Tutorial 1 Questions

(1) Assume that we are given a set of features $\{(x_i, y_i) \,|\, i \in \{1, 2, ..., N\}\}$ with $x_i \in R^d$, $y \in \{-1, +1\}$. We wish to train a function $h : R^d \to R$, so that $\text{Sign}(h(x)) = y$. To that aim, we seek to solve the following:

$$\underset{h\in}{\text{minimize}} \sum_{i=1}^{N} [\text{Sign}(h(x_i)) \neq y_i]$$

Moreover, H is the set of all functions that map from R^d to R. This problem is hard to solve in general. That is why, we resort to several approximations. In the following, mark and explain which ones are good approximator of $I[\text{Sign}(h(x_i)) \neq y_i]$ in the above equation.

$$(i) \quad \max\{0, 1 - y_i \cdot h(x_i)\} \quad \text{(Yes/No)}$$
$$(ii) \quad \min\{0, 1 - y_i \cdot h(x_i)\} \quad \text{(Yes/No)}$$
$$(iii) \quad \frac{\exp(-y_i \cdot h(x_i))}{1 + \exp(-y_i \cdot h(_i))} \quad \text{(Yes/No)}$$
$$(iv) \quad \frac{1}{1 + \exp(-y_i \cdot h(x_i))} \quad \text{(Yes/No)}$$

(2) Suppose we restrict $h(x) = w^T x + b$, *i.e.,* $h(x)$ is a linear function. Then write the approximation of the optimization problem defined in the above question in terms of any (correct) one approximation in the previous question. Specifically, fill up the gaps

$$\text{minimize} \sum_{i=1}^{N} ??$$

(3) Suppose $h(x) = \text{sign}(f(x))$ where $h(x) : R^d \to \{+1, -1\}$. We now cosider a loss function defined as $\sum_{(x_i,y_i)} \ell(y_i f(x_i))$. *i.e.,* ℓ is a function of $yf(x)$. Given below are some graphs with x axis as $yf(x)$ and y axis as the loss ℓ value. Identify the graphs that are adept

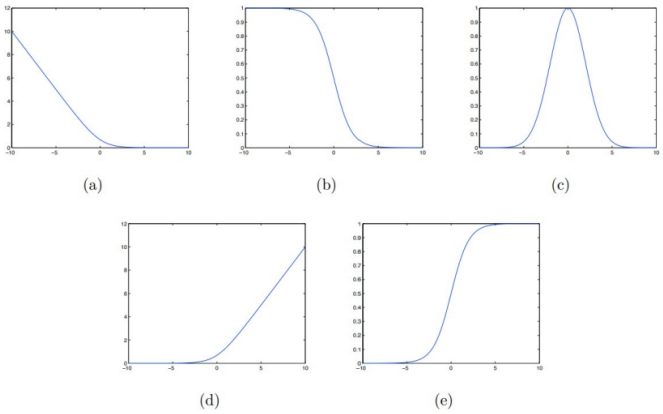


Figure 1. Loss Plots

for classification task.

(4) Consider a Binary classification problem where the dataset D_{Train} is imbalanced. We have 90% examples that belong to class $+1$ and the remaining examples with class -1.

- What is your guess for the best $h \in$ All constant model?
- Compute $Error(h^*) - Error(\hat{h})$ for your guess. Assume that the test set is well-balanced.

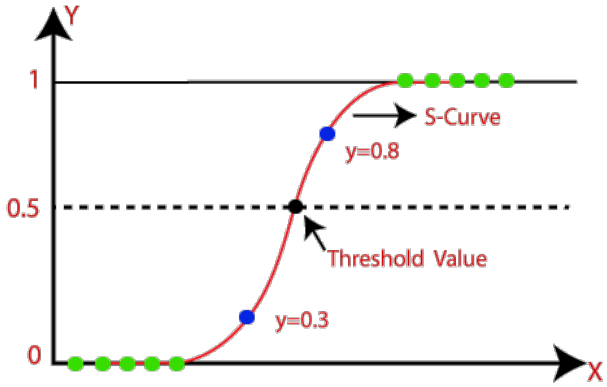
(5) Now, let us consider a weighted loss function given by:

$$\{w^*, b^*\} = \arg\min_{w,b} \sum_{i=1}^{M} r_i \max\left(0, \left(\frac{1}{2} - f(x_i)\right) y_i\right)$$

where $r_i > 0$ are weights associated with loss of each example. Can you propose a weighting scheme for r_i and justify your choice?

Repeat the exercise for the case when test set is also imbalanced with 60% test set examples that belong to class $+1$

(6) Recall that Logistic Regression model is given by: $h(x) = \frac{1}{1+e^{-w^T x}}$ where the labels are binary $\mathcal{Y} = \{0, 1\}$



And the loss that we minimize is called *cross-entropy* loss

$$\sum_{(x_j, y_j) \in D_{Train}} -\{y_i \log h(x_i) + (1 - y_i) \log(1 - h(x_i))\}$$

Finally the decision rule is given by $h(x_i) > 0.5$

- Argue that cross entropy loss is a valid loss function.
- What is $||w||$ when training loss is 0. Assume that all features have unit norm $||x|| = 1$
- Is it wrong, if we take $h(x) = \frac{1}{1+e^{-w^T x}}$. Can you tell verbatim, what interpretations change now?

(7) Now given D_{Test}, the instructor allows you to change the model by modifying the decision rule as $h(x_i) > \tau$ where $\tau \in [0, 1]$. You are free to cheat by inspecting the test set and choosing a τ of your choice. However, you cannot change $\hat{w}, \hat{b}$. Let us evaluate the choices made by the following students:

- Naive student 1: Choose $\tau = 0$
- Naive Student 2: choose $\tau = 1$
- Millennial: choose $\tau = 0.5$
- What would the class choose? Can you pose it as an optimization problem by proposing a loss function and picking τ^* by means of minimizing it?

(8) A function $f(x)$ is said to be linear in x if it satisfies the following two properties
   (a) $f(x+y) = f(x) + f(y)$
   (b) $f(\alpha x) = \alpha f(x)$
   Are the following equations linear. If yes, then with respect to what parameters?

   (a) $f(x) = w_1 * x_1 + w_2 * x_2$
   (b) $f(x) = w_1 * x_1^2 + w_2 * x_2^3$
   (c) $f(x) = w_1 * \ln x_1 + w_2 * e^{x_2}$
   (d) $f(x) = x_1 * \ln w_1 + x_2 * e^{w_2}$
   (e) $f(x) = w^T x \quad w, x \in \mathbb{R}^d$
   (f) $f(x) = w^T x + b \quad w, x \in \mathbb{R}^d \quad b \in \mathbb{R}$

(9) **L-2 Loss** in case of linear regression was defined as follows

   $\mathcal{L}_2(w) = \sum_{i=1}^{N} (y_i - wx_i - b)^2$

   $x_i \in$R, $w \in$R, $b \in$R

   The interesting thing about linear regression is there exist a closed form solution. This means that the solution can be calculated by minimizing the above function.

   Take a gradient of the loss function stated above and prove that the solutions for 1-dimensional case are

   $\widehat{w} = \sum_{i=1}^{N} \frac{(x_i - \overline{x})(y_i - \overline{y})}{(x_i - \overline{x})^2}$

   $\widehat{b} = \overline{y} - \widehat{w}\overline{x}$

(10) **L-2 Loss** in case of linear regression was defined as follows

   $\mathcal{L}_2(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2$

   This loss can be neatly written with the help of design matrix X and label vector Y

   Prove that : $\mathcal{L}_2(w) = ||Xw - Y||^2$

   Now we can take the gradient of the loss function stated above and prove that the solutions for general case. However while taking the gradient a little bit of matrix calculus will be used. We can then finally show that taking the gradient of $\mathcal{L}_2(w)$ and putting it to zero leads us to the normal equations

   Derive $X^T X w = X^T Y$

(11) **Design Matrix** $X \in \mathbb{R}^{nXd}$ is a matrix where all samples of the dataset are stacked one below the other. More specifically

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & . & . & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & . & . & x_d^{(2)} \\ x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & . & . & x_d^{(3)} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_1^{(n)} & x_2^{(n)} & x_3^{(n)} & . & . & x_d^{(n)} \end{bmatrix}$$

Here $x_k^{(i)}$ is the k^{th} feature of i^{th} datapoint vector

Recall that the closed form solution of L-2 regression is $(X^TX)^{-1}X^TY$

Prove that the inverse of X^TX exist.

(12) Although $(X^TX)^{-1}$ does not always exist. $(X^TX + \lambda I)^{-1}$ however does exist. To prove this we will need to understand the definition of positive definite matrices

Given a $n \times n$ matrix M The condition for positive definiteness is

M positive-definite $\iff$ $\mathbf{v}^\mathsf{T} M\mathbf{v} > 0$ for all $\mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$

A positive definite matrix has a non zero determinant. Therefore its inverse always exists.

Can you prove that $(X^TX + \lambda I)$ is positive definite

(13) The Linear regression problem can be modelled in a probabilistic way under the assumptions

$$Y_i = w^T x_i + \epsilon_i,$$

$\epsilon_i \sim$ N(0,,σ^2)

$Y_i \sim$ N(w$^T x_i, , \sigma^2$)

Prove that the maximising the Likelihood of Data

$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^n$

is equivalent to minimizing the l2-loss that we proposed earlier for the standard regression problem