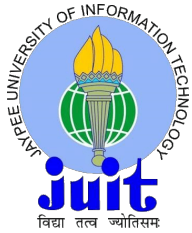

Machine Learning

(Course Code: 18B1WCI634 / 18B11BI611)

Loss Function

Mr. Sandeep Kumar Patel

Assistant Professor

Jaypee University of Information Technology

Waknaghat, Solan, HP-173234

What is a Loss Function?

- A loss function measures the error between:
  - True value y
  - Predicted value $\hat{y}$
- Training objective:

$$\min_{\theta} \ \mathcal{L}(h(x_i), y_i)$$

Loss vs Cost

- **Loss**: error for a single example

$$L\big(h(x_i), y_i\big)$$

- **Cost (Risk)**: average loss

$$\frac{1}{N}\sum_{i=1}^{N} L\big(h(x_i), y_i\big)$$

Regression Loss Functions

Used when:

$$y_i \in \mathbb{R}$$

Common losses:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Huber Loss

Mean Squared Error (MSE)

The loss for a single data point using Mean Squared Error is:

$$L(h(x_i), y_i) = (h(x_i) - y_i)^2$$

The Mean Squared Error over n data points is defined as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (h(x_i) - y_i)^2$$

- Strongly penalizes large errors
- Smooth and differentiable
- Sensitive to outliers

Mean Absolute Error (MAE)

The loss for a single data point using Mean Absolute Error is:

$$L(h(x_i), y_i) = |h(x_i) - y_i|$$

The Mean Absolute Error over n data points is defined as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |h(x_i) - y_i|$$

- Robust to outliers
- Non-differentiable at zero

Root Mean Square Error (RMSE)

The loss for a single data point using Root Mean Squared Error is:

$$L(h(x_i), y_i) = \sqrt{(h(x_i) - y_i)^2}$$

The Root Mean Squared Error over n data points is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2}$$

Huber Loss

$$L\big(h(x_i), y_i\big) = \begin{cases} \frac{1}{2}\big(h(x_i) - y_i\big)^2, & |h(x_i) - y_i| \leq \delta \\ \delta|h(x_i) - y_i| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases}$$

- MSE for small errors
- MAE for large errors

Classification Setting

$$y_i \in \{1, \ldots, K\}$$

Model output:

$$h(x_i) = \big(h_1(x_i), \ldots, h_K(x_i)\big)$$

- $h_k(x_i)$ is predicted probability of class k

Binary Cross-Entropy

Binary labels:

$$y_i \in \{0, 1\}, \quad h(x_i) \in (0, 1)$$

$$L\big(h(x_i), y_i\big) = -\Big[y_i \log h(x_i) + (1 - y_i) \log(1 - h(x_i))\Big]$$

- Used with sigmoid output
- Probabilistic interpretation

Categorical Cross-Entropy

One-hot labels:

$$y_i = (y_{i1}, \ldots, y_{iK})$$

$$L\big(h(x_i), y_i\big) = -\sum_{k=1}^{K} y_{ik} \log h_k(x_i)$$

- Used with softmax output

Probabilistic View

- True distribution: $P(y \mid x_i)$
- Model distribution: $Q(y \mid x_i) = h(x_i)$

Cross-Entropy loss:

$$L\big(h(x_i), y_i\big) = -\sum_{y} P(y \mid x_i) \log Q(y \mid x_i)$$

KL Divergence

$$\mathrm{KL}(P\|Q) = \sum_{y} P(y \mid x_i) \log \frac{P(y \mid x_i)}{Q(y \mid x_i)}$$

$$\mathrm{KL}(P\|Q) = H(P, Q) - H(P)$$

Key Relationship

$$\boxed{H(P, Q) = H(P) + \mathrm{KL}(P \| Q)}$$

- $H(P)$ is constant
- Minimizing cross-entropy minimizes KL divergence