Introduction to Machine Learning (GIF-7015)
Département de génie électrique et de génie informatique
Fall 2022

EXAM

Instructions: – Identify yourself on the titlepage;
– Provide your answers directly in the question sheet;
– One double-sided handwritten cheatsheet is allowed;
– Exam duration: 1 h 50.

Weighting: This exam weight for 20% of the final grade.

Firstname: ____________________

Lastname: ____________________

NI: ____________________

Signature: ____________________

GIF-7015

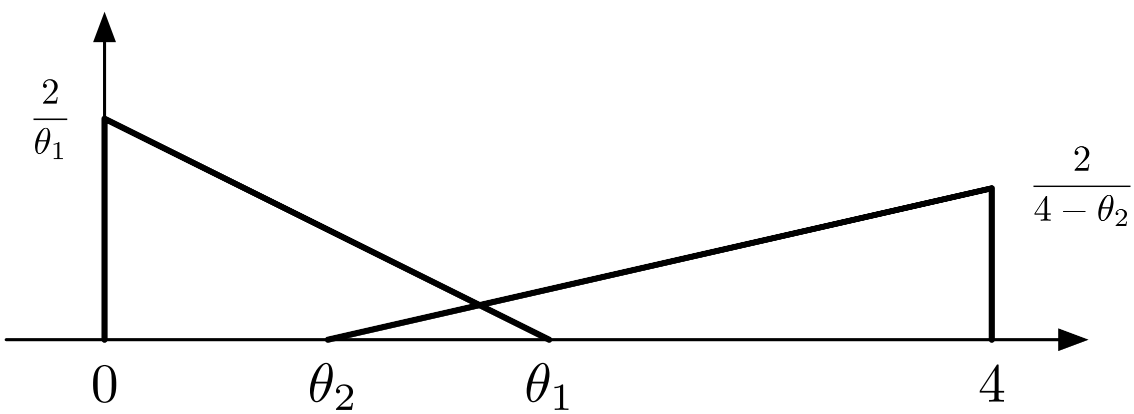
Question 1 (24 points over 100)

Let us consider a two-class parametric classification system with one input variable. The modelling of the distributions for each class is given by the following equations:

$$p(x|C_1) = \begin{cases} \frac{-2\,(x-\theta_1)}{(\theta_1)^2} & \text{if } x \in [0, \theta_1] \\ 0 & \text{otherwise} \end{cases} ,$$

$$p(x|C_2) = \begin{cases} \frac{2\,(x-\theta_2)}{(4-\theta_2)^2} & \text{if } x \in [\theta_2, 4] \\ 0 & \text{otherwise} \end{cases} .$$

Thus, the parameterization of the distribution of the class C_1 is given by θ_1, while that of the class C_2 is given by θ_2. It is also assumed that $0 \leq \theta_2 \leq \theta_1 \leq 4$. The following figure shows the plot of these class distributions.



(8 pts) (a) Suppose $\theta_1 = 3$ and $\theta_2 = 2$, give the function $\mathrm{h}(x)$ corresponding to the decision function for the classification of data according to the value of $x \in [0, 4]$. Assume that the prior probabilities of the classes are equal, i.e. $P(C_1) = P(C_2) = 0.5$. Also assume equal loss for the different types of errors. Give the developments leading to your decision function.

(8 pts) (b) Compute the optimal Bayesian error rate that is obtained with the classifier computed in the previous point. The optimal Bayesian error rate is the error rate obtained when the classified data perfectly follow the estimated distributions for classification.

(8 pts) (c) Let us now assume that the loss function is variable depending on the type of error our classifier makes. More precisely, if a particular instance is classified as being in class C_2 but actually belongs to class C_1, the loss is $\mathcal{L}(\alpha_2,C_1) = 1$, whereas the loss for an instance classified as being in class C_1, but actually belonging to class C_2 is $\mathcal{L}(\alpha_1,C_2) = 0.5$. Compute the new function $\mathrm{h}(x)$ corresponding to the decision function for data classification according to this loss function in the domain $x \in [0, 4]$. Assume that the other parameters are the same as in the previous points, i.e. $\theta_1 = 3$, $\theta_2 = 2$ and $P(C_1) = P(C_2) = 0.5$. Give the developments leading to your decision function.

Question 2 (32 points over 100)

Let us consider an RBF neural network for two classes, composed of a hidden layer of R Gaussian neurons, followed by an output layer of a neuron with linear transfer function. The output value for such a neural network for an input value $\mathbf{x}$ is given by the following equation,

$$\mathrm{h}(\mathbf{x}) = \sum_{i=1}^{R} w_i \phi_i(\mathbf{x}) + w_0 = \sum_{i=1}^{R} w_i \exp\left[-\frac{\|\mathbf{x}-\mathbf{m}_i\|^2}{2s_i^2}\right] + w_0,$$

where:

- $\mathbf{m}_i$ is the value of the centre of the i-th Gaussian neuron of the hidden layer;
- s_i is the spread of the i-th Gaussian neuron;
- w_i is the weight connecting the i-th Gaussian neuron of the hidden layer to the output neuron;
- w_0 is the bias weight of the output neuron.

Suppose we set the spread s_i to predetermined values and want to learn the values w_i, w_0 and $\mathbf{m}_i$ by gradient descent, using the mean square error as a criterion,

$$E = \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} (e^t)^2 = \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} [r^t - \mathrm{h}(\mathbf{x}^t)]^2,$$

where:

- r^t is the desired value for the output neuron of the network;
- $\mathcal{X}$ is the set of N training data.

(16 pts) (a) Develop the equations to update the weights w_i and w_0 of the output neuron by gradient descent, using the mean square error criterion.

(16 pts) (b) Develop the equations to update the values of the $\mathbf{m}_i$ centres of the hidden layer Gaussian neurons by gradient descent, using the mean square error criterion.

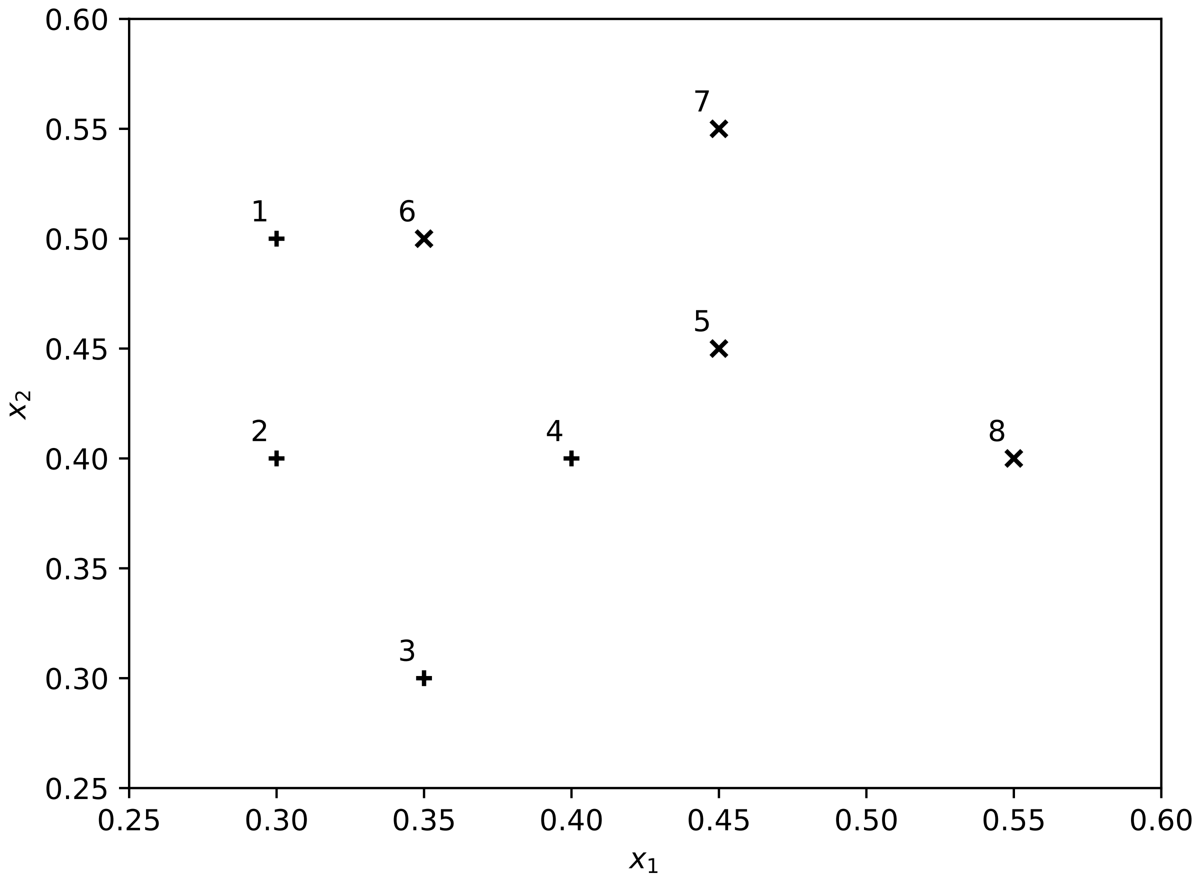
Question 3 (24 points over 100)

Let us consider the following data set, in two dimensions:

$$\mathbf{x}^1 = [0.3\ 0.5]^\top, \quad \mathbf{x}^2 = [0.3\ 0.4]^\top, \quad \mathbf{x}^3 = [0.35\ 0.3]^\top, \quad \mathbf{x}^4 = [0.4\ 0.4]^\top,$$
$$\mathbf{x}^5 = [0.45\ 0.45]^\top, \quad \mathbf{x}^6 = [0.35\ 0.5]^\top, \quad \mathbf{x}^7 = [0.45\ 0.55]^\top, \quad \mathbf{x}^8 = [0.55\ 0.4]^\top.$$

The labels for these data are $r^1 = r^2 = r^3 = r^4 = -1$ and $r^5 = r^6 = r^7 = r^8 = 1$.

The graph below shows the plot of these data.



We obtain the following result by training a linear SVM with **soft margin** with these data, using as regularization parameter value $C = 200$:

$\alpha^1 = 180, \quad \alpha^2 = 0, \quad \alpha^3 = 0, \quad \alpha^4 = 200, \quad \alpha^5 = 180, \quad \alpha^6 = 200, \quad \alpha^7 = 0, \quad \alpha^8 = 0,$
$w_0 = -11.6.$

(8 pts) (a) Compute the values of the vector $\mathbf{w}$ of the separating hyperplane of this classifier.

(8 pts) (b) Determine which data instances are support vectors as well as which instances are in the margins or misclassified.

(8 pts) (c) Now suppose that we want to process a data $\mathbf{x} = [0.37\ 0.45]^\top$ with this SVM. Compute the corresponding $\mathrm{h}(\mathbf{x})$ value (real value before thresholding the output).

Question 4 (20 points over 100)

Using the data from the previous question (question 3), in two dimensions, answer the following questions.

(10 pts) (a) Compute the classification error rate using a *leave-one-out* approach with a k-nearest neighbour classifier, using $k = 1$ neighbours and the distance D_∞. Give details on the procedure leading to the calculation of the error rate.

(10 pts) (b) Perform Wilson editing of this dataset, using one neighbour ($k = 1$) and a Euclidean distance. Process the data in their index order, i.e., in the order x^1,x^2,x^3, . . . ,x^8. Explain your approach and report the data making up the prototype set after editing.