Machine Learning – Lecture Notes

Instructor: Dr. Ali Hassan

1 Empirical Risk Minimization

Empirical risk minimization (ERM) is a principle in statistical learning theory that is used to give theoretical bounds on the performance of learning algorithms. Consider the following situation, which is a general setting of many supervised learning problems. We have two spaces of objects $\boldsymbol{X}$ and $\boldsymbol{Y}$ and we would like to learn a function $h : \boldsymbol{X} \to Y$ which outputs an object $y \in Y$, given $x \in X$. To do so, we have at our disposal a training set of a few examples $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in X$ is an input and $y_i \in Y$ is the corresponding response that we wish to get from $h(x_i)$.

To put it more formally, we assume that there is a joint probability distribution $P(x, y)$ over $\boldsymbol{X}$ and $\boldsymbol{Y}$, and that the training set consists of m instances $(x_1, y_1), \ldots, (x_m, y_m)$ drawn identically and independently drawn (i.i.d) from $P(x, y)$. Note that the assumption of a joint probability distribution allows us to model uncertainty in predictions (e.g. from noise in data) because y is not a deterministic function of x, but rather a random variable with conditional distribution $P(y|x)$ for a fixed x.

We also assume that we are given a non-negative real-valued loss function $L(\hat{y}, y)$ which measures how different the prediction$\hat{y}$ of a hypothesis is from the true outcome y. The risk associated with hypothesis $h(x)$ is then defined as the expectation of the loss function:

$$R(h) = \boldsymbol{E}[L(h(x), y)] = \int L(h(x), y)\, dP(x, y). \tag{1}$$

A loss function commonly used in theory is the 0-1 loss function:$L(\hat{y}, y) = I(\hat{y} \neq y)$, where $I(...)$ is the indicator notation. There are many other types of loss functions that can be used but we will discuss them later.

The ultimate goal of a learning algorithm is to find a hypothesis h^* among a fixed class of functions $\mathcal{H}$ for which the risk $R(h)$ is minimal:

$$h^* = \arg\min_{h \in \mathcal{H}} R(h). \tag{2}$$

1.1 Empirical

The reason that this method is called *empirical* risk minimization is because in general, the risk $R(h)$ cannot be computed because the distribution $P(x, y)$ is unknown to the learning algorithm. However, we can compute an approximation, called empirical risk, by averaging the loss function on the training set:

$$R_{\text{emp}}(h) = \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), y_i). \tag{3}$$

1.2 Commonly Used Binary Classification Loss Functions

Here are some binary loss function for classification of $y \in \{-1, +1\}$ Here we are doing to define the loss function defined as $\ell(h_\theta(\mathbf{x}_i, y_i))$. Remember over here that the hypothesis is parametrized by the θ given by the equation:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \ldots + \theta_n x_n \quad (4)$$
$$h_\theta(x) = \boldsymbol{\theta}^T \boldsymbol{x} \quad (5)$$

Remember that in some book related to statistics or mathematics, they use w instead of θ to parametrize the hypothesis. This is only the difference in the notation but they exactly the same.

1.2.1 Zero-One Loss Function

The zeros-one loss function is defined by the following function:

$$\delta(\text{sign}(h_\theta(\mathbf{x}_i)) \neq y_i) \quad (6)$$

This is a step function hence it is non continuous and is usually used as the actual classification loss to check if the prediction was correct or not.

1.2.2 Log-Loss

This type of loss is one of the most popular loss function in machine learning and is very commonly used in logistic regression. It is defined by the following equation:

$$log(1 + e^{-h_\theta(\mathbf{x}_i) y_i}) \quad (7)$$

Plot of these loss functions are:

1.3 Commonly Used Regression Loss Functions

Commonly used regression loss functions where $y \in \mathbb{R}$ are the squared and absolute loss. Both of them can be used but have different properties:

1.3.1 Square Loss Function

This loss function is the most commonly used function. It is defined by the following equation:

$$(h(\mathbf{x}_i) - y_i)^2 \quad (8)$$
$$(\boldsymbol{X}^T \boldsymbol{\theta} - \boldsymbol{y})^2 \quad (9)$$

The squared loss function has following properties:

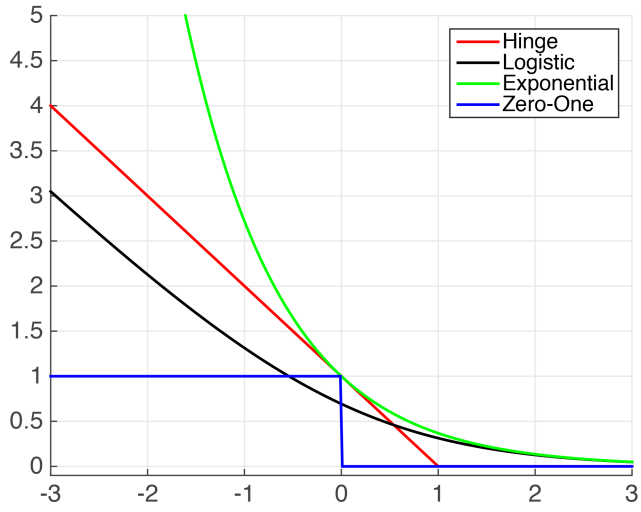

Figure 1: Plots of Common Classification Loss Functions - x-axis: $h(\mathbf{x}_i)y_i$, or "correctness" of prediction; y-axis: loss value.

- Most popular regression loss function
- Estimates Mean Label
- ADVANTAGE: Differentiable everywhere
- DISADVANTAGE: Somewhat sensitive to outliers/noise
- Also known as Ordinary Least Squares (OLS)

1.3.2 Absolute Loss Function

The absolute loss function is defined by the following equation:

$$|h(\mathbf{x}_i) - y_i| \tag{10}$$
$$|\boldsymbol{X}^T\boldsymbol{\theta} - \boldsymbol{y}| \tag{11}$$

- Also a very popular loss function
- Estimates Median Label
- ADVANTAGE: Less sensitive to noise
- DISADVANTAGE: Not differentiable at 0

Plots of Common Regression Loss Functions are:

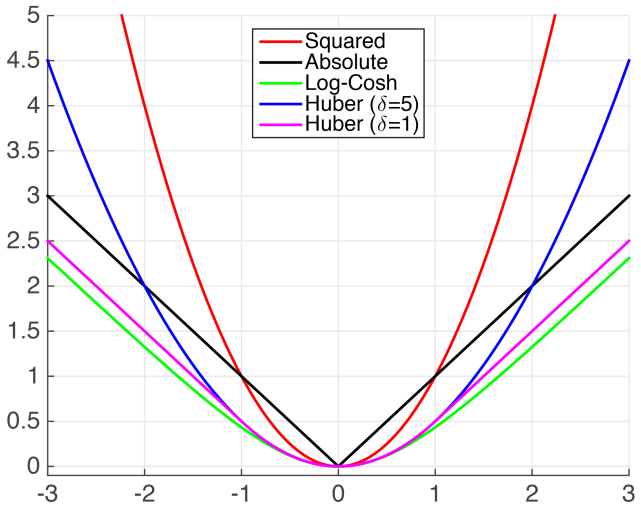

Figure 2: Plots of Common Regression Loss Functions - x-axis: $h(\mathbf{x}_i)y_i$, or "error" of prediction; y-axis: loss value.

Table 1: Special Casses for the Loss Functions.

Loss function and Regularizer	**Function**	**Classification**	**Solutions**
Ordinary Least Square	$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{w}^\top x_i - y_i)^2$	Squared Loss, No Regularization	$\mathbf{w} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}^\top$.
Ridge Regression	$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{w}^\top x_i - y_i)^2 + \lambda \|w\|_2^2$	Squared loss, l_2 regularization	$w = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbb{I})^{-1}\mathbf{X}^\top \mathbf{y}^\top$
Logistic Regression	$\min_{\mathbf{w},b} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i(\mathbf{w}^\top \mathbf{x}_i + b)}\right)$	often l_1 or l_2 regularized	$\Pr(y = +1\|x) = \frac{1}{1+e^{-y(\mathbf{w}^\top x + b)}}$

1.4 Famous Special Cases

This section gives the details of some well known loss function used in several well known optimization problems: