**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

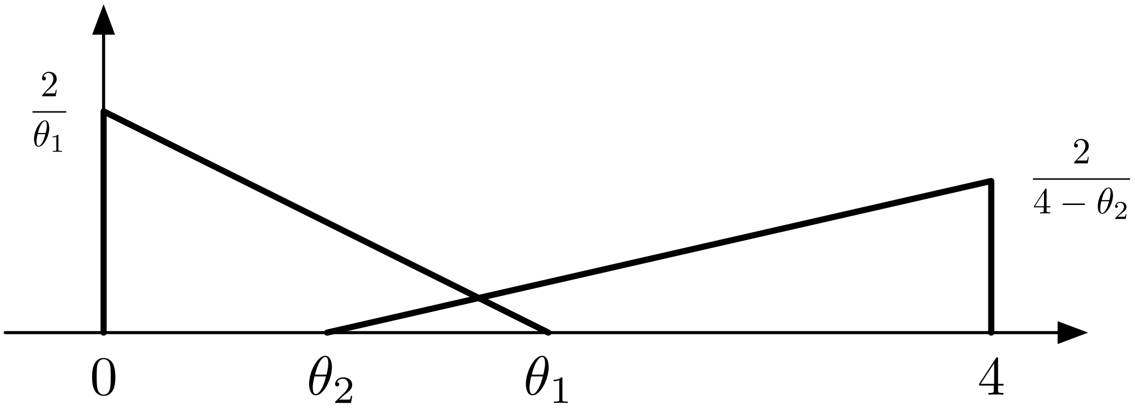
# Question 1 (24 points over 100)

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 $\theta_1$, while that of the class $C_2$ is given by $\theta_2$. 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.

**Solution:** The decision is made according to the maximum value of the classification posterior probabilities, that is:

$$\mathrm{h}(x) = \operatorname*{argmax}_{C_i \in \{C_1, C_2\}} P(C_i|x).$$

Since the evidences $p(x)$ and the prior probabilities are the same for both classes, the decision can be made directly from the class likelihoods, that is:

$$\mathrm{h}(x) = \operatorname*{argmax}_{C_i \in \{C_1, C_2\}} p(x|C_i).$$

Since the class likelihoods are linear functions, it is sufficient to determine the point where the two distributions are equal in the interval $[\theta_2,\theta_1]$:

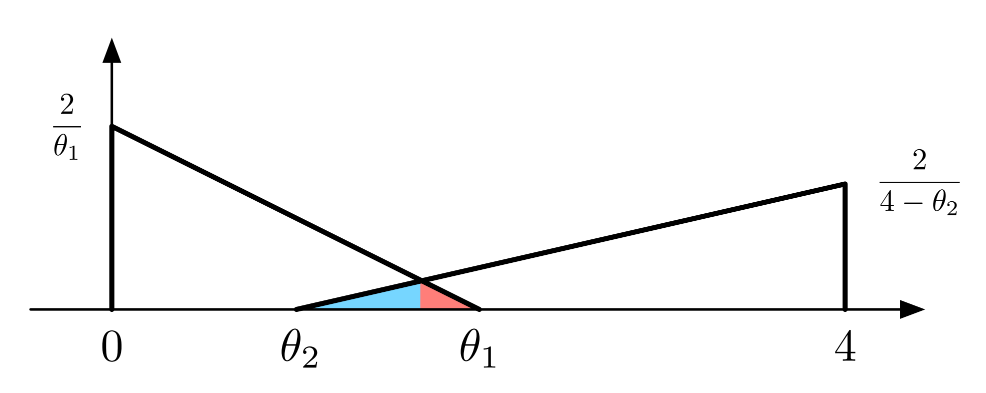
$$\begin{aligned}
p(x|C_1) &= p(x|C_2), \\
\frac{-2\,(x-\theta_1)}{(\theta_1)^2} &= \frac{2\,(x-\theta_2)}{(4-\theta_2)^2}, \\
\frac{-2\,(x-3)}{(3)^2} &= \frac{2\,(x-2)}{(4-2)^2}, \\
\frac{-x+3}{9} &= \frac{x-2}{4}, \\
(0.111+0.25)x &= (0.333+0.5), \\
x &= 2.3077.
\end{aligned}$$

So, based on the figure in the question statement, we get this as the decision making function:

$$\mathrm{h}(x) = \begin{cases} C_1 & \text{for } x \in [0, 2.3077] \\ C_2 & \text{for } x \in [2.3077, 4] \end{cases}.$$

(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.

**Solution:** Errors occur when a data item in class $C_2$ has a value $x < 2.3077$ or a data item in class $C_1$ has a value $x > 2.3077$. The following figure shows the given distributions according to the classes, with the regions of the distributions in red and blue where a classification according to these distributions results in an error.



So, to estimate the classification error in this case, we need to calculate the area of the distributions where a data will be misclassified. In the case of class $C_1$, the error corresponds to the red triangle. First, we need to calculate the height of this triangle:

$$H_1 = p(x = 2.3077|C_1) = \frac{-2\,(x-\theta_1)}{(\theta_1)^2} = \frac{-2\,(2.3077-3)}{3^2} = 0.15384.$$

Next, the length of the triangle is calculated as

$$L_1 = \theta_1 - 2.3077 = 3 - 2.3077 = 0.6923.$$

The area of a right triangle is then calculated as the product of the length and height of the triangle divided by two:

$$A_1 = \frac{H_1 \times L_1}{2} = \frac{0.15384 \times 0.6923}{2} = 0.053252.$$

Similarly, for the class $C_2$ the error corresponds to the blue triangle in the figure and is calculated as follows:

$$\begin{aligned} H_2 &= p(x = 2.3077|C_2) = \frac{2\,(x-\theta_2)}{(4-\theta_2)^2} = \frac{2\,(2.3077-2)}{(4-2)^2} = 0.15384, \\ L_2 &= 2.3077 - \theta_2 = 2.3077 - 2 = 0.3077, \\ A_2 &= \frac{H_2 \times L_2}{2} = \frac{0.15384 \times 0.3077}{2} = 0.023668. \end{aligned}$$

Therefore, the total error is equal to the sum of the two areas multiplied by their respective a priori probabilities, i.e.:

$$E = P(C_1)\,A_1 + P(C_2)\,A_2 = 0.5 \times 0.053252 + 0.5 \times 0.023668 = 0.03846.$$

The optimal Bayesian error rate is therefore 3.85 %.

(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.

**Solution:** With a loss function, the decision making is based on minimizing the classification risk:

$$\mathrm{h}(x) = \underset{C_i \in \{C_1,C_2\}}{\operatorname{argmin}} R(C_i|x),$$

where:

$$R(C_i|x) = \sum_{C_j \in \{C_1,C_2\}} \mathcal{L}(C_i,C_j)\, P(C_j|x).$$

Since $p(x)$ does not change by class and $P(C_1) = P(C_2)$, we can simplify the risk by:

$$R(C_i|x) = \sum_{C_j \in \{C_1,C_2\}} \mathcal{L}(C_i,C_j)\, p(x|C_j).$$

In this case, the risks for classes $C_1$ and $C_2$ are thus:

$$\begin{aligned} R(C_1|x) &= \mathcal{L}(\alpha_1,C_2)\, p(x|C_2) = 0{,}5\, p(x|C_2), \\ R(C_2|x) &= \mathcal{L}(\alpha_2,C_1)\, p(x|C_1) = p(x|C_1). \end{aligned}$$

Since the functions $p(x|C_1)$ and $p(x|C_2)$ are linear equations, which implies that $R(C_1|x)$ and $R(C_2|x)$ are also linear equations, it is sufficient to determine the point where the two lines $R(C_1|x)$ and $R(C_2|x)$ intersect:

$$\begin{aligned} R(C_1|x) &= R(C_2|x), \\ 0{,}5\, p(x|C_2) &= p(x|C_1), \\ 0{,}5 \frac{2\,(x-\theta_2)}{(4-\theta_2)^2} &= \frac{-2\,(x-\theta_1)}{(\theta_1)^2}, \\ 0{,}5 \frac{2x-4}{(4-2)^2} &= \frac{-2x+6}{(3)^2}, \\ \frac{x}{4} + \frac{2x}{9} &= \frac{1}{2} + \frac{6}{9}, \\ 0{,}47222\, x &= 1{,}16667, \\ x &= 2{,}4706. \end{aligned}$$

So, based on the figure in the question statement, we get as a decision making function this:

$$\mathrm{h}(x) = \begin{cases} C_1 & \text{pour } x \in [0, 2{,}4706] \\ C_2 & \text{pour } x \in [2{,}4706, 4] \end{cases}.$$

# 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.

**Solution:**

$$\begin{aligned}
e^t &= r^t - \mathrm{h}(\mathbf{x}^t) = r^t - \left[\sum_{j=1}^{R} w_j \phi_j(\mathbf{x}^t) + w_0\right] \\
\frac{\partial E}{\partial w_i} &= \frac{\partial}{\partial w_i} \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} (e^t)^2 = \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} 2e^t \frac{\partial}{\partial w_i} \left( r^t - \left[\sum_{j=1}^{R} w_j \phi_j(\mathbf{x}^t) + w_0\right]\right) \\
&= -\frac{1}{N} \sum_{\mathbf{x}^t \in \mathcal{X}} e^t \phi_i(\mathbf{x}^t) \\
\Delta w_i &= -\eta \frac{\partial E}{\partial w_i} = \frac{\eta}{N} \sum_{\mathbf{x}^t \in \mathcal{X}} e^t \phi_i(\mathbf{x}^t) \\
\frac{\partial E}{\partial w_0} &= \frac{\partial}{\partial w_0} \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} (e^t)^2 = \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} 2e^t \frac{\partial}{\partial w_0} \left( r^t - \left[\sum_{j=1}^{R} w_j \phi_j(\mathbf{x}^t) + w_0\right]\right) \\
&= -\frac{1}{N} \sum_{\mathbf{x}^t \in \mathcal{X}} e^t \\
\Delta w_0 &= -\eta \frac{\partial E}{\partial w_0} = \frac{\eta}{N} \sum_{\mathbf{x}^t \in \mathcal{X}} e^t \\
w_i &= w_i + \Delta w_i, \quad i = 0, \ldots, R
\end{aligned}$$

(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.

**Solution:**

$$\begin{aligned}
\frac{\partial \phi_i(\mathbf{x}^t)}{\partial m_{i,j}} &= \frac{\partial}{\partial m_{i,j}} \exp\left[-\frac{\|\mathbf{x}^t-\mathbf{m}_i\|^2}{2s_i^2}\right] \\
&= \exp\left[-\frac{\|\mathbf{x}^t-\mathbf{m}_i\|^2}{2s_i^2}\right] \frac{\partial}{\partial m_{i,j}} \left[-\frac{\|\mathbf{x}^t-\mathbf{m}_i\|^2}{2s_i^2}\right] \\
&= \frac{(x_j^t - m_{i,j})}{s_i^2} \exp\left[-\frac{\|\mathbf{x}^t-\mathbf{m}_i\|^2}{2s_i^2}\right] = \frac{x_j^t - m_{i,j}}{s_i^2}\phi_i(\mathbf{x}^t) \\
\frac{\partial E}{\partial m_{i,j}} &= \frac{\partial}{\partial m_{i,j}} \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} (e^t)^2 = \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} 2e^t \frac{\partial}{\partial m_{i,j}} \left( r^t - \left[ \sum_{l=1}^{R} w_l \phi_l(\mathbf{x}^t) + w_0 \right] \right) \\
&= \frac{1}{N} \sum_{\mathbf{x}^t \in \mathcal{X}} e^t (-1) w_i \frac{\partial \phi_i(\mathbf{x}^t)}{\partial m_{i,j}} = -\frac{w_i}{N s_i^2} \sum_{\mathbf{x}^t \in \mathcal{X}} e^t (x_j^t - m_{i,j}) \phi_i(\mathbf{x}^t) \\
\Delta m_{i,j} &= -\eta \frac{\partial E}{\partial m_{i,j}} = \frac{\eta w_i}{N s_i^2} \sum_{\mathbf{x}^t \in \mathcal{X}} e^t (x_j^t - m_{i,j}) \phi_i(\mathbf{x}^t) \\
m_{i,j} &= m_{i,j} + \Delta m_{i,j}, \quad i = 1, \ldots, R, \quad j = 1, \ldots, D
\end{aligned}$$

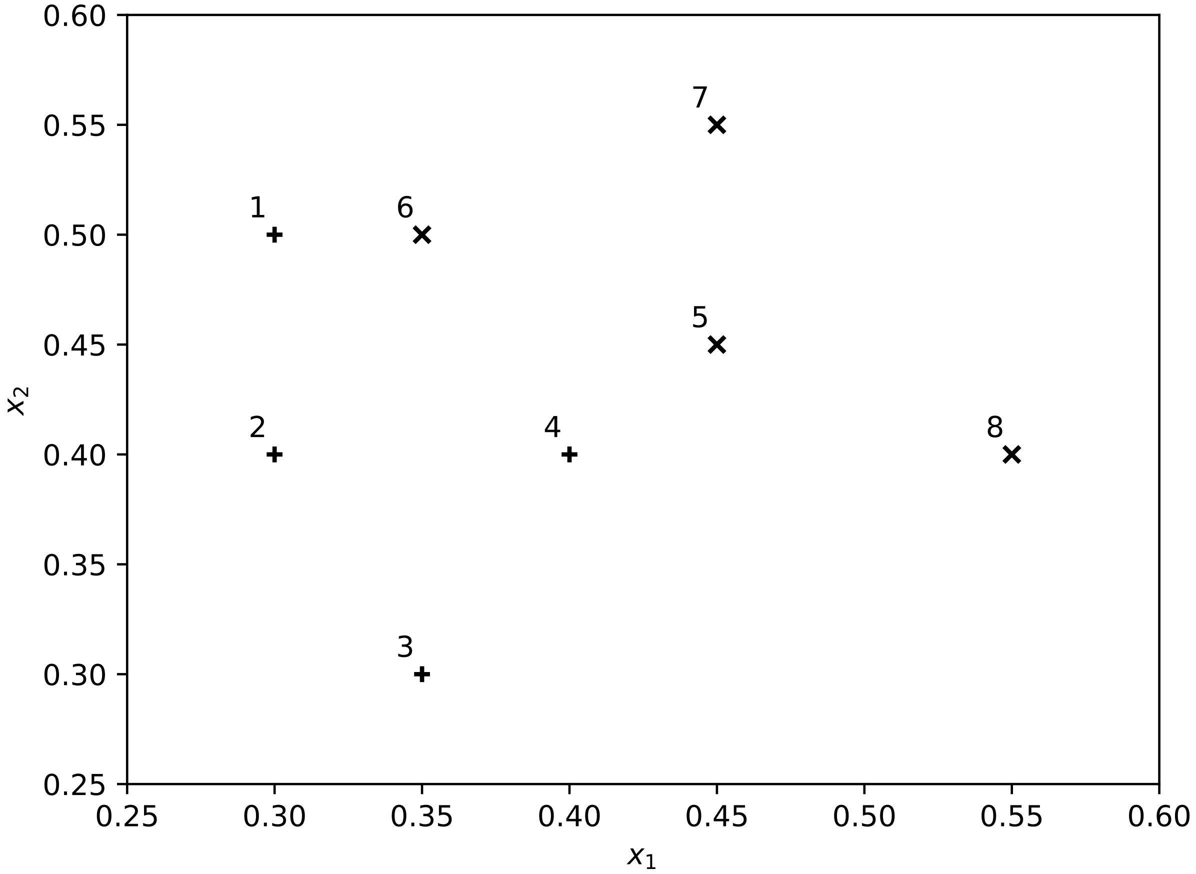
# Question 3 (24 points over 100)

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

$$\begin{array}{llll} \mathbf{x}^1 = [0.3\;0.5]^\top, & \mathbf{x}^2 = [0.3\;0.4]^\top, & \mathbf{x}^3 = [0.35\;0.3]^\top, & \mathbf{x}^4 = [0.4\;0.4]^\top, \\ \mathbf{x}^5 = [0.45\;0.45]^\top, & \mathbf{x}^6 = [0.35\;0.5]^\top, & \mathbf{x}^7 = [0.45\;0.55]^\top, & \mathbf{x}^8 = [0.55\;0.4]^\top. \end{array}$$

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.

**Solution:** The values of the vector $\mathbf{w}$ are calculated according to the following equation:

$$\mathbf{w} = \sum_t \alpha^t \, r^t \, \mathbf{x}^t.$$

In this case, the values of the vector are $\mathbf{w} = [17\ 11]^\top$.

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

**Solution:** The data instances $\mathbf{x}^1$, $\mathbf{x}^4$, $\mathbf{x}^5$ and $\mathbf{x}^{\mathbf{6}}$ represent the support vectors of the classifier, as their respective $\alpha^t$ is non-zero.

Instances in the margin or misclassified have a value of $\alpha^t$ corresponding to the regularization parameter $C$. Thus, the data $\mathbf{x}^4$ and $\mathbf{x}^6$ are in the margin or misclassified, with $\alpha^4 = \alpha^6 = C = 200$.

(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).

**Solution:** We calculate the value of $\mathrm{h}(\mathbf{x})$ according to the following equation:

$$\mathrm{h}(\mathbf{x}) = \sum_t \alpha^t r^t (\mathbf{x}^t)^\top \mathbf{x} + w_0.$$

In this case, with $\mathbf{x} = [0.37\ 0.45]^\top$, the corresponding output of the classifier is $\mathrm{h}(\mathbf{x}) = -0.36$. Thus, the data is assigned to the negative data class ($r = -1$).

## 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 $D_\infty$. Give details on the procedure leading to the calculation of the error rate.

**Solution:**

| Instace | 1-NN-LOO | Error? |
|---|---|---|
| $\mathbf{x}^1$ | $\mathbf{x}^6$ | Yes |
| $\mathbf{x}^2$ | $\{\mathbf{x}^1, \mathbf{x}^3, \mathbf{x}^4\}$ | No |
| $\mathbf{x}^3$ | $\{\mathbf{x}^2, \mathbf{x}^4\}$ | No |
| $\mathbf{x}^4$ | $\mathbf{x}^5$ | Yes |
| $\mathbf{x}^5$ | $\mathbf{x}^4$ | Yes |
| $\mathbf{x}^6$ | $\mathbf{x}^1$ | Yes |
| $\mathbf{x}^7$ | $\mathbf{x}^5$ | No |
| $\mathbf{x}^8$ | $\mathbf{x}^5$ | No |

Thus, there are four errors made out of the eight instances, for a classification error rate of $50\,\%$.

(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.

**Solution:**

| Instance | Prototypes | NN-LOO | Error |
|---|---|---|---|
| $\mathbf{x}^1$ | $\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$ | $\mathbf{x}^6$ | Yes |
| $\mathbf{x}^2$ | $\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$ | $\mathbf{x}^4$ | No |
| $\mathbf{x}^3$ | $\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$ | $\{\mathbf{x}^3, \mathbf{x}^4\}$ | No |
| $\mathbf{x}^4$ | $\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$ | $\mathbf{x}^5$ | Yes |
| $\mathbf{x}^5$ | $\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$ | $\mathbf{x}^7$ | No |
| $\mathbf{x}^6$ | $\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$ | $\mathbf{x}^2$ | Yes |
| $\mathbf{x}^7$ | $\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^7,\mathbf{x}^8\}$ | $\mathbf{x}^5$ | No |
| $\mathbf{x}^8$ | $\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^7,\mathbf{x}^8\}$ | $\mathbf{x}^5$ | No |

Thus, the set of prototypes resulting from Wilson edition is $\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^7,\mathbf{x}^8\}$.