EXERCISE 1

Consider a binary classification problem $X \to \{T, F\}$, with $X = \{T, F\}^3$, i.e. $(x_1, x_2, x_3) \in X$ and $x_i \in \{T, F\}$, and the dataset $D = \{\langle (F,F,F), F\rangle, \langle (F,T,T), T\rangle, \langle (T,T,F), T\rangle, \langle (T,F,T), T\rangle\}$. Consider the two hypothesis $h_1 = (x_1 \wedge \neg x_2 \wedge x_3) \vee x_2$ and $h_2 = (\neg x_1 \wedge x_2 \wedge x_3) \vee x_1$.

1. Determine whether h_1 and h_2 are consistent with D, showing all the passages needed to answer.
2. Assuming the likelihood probabilities $P(D|h_1) = 0.6$ and $P(D|h_2) = 0.8$ and the prior probabilities $P(h_1) = 0.2$ and $P(h_2) = 0.1$, determine the higher a poteriori hypothesis between h_1 and h_2.

both are consistent, For h_1

$h_1(F\ F\ F) = (0 \wedge 1 \wedge 0) \vee 0 = 0$
$h_1(F\ T\ T) = (0 \wedge 0 \wedge 1) \vee 1 = 1$ $\Rightarrow$ same procedure
$h_1(T\ T\ F) = (1 \wedge 0 \wedge 0) \vee 1 = 1$ for h_2.
$h_1(T\ F\ T) = (1 \wedge 1 \wedge 1) \vee 1 = 1$

The max. at priori is: $h_{MAP} = \arg\max_h \mathbb{P}(D|h)\mathbb{P}(h) \Rightarrow \begin{cases} \mathbb{P}(D|h_1)\mathbb{P}(h_1) = \frac{3}{25} \\ \mathbb{P}(D|h_2)\mathbb{P}(h_2) = \frac{2}{25} \end{cases} \Rightarrow h_{MAP} = h_1$

EXERCISE 2

Consider the problem of estimating the function $f : \Re^3 \to \Re$, with dataset $\mathcal{D} = \{(\mathbf{x}_1^T, t_1), \ldots, (\mathbf{x}_N^T, t_N)\}$ and using a feed-forward network.

1. Explain what is a suitable choice for the loss function used for training the network and write the corresponding mathematical expression.
2. Assuming that the gradients of the loss with respect to the parameters are available, describe an algorithm for training the parameters of the network. What are the hyper-parameters of the training algorithm (if any)?

Since it's regression, we consider tanh as a loss function for the hidden layers (the net will be differentiable) and a linear act. for the output.

The net: $\mathcal{S}(x) = W_{(2)}^T \tanh(W_{(1)}x + b_{(1)}) + b_{(2)}$ where $\tanh(x) = 2\sigma(x) - 1$ and
($W_{(2)} \in \mathbb{R}^{10}$, $W_{(1)} \in Mat(10\times 3)$, $b_{(1)} \in \mathbb{R}^3$, $b_{(2)} \in \mathbb{R}$)
$\sigma(x) = (1 + \exp(-x))^{-1}$

An algorithm to train is the SGD:

```
SGD {
    W ← initialize with random param
    while (term. cond):
        S ⊂ D random K samples
        W ← W - η∇_S E  gradient of loss E on S
    }
}
```

Since it's regression, the loss is the MSE, equivalent for find the max likelihood solution under the assumption of zero mean Gaussian noise.

$$E(w) = \frac{1}{2}\sum_{i=1}^{N}(t_i - \mathcal{S}(x_i))^2$$

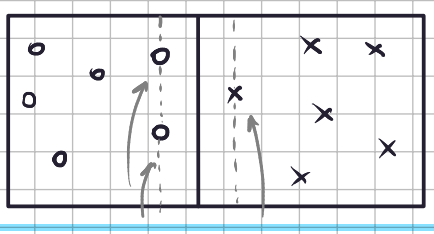
The hyper parameters are: step size η, batch size K, termination condition.

EXERCISE 3

1. Describe the principle of maximum margin used by SVM classifiers through its formal mathematical definition.
2. Draw a linearly separable dataset for 2D binary classification. Draw a possible solution obtained by SVM and highights the margin and the support vectors.
3. Discuss why the maximum margin solution is preferred for the classification problem.

margin for w: $\min_x \frac{|\zeta(x)|}{\|w\|}$

we maximize the margin: $\arg\max_w \min_x \frac{1}{\|w\|}|\zeta(x)|$ s.t. $t_i\zeta(x_i) \geq 0\ \forall i$



is preferred since have less probability of missclassify new samples.

EXERCISE 4

Consider a dataset D for the binary classification problem $f : \Re^3 \mapsto \{Y, N\}$.

1. Describe a probabilistic generative model for such a classification problem, assuming Gaussian distributions.
2. Identify the parameters of the model and determine the size of the model (i.e., the number of independent parameters).

the generative model assumes:

$\mathbb{P}(x \mid Y) = \mathcal{N}(x; \mu_1, \Sigma)$ and $\mathbb{P}(x \mid N) = \mathcal{N}(x; \mu_2, \Sigma)$

$\mathbb{P}(Y) = p$ and $\mathbb{P}(N) = 1-p$

so the ind. param. are

Σ : 3×3 matrix
μ_1, μ_2 : $\mathbb{R}^3$ vectors
$p \in \mathbb{R}$
} size : 16

It can be shown that $\mathbb{P}(Y \mid x, D) = \sigma(w^T x + w_0)$ where $w \in \mathbb{R}^3$, $w_0 \in \mathbb{R}$ depends From Σ, μ_1, μ_2, p.

EXERCISE 5

Consider the following dataset, containing the samples of a function f:

x_1	x_2	f
0.0	0.0	1.0
0.6	0.3	2.2
1.2	0.4	3.0
1.5	0.5	3.5

1. Based on the available data, select a reasonable model for learning f, explicitly indicating its parameters.
2. Show an optimal and a non-optimal solution, explicitly indicating, for each of them, the corresponding value of the loss function.

is a regression problem, so we consider

$f(x_1, x_2) = w_1 x_1 + w_2 x_2 + w_0$, we consider the error $E(w) = \frac{1}{2}\sum_{i=1}^{4} (t_i - y(x_i))^2$

$\Rightarrow \frac{\partial E}{\partial w_j} = -\sum (t_i - y(x_i)) x_{ij}$. We update $w \leftarrow w - \gamma \nabla E$ until convergence.

We consider $w_0 = 1$, $w_1 = 1$, $w_2 = 2$:

$t_1 - (w_1 0 + w_1 0 + w_0) = 0$

$t_2 - (w_1 0.6 + w_2 0.3 + w_0) = 0$

$t_3 - (w_1 \cdot 1.2 + w_2 0.4 + w_0) = 0$

$t_4 - (w_1 \cdot 1.5 + w_2 \cdot 0.5 + w_0) = 3.5 - (1.5 + 1 + 1) = 0$

$\Rightarrow E(w_1 = 1, w_2 = 2, w_0 = 0) = 0$

This is an optimal model. A non optimal model is $w_1 = w_2 = 0$, $w_0 = 100$

$\Rightarrow$ $t_1 - y(x_1) = -99$
$t_2 - y(x_2) = -97.8$
$t_3 - y(x_3) = -97$
$t_4 - y(x_4) = -96.5$

$\Rightarrow E \approx 19043$

EXERCISE 6

1. Describe the concept of bagging in the definition of an ensemble model. Describe precisely the training procedure for such a model and the final formula used for prediction.
2. Discuss the difference between bagging and voting, highlighting in particular the use of different types of models.

In bagging we have M learned $y_1 \cdots y_M$ (comes From the same model) trained on M different set of points $D_1 \cdots D_M$, each sampled from the global dataset D. A new sample is predicted as $y(x) = \frac{1}{M}\sum_{i=1}^{M} y_i(x)$. In voting, we have M different models trained on the same dataset, with predictions: $y(x) = \sum_{i=1}^{M} \alpha_i y_i(x)$ where α_i weights how reliable a model is.