Introduction to Machine Learning - HW1

Professors: Abolghasemi & Erabi
Student: Mohamad Mahdi Samadi
Student ID: 810101465

Q2.

a. Explain L1 and L2 regularization and point out their differences.

The two named techniques are used to prevent overfitting by adding an extra penalty term to the loss function. We do it to encourage the model to learn a more simpler and generalizable representation of the provided data.

- L1 regularization: Also called Lasso, adds a penalty term proportional to the absolute values of the model coefficients. Produces sparse models by driving some parameters to zero. It helps to select features based on non-zero coefficients.

  $$L1\ loss\ =\ Original\ Loss\ +\ \lambda \sum_{i=1}^{p} |\beta_i|$$

- L2 regularization: Also known as Ridge, adds a penalty term proportional to the squared values of parameters. Prevents the model to assign high weights to the correlated features.

  $$L2\ loss\ =\ Original\ Loss\ +\ \lambda \sum_{i=1}^{p} \beta_i^2$$

In conclusion, Lasso zero-outs coefficients, on the other hand Ridge shrinks them but not zero them.

b. Given a training set of data ($y_i \in R$, $x_i \in R^d$) and using $L(w)$ as loss function find the optimal value of vector w.

$$L(w) = \sum_{i=1}^{n} (w^T . x_i - y_i)^2 + \sum_{j=1}^{p} (w_j \lambda)^2$$

Let's write it in matrix form.

$$L(w) = (y - XW)^T (y - XW) - \lambda W^T W$$

Where X is the matrix with x_i as rows, y is the vector of outputs, W is the vector of weights and λ is the constant parameter.

Knowing below facts about transpose, we'll continue to solve the problem:

- $A^T B = B^T A$
- $(AB)^T = B^T A^T$

$$L(w) = (y^T - W^T X^T)(y - XW) + \lambda W^T W = y^T y - X^T W^T y - y^T XW + W^T X^T XW + \lambda W^T W$$

$$X^T W^T y = (X^T W^T) y = y^T (X^T W^T)^T = y^T W X \quad (I)$$

$$(I) \rightarrow L(w) = y^T y - 2 y^T X W + W^T X^T X W + \lambda W^T W = y^T y - 2 W^T y X^T + W^T X^T X W + \lambda W^T W$$

$$\frac{\partial L(w)}{\partial w} = -2 y X^T + 2 X^T X W + 2 \lambda W = 0$$

$$\rightarrow - y X^T + X^T X W + \lambda W = 0 \rightarrow y X^T = X^T X W + \lambda W$$

$$y X^T = (X^T X + \lambda I) W \rightarrow W = (X^T X + \lambda I)^{-1} y X^T$$