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.
2. For which value of p is this expected misclassification error the smallest?
3. Assuming that the dataset is rather large, what is the precision approximately equal to ? Provide a mathematical justification.

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 $\widehat{R}_{\boldsymbol{\gamma}}(f_{\boldsymbol{w},\Pi})$?
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.
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.
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?

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.

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

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

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 $h_1(x)$.

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.

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.

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

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.

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.

Problem D (9 points)

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

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 $\mathcal{H}$.

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

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

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 $\mathcal{H}$.

6. Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a continuous piecewise linear function, such that $f(0) = 0$, with exactly d breakpoints $x_1 < \ldots < x_d$ 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.

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 $1 \le j \le d$.

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

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).