Loss Formulations

ENEE 691 – Machine Learning and Photonics @ UMBC

Ergun Simsek, Ph.D. and Masoud Soroush, Ph.D.

April 17, 2023

1 Loss Formulations

1.1 Regression

Assume x and y are D dimensional vectors, and x_i denotes the value on the ith dimension of x.

Mean Absolute Error (MAE)

$$\sum_{i=1}^{D} |x_i - y_i| \tag{1}$$

Mean Squared Error (MSE)

$$\sum_{i=1}^{D} (x_i - y_i)^2 \tag{2}$$

Huber Loss (less sensitive to outliers than the MSE as it treats error as square only inside an interval)

$$L_\delta = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & if \, |(y - \hat{y})| < \delta \\ \delta((y - \hat{y}) - \frac{1}{2}\delta) & otherwise \end{cases} \tag{3}$$

1.2 Classification

Cross Entropy

Cross-entropy is a measure that quantifies the difference between two probability distributions for a given random variable or set of events. In information theory, one measures the amount of information required to encode and transmit an event in bits. The amount of information conveyed by an event is inversely proportional to its probability, with lower probability events having more information and higher probability events having less information.

In information theory, the concept of "surprise" is used to describe the unexpectedness of an event. A less likely event is more surprising, and therefore contains more information. One can calculate the information $h(x)$ of an event x, given its probability $P(x)$, using the formula $h(x) = -\log(P(x))$.

Entropy, on the other hand, is a measure of the number of bits required to transmit a randomly selected event from a probability distribution. A skewed distribution, where certain events are more likely than others, has a low entropy, while a distribution where events have equal probability has a higher entropy.

A skewed probability distribution has less surprise and, consequently, a low entropy because likely events dominate. In contrast, a balanced distribution is more surprising and has higher entropy because events are equally likely. The entropy $H(x)$ of a random variable with a set of x in N discrete states and their probability $P(x)$ can be calculated using the formula $H(N) = -\sum_{x=1}^{N} P(x) \log(P(x))$.

Cross-entropy extends the idea of entropy from information theory and calculates the number of bits required to represent or transmit an average event from one distribution compared to another distribution. Basically, the cross entropy is the average number of bits needed to encode data coming from a source with distribution p when we use model q.

- In binary classification, where the number of classes M equals 2, Binary Cross-Entropy(BCE) can be calculated as:
$$-(y \log(p) + (1-y)\log(1-p)) \tag{4}$$
- If $M > 2$ (i.e. multiclass classification), we calculate a separate loss for each class label per observation and sum the result.
$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{5}$$

where
M - number of classes
log - the natural log
y - binary indicator (0 or 1) if class label c is the correct classification for observation o
p - predicted probability observation o is of class c

Negative Log-likelihood

$$NLL(y) = -\log(p(y)) \tag{6}$$

Minimizing negative loglikelihood

$$\min_{\theta} \sum_{y} -\log(p(y;\theta)) \tag{7}$$

is equivalent to Maximum Likelihood Estimation(MLE).

$$\max_{\theta} \prod_{y} p(y;\theta) \tag{8}$$

Here $p(y)$ is a scaler instead of vector. It is the value of the single dimension where the ground truth y lies. It is thus equivalent to cross entropy.

Hinge loss (Used in SVMs)

$$max(0, 1 - y \cdot \hat{y}) \tag{9}$$

KL/JS divergence

$$KL(\hat{y}||y) = \sum_{c=1}^{M} \hat{y}_c \log \frac{\hat{y}_c}{y_c} \tag{10}$$

$$JS(\hat{y}||y) = \frac{1}{2}(KL(y||\frac{y+\hat{y}}{2}) + KL(\hat{y}||\frac{y+\hat{y}}{2})) \tag{11}$$

the KL divergence is often referred to as the "relative entropy" because KL Divergence is basically the average number of extra bits to represent an event from Q instead of P, while the cross-entropy is the average number of total bits to represent an event from Q instead of P.

1.3 Regularization

The *Error* below can be any of the above loss.

L1 regularization: A regression model that uses L1 regularization technique is called Lasso Regression.

$$Loss = Error(Y - \widehat{Y}) + \lambda \sum_{1}^{n} |w_i| \tag{12}$$

L2 regularization: A regression model that uses L1 regularization technique is called Ridge Regression.

$$Loss = Error(Y - \widehat{Y}) + \lambda \sum_{1}^{n} w_i^2 \tag{13}$$

1.4 Some More Classification Metrics

Some of them overlaps with loss, like MAE, KL-divergence.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{17}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{18}$$

$$Specificity = \frac{TN}{FP + TN} \tag{19}$$

AUC is calculated as the Area Under the *Sensitivity*(TPR)-$(1 - Specificity)$(FPR) Curve.

1.5 Similarity/Relevance

$$Cosine(x, y) = \frac{x \cdot y}{|x||y|} \tag{20}$$

Jaccard (Similarity of two sets U and V.)

$$Jaccard(U, V) = \frac{|U \cap V|}{|U \cup V|} \tag{21}$$

Pointwise Mutual Information (PMI) (Relevance of two events x and y)

$$PMI(x; y) = \log \frac{p(x, y)}{p(x)p(y)} \tag{22}$$

For example, $p(x)$ and $p(y)$ is the frequency of word x and y appearing in corpus and $p(x, y)$ is the frequency of the co-occurrence of the two.