Machine Learning Fundamentals - Homework 3

Milena Markovic (milena.markovic@grenoble-inp.org)

Adaboost algorithm

Question 1: Explain the Adaboost algorithm seen in the course.

The Adaboost algorithm utilizes a set of weak learners, which when combined, produce an efficient final classifier. In the algorithm, we start by assigning all training points equal weights. We go through T iterations where we first train a weak classifier, pick a step size for the current iteration based on the current distribution of misclassified points, and then based on those observations, we update the weights of the training points - these weights will influence the weak learner in the next iteration to classify misclassified points correctly.

Question 2: What is the role of the distribution Dt ?

The distribution assigns weights to training points. In each iteration, if the point was misclassified, it's weight increases. If it was classified correctly, it's weight in the distribution decreases. When a point's weight is increased, the classifier in the next iteration is more probable to classify it correctly.

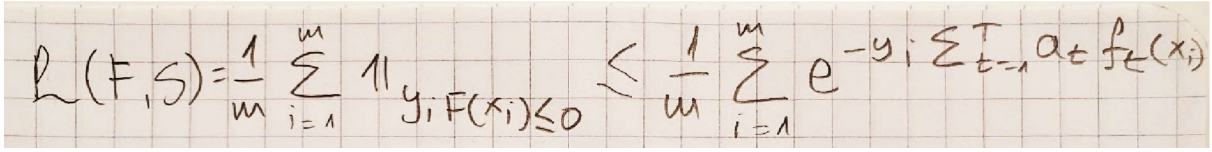
Question 3-4: After T rounds, the algorithm will learn T weak-classifiers (ft)1≤t≤T , with their associated weights (αt)1≤t≤T where the output of each weak classifier is binary in the set {−1, +1}. How is the final classifier F obtained?

The final classifier F is obtained as $F(x) = sign\left(\sum_{t=1}^{T} \alpha_t f_t(x)\right)$ where α_t is the step value of iteration t and f_t is the classification of the weak classifier t.

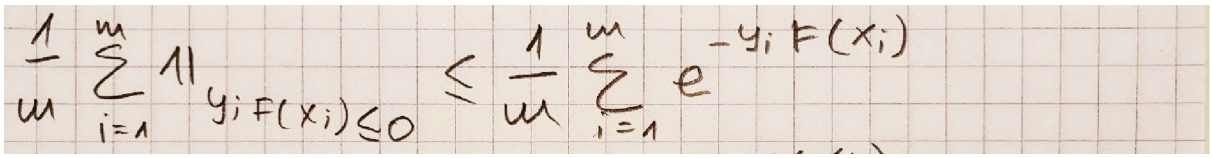
Question 5: Explain why the empirical error of the final classifier F on a training set of size m; $S = \{(xi, yi) \mid i \in \{1, \ldots, m\}\}$ is bounded by the following surrogate loss:

$$\mathcal{L}(F,S) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{y_i F(\mathbf{x}_i)\leq 0} \leq \frac{1}{m}\sum_{i=1}^{m} e^{-y_i \sum_{t=1}^{T} \alpha_t f_t(\mathbf{x}_i)}$$

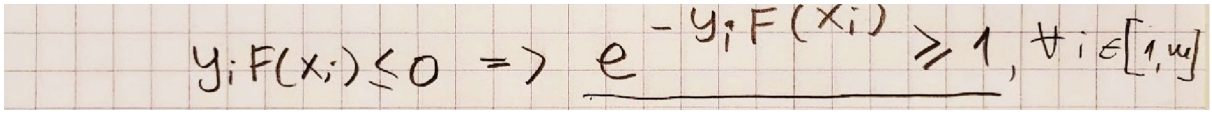
where, 1_π if the predicate π is true; and 0 otherwise.



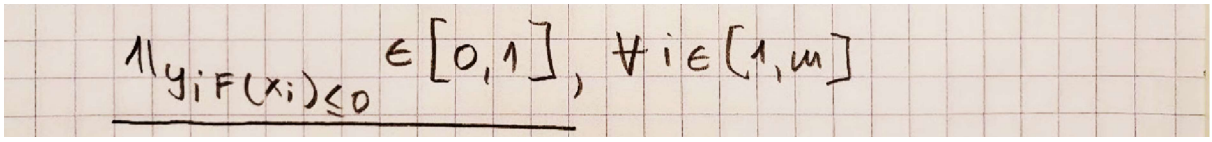
As per the definition of the final classifier F, the following statement can be transformed as:



From the constraint of the left hand side of the equation, we can determine the codomain of the expression in the sum on the right hand side:



As the expression from the sum on the left hand side can only be 1 or 0, and the full expressions on both sides are average values on the domain $\forall\, i \in [1, m]$, we can conclude that the statement that the empirical error is bounded by loss is true.

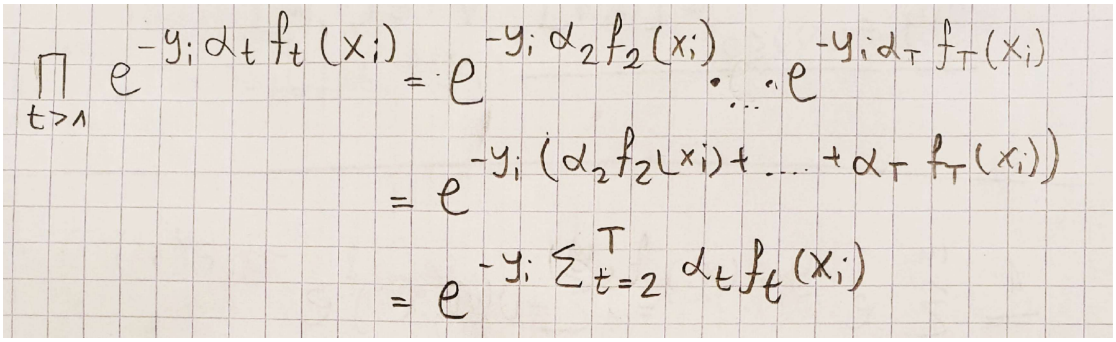


Question 6: Show that

$$\frac{1}{m}\sum_{i=1}^{m} e^{-y_i F(\mathbf{x}_i)} = \sum_{i=1}^{m} Z_1 D_2(i) \prod_{t>1} e^{-y_i \alpha_t f_t(\mathbf{x}_i)} \tag{1}$$

$$\text{where, } \forall t,\ Z_t = \sum_{i=1}^{m} D_t(i) e^{-\alpha_t y_i f_t(\mathbf{x}_i)} \tag{2}$$

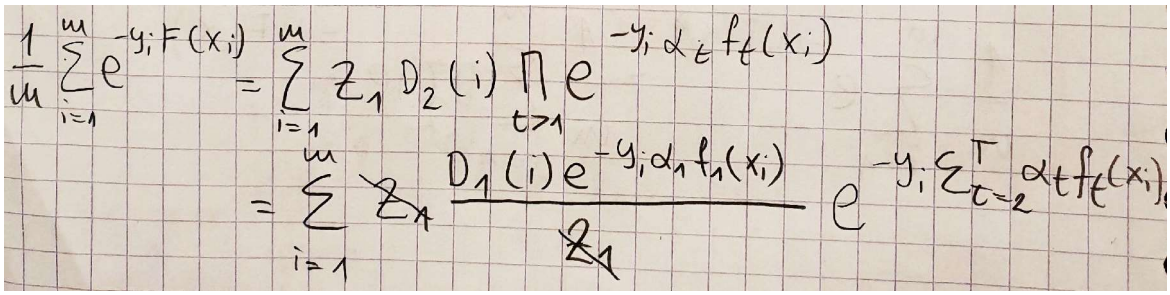
We will start off by transforming the product on the right hand side of the equation:



Next, using the formula for the distribution

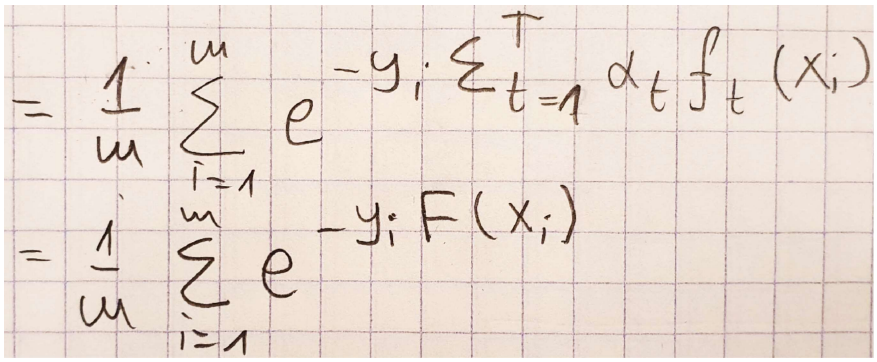
$$D_{t+1}(i) = \frac{D_t(i) e^{-\alpha_t y_i f_t(x_i)}}{Z_t}$$

we will start transforming the equation by substituting the product and D_2 :



Because of the initial value of D_1 given by the algorithm definition, we can continue

$$\forall i \in \{1, \ldots, m\}, D_1(i) = \frac{1}{m}$$

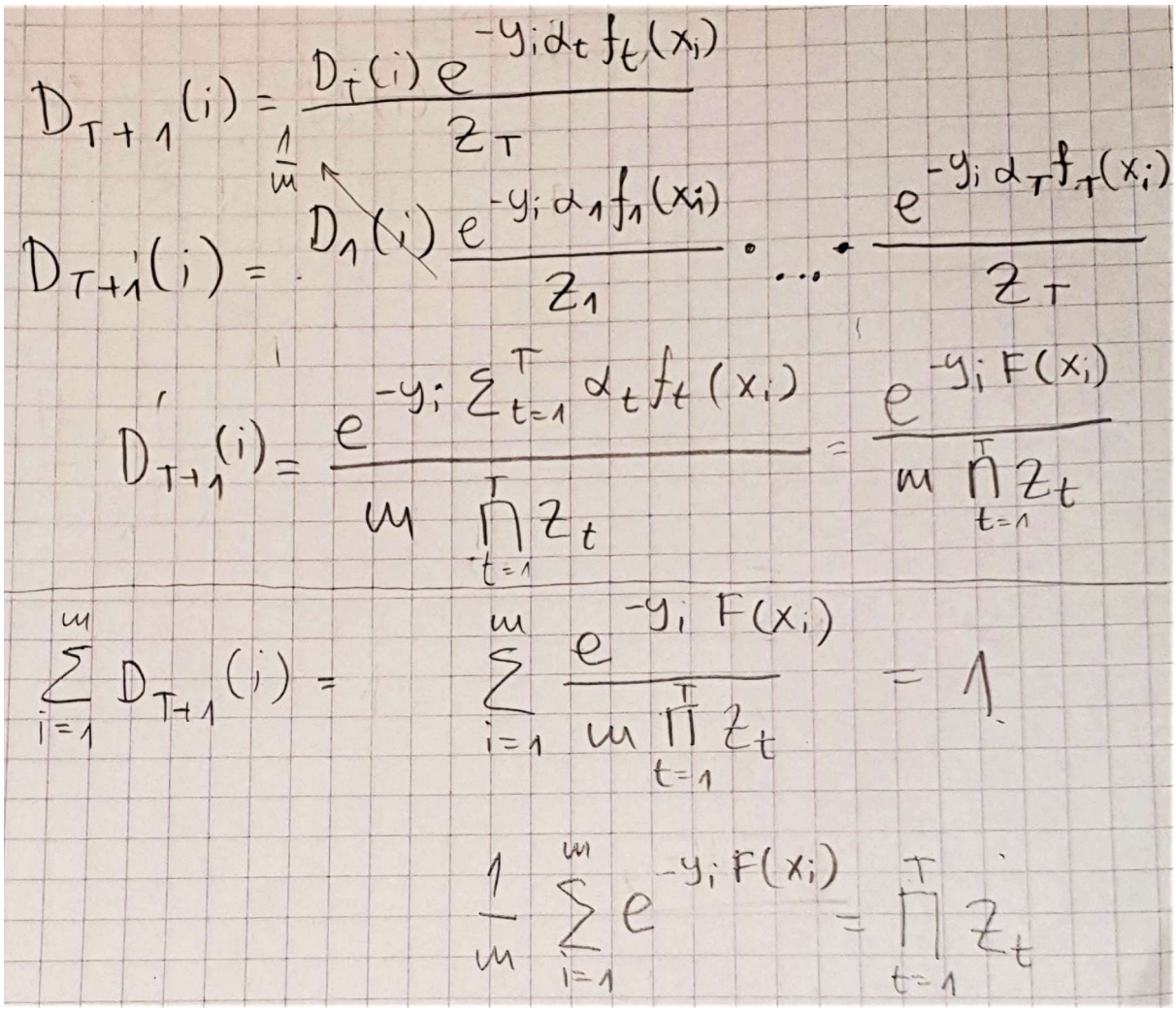


which proves the equation (1).

Question 7: By induction deduce that :

$$\frac{1}{m}\sum_{i=1}^{m} e^{-y_i F(\mathbf{x}_i)} = \prod_{t=1}^{T} Z_t$$

As the normalization terms are all positive, the minimization of the surrogate loss (Eq. 1) is then equivalent to the minimization of the normalization factors Zt , at each iteration.



Question 8: Considering the equation (Eq. 2); for which value of α_t – expressed with respect to the error $\epsilon_t = \sum_{i:y_i \neq f_t(\mathbf{x}_i)} D_t(i)$ – the factor Z_t is minimized?

We will start by transforming the definition of Z_t through the error ϵ_t.

$$\epsilon_t = \sum_{i:y_i \neq f_t(x_i)} D_t(i) \Rightarrow \sum_{i:y_i = f_t(x_i)} D_t(i) = 1 - \epsilon_t$$

$$Z_t = \sum_{i=1}^{m} D_t(i)\, e^{-\alpha_t y_i f_t(x_i)}$$

$$= \sum_{i=1}^{m} \left(\mathbb{1}_{y_i \neq f_t(x_i)} D_t(i)\, e^{-\alpha_t \cdot (-1)} + \mathbb{1}_{y_i = f_t(x_i)} D_t(i)\, e^{-\alpha_t (+1)} \right)$$

$$Z_t = \epsilon_t \cdot e^{\alpha_t} + (1 - \epsilon_t)\, e^{-\alpha_t}$$

In order to minimize Z_t, we want the difference to be zero. After expressing α_t through ϵ_t, we get the same step size formula that is used in the algorithm.

$$\epsilon_t e^{\alpha_t} = (1 - \epsilon_t) e^{-\alpha_t} \quad | \cdot e^{\alpha_t}$$

$$\epsilon_t e^{2\alpha_t} = 1 - \epsilon_t$$

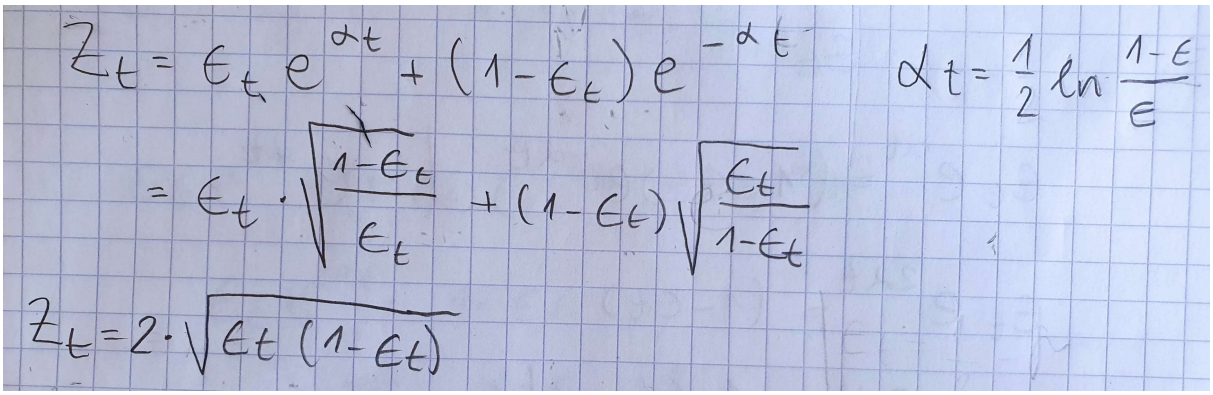
$$e^{2\alpha_t} = \frac{1 - \epsilon_t}{\epsilon_t} \quad | \ln$$

$$2\alpha_t = \ln \frac{1 - \epsilon_t}{\epsilon_t}$$

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t}$$

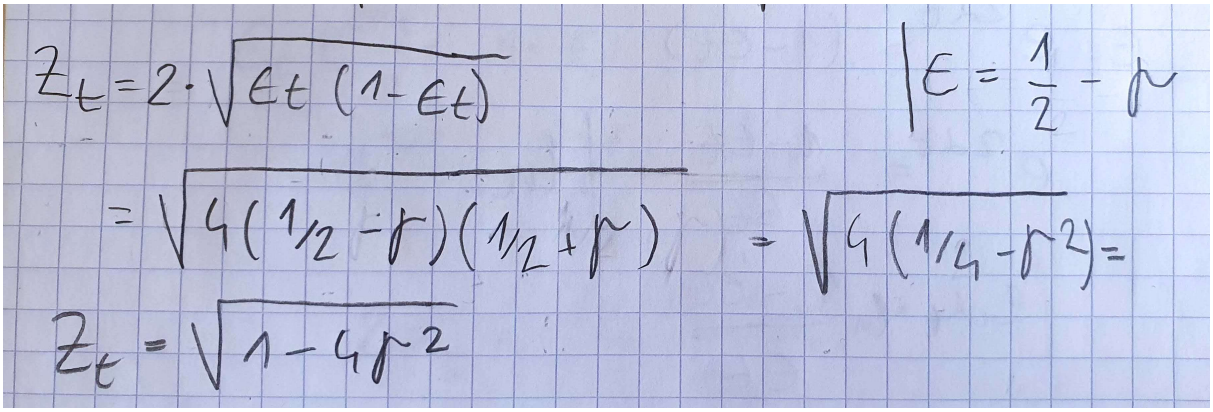
Question 9: For this particular value of αt , what is the minimum value of Z_t ?

We can get the minimum value of Zt by using the value of αt acquired in question 8.



Question 10: Considering the following variable change $\gamma_t = \frac{1}{2} - \epsilon_t$; show that :

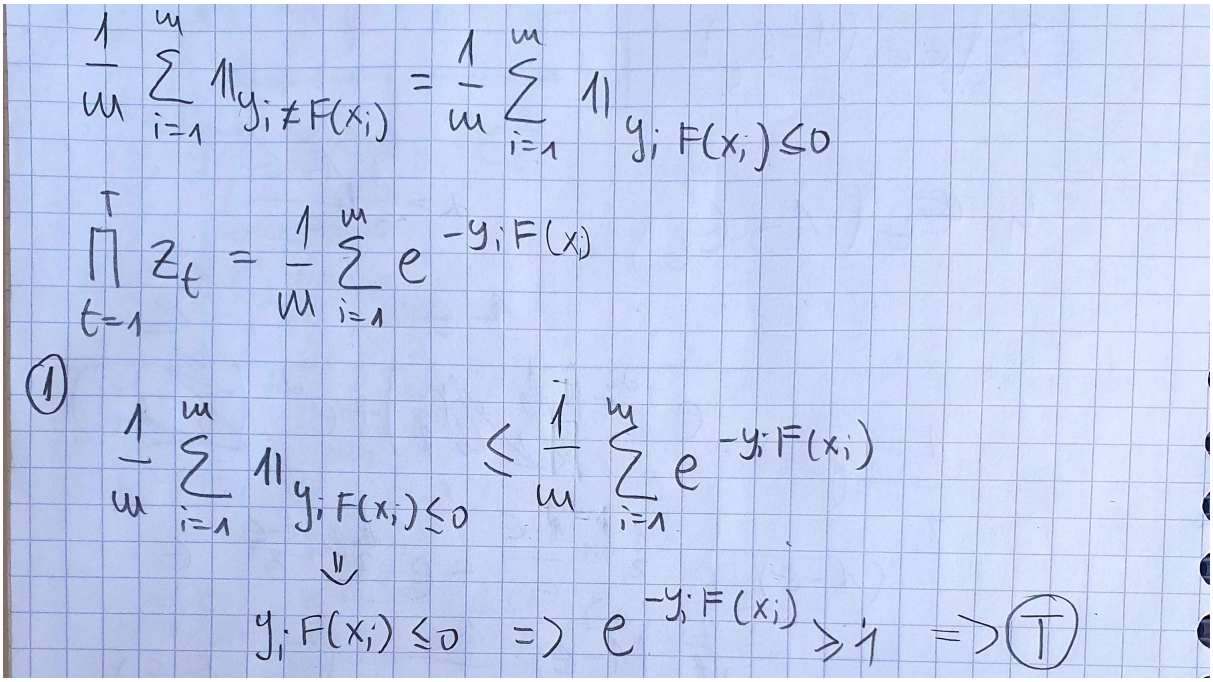
$$\forall t, Z_t = \sqrt{1 - 4\gamma_t^2}$$



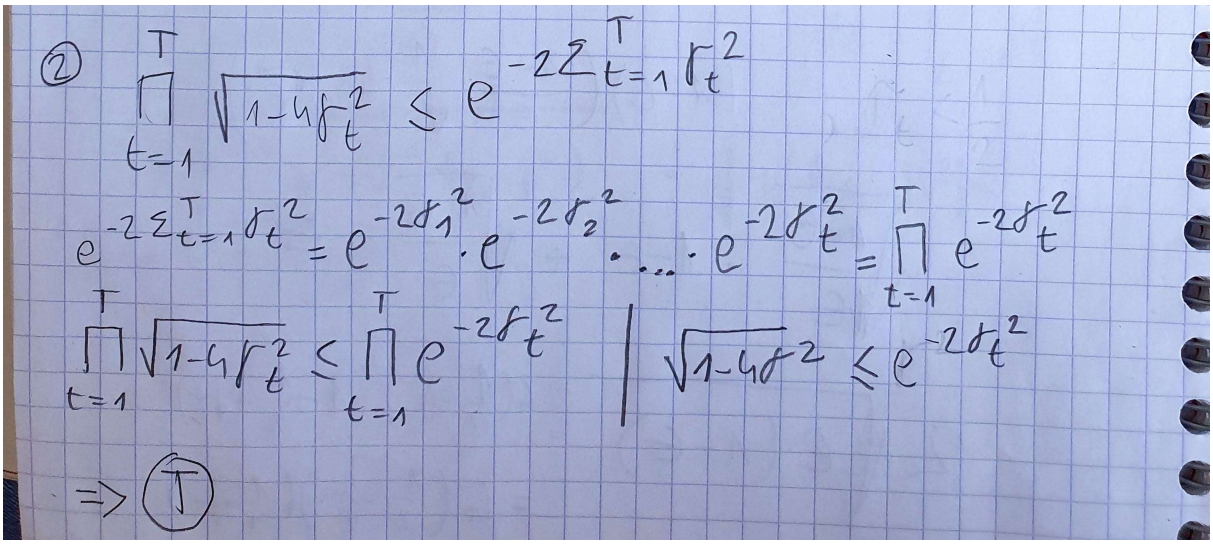
Question 11: For γt < 1/2 , we have $\sqrt{1-4\gamma_t^2} \leq e^{-2\gamma_t^2}$. In this case show that :

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{y_i \neq F(x_i)} \leq \prod_{t=1}^{T} Z_t \leq e^{-2\sum_{t=1}^{T}\gamma_t^2}$$

We will split this expression in two parts. First off, we will prove the left part the same way as we have in question 5.



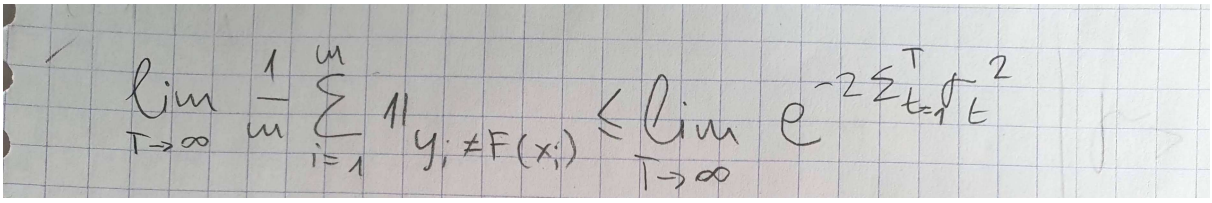
Second, we will transform the right-most part of the expression into a product:



When we take the condition from the question, we see that the expression is valid.

Question 12: Explain why the misclassification error of the final classifier F is ensured to converge to 0 when the number of iterations T tends to infinity.

From the equation in the previous question, we can establish that the limit of the misclassification is bounded by the following:



To find the limit of the left side, we need to establish that the value of γ is positive. As the limit of the left size is 0, the limit of the right side has to be less or equal to 0 as well which means the error does converge to 0.