BAUHAUS UNIVERSITÄT WEIMAR

INTRODUCTION TO MACHINE LEARNING

SUBMITTED BY

Anjana MuraleedharanNair(125512)

Isabel Maria Binu (125514)

Vishal Sanjay Shivam (125353)

Sharat Anand (125404)

Exercise 1: Linear Models

(a) l(c, y(x)) - pointwise loss

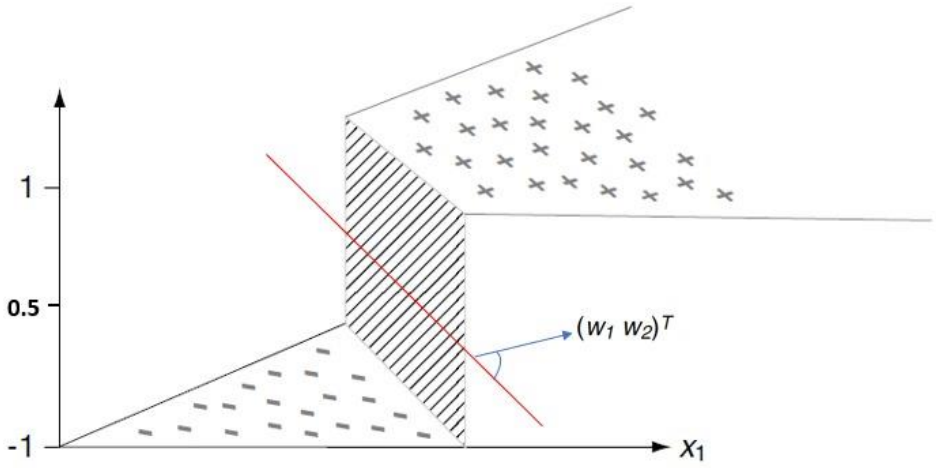
L(w) - Global loss function

L(w) - Error function

w - the entire parameter vector

$\vec{w}$ -direction vector excluding w_0.

(b)



(c) Both linear regression and logistic regression involve optimising w through a loss minimization problem, where "loss" refers to the interpretation of residuals. Since the residuals in linear regression are squared, outliers are given a high weight. However, logistic regression models an increasing confidence in class membership probability with increasing hyperplane distance. This difference in interpretation typically leads to a different parameter vector w and, consequently, a different decision boundary for logistic regression compared to linear regression.

(d) Lasso regression includes a penalty term based on the L1 norm of the weight vector (||w||1). This penalty encourages sparsity, meaning it tends to drive some of the coefficients to exactly zero during optimization. This is due to the geometric shape of the L1 norm, which has corners at the axes, making it more likely for the optimization process to hit exactly zero when minimizing the combined loss and penalty term. In contrast, the L2 penalty used in ridge regression (||w||2) tends to shrink coefficients towards but not exactly to zero, as it penalizes large values without promoting true sparsity.

(e) Gradient descent can't be directly applied to $L0/1$ regularization due to non-differentiability at points where vectors have zero elements.

Exercise 2: Pointwise Loss Functions

(a) $l\sigma(0,y(x)) = \log(1+e^{w^T}x)$

Sigmoid function -> $\sigma(z)=1/1+e^z)$

$\sigma(w^Tx) = 1/1+e^{w^T}x$

$\sigma(-a) = 1-\sigma(a)$ (given)

$\sigma(-w^Tx)=1- \sigma(w^Tx)$

$\sigma(w^Tx) = 1- \sigma(-w^Tx)$

$\sigma(w^Tx) = 1- (1/1+e^{w^T}x) = (1+e^{w^T}x -1)/(1+e^{w^T}x)$

$\sigma(w^Tx) = e^{w^T}x/1+e^{w^T}x$

Taking natural logarithm on both sides

$\log(\sigma(w^Tx)) = \log(e^{w^T}x/1+e^{w^T}x)$

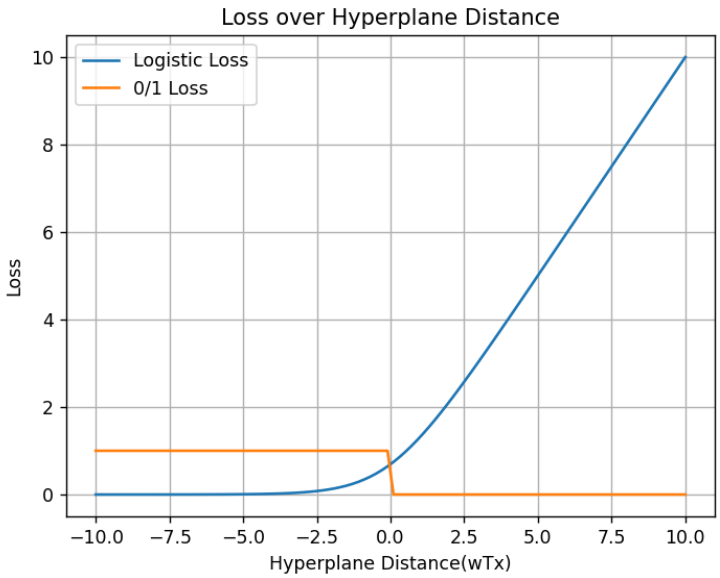
$=\log(e^{w^T}x) - \log(1+e^{w^T}x)$ $\log(a/b) = \log(a) - \log(b)$

$= w^Tx - \log(1+e^{w^T}x)$ $(\log(e^a)=a)$

Substituting $w^Tx = 0$

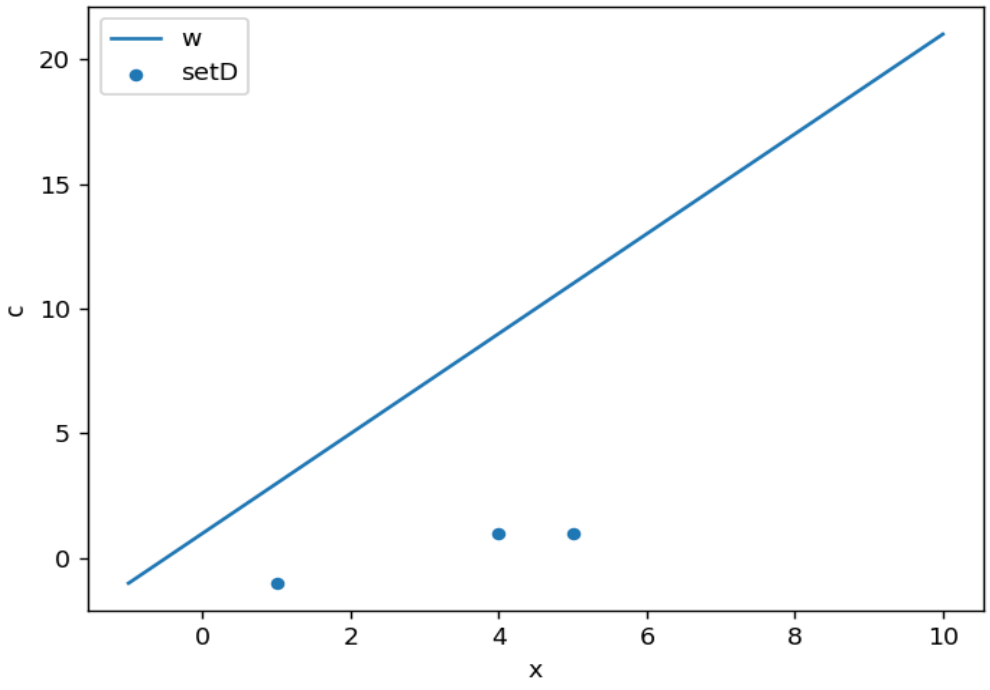
$l\sigma(0,y(x)) = \log(1+e^{w^T}x)$

(b)



Exercise 3 : Gradient Descent

(a)



(b) For x_1=(4,1), c=1

$y(x_1) = w^T x_1 = 9$

$l_2(c, y(x1)) = \frac{1}{2} * (c - y(x1))^2 = ½*(1-9)^2=32$

(c)

$l_2 = \frac{1}{2} * (c - w^T x)^2$

$\partial l_2/\partial w0 = -(c-w^T x) * \partial/\partial w0 (w^T x)$ ($w^T x = w_0 x_0 + w_1 x_1$)

$= -(c-w^T x) x_0$

$\partial l_2/\partial w1 = -(c-w^T x) * \partial/\partial w1 (w^T x)$ ($w^T x = w_0 x_0 + w_1 x_1$)

$= -(c-w^T x) x_1$

$(\partial l_2/\partial w0, \partial l_2/\partial w1)^T = \begin{bmatrix} -(c-w^T x)x_0 \\ -(c-w^T x)x_1 \end{bmatrix}$

$= -\delta . x$ where $\delta = -(c-w^T x)$ and $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$

(d) $\partial l_2/\partial w = -(c-w^T x) * \partial/\partial w (w^T x)$

$= -(c-w^T x) x$

$\partial l_2/\partial x = -(c-w^T x) * \partial/\partial w (w^T x)$

$= -(c-w^T x) w$

Loss gradient for x1 = -(1-9).1 = 8

Loss gradient for w = -(1-9).4 = 32

(e)

$\Delta w = -\eta * \nabla L_2(w)$

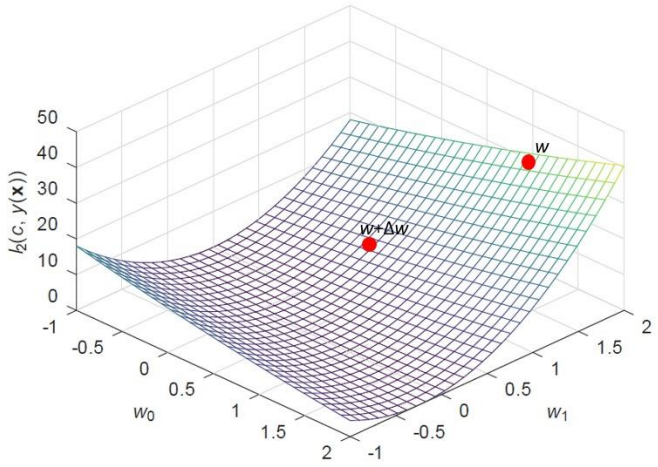
$\nabla L_2 = -\Sigma (c-w^T x) * x$

$\Delta w = -\eta * (\partial l_2/\partial w0, \partial l_2/\partial w1) = -0.03 * \begin{matrix} 8 \\ 32 \end{matrix}$

$= \begin{matrix} -0.24 \\ -0.96 \end{matrix}$

$w+\Delta w = \begin{matrix} 0.76 \\ 1.04 \end{matrix}$

(f)



Exercise 4 : Regularization

(a) Option A - steadily increase

$L(w) = RSS_{tr}(w) + \lambda w^T w$

$RSS_{tr}(\mathbf{w}) = \sum_{(\mathbf{xi},\mathbf{yi})\in \mathbf{Dtr}} (y_i - \mathbf{w}^T x_i)^2$

Here λ is increased from 0.

If λ=0, then $L(w)=RSS_{tr}(w)$

Λ controls the impact of $w^T.w$

So, when λ increases, the $RSS_{tr}(w)$ value also steadily increases.

(b) Option E - decrease initially, then eventually start increasing in a U shape.

When λ equals zero, the model effectively captures the training data. Therefore, we adjust the λ value to prevent overfitting. However, the λ parameter's variation can cause the model to fit the D_{test} only up to a certain threshold. Consequently, an initial increase in λ leads to reduced error on D_{test}. Nevertheless, beyond a certain threshold, the model ceases to fit the D_{test}, resulting in elevated $RSS_{test}(w)$

So, $RSS_{test}(w)$ will decrease initially, then eventually start increasing in a U shape.