MATH412 - Statistical Machine Learning
Fall semester 2022 - Final Exam - Duration: 3 hours.

Problems A, B, C and D are independent. Please use different sheets of papers for the different exercises and number all the pages. Don't forget to put your name on all sheets of paper.

Problem A (3 points)

We consider a purely random binary classifier, which predicts class 1 with probability p regardless of the input. We apply this classifier to a dataset with 30% of positives.

1. Express the expected misclassification error as a function of p.

   We can use the expression of the misclassification error as a function of α the false positive rate and β the false negative rate:
   $$\widehat{\mathcal{R}}_{0\text{-}1} = \pi\,\mathrm{fnr} + (1-\pi)\,\mathrm{fpr},$$
   where fnr and fpr are respectively the false negative rate and the false positive rate, and where $\pi = 0.3$ is the proportion of positives. We also have $\mathbb{E}[\mathrm{fpr}] = p$, and $\mathbb{E}[\mathrm{fnr}] = (1-p)$. So, taking expectation of the previous misclassification error, we get,
   $$\mathbb{E}[\widehat{\mathcal{R}}_{0\text{-}1}] = \pi(1-p) + (1-\pi)p = \pi + p - 2\pi p = 0.3 + 0.4p.$$

2. For which value of p is this expected misclassification error the smallest?

   In general, since this is a linear function of p, the minimum is obtained for $p \in \{0, 1\}$. It is then immediate to see that $p = 0$ is best if $\pi < 0.5$ and $p = 1$ is best if $\pi > 0.5$. So in this case, this is for $p = 0$.

3. Assuming that the dataset is rather large, what is the precision approximately equal to ? Provide a mathematical justification. If we let fp and tp denote respectively the number of false positives and the number of true positives and if P and N are the respective number of positives and negatives, then we have
   $$\text{Precision} = \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}} = \frac{\frac{\mathrm{tp}}{P}}{\frac{\mathrm{tp}}{P} + \frac{N}{P}\frac{\mathrm{fp}}{N}}$$
   So
   $$\text{Precision} = \frac{\mathrm{tpr}}{\mathrm{tpr} + \frac{N}{P}\mathrm{fpr}} \approx \frac{p}{p + \frac{N}{P}p} = \pi,$$
   because $\mathrm{tpr} \approx \mathbb{E}[\mathrm{tpr}] = p$ and $\mathrm{fpr} \approx \mathbb{E}[\mathrm{fpr}] = p$ by the LLN, so the Precision is close to $\pi = 30\%$.

Problem B (4 points)

Given a training set $D = \big((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\big)$, and an associated vector of non-negative weights $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ such that $\sum_{i=1}^n \gamma_i = 1$, we consider the problem of constructing a regression tree learning algorithm for the weighted empirical risk
$$\widehat{R}_{\boldsymbol{\gamma}}(f) = \sum_{i=1}^n \gamma_i\, \ell(f(\boldsymbol{x}_i), y_i),$$
in the particular case of the square loss $\ell(a, y) = (a-y)^2$. We assume for simplicity in the rest of this exercise that $\forall i,\ \boldsymbol{x}_i \in [0,1]^d$.

1. Consider a regression tree of the form $f_{\boldsymbol{w},\Pi}(\boldsymbol{x}) = \sum_{j=1}^d w_j\, 1_{\{\boldsymbol{x}\in R_j\}}$, where $\Pi = \{R_1, \ldots, R_d\}$ is a fixed partition of $[0,1]^d$ into hyper-rectangles $R_j \subset [0,1]^d$ obtained by recursive splitting on the value of one of the variables each time, as in the algorithm for decision or regression tree learning seen in the course. For

a fixed partition Π, what is the value of $\boldsymbol{w} = (w_1, \ldots, w_d)$ which minimizes $\widehat{R}_{\boldsymbol{\gamma}}(f_{\boldsymbol{w},\Pi})$? We can rewrite the risk as follows

$$\widehat{R}_{\boldsymbol{\gamma}}(f_{\boldsymbol{w},\Pi}) = \sum_{j=1}^{d} \Gamma_j \sum_{i:\boldsymbol{x}_i \in R_j} \frac{\gamma_i}{\Gamma_j} (w_j - y_i)^2,$$

with $\Gamma_j := \sum_{i=1}^n \gamma_j 1_{\{\boldsymbol{x}_i \in R_j\}}$. If we let $\tilde{\gamma}_i = \frac{\gamma_i}{\Gamma_j}$, we find the objective is separable, so that

$$\hat{w}_j = \arg\min_{w_j} \sum_{i:\boldsymbol{x}_i \in R_j} \tilde{\gamma}_i (w_j - y_i)^2, \quad \text{and so} \qquad \hat{w}_j = \sum_{i:\boldsymbol{x}_i \in R_j} \tilde{\gamma}_i y_i.$$

2. Show that

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \widehat{R}_{\boldsymbol{\gamma}}(f_{\boldsymbol{w},\Pi}) = \sum_{j=1}^{d} \hat{\pi}_j \, h(D, \boldsymbol{\gamma}, R_j),$$

for $\hat{\pi}_j = \sum_{i=1}^n \gamma_i 1_{\{x_i \in R_j\}}$, and a function $h : (D, \boldsymbol{\gamma}, R) \mapsto h(D, \boldsymbol{\gamma}, R)$ to be defined.
Once we inject the form of $\hat{w}_j$ into the weighted empirical risk we obtain

$$\widehat{R}_{\boldsymbol{\gamma}}(f_{\hat{\boldsymbol{w}},\Pi}) = \sum_{j=1}^{d} \Gamma_j \sum_{i:\boldsymbol{x}_i \in R_j} \tilde{\gamma}_i \Big(y_i - \sum_{k:\boldsymbol{x}_k \in R_j} \tilde{\gamma}_k \, y_k \Big)^2.$$

We can identify $\hat{\pi}_j = \Gamma_j$ and

$$h(D, \boldsymbol{\gamma}, R) = \sum_{i:\boldsymbol{x}_i \in R} \frac{\gamma_i}{\sum_{i'=1}^n \gamma_{i'} 1_{\{\boldsymbol{x}_{i'} \in R\}}} \Big(y_i - \sum_{k:\boldsymbol{x}_k \in R} \frac{\gamma_k}{\sum_{i'=1}^n \gamma_{i'} 1_{\{\boldsymbol{x}_{i'} \in R\}}} y_k \Big)^2.$$

3. Show that $h(D, \boldsymbol{\gamma}, R_j)$ can be interpreted as the variance of a discrete distribution putting mass only over the set $\{y_1, \ldots, y_n\}$. Specify which one. We consider the probability distribution

$$P_{n,\gamma} = \sum_{i=1}^{n} \gamma_i \delta_{(\boldsymbol{x}_i, y_i)},$$

where $\delta_{(\boldsymbol{x}_i, y_i)}$ is the Dirac mass at $(\boldsymbol{x}_i, y_i)$. We can then check that we have $\Gamma_j = \mathbb{P}_{n,\gamma}(X \in R_j)$ and $\tilde{\gamma}_i = \mathbb{P}_{n,\gamma}(X = \boldsymbol{x}_i \mid X \in R_j)$, and as a consequence

$$\hat{w}_j = \sum_{i:\boldsymbol{x}_i \in R_j} \tilde{\gamma}_i \, y_i = \mathbb{E}_{n,\gamma}[Y | X \in R_j]$$

and

$$h(D, \boldsymbol{\gamma}, R) = \mathrm{Var}_{n,\gamma}(Y \mid X \in R) := \mathbb{E}_{n,\gamma}\big[\big(Y - \mathbb{E}_{n,\gamma}[Y | X \in R] \big)^2 \mid X \in R \big].$$

4. Following the same logic as for classical regression tree learning, what is the measure of impurity reduction that should be maximized to determine the next best split of a region R_j?
If we consider splitting a region R_j into two R_j^+ and R_j^-, let $\Gamma_j^+ = \sum_{i=1}^n \gamma_i 1_{\{\boldsymbol{x}_i \in R_j^+\}}$ and $\Gamma_j^- = \sum_{i=1}^n \gamma_i 1_{\{\boldsymbol{x}_i \in R_j^-\}}$, the decrease of the empirical risk after reminimization over $\boldsymbol{w}$ is

$$\Gamma_j^+ \, h(D, \boldsymbol{\gamma}, R_j^+) + \Gamma_j^- \, h(D, \boldsymbol{\gamma}, R_j^-) - \Gamma_j \, h(D, \boldsymbol{\gamma}, R_j).$$

This is the quantity that should be used a measure of impurity to find the best split.

Problem C (9 points)

We consider the logistic loss $\ell(a, y) = \log(1 + \exp(-a(2y - 1)))$, and an input-output pair (X, Y) following a distribution $P_{(X,Y)}$, with X and Y taking respectively values in $\mathbb{R}^p$ and $\{0, 1\}$. Let $\mathcal{R}(f) = \mathbb{E}[\ell(f(X), Y)]$ the associated risk of a decision function f. It is possible to show that the target function is the log odds

$$f^*(x) = \log \frac{\eta(x)}{1 - \eta(x)} \quad \text{where} \quad \eta(x) = \mathbb{P}(Y = 1 \mid X = x).$$

We assume that $P_{(X,Y)}$ is specified via the following quantities: the proportions of positives and negative examples are $\pi_1 := \mathbb{P}(Y = 1)$ and $\pi_0 := 1 - \pi_1$, and the probability density functions of the positive and negative examples are $h_1(x) := p(x \mid Y = 1)$ and $h_0(x) := p(x \mid Y = 0)$ respectively.

1. Explain why the logistic loss introduced in the exercise is equivalent to the form of the logistic loss seen in class.

   In class, we have seen a similar expression when the label took the values in $\{-1, 1\}$, while now the labels are 0 and 1. The transformation $\zeta : y \mapsto 2y - 1$ maps 1 to 1 and 0 to -1. With that transformation we recover the same expression as in the course.

2. Express $f^*(x)$ as a function of $\pi_0, \pi_1, h_0(x)$ and $h_1(x)$.

   By definition,

$$f^*(x) = \log \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)} = \log \frac{p(x|Y = 1)\mathbb{P}(Y = 1)}{p(x|Y = 0)\mathbb{P}(Y = 0)} = \log \left[\frac{\pi_1}{\pi_0} \frac{h_1(x)}{h_0(x)}\right].$$

3. If you have to assign labels to new data based on f^* how would you proceed?

   We have seen in class that $f^*(x)$ is positive if and only if $\mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x)$, i.e., if and only if the decision minimizing the 0-1 loss is $\hat{y} = 1$. So we would predict class 1 if and only if $f^*(x) > 0$ (or $f^*(x) \geq 0$).

4. We consider in the rest of this exercise the situation where the data from the two classes have the same distributions as before, but we only have partial information about the labels. More precisely, we assume that a certain fraction of the positive examples have been identified and labelled as such and that the labels of the rest are unknown. We further assume that the positive examples that are labelled are taken at random from the population of positive examples. So we have new labels $\tilde{y}_i$ where $\tilde{y}_i = 1$ implies $y_i = 1$ and $\tilde{y}_i = 0$ means "we don't know y_i". We call the two corresponding classes the labelled class and the unlabelled class. The proportions of the labelled and unlabelled classes are respectively $\pi_1' := \mathbb{P}(\tilde{Y} = 1) < \pi_1$ and $\pi_0' = 1 - \pi_1'$. And these two classes have the densities

$$p(x \mid \tilde{Y} = 1) = h_1(x) \quad \text{and} \quad p(x \mid \tilde{Y} = 0) = \frac{\pi_0}{\pi_0'} h_0(x) + \frac{\pi_1 - \pi_1'}{\pi_0'} h_1(x).$$

   Suppose that we decide to simply learn a classifier that tries to predict the new labels $\tilde{y}_i$ and that we still use the logistic loss. Express the new target function $\tilde{f}^*$ as a function of $\pi_0, \pi_1, \pi_1', h_0(x)$ and $h_1(x)$.

   As before,

$$\tilde{f}^*(x) = \log \frac{p(x|\tilde{Y} = 1)\, t\mathbb{P}(\tilde{Y} = 1)}{p(x|\tilde{Y} = 0)\, \mathbb{P}(\tilde{Y} = 0)} = \log \frac{\pi_1' h_1(x)}{\pi_0\, h_0(x) + (\pi_1 - \pi_1')\, h_1(x)}.$$

5. We define $\rho^*(x) := \exp(-f^*(x))$ and $\tilde{\rho}^*(x) := \exp(-\tilde{f}^*(x))$. Show that $\tilde{\rho}^*(x)$ is an affine function of $\rho^*(x)$. Deduce from this that, provided the fraction of positive data which has been labelled $\theta := \frac{\pi_1'}{\pi_1}$ is known, it is possible to solve the initial classification problem based on $\tilde{f}^*(x)$. Explain how.

   We have

$$\tilde{\rho}^*(x) = \frac{\pi_0\, h_0(x) + (\pi_1 - \pi_1')\, h_1(x)}{\pi_1' h_1(x)} = \frac{\pi_0\, h_0(x)}{\pi_1' h_1(x)} + \frac{\pi_1 - \pi_1'}{\pi_1'} = \frac{1}{\theta}\rho^*(x) + \frac{1}{\theta} - 1.$$

   Given that there is a linear relationship between $\tilde{\rho}^*$ and ρ^* with known coefficients, we can certainly compute one if we have the other. Also, note that the region corresponding to a positive prediction is $\rho^*(x) < 1$ which is equivalent to

$$\tilde{\rho}^*(x) < \frac{2}{\theta} - 1.$$

6. Based on the answer to the previous question, explain how you could use a logistic regression based on the training data $\tilde{D}_n := \{(x_1, \tilde{y}_1), \ldots, (x_n, \tilde{y}_n)\}$ to solve the initial classification problem. In particular, express as a function of θ and of the value of the decision function obtained from logistic regression, how you would assign a point to the positive or to the negative class.

   We would train a logistic regression model with $\hat{f}$ of the form $\hat{f}(x) = w^\top x + b$ from the $\tilde{D}_n$ and then predict class 1 if $e^{-\hat{f}(x)} < \frac{2}{\theta} - 1$.

7. Under which conditions would the approach you proposed in the previous question succeed in practice?

   Given the fact that the reasoning is based on relations between two target functions, the logistic regression would work well if it has chances to provide a reasonable approximation to the log-odds, which is assuming that the log-odds are relatively close to linear. As a remark, it would also be possible to use a kernelized version of logistic regression.

8. Would it be possible to solve the same problem by training a Random Forest estimator on $\tilde{D}_n$? Describe how you would do it or why it is not possible.

   The decision function learned by a random forest is also trying to approximate the target function associated either with the square loss (if Gini entropy is used) or with the log-loss (if the the Shannon entropy is used). In both cases, this target function is $\eta(x) = \mathbb{P}(Y = 1|X = x)$. We can thus get from it an approximation of $\tilde{f}^*$ or $\tilde{\rho}^*$ and assign the data to class one or zero based on the same threshold as for logistic regression.

   To add a further comment on whether it can be a good idea to use Random-Forest in this context: given that it is non-parametric and can possibly model very non-linear functions, RF can be a good choice, but the RF model makes it easy to have sharp transitions aligned with the axes, less easy for transitions in general position.

9. Would it be possible to solve the same problem by training a Nadaraya-Watson estimator on $\tilde{D}_n$? Describe how you would do it or why it is not possible.

   The decision function learned by the Nadaraya-Watson estimator is trying to approximate the target function associated either with the square loss. This target function is $\mathbb{E}[Y = 1|X = x] = \mathbb{P}(Y = 1|X = x) = \eta(x)$. We can thus get from it an approximation of $\tilde{f}^*$ or $\tilde{\rho}^*$ and assign the data to class one or zero based on the same threshold as for logistic regression.

   To add a further comment on whether it can be a good idea to use Nadaraya-Watson in this context: given that Nadarya-Watson is non-parametric it can flexibly approximate any decision function, but it will only work well if the input dimension is moderate and a lot of data are available.

Problem D (9 points)

1. Compute $\int_0^\infty 1_{\{t\le a\}} 1_{\{t\le b\}} dt$.

   This is equal to $\displaystyle\int_0^\infty 1_{\{t\le \min(a,b)\}} dt = \min(a,b)$.

2. Show that the function $K : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined by $K(x,y) = \min(x,y)$ is the reproducing kernel of an RKHS $\mathcal{H}$.

   It suffices to show that K is a positive definite function, i.e., that for any collection $x_1, \dots, x_n \in \mathbb{R}_+$ and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, we have $\sum_{1\le i,j\le n} \alpha_i \alpha_j K(x_i, x_j) \ge 0$. But

$$\sum_{1\le i,j\le n} \alpha_i\alpha_j K(x_i,x_j) = \int_0^\infty \sum_{1\le i,j\le n} \alpha_i\alpha_j 1_{\{t\le x_i\}} 1_{\{t\le x_j\}} dt = \int_0^\infty \Big(\sum_{i=1}^n \alpha_i 1_{\{t\le x_i\}}\Big)^2 dt \ge 0.$$

3. Show that for any function $f \in \mathcal{H}$ such that $\|f\|_{\mathcal{H}} \le c$, we have $|f(x) - f(y)| \le c\sqrt{|x-y|}$.

   We have seen in class that by Cauchy-Schwarz

$$|f(x) - f(y)| = \langle K(x,\cdot) - K(y,\cdot), f\rangle_{\mathcal{H}} \le \|f\|_{\mathcal{H}} \|K(x,\cdot) - K(y,\cdot)\|_{\mathcal{H}},$$

   but

$$\|K(x,\cdot) - K(y,\cdot)\|_{\mathcal{H}}^2 = K(x,x) + K(y,y) - 2K(x,y) = x + y - 2\min(x,y) = |x-y|,$$

   because $x + y = \max(x,y) + \min(x,y)$. Hence the result.

4. Deduce from the previous questions that non-zero linear functions on $\mathbb{R}_+$ do not belong to $\mathcal{H}$.

   A linear function is of the form $f(x) = ax$. If the function was in $\mathcal{H}$ then there would have to be a constant c such that $|a|x = |f(x) - f(0)| \le c\sqrt{x}$ but this is not possible.

5. Show that for all $a, h \in \mathbb{R}_+$, the function $x \mapsto (x-a)1_{\{a\le x<a+h\}} + h\,1_{\{x\ge a+h\}}$ belongs to $\mathcal{H}$. The functions $K(a,\cdot)$ and $K(a+h,\cdot)$ belong to $\mathcal{H}$ by construction. And since $\mathcal{H}$ is a vector space, so does $K(a+h,\cdot) - K(a,\cdot)$. It is easy to check that on $[0,a]$, $(a,a+h]$, and $(a+h,\infty)$ it is equal to the proposed function.

6. Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a continuous piecewise linear function, such that $f(0) = 0$, with exactly d breakpoints $x_1 < \dots < x_d$ and such that f is constant for all $x > x_d$. Show that there exists $\alpha, \beta \in \mathbb{R}$ such that $f + \alpha K(x_{d-1},\cdot) + \beta K(x_d,\cdot)$ has exactly $d-1$ breakpoints.

Defining

$$\Delta_d(\cdot) := \frac{f(x_d) - f(x_{d-1})}{x_d - x_{d-1}}\big[K(x_d, \cdot) - K(x_{d-1}, \cdot)\big],$$

the function $\Delta_d(\cdot) + f(x_{d-1})$, is exactly equal to f on $[x_{d-1}, \infty)$. As a consequence the function $f_{d-1} := f - \Delta_d$ is a piecewise linear function with $d-1$ pieces. Furthermore, note that if $f(0) = 0$ then $f_{d-1}(0) = 0$.

7. Let f be as in the previous question. Show that f belongs to $\mathcal{H}$, and express it as a linear combination of the functions $K(x_j, \cdot)$, for $1 \leq j \leq d$.

   Since we have a piecewise linear function with one less piece, we can recursively remove one after the other all the pieces of f which shows that

$$f = \sum_{j=1}^{d} \frac{f(x_j) - f(x_{j-1})}{x_j - x_{j-1}}\big[K(x_j, \cdot) - K(x_{j-1}, \cdot)\big],$$

   with $x_0 := 0$. This proves that $f \in \mathcal{H}$.

8. Show that, if $x \leq y \leq z \leq t$, then $\langle K(t, \cdot) - K(z, \cdot), K(y, \cdot) - K(x, \cdot)\rangle_{\mathcal{H}} = 0$.

   The above quantity is equal to

$$\begin{aligned}
&\langle K(t, \cdot) - K(z, \cdot), K(y, \cdot)\rangle_{\mathcal{H}} - \langle K(t, \cdot) - K(z, \cdot), K(x, \cdot)\rangle_{\mathcal{H}} \\
=&\min(t, y) - \min(z, y) - (\min(t, x) - \min(z, x)) \\
=&y - y - (x - x) = 0.
\end{aligned}$$

9. Show that, if $f \in \mathcal{H}$ is a continuous piecewise linear function with a finite number of breakpoints, then we have $\|f\|_{\mathcal{H}}^2 = \int_0^\infty (f'(x))^2 dx$, where f' is the derivative of f (which is defined everywhere except at a finite number of points).

Let $c_j = \frac{f(x_j) - f(x_{j-1})}{x_j - x_{j-1}}$. Given what was proven in the previous question we have

$$\|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{d} c_j^2 \|K(x_j, \cdot) - K(x_{j-1}, \cdot)\|_{\mathcal{H}}^2 = \sum_{j=1}^{d} c_j^2 (x_j - x_{j-1})$$

because in the solution of question 3, we have shown that $\|K(x_j, \cdot) - K(x_{j-1}, \cdot)\|_{\mathcal{H}}^2 = |x_j - x_{j-1}|$ and $x_j > x_{j-1}$. Now $\forall x \in (x_{j-1}, x_j)$, the derivative f' is well defined and $f'(x) = c_j$, so

$$c_j^2 (x_j - x_{j-1}) = \int_{x_{j-1}}^{x_j} c_j^2 dx = \int_{x_{j-1}}^{x_j} f'(x)^2 dx,$$

and

$$\|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{d} \int_{x_{j-1}}^{x_j} f'(x)^2 dx = \int_0^{x_d} f'(x)^2 dx = \int_0^{\infty} f'(x)^2 dx,$$

where the last equality is because $f'(x) = 0$, for all $x > x_d$.