

# Mohamad Lakkis

February, 2024

## 1 Analysis of Training and Test Error using different Loss Functions in Binary Classification: Theory and Application.

### L2-Loss

- a) Obtaining  $f^*$ , the population minimizer of the corresponding population risk with L2-Loss:

$$f^* = \arg \min_f E_{X,Y}[L(Y, f(X))]$$

Notice that since  $\mathbf{X}$  is fixed at  $\mathbf{X} = x$ , the expected value is with respect to  $Y$  only. And since  $Y \sim \text{Ber}(p)$  with values  $\{-1, 1\}$ , and working locally i.e. on a particular  $f$  at a certain point, which when we minimize each one of these  $f$  we will get the overall  $f^*$  the minimizer

$$f^* = \arg \min_f [L(1, f) \cdot p + L(-1, f) \cdot (1 - p)]$$

Now since this is an L2-loss, thus,

$$f^* = \arg \min_f [(f - 1)^2 \cdot p + (f + 1)^2 \cdot (1 - p)] = \arg \min_f (f^2 - 4fp + 2f - 1)$$

Now take the derivative and set it equal to 0 to get

$$f^* = 2p - 1$$

and notice that it is not the case where  $f^*$  maps  $X$  to  $\{-1, 1\}$ , since it doesn't put a label as a prediction yet! Now we only care about minimizing its value, we can then put say for example if  $f$  at a certain point  $x_i$   $f(x_i) > 1/2$  then 1 else -1 or any other criterion depending on the situation

Notice that even now  $f^*$  is still unknown, since  $p_i$  for each  $Y_i$  is unknown, and so we will estimate it  $p_i \approx \hat{p}_i$ , in order to get  $\hat{f} \approx f^*$  And so,  $\hat{f} = 2\hat{p} - 1$  And so, with

$$\hat{p} = \frac{1 + ay}{2}$$

and on a particular point  $(x_i, y_i)$

$$\hat{f}(x_i) = ay_i$$

Obviously we see that the larger the  $a$  is the more weight we add to the training set, if we set  $a = 0$ , then we put 0 weights on the training set and similar to the notes, in this particular case we can predict  $y$  by a flip of a fair coin! And that is what is meant by when we say  $a$  is a parameter that controls the degree of fit to the training data.

b.1) Computing  $\hat{R}_{tr}$ , the training error

This means that we are approximating  $Err(T)$  by  $\hat{R}_{tr}$ , at  $X=x$

$$\hat{R}_{tr} = \frac{1}{n} \sum_{i=1}^n L(Y_i, \hat{f}(x_i)),$$

Now, at a particular  $Y_i = y_i$ , we get

$$\begin{aligned} \hat{r}_{tr} &= \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}(x_i)), \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - y_i)^2, \\ &= \frac{1}{n} \sum_{i=1}^n (a^2 y_i^2 - 2a y_i^2 + y_i^2), \\ &= \frac{(a-1)^2}{n} \sum_{i=1}^n y_i^2, \end{aligned}$$

Notice that

$$\sum_{i=1}^n y_i^2 = n$$

Therefore,

$$\hat{r}_{tr} = (a-1)^2$$

Therefore, we can see that as  $a$  gets closer to one,  $\hat{r}_{tr}$  becomes smaller and smaller (i.e. the training error) which is logical since  $a$  is the degree of fit to the training set. So for example when  $a=1$ ,  $\hat{r}_{tr}$  becomes equal to 0 which is consistent with our findings since  $\hat{f}(x_i)$  would be exactly equal to  $y_i$ , (i.e. 0 Loss)

b.2) Computing  $R = Err$ , the average test error

Now let's derive  $R = Err$ , the average test error. Note that usually we would go to get an estimation for  $Err(T)$  but sometimes that is hard so

we try estimating Err over the rule  $\hat{F}$ .

$$Err = \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} E_{T, Y^*, X^*} [L(y_i^*, \hat{f}(x_i^*))]$$

But since  $X^* = X$  and  $X$  is fixed at a particular value  $X_i = x_i$ , and the loss is an L2-Loss, then

$$Err = \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} E_{T, Y^*} [(y_i^* - \hat{f}(x_i))^2]$$

with  $\hat{f}(x_i) = ay_i$ ,

$$Err = \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} E_{T, Y^*} [(y_i^*)^2 - 2ay_i y_i^* + a^2 y_i^2]$$

Now since  $Y_i^* \sim \text{Ber}(p_i)$ , then the expected value with respect to  $Y^*$  becomes,

$$Err = \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} E_T [p_i(1 - 2ay_i + a^2 y_i^2) + (1 - p_i)(1 + 2ay_i + a^2 y_i^2)]$$

$$Err = \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} E_{Y_i} [p_i(1 - 2ay_i + a^2) + (1 - p_i)(1 + 2ay_i + a^2)]$$

And similarly,  $Y_i \sim \text{Ber}(p_i)$ , then the expected value with respect to  $Y_i$  becomes with some basic mathematical formulas and using the fact that  $E(h(Y)) = h(-1)P(Y = -1) + h(1)P(Y = 1)$ , if  $Y \sim \text{Ber}$  with values  $\{-1, 1\}$ ,

$$\begin{aligned} Err &= 4a \frac{2}{n_{ts}} \sum_{i=1}^{n_{ts}} [p_i(1 - p_i)] + (a - 1)^2 \\ &= 4a\bar{e} + (a - 1)^2 \end{aligned}$$

Now in order to get the  $a^* = a$  that minimizes the Err, we need to do the derivative with respect to  $a$ , and thus we get,

$$a = 1 - 2\bar{e}$$

and since  $0 \leq \bar{e} \leq \frac{1}{2}$ , we get  $0 \leq a \leq 1$ , which is consistent with the condition of  $a$  from the question !

We can easily observe that as  $a$  gets closer to 1, the Err derived from the test error gets larger and larger which is overfitting the model on training set ( since we are making  $a$  bigger this means that we are giving more weights to the training set )

Now in order to get, the mean  $= E_T \hat{f}(x_i, T)$ , we will use to our advantage that  $T$  and  $Y_i$  have the same distribution since  $X_i = x_i$  is fixed !

$$\begin{aligned} E_T \hat{f}(x_i, T) &= E_{Y_i} \hat{f}(x_i, T) \\ E_{Y_i} \hat{f}(x_i, T) &= a E_{Y_i} [y_i] \\ &= 2ap_i - a \\ &= a(2p_i - 1) \end{aligned}$$

Now as for the variance we will use similar analysis,  $E_T[(\hat{f}(x_i, T) - E_T \hat{f}(x_i, T))^2]$

$$\begin{aligned} Var &= E_T[(ay_i - 2ap_i + a)^2] \\ &= a^2 E_T[(y_i - 2p_i + 1)^2] \end{aligned}$$

And now using similar analysis as before we get,

$$Var = 4a^2 p_i(1 - p_i)$$

### L1-Loss

- a) Obtaining  $f^*$ , the population minimizer of the corresponding population risk with L1-Loss:

$$f^* = \arg \min_f E_{X,Y}[L(Y, f(X))]$$

Notice that since  $\mathbf{X}$  is fixed at  $\mathbf{X} = x$ , the expected value is with respect to  $Y$  only. And since  $Y \sim \text{Ber}(p)$  with values  $\{-1, 1\}$ , and working locally i.e. on a particular  $f$  at a certain point  $x_i$ , which when we minimize each one of these  $f$  we will get the overall  $f^*$  the minimizer

$$f^* = \arg \min_f [L(1, f) \cdot p + L(-1, f) \cdot (1 - p)]$$

Now since this is an L1-Loss  $L(y, f) = |y - f|$ , thus,

$$f^* = \arg \min_f [|1 - f|p + |1 + f|(1 - p)]$$

We have seen in class that in this case ,

$$f^*(x) = \text{median}(Y|X = x)$$

, And thus, one classifier that works is:

$$f^*(x) = \begin{cases} 1 & \text{if } p \geq \frac{1}{2}, \\ -1 & \text{otherwise.} \end{cases}$$

and notice that in this case  $f^*$  immediatly maps to  $\{-1, 1\}$

Notice that even now  $f^*$  is still unknown, since  $p_i$  for each  $Y_i$  is unknown, and so we will estimate it  $p_i \approx \hat{p}_i$ , in order to get  $\hat{f} \approx f^*$ . And so, with

$$\hat{p} = \frac{1 + ay}{2}$$

and on a particular point  $(x_i, y_i)$ ,

$$\hat{f}(x_i) = \begin{cases} 1 & \text{if } \hat{p}_i > \frac{1}{2}, \\ -1 & \text{otherwise.} \end{cases}$$

$$\hat{f}(x_i) = \begin{cases} 1 & \text{if } ay_i > 0, \\ -1 & \text{otherwise.} \end{cases}$$

Which now we can say that

$$\hat{f}(x_i) = \text{sign}(ay_i)$$

But since  $a > 0$ , thus we get:

$$\hat{f}(x_i) = \text{sign}(y_i)$$

In the case where  $a = 0$  we can decide with a fair flip of a coin!

Notice that in this particular loss we are not really paying attention to the value of a  $a$  only in the case where  $a$  is 0, but in all other cases we are fitting the model to exactly the training data ( like in the previous losses when  $a = 1$  ), so we can expect that the training error to be exactly 0.

- b.1) Computing  $\hat{R}_{tr}$ , the training error. And now to test our hypothesis (Hope it is correct)

$$\hat{R}_{tr} = \frac{1}{n} \sum_{i=1}^n L(Y_i, \hat{f}(x_i)),$$

Now, at a particular  $Y_i = y_i$ , we get

$$\begin{aligned} \hat{r}_{tr} &= \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}(x_i)), \\ &= \frac{1}{n} \sum_{i=1}^n (|y_i - \text{sign}(y_i)|) = 0 \end{aligned}$$

- b.2) Computing  $R = Err$ , the average test error

$$Err = \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} E_{T, Y^*, X^*} [L(y_i^*, \hat{f}(x_i^*))]$$

But since  $X^* = X$  and  $X$  is fixed at a particular value  $X_i = x_i$ , and the loss is an Exp-Loss, then

$$Err = \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} E_{T,Y^*} [|y_i^* - \text{sign}(y_i)|],$$

with some similar computations as before we get

$$Err = 2\bar{e},$$

Here there is no min value of  $a = a^*$  available since  $Err$  is independent of  $a$ !

Now in order to get, the mean  $= E_T \hat{f}(x_i, T)$ , we will use to our advantage that  $T$  and  $Y_i$  have the same distribution since  $X_i = x_i$  is fixed !

$$E_T \hat{f}(x_i, T) = E_{Y_i} \hat{f}(x_i, T)$$

$$E_{Y_i} \hat{f}(x_i, T) = 2p_i - 1$$

Now as for the variance we will use similar analysis,  $E_T[(\hat{f}(x_i, T) - E_T \hat{f}(x_i, T))^2]$

$$\begin{aligned} Var &= E_T[(ay_i - 2ap_i + a)^2] \\ &= a^2 E_T[(y_i - 2p_i + 1)^2] \end{aligned}$$

And now using similar analysis as before we get,

$$Var = 4 + 4p_i(1 - p_i)$$

## Exponential

- a) Obtaining  $f^*$ , the population minimizer of the corresponding population risk with Exp-Loss:

$$f^* = \arg \min_f E_{X,Y} [L(Y, f(X))]$$

Notice that since  $\mathbf{X}$  is fixed at  $\mathbf{X} = x$ , the expected value is with respect to  $Y$  only. And since  $Y \sim \text{Ber}(p)$  with values  $\{-1, 1\}$ , and working locally i.e. on a particular  $f$  at a certain point  $x_i$ , which when we minimize each one of these  $f$  we will get the overall  $f^*$  the minimizer

$$f^* = \arg \min_f [L(1, f) \cdot p + L(-1, f) \cdot (1 - p)]$$

Now since this is an Exponential Loss  $L(y, f) = e^{-yf}$ , thus,

$$f^* = \arg \min_f [e^{-f} p + e^f (1 - p)] = \arg \min_f [e^{-f} p + e^f - e^f p]$$

Now take the derivative and set it equal to 0 to get

$$f^* = \frac{1}{2} \log\left(\frac{p}{1-p}\right)$$

and notice that it is not the case where  $f^*$  maps  $X$  to  $\{-1, 1\}$ , since it doesn't put a label as a prediction yet! Now we only care about minimizing its value, we can then put say for example if  $f$  at a certain point  $x_i$ ,  $f(x_i) > 1/2$  then 1 else -1 or any other criterion depending on the situation

Notice that even now  $f^*$  is still unknown, since  $p_i$  for each  $Y_i$  is unknown, and so we will estimate it  $p_i \approx \hat{p}_i$ , in order to get  $\hat{f} \approx f^*$

And so, with

$$\hat{p} = \frac{1 + ay}{2}$$

and on a particular point  $(x_i, y_i)$ ,

$$\hat{f}(x_i) = \frac{1}{2} \log\left(\frac{1 + ay_i}{1 - ay_i}\right)$$

Obviously we see that the larger the  $a$  is the more weight we add to the training set, if we set  $a = 0$ , then we put 0 weights on the training set and similar to the notes  $\hat{f}$  will become 0, so in this particular case we can predict  $y$  by a flip of a fair coin! And that is what is meant by when we say  $a$  is a parameter that controls the degree of fit to the training data.

b.1) Computing  $\hat{R}_{tr}$ , the training error

This means that we are approximating  $\text{Err}(T)$  by  $\hat{R}_{tr}$ , at  $X=x$

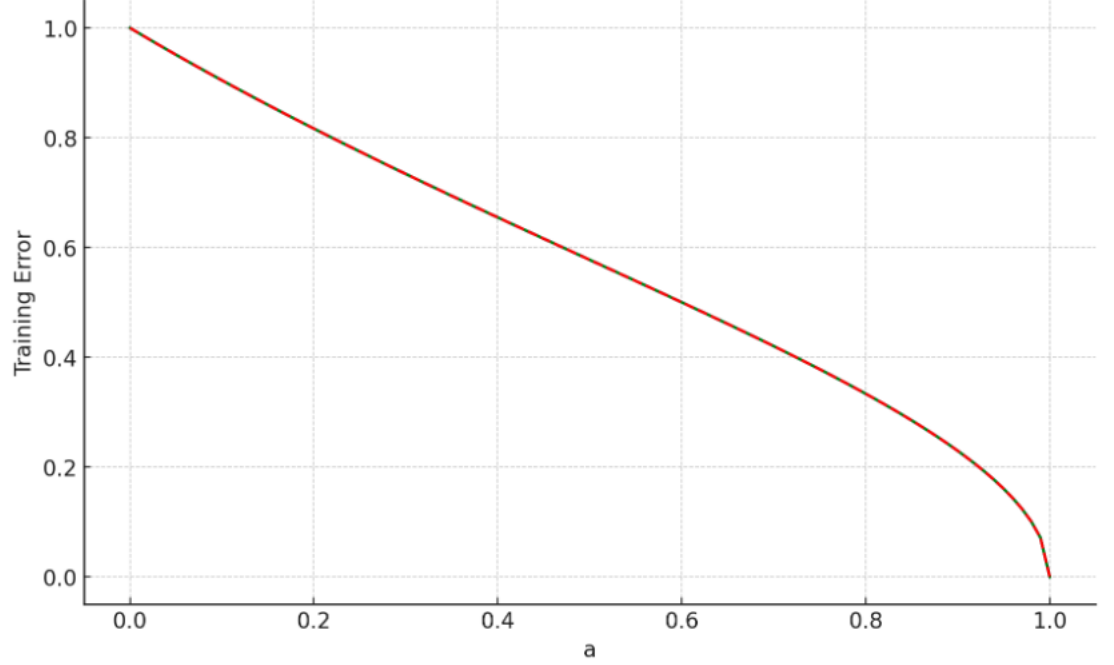
$$\hat{R}_{tr} = \frac{1}{n} \sum_{i=1}^n L(Y_i, \hat{f}(x_i)),$$

Now, at a particular  $Y_i = y_i$ , we get

$$\begin{aligned} \hat{r}_{tr} &= \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}(x_i)), \\ &= \frac{1}{n} \sum_{i=1}^n (e^{-y_i \hat{f}(x_i)}), \\ &= \frac{1}{n} \sum_{i=1}^n (e^{-\frac{1}{2} y_i \log(\frac{1+ay_i}{1-ay_i})}), \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{1 - ay_i}{1 + ay_i}\right)^{\frac{y_i}{2}}, \end{aligned}$$

And Now in order to understand how the  $\hat{r}_{tr}$  changes with  $a$ , let's draw a plot, ( notice that for both  $y_i = 1$  or  $-1$  the  $\hat{r}_{tr}$  has the same formula and thus the same graph as we will see now), in this plot I will fix  $y_i = 1$  ( it

works with -1 as explained since same formula), after fixing it, I will vary  $a$  and see how the training error changes.



So as we see from the graph as  $a$  increase as the  $\hat{r}_{tr}$  decreases which is logical since  $a$  is the degree of fit to the training data, and thus at  $a = 1$  we can see that the training error becomes 0, because we are fitting exactly like the training data (this tends to have a negative effect on the test error because of concept of overfitting) !

b.2) Computing  $R = Err$ , the average test error

Now let's derive  $R = Err$ , the average test error. Note that usually we would go to get an estimation for  $Err(T)$  but sometimes that is hard so we try estimating  $Err$  over the rule  $\hat{F}$ .

$$Err = \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} E_{T, Y^*, X^*} [L(y_i^*, \hat{f}(x_i^*))]$$

But since  $X^* = X$  and  $X$  is fixed at a particular value  $X_i = x_i$ , and the loss is an Exp-Loss, then

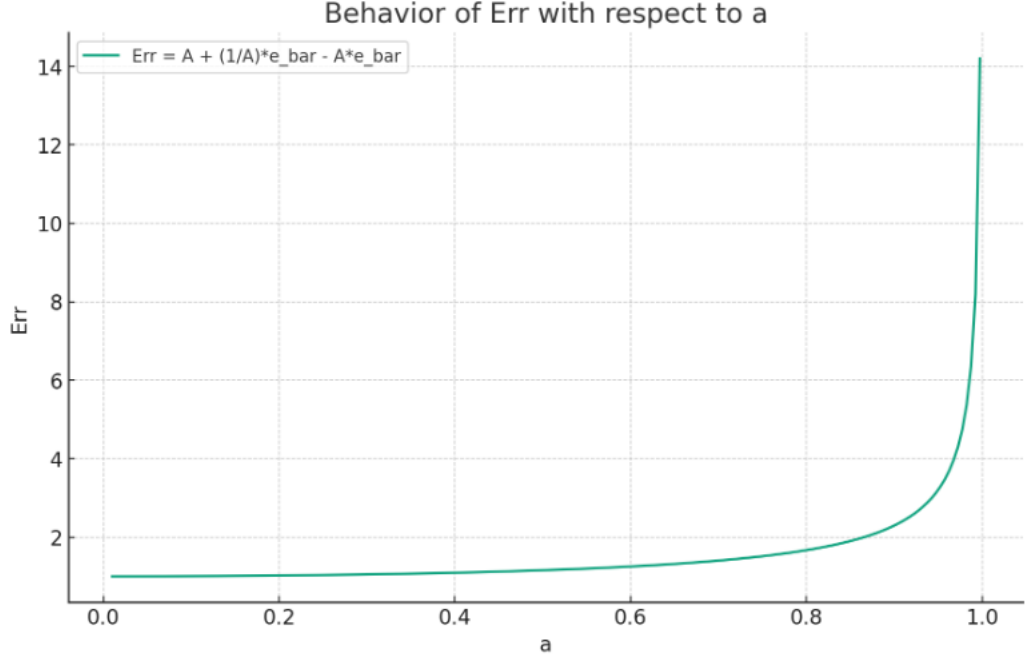
$$\begin{aligned} Err &= \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} E_{T, Y^*} [e^{-\frac{1}{2} y_i^* \log(\frac{1+ay_i}{1-ay_i})}] \\ &= \frac{1}{n_{ts}} \sum_{i=1}^{n_{ts}} E_{T, Y^*} [(\frac{1-ay_i}{1+ay_i})^{\frac{y_i^*}{2}}] \end{aligned}$$



After applying similar steps as before, we get finally

$$Err = A + \frac{1}{A}\bar{e} - A\bar{e}$$

with  $A = (\frac{1-a}{1+a})^{\frac{1}{2}}$  Now We will explore how does the  $Err$  varies with  $a$ ,  
with  $\bar{e} = \frac{1}{2}$



Now in order to get, the mean  $= E_T \hat{f}(x_i, T)$ , we will use to our advantage that  $T$  and  $Y_i$  have the same distribution since  $X_i = x_i$  is fixed !

$$E_T \hat{f}(x_i, T) = E_{Y_i} \hat{f}(x_i, T)$$

$$E_{Y_i} \hat{f}(x_i, T) = \frac{1}{2}(p_i \log(\frac{1}{A}) + \log(A) - p_i \log(A))$$

with  $A = \frac{1-a}{1+a}$

Now as for the variance we will use similar analysis,  $E_T[(\hat{f}(x_i, T) - E_T \hat{f}(x_i, T))^2]$

$$Var = E_T[(\frac{1}{2} \log(\frac{1+ay_i}{1-ay_i}) + \frac{1}{2}(p_i \log(\frac{1}{A}) + \log(A) - p_i \log(A)))^2]$$

And now using similar analysis as before we get,

$$Var = p_i(\log(\frac{1}{A}))^2 + \frac{1}{4}(1 - p_i)(-\log(A))^2 + p_i^2(\log(\frac{1}{A}))^2$$