MATH412 - Statistical Machine Learning
Fall semester 2019 - Final Exam
Duration: 3 hours
Solutions

- The proposed *very detailed* answers are typed in blue. In most parts of the exam, significantly less detailed answers were expected from the students. Problem 4 is an exception, because the reasoning has to be precise and thus the answers have to be more detailed.
- Additional comments are provided in green.
- Two typos are corrected in red in 4.3 and 4.4: $b \leq -1$ should have been $b < -1$.

1 Practical ML [10 points]

We consider a binary classification problem for which a predictor is being considered. A validation set which contains 20% of positive examples is used to evaluate the classifier. On this set the recall is at 90% and the false positive rate is of 5%. What is the misclassification error? Please detail your reasoning and calculation.

Let P, N, FP, FN, TP denote respectively the number of positives, of negatives, of false positives, false negatives and true positives. Let $n = N + P$ be the total number of datapoints in the validation set. By definition, the rate of false positives is $rFP = \frac{FP}{N}$ and the rate of true positives, aka recall, is $rTP = \frac{TP}{P} = 1 - \frac{FN}{P}$. Let $\pi = \frac{P}{n}$.

By definition, the misclassification error is

$$\widehat{\mathcal{R}}_{0\text{-}1} = \frac{FP}{n} + \frac{FN}{n} = (1-\pi)\frac{FP}{N} + \pi\frac{FN}{P} = (1-\pi)\,rFP + \pi\,(1-rTP).$$

With the number provided:

$$\widehat{\mathcal{R}}_{0\text{-}1} = (1-0.2)\cdot 0.05 + 0.2\cdot(1-0.9) = 0.8\cdot 0.05 + 0.2\cdot 0.1 = 0.06.$$

The misclassification error is thus of 6%.

2 Kernelized k-means [20 points]

The goal of this exercise is to construct a "kernelized" version of the k-means algorithm, that is a version of the k-means algorithm in which the distance between any pair of points used is the one derived from a Mercer kernel K associated with a Hilbert space $\mathcal{H}$. To reason about the problem we will assume that $\phi : \mathcal{X} \to \mathcal{H}$ is a feature map associated with the kernel K in the sense that $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = K(x, y)$. However, eventually we would like to design an algorithm that computes only values of the kernel function K and that never has to compute $\phi(x)$ because this element could be infinite dimensional. We consider the problem of clustering a dataset composed of n datapoints $\boldsymbol{x}_i \in \mathcal{X}$

1. Let C be a subset of $\{1, \ldots, n\}$ and let f be an element of $\mathcal{H}$ defined by $f = \frac{1}{|C|}\sum_{i\in C}\phi(x_i)$, where $|C|$ is the cardinality of C. Show that for any x we can express $\langle \phi(x), f \rangle_{\mathcal{H}}$ as a function of the $(K(x, x_i))_{i\in C}$.

$$\langle \phi(x), f \rangle_{\mathcal{H}} = \langle \phi(x), \frac{1}{|C|} \sum_{i \in C} \phi(x_i) \rangle_{\mathcal{H}} = \frac{1}{|C|} \sum_{i \in C} \langle \phi(x), \phi(x_i) \rangle_{\mathcal{H}} = \frac{1}{|C|} \sum_{i \in C} K(x, x_i).$$

2. Similarly show that the distance $\|\phi(x_j) - f\|_{\mathcal{H}}^2$ can be expressed as a function of the $(K(x_i, x_{i'}))_{i,i' \in C \cup \{j\}}$.

$$\|\phi(x_j) - f\|_{\mathcal{H}}^2 = \langle \phi(x_j), \phi(x_j) \rangle_{\mathcal{H}} - 2\langle \phi(x_j), f \rangle_{\mathcal{H}} + \langle f, f \rangle_{\mathcal{H}}.$$

Now $\langle f, f \rangle_{\mathcal{H}} = \frac{1}{|C|^2} \sum_{i,i' \in C} \langle \phi(x_i), \phi(x_{i'}) \rangle_{\mathcal{H}} = \frac{1}{|C|^2} \sum_{i,i' \in C} K(x_i, x_{i'})$.

So combining all, and using the shorthand $K_{i,j} = K(x_i, x_j)$, we have

$$\|\phi(x_j) - f\|_{\mathcal{H}}^2 = K_{j,j} - 2\frac{1}{|C|} \sum_{i \in C} K_{i,j} + \frac{1}{|C|^2} \sum_{i,i' \in C} K_{i,i'}.$$

3. Propose an algorithm written under the form of pseudo-code that will execute the k-means algorithm implicitly in the Hilbert space and return the obtained clusters.

The only difference between a classical k-means algorithm and a kernelized version is that the barycenter of a cluster cannot be computed explicitly since the average of the feature maps of a cluster of points is not the feature map of some a point of the input space in generak. So, instead of having two separate steps that compute first all barycenters and then the distance of each point to the barycenter, one has to compute the distance directly. However for efficiency, it is better to precompute the squared norm of each of the barycenters (this is ν_ℓ in the algorithm below) at each iteration to save computation.

Algorithm 1 Kernelized k-means

Input: Kernel matrix $\mathbf{K}$ and number of clusters k.
Draw k datapoints without replacement and let their indices be $i_1, \ldots, i_k$.
for $\ell = 1$ to k **do**
  $C_\ell \leftarrow \{i_\ell\}$
end for
`NOT_DONE_FLAG` $\leftarrow$ `TRUE` ▷ Boolean flag to terminate the algorithm
while `NOT_DONE_FLAG` **do**
  for $\ell = 1$ to k **do**
    $\nu_\ell \leftarrow \frac{1}{|C_\ell|^2} \sum_{j,j' \in C_\ell} K_{j,j'}$ ▷ Calculating the squared norm of the barycenter of cluster ℓ
  end for
  for $i = 1$ to n and $\ell = 1$ to k **do**
    $d_{i\ell}^2 \leftarrow K_{i,i} - \frac{2}{|C_\ell|} \sum_{j \in C_\ell} K_{i,j} + \nu_\ell$ ▷ $d_{i\ell}^2 = \left\|\phi(x_j) - \frac{1}{|C|} \sum_{j \in C} \phi(x_j)\right\|_{\mathcal{H}}^2$ (cf. question 2)
  end for
  for $\ell = 1$ to k **do**
    $C_\ell^{\text{old}} \leftarrow C_\ell$ ▷ Keep a copy of the previous clustering to see if it changes
    $C_\ell \leftarrow \varnothing$
  end for
  for $i = 1$ to n **do**
    $\ell_i \leftarrow \arg\min_{\ell : 1 \leq \ell \leq k} d_{i\ell}^2$ ▷ Assign datapoint i to the closest centroid
    $C_{\ell_i} \leftarrow C_{\ell_i} \cup \{i\}$
  end for
  `NOT_DONE_FLAG` $\leftarrow$ `BOOLEAN`$(\exists \ell, C_\ell^{\text{old}} \neq C_\ell)$ ▷ Continue to iterate as long as some clusters change
end while
Output: $(C_\ell)_{1 \leq \ell \leq k}$

For a practical implementation in R or Python, its would be more appropriate to encode $(C_\ell)_{1 \leq \ell \leq k}$ with a sparse binary matrix. In that case the calculations of ν_ℓ and d_{il}^2 can be done with efficient sparse matrix products.

3 Tikhonov regularization to estimate a mean [35 points]

Assume that we observe n i.i.d. datapoints $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with $\boldsymbol{X}_i \in \mathbb{R}^d$ and with

$$\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{\mu}_0 \qquad \text{and} \qquad \mathbb{E}[(\boldsymbol{X}_i - \boldsymbol{\mu}_0)(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^\top] = \sigma^2 \boldsymbol{I}_d$$

where $\boldsymbol{I}_d$ is the $d \times d$ identity matrix, and where $\boldsymbol{\mu}_0$ and σ^2 are unknown.

We consider the problem of choosing a predictor $\boldsymbol{\theta}$ that minimizes the ℓ_2 risk $\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}[\ell(\boldsymbol{\theta}, \boldsymbol{X})]$ with

$$\ell(\boldsymbol{\theta}, \boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{\theta}\|_2^2 = \sum_{j=1}^{d} (x_j - \theta_j)^2 \tag{1}$$

on a future datapoint $\boldsymbol{X}$ where $\boldsymbol{X}$ is assumed to follow the same distribution as the observed $\boldsymbol{X}_i$. Note that as compared with the supervised learning setting seen in class, here $\boldsymbol{X}$ plays the role of the output variable, and there is no input variable, which is the reason why the predictor $\boldsymbol{\theta}$ is a constant.

1. What is the value of the oracle predictor (aka target function) $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathcal{R}(\boldsymbol{\theta})$ (justify briefly your answer).

   By definition $\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}[\ell(\boldsymbol{\theta}, \boldsymbol{X})] = \mathbb{E}[\|\boldsymbol{X} - \boldsymbol{\theta}\|^2]$ and $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}[\|\boldsymbol{X} - \boldsymbol{\theta}\|^2] = \mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}_0$.

   No further justifications were expected, but if we had to explain further, to obtain this we can either use

   - the fact that the objective is a differentiable convex function, which entails, by Fermat's theorem on differentiable convex functions, that the minima are the stationary points, i.e., the points whose gradient is equal to 0 and the fact that $\nabla\mathcal{R}(\boldsymbol{\theta}) = 2(\boldsymbol{\theta} - \boldsymbol{\mu}_0)$.
   - or, the fact that the problem separates over all dimensions, and that for each dimension the target function is the conditonal expectation of the output variable given the input, but since there is not input this is just $\mathbb{E}[X_j] = \mu_{j0}$.

2. What is the value of the excess risk $\mathcal{R}(\boldsymbol{\theta}) - \mathcal{R}(\boldsymbol{\theta}^*)$ (no proof needed).

   $\mathcal{R}(\boldsymbol{\theta}) - \mathcal{R}(\boldsymbol{\theta}^*) = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$

   This is because $\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}[\|\boldsymbol{X} - \boldsymbol{\theta}\|^2] = \mathbb{E}[\|\boldsymbol{X} - \boldsymbol{\mu}_0 + \boldsymbol{\mu}_0 - \boldsymbol{\theta}\|^2] = \mathbb{E}[\|\boldsymbol{X} - \boldsymbol{\mu}_0\|^2] + \|\boldsymbol{\mu}_0 - \boldsymbol{\theta}\|^2 = \mathcal{R}(\boldsymbol{\theta}^*) + \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|^2$, using the same proof as the one presented in class to identify the target function for the square loss (but with no input and a multivariate output).

3. We assume that we have a sample of observations (or training data) $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. What is the predictor obtained from the empirical risk minimization principle?

   $\arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \|\boldsymbol{X}_i - \boldsymbol{\theta}\|^2 = \bar{\boldsymbol{X}}$. This is well know and can be rederived using again Fermat's theorem.

4. We now consider predictors of the form $\hat{\boldsymbol{\theta}} = c\bar{\boldsymbol{X}}$ with $\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i$. We call $\mathbb{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\mu}_0$ and $\mathbb{E}[\|\hat{\boldsymbol{\theta}} - \mathbb{E}[\hat{\boldsymbol{\theta}}]\|^2]$ respectively the bias and the total variance of the predictor $\hat{\boldsymbol{\theta}}$. Express both the bias and the variance as a function of $c, \boldsymbol{\mu}_0, \sigma^2, n$ and the dimension d?

   $\text{bias}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[c\bar{\boldsymbol{X}}] - \boldsymbol{\mu}_0 = (c-1)\boldsymbol{\mu}_0$.

   $\text{Var}(\hat{\boldsymbol{\theta}}) = \mathbb{E}\big[\|c\bar{\boldsymbol{X}} - \mathbb{E}[c\bar{\boldsymbol{X}}]\|^2\big] = c^2\mathbb{E}\big[\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0\|^2\big] = c^2\text{tr}\big(\mathbb{E}\big[(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top\big]\big) = c^2\text{tr}\big(\frac{1}{n}\sigma^2\boldsymbol{I}_d\big) = c^2\frac{d}{n}\sigma^2$.

5. Show that $\mathbb{E}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_0\|_2^2]$ can be expressed as a sum of two terms which depend respectively on the bias and on the total variance of the predictor.

   By classical bias-variance decomposition:

   $\mathbb{E}\big[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_0\|^2\big] = \mathbb{E}\big[[\|\hat{\boldsymbol{\theta}} - \mathbb{E}[\hat{\boldsymbol{\theta}}] + \mathbb{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\mu}_0\|^2\big] = \mathbb{E}\big[\|\hat{\boldsymbol{\theta}} - \mathbb{E}[\hat{\boldsymbol{\theta}}]\|^2\big] + \mathbb{E}\big[\|\mathbb{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\mu}_0\|^2\big] = \text{Var}(\hat{\boldsymbol{\theta}}) + \|\text{bias}(\hat{\boldsymbol{\theta}})\|^2$.

6. Explain why, when changing c, we trade-off between these two terms? Is this related to the "approximation error-estimation error" tradeoff that we have seen in class? Why?

Note that $\mathbb{E}\big[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_0\|^2\big]$ is actually the expected excess risk $\mathbb{E}[\mathcal{R}(\hat{\boldsymbol{\theta}})] - \mathcal{R}(\boldsymbol{\mu}_0)$ of the estimator $\hat{\boldsymbol{\theta}}$ for the excess risk considered in question 2. Using the expression from the two previous questions, we have:

$$\mathbb{E}[\mathcal{R}(\hat{\boldsymbol{\theta}})] - \mathcal{R}(\boldsymbol{\mu}_0) = \mathbb{E}\big[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_0\|^2\big] = c^2 \frac{d}{n}\sigma^2 + (c-1)^2\|\boldsymbol{\mu}_0\|^2.$$

For $c \leq 1$, we see that, when c decreases, the variance term decreases and the bias term increases. (Taking $c > 1$ is certainly not interesting since it increases both terms.) This is a classical bias-variance tradeoff which is encountered for many estimators in statistics (one classical example is the maximum likelihood estimator for the variance $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$ which is biased but has lower variance than its unbiased counterpart $\tilde{\sigma} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$). This resembles relatively closely an "expected" approximation-estimation tradeoff, with the *bias* acting like an *approximation error* and the *variance* being comparable to an *expected estimation error*, although in the case of the square loss for instance, the expected estimation error itself usually decomposes into the variance of the predictor and its bias.

Beyond what is expected as an answer to this question, let us say that it is not possible to define a deterministic set of predictors S such that $\hat{\boldsymbol{\theta}} \in S$ and $c\boldsymbol{\mu}_0$ is the best predictor in that class. It could be tempting to try and consider $S = \{\boldsymbol{\mu} \mid \|\boldsymbol{\mu}\| \leq c\|\boldsymbol{\mu}_0\|\}$ but we cannot guarantee that $\hat{\boldsymbol{\theta}}$ will be in that set. (If, instead of sets of predictors, we could choose sets of estimators, that is statistics which are defined as functions of the training data, then we could consider the set of all estimators whose expected value is smaller than $c\|\boldsymbol{\mu}_0\|$ in norm, but this is not how we defined the estimation-approximation tradeoff in class since we always talked about sets of predictors, i.e., deterministic functions of the input data).

7. What is the optimal value of c? Express it as a function of $\boldsymbol{\mu}_0, d, n$ and σ^2.

   To minimize $\mathbb{E}\big[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_0\|^2\big]$ w.r.t. c, we set the derivative to 0 since we can still apply Fermat's theorem, and we get:

$$c^* = \frac{\|\boldsymbol{\mu}_0\|^2}{\|\boldsymbol{\mu}_0\|^2 + \frac{d}{n}\sigma^2} = \frac{1}{1 + \frac{1}{n}\frac{\sigma^2 d}{\|\boldsymbol{\mu}_0\|^2}}.$$

   Note that $\frac{\|\boldsymbol{\mu}_0\|^2}{\sigma^2 d/n}$ can be thought as a squared *signal-to-noise ratio*. Using c^* instead of $c = 1$ is really worth it if the level of the "noise" $\sqrt{\sigma^2 d/n}$ in the classical $\bar{\boldsymbol{X}}$ estimator is large compared to the magnitude $\|\boldsymbol{\mu}_0\|$ of the signal $\boldsymbol{\mu}_0$ it estimates.

8. We now consider the predictor obtained as minimizer of the regularized empirical risk $\widehat{\mathcal{R}}_n(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_2^2$, where $\widehat{\mathcal{R}}_n(\boldsymbol{\theta})$ denotes the empirical risk. Express the obtained predictor as a function of λ and $\bar{\boldsymbol{X}}$. What is the optimal value of λ as function of $\boldsymbol{\mu}_0, \sigma^2, n$ and the dimension d?

   Still applying Fermat's theorem, since the gradient of $\widehat{\mathcal{R}}_n(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_2^2$ is $\boldsymbol{\mu} - \bar{\boldsymbol{X}} + \lambda\boldsymbol{\mu}$, we obtain a regularized estimator of the form:

$$\hat{\boldsymbol{\mu}}_\lambda = \frac{1}{1+\lambda}\bar{\boldsymbol{X}}.$$

   which is exactly of the form $c\bar{\boldsymbol{X}}$ with $c = \frac{1}{1+\lambda}$. This shows by identification that the best value of λ is $\lambda^* = \frac{1}{n}\frac{\sigma^2 d}{\|\boldsymbol{\mu}_0\|^2}$.

9. The optimal values for λ or c determined so far are "oracle values" in the sense that they depend on quantities which are unknown (and closely related to the quantities that one tries to estimate); as a result they of course cannot be used in practice... In the rest of the problem we investigate how we can estimate c using leave-one out cross-validation. Given a sample of observations $(\boldsymbol{x}_i)_{i=1..n}$, give the formula for the leave-one-out estimate $\hat{\mathcal{R}}_{\mathrm{LOO}}$ of the risk $\mathcal{R}(\hat{\boldsymbol{\theta}})$ for the predictor of question 4 (whose hyperparameter is c).

   Let $\bar{\boldsymbol{x}}_{-i} = \frac{1}{n-1}(n\bar{\boldsymbol{x}} - \boldsymbol{x}_i)$ note the empirical mean of all datapoint except $\boldsymbol{x}_i$. Then,

$$\hat{\mathcal{R}}_{\mathrm{LOO}}(c) = \frac{1}{n}\sum_{i=1}^n \|c\,\bar{\boldsymbol{x}}_{-i} - \boldsymbol{x}_i\|^2.$$

10. Rewrite the previous formula as a function of $(\boldsymbol{x}_i)_{i=1..n}, \bar{\boldsymbol{x}}, c$ and n.

$$\hat{\mathcal{R}}_{\text{LOO}}(c) = \frac{1}{n}\sum_{i=1}^{n} \|c\,\frac{1}{n-1}(n\bar{\boldsymbol{x}} - \boldsymbol{x}_i) - \boldsymbol{x}_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|\,\frac{1}{n-1}(nc\bar{\boldsymbol{x}} - c\boldsymbol{x}_i - (n-1)\boldsymbol{x}_i)\|^2.$$

11. Show that as a function of c the leave one out estimate of the risk can be written as:

$$\hat{\mathcal{R}}_{\text{LOO}}(c) = \Big(1 + \frac{c}{n-1}\Big)^2 \hat{\sigma}^2 d + (1-c)^2\|\bar{\boldsymbol{x}}\|_2^2 \qquad \text{where} \qquad \hat{\sigma}^2 = \frac{1}{nd}\sum_{i=1} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|_2^2.$$

We need to rewrite the previous expression so as to make appear terms of the form $\boldsymbol{x}_i - \bar{\boldsymbol{x}}$ which will contribute to a variance term and a remaining term proportional to $\bar{\boldsymbol{x}}$. But

$$\begin{aligned} nc\bar{\boldsymbol{x}} - c\boldsymbol{x}_i - (n-1)\boldsymbol{x}_i &= nc\bar{\boldsymbol{x}} - c\boldsymbol{x}_i - (n-1)c\boldsymbol{x}_i - (n-1)(1-c)\boldsymbol{x}_i \\ &= nc(\bar{\boldsymbol{x}} - \boldsymbol{x}_i) - (1-c)(n-1)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) + (1-c)(n-1)\bar{\boldsymbol{x}} \\ &= (1-c)(n-1) + nc)(\bar{\boldsymbol{x}} - \boldsymbol{x}_i) + (1-c)(n-1)\bar{\boldsymbol{x}} \\ &= (n-1)\big[(1 + \tfrac{c}{(n-1)})(\bar{\boldsymbol{x}} - \boldsymbol{x}_i) + (1-c)\bar{\boldsymbol{x}}\big] \end{aligned}$$

Replacing in the expression found in question 10 we get:

$$\begin{aligned} \hat{\mathcal{R}}_{\text{LOO}}(c) &= \frac{1}{n}\sum_{i=1}^{n} \|(1 + \tfrac{c}{(n-1)})(\bar{\boldsymbol{x}} - \boldsymbol{x}_i) + (1-c)\bar{\boldsymbol{x}}\|^2 \\ &= (1 + \tfrac{c}{(n-1)})^2 \frac{1}{n}\sum_{i=1}^{n} \|\bar{\boldsymbol{x}} - \boldsymbol{x}_i\|^2 + (1-c^2)\|\bar{\boldsymbol{x}}\|^2, \end{aligned}$$

since the cross-term in the expansion of the square is equal to 0. This establishes the announced formula. (This question is a bit more technical but it is possible to do the rest of the problem without answering it.)

12. Compute $\mathbb{E}[\|\bar{\boldsymbol{X}}\|_2^2]$.

$$\mathbb{E}[\|\bar{\boldsymbol{X}}\|^2] = \mathbb{E}[\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0 + \boldsymbol{\mu}_0\|^2] = \mathbb{E}[\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0\|_2^2] + \|\boldsymbol{\mu}_0\|_2^2 = \frac{d}{n}\sigma^2 + \|\boldsymbol{\mu}_0\|^2,$$

by the same reasoning as the variance calculation in question 4.

13. Calculate the optimal value $\hat{c}$ of c according to $\hat{\mathcal{R}}_{LOO}$.

If we minimize the expression of $\hat{\mathcal{R}}_{\text{LOO}}(c)$ from question 11 with respect to c we get

$$c_{\text{LOO}} = \frac{\|\bar{\boldsymbol{x}}\|^2 - \frac{\hat{\sigma}^2 d}{n-1}}{\|\bar{\boldsymbol{x}}\|^2 + \frac{\hat{\sigma}^2 d}{(n-1)^2}}$$

14. Given that $\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$, argue informally that $\hat{c}$ is close to the optimal value of c determined in question 7.

We consider the numerator and the denominator in the expression of c:
The expected value of the numerator is

$$\mathbb{E}\Big[\|\bar{\boldsymbol{x}}\|^2 - \frac{\hat{\sigma}^2 d}{n-1}\Big] = \|\boldsymbol{\mu}_0\|^2 + \frac{d}{n}\sigma^2 - \frac{d}{n}\sigma^2 = \|\boldsymbol{\mu}_0\|^2.$$

The expected value of the denominator is

$$\mathbb{E}\Big[\|\bar{\boldsymbol{x}}\|^2 + \frac{\hat{\sigma}^2 d}{(n-1)^2}\Big] = \|\boldsymbol{\mu}_0\|^2 + \frac{d}{n}\sigma^2 + \frac{d}{n(n-1)}\sigma^2.$$

We see that the ratio of these expectations is

$$\frac{\|\boldsymbol{\mu}_0\|^2}{\|\boldsymbol{\mu}_0\|^2 + \frac{d}{n}\sigma^2 + \frac{d}{n(n-1)}\sigma^2},$$

which is quite similar to the expression of c^*...
Unfortunately, this is not enough to conclude rigorously, because the random variables that we consider converge themselves towards their expected value at a speed which is $O(\frac{1}{n})$. So this is suggestive more than conclusive...
To make a more rigorous analysis (which was not expected as an answer to this question), we would have to transform the expression of c_{LOO} into

$$c_{\text{LOO}} = 1 - \frac{1}{\frac{(n-1)^2}{n}\frac{\|\bar{\boldsymbol{x}}\|^2}{\hat{\sigma}^2 d} + \frac{1}{n}}$$

and to transform similarly c^* to get

$$c^* = 1 - \frac{1}{n\frac{\|\boldsymbol{\mu}_0\|^2}{\sigma^2 d} + 1}$$

This shows clearly that $c^* = 1 + O(\frac{1}{n})$. So to show that c_{LOO} is a good approximation of c^* we would need to show that $c_{\text{LOO}} - c^* = o(\frac{1}{n})$. (To be rigorous since c_{LOO} is random, I would have to use o_p and O_P notations instead of the classical Landau notations, which are beyond the scope of this course.)

15. **Going further...**
Even though we did not formally prove it, the ideas in this problem suggest that there exist estimators for the expected value $\boldsymbol{\mu}_0$ of a random variable, which have a lower risk for the square loss than the empirical average $\bar{\boldsymbol{X}}$. In this course, we have seen right from the beginning that using regularization is beneficial, so this might come to no surprise to many of you that using regularization can improve the estimator for the expected value $\boldsymbol{\mu}_0$. However, the square loss is also the negative log-likelihood for the Gaussian distribution when the parameter of interest is the expected value $\boldsymbol{\mu}_0$, and, in that case, $\bar{\boldsymbol{X}}$ is also the maximum likelihood estimator. Basically, what we are showing is that there exist estimators which always have higher expected log-likelihood than the maximum likelihood estimator, although the latter seems darn simple and natural... Of course this has to do with the optimism of empirical risk minimization that we talked about in the course on model evaluation.

Statisticians of the XXth century have worked quite hard to show a number of theoretical properties of the maximum likelihood estimator, and a number of these results are optimality results. For example, one of the well-know result is that the maximum likelihood estimator is asymptotically unbiased (statisticians say that it is *consistent* which means that it converges a.s. to the parameter it estimates) and it has asymptotically minimal variance (it achieves the Cramér-Rao lower bound). A well know instance of an estimator of the form $c\bar{\boldsymbol{X}}$ is the *James-Stein estimator*, which has a better quadratic risk than $\bar{\boldsymbol{X}}$ for any μ_0, σ^2, n and any $d \geq 3$. This came as a shock for statisticians in the 60ies (especially frequentist statisticians) who held the maximum likelihood estimator as the "right way" to estimate. This is actually leads to *Stein's paradox* which is that it is possible to couple several unrelated estimation problem to obtain better estimates. The James-Stein estimator is always superior to the empirical mean $\bar{\boldsymbol{X}}$. As a consequence $\bar{\boldsymbol{X}}$ is said to be *inadmissible* (an estimator is *admissible* if there exists at least one particular setting under which it achieves lower expected risk than all other estimators).

4 Distance learning [35 points]

In this problem, given a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$, we will write "$\boldsymbol{\theta} > 0$" to mean that $\forall j \in \{1, \ldots, p\}$, $\theta_j > 0$.

We consider the problem of learning a distance function between datapoints such that points that are similar are closer together and point that are dissimilar are further apart. Specifically, given a set of points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in $\mathbb{R}^p$, we assume that we are given a set S of similar pairs of points, which should be close from each other and a set D of dissimilar points which should be far from each other. More precisely, we would like to learn a distance d such that

$$\forall (\{i,j\}, \{i',j'\}) \in S \times D, \qquad d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 + 2 \leq d(\boldsymbol{x}_{i'}, \boldsymbol{x}_{j'})^2. \tag{2}$$

We choose to learn a Mahalanobis distance function: The Mahalanobis distance function associated with the positive definite matrix $\boldsymbol{M}$ is the Euclidean distance $d_{\boldsymbol{M}}$ defined on $\mathbb{R}^p$ by

$$\forall x, y \in \mathbb{R}^p, \quad d_{\boldsymbol{M}}(\boldsymbol{x}, \boldsymbol{y})^2 = (\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{M} (\boldsymbol{x} - \boldsymbol{y}).$$

To simplify, we focus on the case where $\boldsymbol{M}$ is the diagonal matrix $\boldsymbol{M} = \mathrm{Diag}(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \mathbb{R}^p$. Note that $d_{\mathrm{Diag}(\boldsymbol{\theta})}$ is a distance if and only if $\boldsymbol{\theta} > 0$. For all $\{i, j\} \in S \cup D$, let $y_{ij} = -1$ if $\{i, j\} \in S$ and $y_{ij} = 1$ if $\{i, j\} \in D$.

1. By considering the pair $(\{i, j\}, \{i', j'\}) \in S \times D$ that attains the smallest distance difference, show that (2) is satisfied if and only if there exist $\gamma_-, \gamma_+ > 0$ such that $\gamma_+ - \gamma_- = 2$ and the following set of inequalities hold

$$\begin{cases} \forall \{i, j\} \in S, & d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 \leq \gamma_-, \\ \forall \{i, j\} \in D, & d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 \geq \gamma_+. \end{cases} \tag{$\star$}$$

   We first prove the "if" statement:
   Assume that (2) holds. Let $\gamma'_- = \max_{\{i,j\} \in S} d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2$ and $\gamma_+ = \min_{\{i,j\} \in D} d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2$. Since (2) holds, we have $\gamma'_- + 2 \leq \gamma'_+$. And so if we define $\gamma_- = \gamma_+ - 2$ then $\gamma_- \geq \gamma'_- \geq d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 \geq 0$ for all $\{i, j\} \in S$. If S contains at least a non-trivial pair then we have $0 < \gamma_- \leq \gamma_+$.
   We then prove the "only if" statement:
   Assume ($\star$) holds, then for all $(\{i, j\}, \{i', j'\}) \in S \times D$, $d(\boldsymbol{x}_{i'}, \boldsymbol{x}_{j'})^2 - d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 \geq \gamma_+ - \gamma_- = 2$.

2. Propose a mapping $\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^p$ such that: given $\gamma_-, \gamma_+ > 0$ such that $\gamma_+ - \gamma_- = 2$, and letting $b = -\frac{\gamma_+ + \gamma_-}{2}$, there exists a Mahalanobis distance of the form $d_{\mathrm{Diag}(\boldsymbol{\theta})}$ satisfying (2) if and only if there exists $\boldsymbol{\theta} > 0$ such that we have

$$\forall \{i, j\} \in S \cup D, \qquad y_{ij} \big(\boldsymbol{\theta}^\top \phi(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \big) \geq 1. \tag{$\star\star$}$$

   Let $\phi(x_i, x_j) = \big((x_{im} - x_{jm})^2\big)_{1 \leq m \leq p}$. We have $\phi(x_i, x_j)^\top \boldsymbol{\theta} = d_{\boldsymbol{M}}(x_i, x_j)^2$ with $\boldsymbol{M} = \mathrm{Diag}(\boldsymbol{\theta})$.
   Now, note that ($\star$) is equivalent to

$$\begin{cases} \forall \{i, j\} \in S, & d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 + \frac{\gamma_+ - \gamma_-}{2} \leq \gamma_- + \frac{\gamma_+ - \gamma_-}{2} \\ \forall \{i, j\} \in D, & d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 - \frac{\gamma_+ - \gamma_-}{2} \geq \gamma_+ - \frac{\gamma_+ - \gamma_-}{2}. \end{cases}$$

   But given that $\frac{\gamma_+ - \gamma_-}{2} = 1$ and $\gamma_- + \frac{\gamma_+ - \gamma_-}{2} = \gamma_+ - \frac{\gamma_+ - \gamma_-}{2} = -b$, the above is still equivalent to

$$\begin{cases} \forall \{i, j\} \in S, & d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 + 1 \leq -b \\ \forall \{i, j\} \in D, & d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 - 1 \geq -b. \end{cases}$$

   which is equivalent to

$$\begin{cases} \forall \{i, j\} \in S, & d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 + b \leq -1 \\ \forall \{i, j\} \in D, & d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 + b \geq +1. \end{cases}$$

   And imposing these inequalities to $d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 = d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 = \phi(x_i, x_j)^\top \boldsymbol{\theta}$ is equivalent to ($\star\star$) holds for some $\boldsymbol{\theta}$. Last, but not least, $\boldsymbol{\theta}$ has to satisfy $\boldsymbol{\theta} > 0$ otherwise $\mathrm{Diag}(\boldsymbol{\theta})$ is not a positive definite matrix.

3. Show that for $\gamma_-, \gamma_+ > 0$ such that $\gamma_+ - \gamma_- = 2$ we have that $b = -\frac{\gamma_+ + \gamma_-}{2}$ satisfies $b < -1$.

   Given that $\gamma_+ = \gamma_- + 2$ we have $-2b = \gamma_+ + \gamma_- = 2 + 2\gamma_- > 2$ which proves the result.

4. Show that if b satisfies ($\star\star$) and S is not empty then $b < -1$.

   If there exists at least one constraint associated to a (non-trivial) pair in S in ($\star\star$), then for that pair $\{i, j\} \in S$, we have $b < d(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 + b \leq -1$.

5. Assuming that S and D are not empty, explain why the problem of finding a Mahalanobis distance of the form $d_{\mathrm{Diag}(\boldsymbol{\theta})}$ satisfying (2) is equivalent to finding $\boldsymbol{\theta} \in \mathbb{R}^p$, $\boldsymbol{\theta} > 0$ and $b \in \mathbb{R}$ such that ($\star\star$) holds.

($\Rightarrow$) We prove first that if we have a Mahalanobis distance of the form $d_{\text{Diag}(\boldsymbol{\theta})}$ satisfying (2) then it produces $\boldsymbol{\theta} \in \mathbb{R}^p$, $\boldsymbol{\theta} > 0$ and $b \in \mathbb{R}$ such that ($\star\star$) holds:
We proved in question 2 that the set of inequalities of (2) are equivalent to the set of inequalities of ($\star\star$) for a particular value of b given by $b = -\frac{\gamma_+ + \gamma_-}{2}$. Besides the fact that $d_{\text{Diag}(\boldsymbol{\theta})}$ is a distance implies that $\boldsymbol{\theta} > 0$.

($\Leftarrow$) We then prove that if we have $\boldsymbol{\theta} \in \mathbb{R}^p$, $\boldsymbol{\theta} > 0$ and $b \in \mathbb{R}$ that satisfies ($\star\star$) then there exists $\gamma_+, \gamma_- > 0$ with $\gamma_+ - \gamma_- = 2$ such that $d_{\text{Diag}(\boldsymbol{\theta})}$ is a Mahalanobis distance satisfying (2):
Let $(\boldsymbol{\theta}, b) > 0$ with $\boldsymbol{\theta} > 0$ be a pair that solves ($\star\star$), if we set $\gamma_+ = 1 - b$ and $\gamma_- = -b - 1$ then $\gamma_+ - \gamma_- = 2$ and then by question 4, we have $-b > 1$ so that we must also have $\gamma_- = -b - 1 > 1 - 1 = 0$ which shows that this produces a valid pair $\gamma_+, \gamma_- > 0$ for which the equivalence proved in question 2 proves that the equations of ($\star\star$) are equivalent to those of (2).

In conclusion of this question, we have shown that learning (or actually finding) $\gamma_+, \gamma_- > 0$ and $d_{\text{Diag}(\boldsymbol{\theta})}$ such that (2) holds is equivalent to learning $\boldsymbol{\theta}, b$ with $\boldsymbol{\theta} > 0$ such that ($\star\star$) holds.

6. What is the relationship between the problem of learning a Mahalanobis distance of the form $d_{\text{Diag}(\boldsymbol{\theta})}$ satisfying (2) and hard SVMs? Based on ($\star\star$), propose an optimization formulation similar to the hard SVM formulation to solve this distance learning problem.

   ($\star\star$) is exactly the constraint set of a hard-SVM. There is however an additional constraint for the distance learning problem which is that we impose $\boldsymbol{\theta} > 0$. In the hard SVM, the best separating hyperplane is defined as the one maximizing the margin, which is equivalent to imposing that it has minimal ℓ_2 norm. We can extend these ideas to distance learning and use the same regularization here. This leads us to solve,

$$\min_{\boldsymbol{\theta} \geq 0} \|\boldsymbol{\theta}\|_2^2 \qquad \text{s.t.} \qquad (\star\star) \quad \text{holds.}$$

   Note that we need to relax the constraint $\boldsymbol{\theta} > 0$ into $\boldsymbol{\theta} \geq 0$ so have a closed constrained set, otherwise the optimization problem can fail to have a minimum (if the infimum is only attained on the closure of the domain).

7. We now consider the case in which the set of constraints ($\star\star$) is potentially not feasible. Use the ideas from soft-SVMs to propose an optimization problem that allows some of the constraints to be violated but penalizes the total amount of violation.

   We introduce a non-negative slack variable ξ_{ij} for each constraint, and penalize the sum of the violations as in the soft-SVM.

$$\min_{\boldsymbol{\theta} \geq 0} \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + C \sum_{\{i,j\} \in S \cup D} \xi_{ij} \qquad \text{s.t.} \qquad \forall \{i,j\} \in S \cup D, \quad y_{ij}\big(\boldsymbol{\theta}^\top \phi(\boldsymbol{x}_i, \boldsymbol{x}_j) + b\big) \geq 1 - \xi_{ij} \quad \text{and} \quad \xi_{ij} \geq 0.$$

   Note that the constraint $\boldsymbol{\theta} \geq 0$ still makes it different than the classical SVM.

8. Explain how it would also be possible to propose SVM formulations directly inspired by (2). Why would this be less efficient computationnally, especially for large datasets?

   We can rewrite constraint (2) as

$$\forall (\{i,j\}, \{i',j'\}) \in S \times D, \qquad \tfrac{1}{2}\boldsymbol{\theta}^\top \big(\phi(x_{i'}, x_{j'}) - \phi(x_i, x_j)\big) \geq 1$$

   and so if we use this set of constraints to construct an SVM formulation, this leads to

$$\begin{aligned} \min_{\boldsymbol{\theta} \geq 0} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + C \sum_{(\{i,j\},\{i',j'\}) \in S \times D} \xi_{iji'j'} \\ \text{s.t.} \quad & \forall (\{i,j\}, \{i',j'\}) \in S \times D, \quad \begin{cases} \frac{1}{2}\boldsymbol{\theta}^\top \big(\phi(x_{i'}, x_{j'}) - \phi(x_i, x_j)\big) \geq 1 - \xi_{iji'j'}, \\ \xi_{iji'j'} \geq 0. \end{cases} \end{aligned}$$

   but then the number of constraints scales as $|S| \cdot |D|$ as opposed to $|S| + |D|$, so this formulation would not scale as well as the one we proposed in the previous question.