INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

IML

Exam 2019/2020

4^{th} year

G. Gasso, B. Gauzere

- Duration: 2h,
- *Access to handouts and notes is granted*

1 Quantile regression (8 points)

Given a set of samples $\mathcal{D} = \left\{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\right\}_{i=1}^n$ our goal is to learn a linear regression model $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$, with $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, relating an input $\mathbf{x}_i$ to its output y_i. More specifically we aim to estimate a function f robust to outliers (abnormal data) by solving the problem

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \ell_\tau\left(y_i - (\mathbf{w}^\top \mathbf{x}_i + b)\right) \tag{1}$$

where the loss function ℓ_τ is defined as $\ell_\tau(z) = \begin{cases} \tau z & \text{if} \quad z \geq 0 \\ -(1-\tau)z & \text{if} \quad z < 0 \end{cases}$ with $\tau \in [0, 1]$.

Notice that for a random variable $z \in \mathbb{R}$, its τ-quantile is $\mu_\tau = \inf_\mu\{Pr(z \leq \mu)\} = \tau = \operatorname{argmin}_\mu \mathbb{E}\text{sp}\, \ell_\tau(z - \mu)$. The median is for instance the 0.5-quantile and is known to be robust to outliers contrary to the mean estimator.

1. Plot the loss function $\ell_\tau(z)$ for $\tau = 0.1$ and $\tau = 0.5$. Are these functions derivable?

To solve Problem (1), we rather consider the following equivalent formulation

$$\begin{aligned}
\min_{\mathbf{w},b,\{\xi_i\}_{i=1}^n,\{\gamma_i\}_{i=1}^n} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n (\tau\xi_i + (1-\tau)\gamma_i) \\
\text{subject to} \quad & y_i - (\mathbf{w}^\top \mathbf{x}_i + b) \leq \xi_i \quad \forall i = 1, n \\
& (\mathbf{w}^\top \mathbf{x}_i + b) - y_i \leq \gamma_i \quad \forall i = 1, n \\
& \xi_i \geq 0 \quad \forall i = 1, n \\
& \gamma_i \geq 0, \quad \forall i = 1, n
\end{aligned}$$

with $C > 0$, the regularization parameter. ξ_i and γ_i, $i = 1, n$ are the slack variables.

2. How many constraint does the problem have ? Express the lagrangian of this problem.
3. Write the KKT stationary optimality conditions with relation to the primal variables $\mathbf{w}$, b, ξ_k and γ_k, $k = 1, n$. What is the expression of $\mathbf{w}$?
4. Derive the corresponding dual problem. Propose a way to solve it.

2 Bayes's decision rule (7 points)

Consider a binary classification problem with classes $\mathcal{C}_0$ and $\mathcal{C}_1$. The $\mathcal{C}_k, k = 0, 1$ are characterized by the respective conditional densities:

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{\theta_k e^{-x}}{(1+e^{-x})^{\theta_k+1}}, \qquad \text{with } x \in \mathbb{R}. \tag{2}$$

$\theta_k, k = 0$ and $k = 1$ are respectively the parameters of the density functions $p(\mathbf{x}|\mathcal{C}_0)$ and $p(\mathbf{x}|\mathcal{C}_1)$.

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R} \times \{0, 1\}\}_{i=1}^{n}$ be the training dataset. $\mathcal{D}$ includes n_0 samples from class $\mathcal{C}_0$ and n_1 points from the second class.

1. Give an estimation of $\mathbb{P}(\mathcal{C}_k)$ the prior probability of $\mathcal{C}_k, k = 0, 1$.
2. We want to estimate the parameters $\theta_k \in \mathbb{R}, k = 0, 1$ of the conditional densities by maximizing the likelihood of the training data. For each class $\mathcal{C}_k, k = 0, 1$:
   (a) Give the expression of the log-likelihood.
   (b) Deduce the estimation of θ_k by maximum likelihood estimation.

We want to design a discrimination function of the samples using Bayesian theory. The cost of a good decision is 0 and a bad decision costs λ_s.

3. Give the expression of the conditionals risks $R(\mathcal{C}_k/x), k = 0, 1$.
4. Deduce that the minimum risk is attained by deciding $\mathcal{C}_k$ if $\mathbb{P}(\mathcal{C}_k|x) > \mathbb{P}(\mathcal{C}_\ell|x)\ \ \forall \ell \neq k$.
5. Give the explicit expression of the decision function knowing that the $p(x|\mathcal{C}_k)$ are given by Equation (2).

Student name : ____________________

3 Learning principles (5 points)

For this exercise, mark your answers ON THE EXAM ITSELF. Fill in the bubbles that represent the best answer(s) to the question.

1. Let $k(\mathbf{x}_i, \mathbf{x}_j) = \exp^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$ a gaussian kernel. Let the samples $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ such that $\mathbf{x}_1$ is geometrically close to $\mathbf{x}_2$ and far from $\mathbf{x}_3$. What will be the value of $k(\mathbf{x}_2, \mathbf{x}_1)$ and $k(\mathbf{x}_3, \mathbf{x}_1)$?
   - ◯ $k(\mathbf{x}_2, \mathbf{x}_1)$ is close to 0 and $k(\mathbf{x}_3, \mathbf{x}_1)$ close to 1
   - ◯ $k(\mathbf{x}_2, \mathbf{x}_1)$ is close to 1 and $k(\mathbf{x}_3, \mathbf{x}_1)$ close to 0
   - ◯ $k(\mathbf{x}_2, \mathbf{x}_1) > 1$ and $k(\mathbf{x}_3, \mathbf{x}_1) < 0$
   - ◯ $k(\mathbf{x}_2, \mathbf{x}_1) < 0$ and $k(\mathbf{x}_3, \mathbf{x}_1) > 1$

2. Let a classification problem. We design two positive definite kernel functions k_0 and k_1. Under which condition their combination $k = a_0 k_0 + a_1 k_1$ is a positive definite kernel?
   - ◯ a_0 is negative and a_1 is positive
   - ◯ a_0 and a_1 are both positive
   - ◯ a_0 is positive and a_1 is negative
   - ◯ a_0 and a_1 are both negative

3. In a non-linear SVM for classification the kernel function is used to
   - ◯ reduce the dimension of the inputs
   - ◯ expand the dimension for a better separability of the inputs
   - ◯ measure the similarity between a point $\mathbf{x}_i$ and its label y_i
   - ◯ to find a non-linear decision function

4. To estimate the best regularized logistic model, you test the following regularization parameters $C = 10^{-2}, 10^{-1}, 1, 10, 10^2$. How many different logistic models do you need to train if you implement a 10-fold validation?

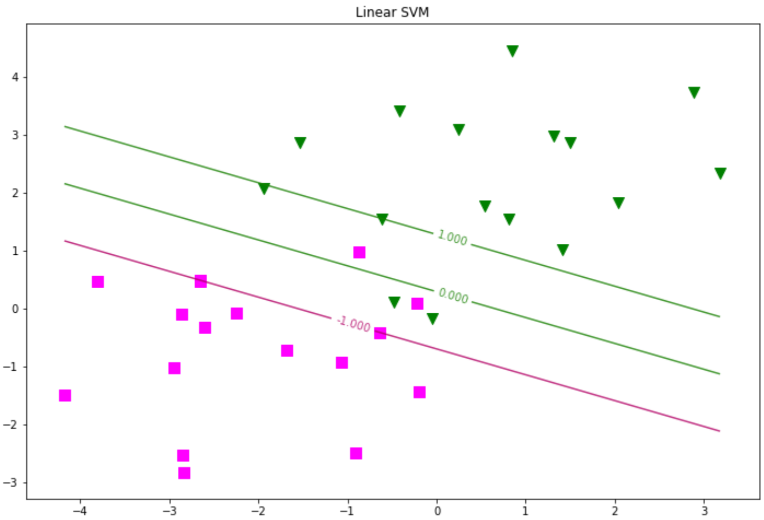
   ◯ 5, ◯ 10, ◯ 25, ◯ 50

σ^2	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}
Train error	35.12	27.18	9.71	2.35	0.00
Validation error	37.52	30.15	11.43	7.93	12.13

5. The model selection procedure of a SVM with gaussian kernel leads to the results in the table above. The optimal value of σ^2 is

   ◯ 10^{-1}, ◯ 10^{-2}, ◯ 10^{-3}, ◯ 10^{-4}.

6. The figure below corresponds to a linear SVM trained with $C = 1$. Mark on the figure the support vectors.



7. t-SNE and PCA are

   ◯ supervised methods for dimension reduction
   ◯ linear approaches for classification
   ◯ unsupervised methods for dimension reduction

8. $k-$means clustering method does not require to specify the number of clusters

   ◯ False ◯ True