CS419M

Lecture 11: Tutorial on Loss, Classification & Regression

17 Feb 2023

Lecturer: Abir De *Scribe: Groups 21 & 22*

This tutorial covers the problem based on loss functions, Regression and Classification.

Question 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.

$$(i) \quad \max\{0, 1 - y_i \cdot h(x_i)\} \quad \text{(Yes/No)}$$

$$(ii) \quad \min\{0, 1 - y_i \cdot h(x_i)\} \quad \text{(Yes/No)}$$

$$(iii) \quad \frac{\exp(-y_i \cdot h(x_i))}{1 + \exp(-y_i \cdot h(_i))} \quad \text{(Yes/No)}$$

$$(iv) \quad \frac{1}{1 + \exp(-y_i \cdot h(x_i))} \quad \text{(Yes/No)}$$

Solution:

$$\textbf{case} : \textbf{1} \quad if \quad y_i \quad and \quad h(x_i) \quad have \quad opposite \quad signs, \quad then$$
$$1 - y_i.h(x_i) \in [1, \infty)$$
$$\textbf{case} : \textbf{2} \quad if \quad y_i \quad and \quad h(x_i) \quad have \quad same \quad signs, \quad then$$
$$1 - y_i.h(x_i) \in (-\infty, 1]$$

The original loss function can take only discrete values 0 and 1. The good approximator of I[Sign $h(x_i) \neq y_i$] will be the one which penalizes based on how far the point is.

y_i	$h(x_i)$	$max\{0, 1 - y_i.h(x_i)\}$	$min\{0, 1 - y_i.h(x_i)\}$	original
H	H	L	H	L
H	L	H	L	H
L	H	H	L	H
L	L	L	H	L

Hence, (i) is a good approximator.

y_i	$h(x_i)$	(iii)	(iv)	original
H	H	L	H	L
H	L	H	L	H
L	H	H	L	H
L	L	L	H	L

(iii) ia also a good approximator.

Question 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

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

Solution:

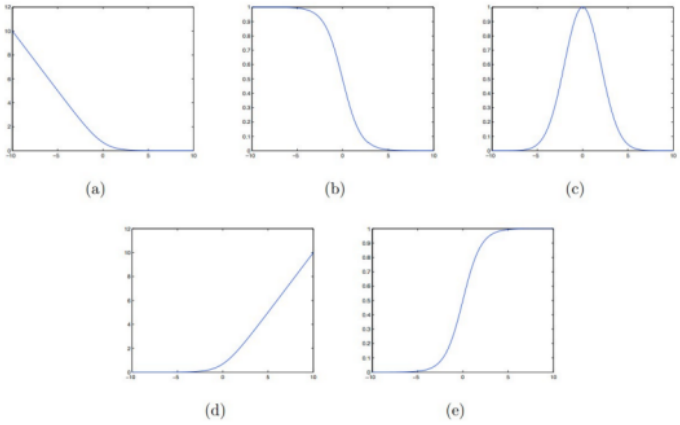
As specified in Q.1 we can conclude that,

$$\min_{w,b} \sum_{i=0}^{n} max\{0, 1 - y_i(w^\top x + b)\}$$

is the approximation of optimization problem.

Question 3

Suppose $h(x) = \text{sign}(f(x))$ where $h(x) : R^d \rightarrow \{+1, -1\}$. 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



Solution:

y	$f(x)$	$loss$	$y.f(x)$
H	H	L	H
H	L	H	L
L	H	H	L
L	L	L	H

Only graphs (a) and (b) satisfy above condition.

Question 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.

Solution:

$h\hat{}(x_i)$ is a model which maximizes accuracy over D_{train} and $h^*(x_i)$ is a model which maximizes accuracy on D_{test}

part (a):

h is a constant model $\implies$ h(x_i) $= C$
90% of the test data contains examples that belong to class +1. Hence, best guess for h would be +1.

part (b):

Test set is well balanced.
$h^*(x_i)$ = +1/-1 with the accuracy of 50%.
$h\hat{}(x_i)$ = +1 with 50% accuracy over test set.

$\therefore Error$(h*) - Error($h\hat{}$) = 0.

Question 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

Solution:

We have 90% examples in +1 and 10% in -1. So we need to choose r_i such that the training set gets balanced. The technique of weighting is utilized to address the bias in the training dataset. When the dataset is highly imbalanced, minimizing the unweighted loss function leads to a bias towards the majority class in the training set. This may be inappropriate if the test set is nearly balanced. In such a scenario, it is beneficial to assign greater weight to the loss from the minority class. One possible scheme could be choosing

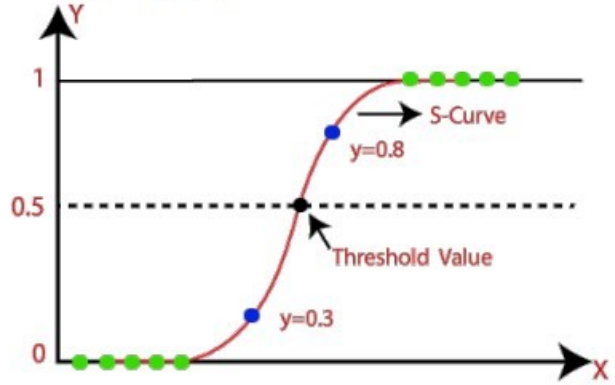
$$r_i = 9 \quad \text{if} \quad y = -1$$
$$r_i = 1 \quad \text{if} \quad y = +1$$

Here r_i values are such that a misclassification of -1 will be penalized 9 times more than a misclassification of +1. Another way to look at it is as if we are replacing the actual dataset with a new dataset where number of -1 is 9 times more than the original dataset to balance the test set. In case of 60% +1 and 40% -1 test set we need to take value of r_i less than 9

$$1 \leq r_i < 9 \quad \text{if} \quad y = -1$$
$$r_i = 1 \quad \text{if} \quad y = +1$$

Question 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?

Solution:

When y_i = +1 we are penalizing $-\log h(x_i)$
When y_i = 0 we are penalizing $-\log(1 - h(x_i))$ where $0 < h(x_i) \leq 1$

Part 1

y	h(x)	loss
H	H	L
H	L	H
L	H	H
L	L	L

H: High
L: Low

Part 2

For the training loss to be zero at y_i = +1, $h(x_i)$ has to be 1 and at y_i = 0, $h(x_i)$ has to be 0. Moreover, each term in the loss function is non-negative. Therefore they have to 0. Which gives $\|w\| = \infty$

Part 3

This can be obtained by a simple rotation about the y-axis, i.e. by putting x as -x, $h(x_i) \rightarrow 1-h(x_i)$. The new loss function will be

$$\sum_{(x_i,y_i)\in\mathcal{D}_{\text{Train}}} -\{y_i \log(1-h(x_i)) + (1-y_i)\log h(x_i)\}$$

Question 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?

Solution:

If τ is chosen to 0 or 1 then all the points will be classified as the same class, i.e. for τ = 0, the model will always classify any point as +1 and for τ = 1 it will classify any point as 0.

For τ = 0.5, it is a good choice when we don't know anything about the test set, it will give the same value as $\hat{h}$. But if we have additional data about the test set then we can do better.
Optimized value of τ can be determined as

$$\tau^* = \text{argmin}_{\tau \in [0,1]} \quad \Sigma_{D_{test}}[\text{sign}(h(x_i) - \tau) \neq y_i]$$

Question 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) $f(x) = w^T x + b \quad w, x \in \mathbb{R}^d \quad b \in \mathbb{R}$

Solution:

Take

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

1.
$$f(x+y) = w^T(x+y) = w^T x + w^T y = f(x) + f(y)$$
$$f(\alpha x) = w^T(\alpha x) = \alpha w^T x = \alpha f(x)$$
$$f(w_1 + w_2) = (w_1 + w_2)^T x = (w_1^T x + w_2^T x) = f(w_1) + f(w_2)$$
$$f(\alpha w) = (\alpha w)^T x = \alpha w^T x = \alpha f(w)$$

So linear in both w and x.

2.
$$f(x+y) = w_1 * (x_1 + y_1)^2 + w_2 * (x_2 + y_2)^3 \neq f(x) + f(y)$$
$$f(\alpha x) = w_1 * \alpha^2 x_1^2 + w_2 * \alpha^3 x_2^3 \neq \alpha f(x)$$

So linear in w but not in x.

3.

$$f(x+y) = w_1 * \ln(x_1 + y_1) + w_2 * e^{(x_2+y_2)} \neq f(x) + f(y)$$
$$f(\alpha x) = w_1 * \ln(\alpha x) + w_2 * e^{\alpha x} \neq \alpha f(x)$$

linear in w not in x.

4. Simply changing x and w in part (3), linear to x but in w.

5. Similar to (1), but more case, linear in both x and w.

6.

$$f(x+y) = w^T(x+y) + b = w^T x + w^T y + b = (w^T x + b) + (w^T y + b) - b \neq f(x) + f(y)$$
$$f(\alpha x) = w^T(\alpha x) + b = \alpha w^T x + b \neq \alpha f(x)$$
$$f(w_1 + w_2) = (w_1 + w_2)^T x + b = (w_1^T x + w_2^T x) + b \neq f(w_1) + f(w_2)$$
$$f(\alpha w) = (\alpha w)^T x + b = \alpha w^T x + b \neq \alpha f(w)$$

If we take w$'$ = [b w_1 w_2 ...]^T and x$'$ = [1 x_1 x_2 ...] then we can write f(x) as $f(x) = w'^T x'$. Which is linear in both w' and x'.

Question 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

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

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

Solution:

$$L_2(w,b) = \sum_{i=1}^{N}(y_i - wx_i - b)^2$$

$$\frac{\partial L_2(w,b)}{\partial w} = 0 \; at \; optimal \; \hat{w} \; and \hat{b}$$

$$\sum_{i=1}^{N} 2(y_i - wx_i - b)x_i = 0$$

$$\sum[(y_i - \bar{y}) - w(x_i - \bar{x})]\,(x_i - \bar{x}) = 0$$

$$\hat{w} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})]}{\sum(x_i - \bar{x})^2}$$

$$\frac{\partial L_2(w,b)}{\partial b} = 0 \; at \; optimal \; w \; and \; b$$

$$\frac{\sum_{i=1}^{N} y_i}{N} - \hat{w}\frac{\sum_{i=1}^{N} x_i}{N} - \frac{\sum_{i=1}^{N} \hat{b}}{N} = 0$$

$$\hat{b} = \bar{y} - \hat{w}\bar{x}$$

Question 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 $X^T Xw = X^T Y$

Solution:

$$L_2(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2$$

$$writing\ this\ in\ matrix\ form:$$

$$L_2(w) = \| \begin{bmatrix} (y_1 - w^T x_1) \\ \vdots \\ (y_n - w^T x_n) \end{bmatrix} \|^2$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

Hence this can be written as:

$$\|Xw - Y\|^2$$

Remember that:

$$\frac{\partial \|X\|^2}{\partial x} = 2X$$

$$\frac{\partial Ax}{\partial x} = A^T$$

So we get:

$$\frac{\partial \|Xw - Y\|^2}{\partial x} = 2X^T(Xw - Y)$$

$$X^T(Xw - Y) = 0$$

$$X^T Xw - X^T y = 0$$

$$X^T Xw = X^T y$$

$$w = (X^T X)^{-1} XY$$

Question 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 $(X^T X)^{-1} X^T Y$
Prove that the inverse of $X^T X$ exist.

Solution:

$X^T X$ is invertible if X has a full column rank

If a matrix A has a full column rank then

$Ax = b$ has only one solution i.e. x = 0

so here we have $A^T A$

$$A^T Ax = 0$$
$$\implies x^T A^T Ax = 0$$
$$\implies \|Ax\| = 0$$
$$\implies Ax = 0$$
$$\implies x = 0$$

Hence we can say $X^T X$ has a full column rank and is invertible.

Question 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

Solution:

we have: $(X^TX + \lambda I)$

$$V^T(X^TX + \lambda I)V$$

$$\implies V^TX^TXV + \lambda V^TV$$

$$\implies \|XV\| + \lambda\|V\| \geq 0$$

Hence $(X^TX + \lambda I)$ is positive definite

Question 13

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

$$Y_i = w^Tx_i + \epsilon_i,$$

$\epsilon_i \sim$ N(0,,σ^2)

$Y_i \sim$ N($\mathrm{w}^Tx_i, , \sigma^2$)

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

Solution:

$$Y_i = w^T x_i + \varepsilon$$

Likelihood of data is given by:

$$P_i = \frac{1}{2} e^{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}}$$

$$\Pi P = \frac{1}{2} e^{-\frac{\sum (y_i - w^T x_i)^2}{2\sigma^2}}$$

we can see that maximizing this would be equivalent to:

$$\min \sum (y_i - w^T x_i)^2$$

References