points.) Using the two previous questions, explain why computing $\mathbf{w}_{\text{ridge}}^*$ is numerically more stable than computing the Least-squares solution \mathbf{w}^* (the solution obtained for $\lambda = 0$), i.e., why the solution can be still reliably computed even if $\mathbf{X}^\top \mathbf{X}$ is numerically singular.

☐ 0 ☐ 1 ☐ 2

Question 39: (3 points.) In the lecture on bias-variance decomposition we have seen that the true error can be decomposed into noise, bias and variance terms. What happens to the three terms for ridge regression when the regularization parameter λ grows? Explain your answer.

☐ 0 ☐ 1 ☐ 2 ☐ 3

To the noise nothing happens as it depends on the data and not the model

- The bias becomes bigger as the model weights move the 2
- The variance decreases as the bias becomes huge.



Kernel PCA

In this exercise, we will see how to combine the Principal Component Analysis (PCA) and the kernel method into an algorithm known as kernel PCA.

We are given n observations in a low dimensional space $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^L$ and we consider a kernel k and its associated features map $\phi : \mathbb{R}^L \mapsto \mathbb{R}^H$ which satisfies:

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathbb{R}^H},$$

where $\langle \cdot, \cdot \rangle_{\mathbb{R}^H}$ is the standard scalar product of \mathbb{R}^H .

We define the empirical covariance matrix and the empirical covariance matrix of the mapped observations as:

$$\Sigma := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \quad \text{and} \quad \Sigma^H := \frac{1}{n} \sum_{i=1}^n \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top.$$

The kernel matrix \mathbf{K} is defined by:

$$\mathbf{K}_{i,j} := k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathbb{R}^H}.$$

We also define the data matrix and the corresponding matrix of the mapped data as:

$$\mathbf{X} := \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} \in \mathbb{R}^{n \times L} \quad \text{and} \quad \Phi := \begin{pmatrix} \phi(\mathbf{x}_1)^\top \\ \vdots \\ \phi(\mathbf{x}_n)^\top \end{pmatrix} \in \mathbb{R}^{n \times H}.$$

Finally we denote the eigenpairs (eigenvalues and eigenvectors) of Σ^H by $\{(\lambda_i, \mathbf{v}_i)\}_{i=1}^H$ and those of \mathbf{K} by $\{(\rho_j, \mathbf{w}_j)\}_{j=1}^n$. We also assume that the vectors \mathbf{v}_i and \mathbf{w}_j are normalized. Thus:

$$\Sigma^H \mathbf{v}_i = \lambda_i \mathbf{v}_i, \quad \|\mathbf{v}_i\|_2 = 1 \quad \text{and} \quad \mathbf{K} \mathbf{w}_j = \rho_j \mathbf{w}_j, \quad \|\mathbf{w}_j\|_2 = 1.$$

Let us remind that we assume in the kernel setting that we can compute $k(\mathbf{x}, \mathbf{y})$ but that we cannot directly compute $\phi(\mathbf{x})$.

What we would like to do is to first map the data into the high-dimensional space using the features map ϕ and then to apply the standard PCA algorithm in the high-dimensional space \mathbb{R}^H . This would amount to:

- Computing the empirical covariance matrix Σ^H of the mapped data $\phi(\mathbf{x}_i)$.
- Computing the eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_N$ associated with the N largest eigenvalues of Σ^H .
- Computing the projection $\Pi(\phi(\mathbf{x}_i)) \in \mathbb{R}^L$ for each data point onto these eigenvectors, where the j -th component of the projection is given by:

$$\Pi_j(\phi(\mathbf{x}_i)) = \langle \phi(\mathbf{x}_i), \mathbf{v}_j \rangle_{\mathbb{R}^H}. \quad (\text{Proj})$$



Question 40: (1 points.) Explain why we cannot directly apply the algorithm explained above.

☐ 0 ☐ 1

We assume that $\phi(x)$ is difficult to compute.

Instead, we will now apply the kernel trick to this problem:

Question 41: (2 points.)

Write the empirical covariance matrices Σ and Σ^H in function of the design matrix \mathbf{X} and the features matrix Φ . What are the sizes of these matrices Σ and Σ^H ?

☐ 0 ☐ 1 ☐ 2

$$\Sigma = \frac{1}{n} \mathbf{X} \mathbf{X}^T$$
$$\Sigma^H = \frac{1}{n} \Phi \Phi^T$$

Question 42: (1 points.)

Write the kernel matrix \mathbf{K} as a function of the features matrix Φ . What is the size of this matrix?

☐ 0 ☐ 1

$$\mathbf{K} = \langle \phi(x_i), \phi(x_j) \rangle = \Phi \Phi^T$$



Question 43: (2 points.) We are given an eigenvector \mathbf{w}_i of the matrix \mathbf{K} with associated eigenvalue ρ_i . Show that the vector $\Phi^\top \mathbf{w}_i$ is an eigenvector of Σ^H . What is the associated eigenvalue?

 ₀ ₁ ₂

$$\Sigma^H \Phi^\top \mathbf{w}_i = \frac{1}{n} \Phi \Phi^\top \Phi \mathbf{w}_i = \frac{1}{n} \Phi \mathbf{K} \mathbf{w}_i = \frac{\rho_i}{n} \Phi \mathbf{w}_i$$

Question 44: (2 points.) Let us define $\tilde{\mathbf{v}}_i := \Phi^\top \mathbf{w}_i$. The vectors $\tilde{\mathbf{v}}_i$ are not normalized (i.e., their norms are not necessarily equal to one). Give a formula for the unit vector \mathbf{v}_i (with respect to the norm defined by the scalar product $\langle \cdot, \cdot \rangle_{\mathbb{R}^H}$) which is aligned with the vector $\tilde{\mathbf{v}}_i$?

 ₀ ₁ ₂

$$\|\mathbf{v}_i\|_F^2 = \|\Phi^\top \mathbf{w}_i\|_F^2 = \mathbf{w}_i^\top \Phi \Phi^\top \mathbf{w}_i = \mathbf{w}_i^\top \mathbf{K} \mathbf{w}_i = g_i \mathbf{w}_i^\top \mathbf{w}_i = g_i \|\mathbf{w}_i\|_2^2 = 1$$

$$\mathbf{v}_i = \frac{\Phi^\top \mathbf{w}_i}{\sqrt{g_i}}$$

$$\|\mathbf{w}_i\|_2^2 = \frac{1}{g_i}$$

$$\|\mathbf{w}_i\| = \frac{1}{\sqrt{g_i}}$$

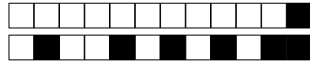
Question 45: (2 points.)

Let us assume that we have selected the N largest eigenvalues of \mathbf{K} and computed the associated N eigenvectors $\mathbf{w}_1, \dots, \mathbf{w}_N$. We are given a vector $\mathbf{x} \in \mathbb{R}^L$. Derive a formula which computes the projection $\Pi(\phi(\mathbf{x}))$ of $\phi(\mathbf{x})$ onto the vectors $\mathbf{v}_1, \dots, \mathbf{v}_N$ using only the kernel and objects defined on \mathbb{R}^L . You should ensure that your final result does not explicitly contain any ϕ or any vector or matrix in \mathbb{R}^H .

 ₀ ₁ ₂

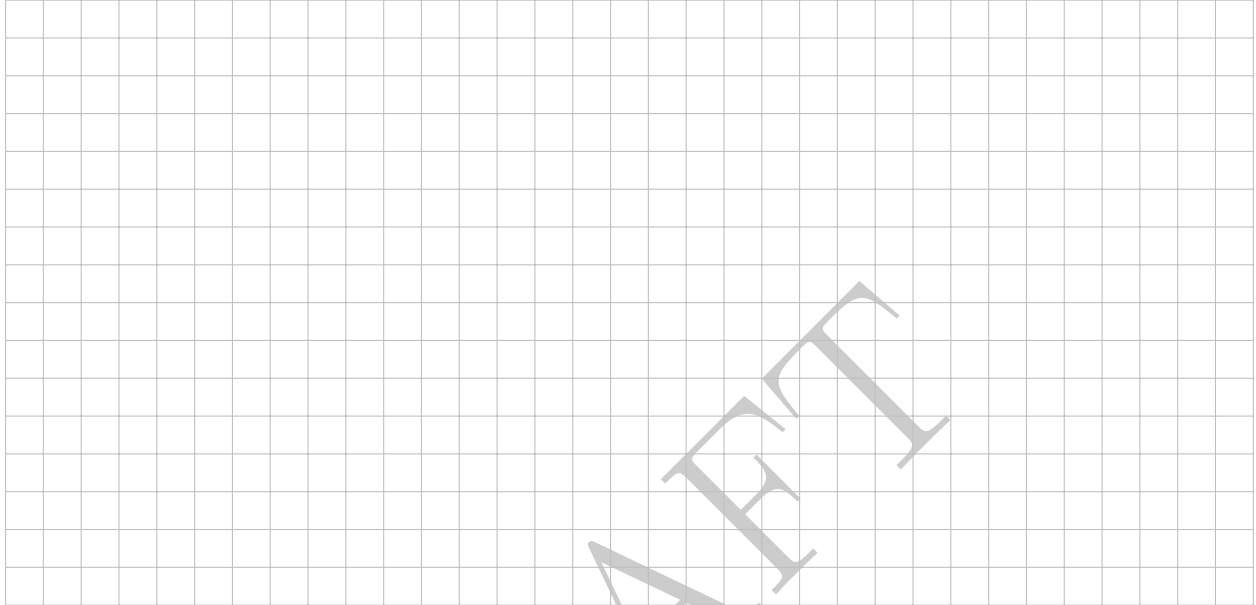
$$\Pi(\phi(\mathbf{x})) = \langle \phi, \mathbf{v}_i \rangle = \Phi \frac{\Phi^\top \mathbf{w}_i}{\sqrt{g_i}} = \frac{\mathbf{K} \mathbf{w}_i}{\sqrt{g_i}}$$

$$= \frac{1}{\sqrt{g_i}} \sum \phi(\mathbf{x}) \Phi^\top \mathbf{w}_i = \frac{1}{\sqrt{g_i}} \mathbf{K}(\mathbf{x}, \mathbf{x}_i) \mathbf{w}_i$$



Question 46: (2 points.) Finally, rewrite the (computationally expensive) algorithm explained at the beginning of the exercise in terms of the derived results, as a method that could actually be implemented. The input should be the samples $(\mathbf{x}_1, \dots, \mathbf{x}_n)$, the kernel k and the number N of principal components. The outputs are the projected data points $(\Pi_1(\phi(\mathbf{x}_i)), \dots, \Pi_N(\phi(\mathbf{x}_i)))^\top$ for $i = 1, \dots, n$ defined in Equation (Proj).

₀ ₁ ₂





DRAFT



DRAFT