Loss Functions in Classification Tasks

Dr. Kevin Koidl
School of Computer Science and Statistic Trinity College Dublin
ADAPT Research Centre
Loss Function

When training a classifier one defines a loss function $L(\hat{y}, y)$, stating the loss of predicting $\hat{y}$ when the true output is $y$.

The training objective is then to minimize the loss across the different training examples.

The loss $L(\hat{y}, y)$ usually assigns a numerical value for the output $\hat{y}$ given the true expected output $y$.

The **loss function should be bounded from below**, with the minimum attained only for cases where the network’s output is correct.
For binary classification problems, the output is a single value $\hat{y}$ and the intended output $y$ is in $\{+1, -1\}$.

The classification rule is $\text{sign}(\hat{y})$, and a classification is considered correct if $y \cdot \hat{y} \geq 0$, meaning that $y$ and $\hat{y}$ share the same sign. The hinge loss, also known as margin loss:

$$L_{\text{hinge(binary)}}(\hat{y}, y) = \max(0, 1 - y \cdot \hat{y})$$

The loss is 0 when $y$ and $\hat{y}$ share the same sign and $|\hat{y}| \geq 1$. Otherwise, the loss is linear.

The binary hinge loss attempts to achieve a correct classification, with a margin of at least 1.
Categorical Cross-Entropy Loss

The categorical cross-entropy loss (negative log likelihood) is used when a probabilistic interpretation of the scores is desired.

Let $y = y_1, \ldots, y_n$ be a vector representing the distribution over the labels $1, \ldots, n$, and let $\hat{y} = \hat{y}_1, \ldots, \hat{y}_n$ be the classifier's output.

The categorical cross entropy loss measures the dissimilarity between the true label distribution $y$ and the predicted label distribution $\hat{y}$, and is defined as cross entropy:

$$L_{\text{cross-entropy}}(\hat{y}, y) = - \sum_i y_i \log(\hat{y}_i)$$
Further Examples

- Zero-one loss: standard loss function in classification

\[ L(y, y') = 1_{y \neq y'}, \quad y, y' \in \mathcal{Y} \]

- Non-symmetric losses: e.g., for spam classification

\[ L(\hat{\text{ham}}, \text{spam}) \leq L(\hat{\text{spam}}, \text{ham}) \]

- Squared loss: standard loss function in regression;

\[ L(y, y') = (y' - y)^2 \]