Introduction to Large Language Model

Lecture 2: Preliminary - Linear and logistic regressoin

Kun Yuan

Peking University

Main contents in this lecture

- Linear regression
- Logistic regression
- Multi-class classification

Motivation

- You consider renting an apartment
- You don't know whether the price the agent offered is good or not
- You collect a dataset

Table: Collected dataset

Size (x_1)	**Location** (x_2)	**Green rate** (x_3)	**Decoration** (x_4)	**Price** (y)
80 m^2	8	20%	6	10000
60 m^2	10	30%	8	9000
100 m^2	5	20%	5	9000
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
70 m^2	10	25%	9	12000

Motivation

- Below is your target apartment's description. What should be the reasonable price for this apartment?

Table: Your target apartment

Size (x_1)	**Location** (x_2)	**Green rate** (x_3)	**Decoration** (x_4)	**Price** (y)
100 m^2	8	35%	8	?

- You need to learn how (x_1, x_2, x_3, x_4) will map to y from your dataset
- This is a typical task in machine learning: linear regression

Linear regression

- Consider a set of data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ where

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id}) \in \mathbb{R}^d$$

  is the feature vector, e.g., $x_{i1} =$ "Size" and $x_{i2} =$ "Location", etc., and y is the label, e.g., $y =$ "Price"

- We assume the mapping between $\boldsymbol{x}_i$ and y_i is in the **linear** form

$$y_i \approx \boldsymbol{x}_i^\top \boldsymbol{w} \tag{1}$$

  where $\boldsymbol{w} \in \mathbb{R}^d$ is the unknown parameter to learn

- If the parameter $\boldsymbol{w}$ is known, given a new feature vector $\boldsymbol{x}$ (e.g., the data for your target apartment), you can estimate its lable y according to (1)

Linear regression

- How to get the parameter $\boldsymbol{w}$? We can calculate it with $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$
- A good $\boldsymbol{w}$ will incur the minimum estimation error

$$\boldsymbol{w}^\star = \arg\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (\boldsymbol{x}_i^\top \boldsymbol{w} - y_i)^2 \right\} \tag{2}$$

  where is called the linear regression problem

- If we introduce

$$X = [\boldsymbol{x}_1^\top; \cdots; \boldsymbol{x}_N^\top] \in \mathbb{R}^{N \times d} \quad y = [y_1; y_2; \cdots; y_N] \in \mathbb{R}^N$$

  problem (2) becomes

$$\boldsymbol{w}^\star = \arg\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2} \|X\boldsymbol{w} - y\|^2 \right\}$$

Solve the linear regression problem

- Consider the linear regression problem

$$\boldsymbol{w}^\star = \arg\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2} \|X\boldsymbol{w} - y\|^2 \right\}$$

- Let $f(\boldsymbol{w}) = \frac{1}{2}\|X\boldsymbol{w} - y\|^2$, the gradient is given by

$$\nabla f(\boldsymbol{w}) = X^\top (X\boldsymbol{w} - y)$$

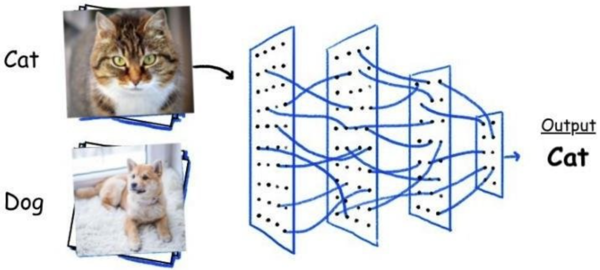
- The gradient descent is

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \gamma X^\top (X\boldsymbol{w}_k - y)$$

A code example

Logistic regression

- Another important machine learning task is classification



Logistic regression

- Again, we collect the dataset

Size (x_1)	**Ear shape** (x_2)	**Tail length** (x_3)	**Color** (x_4)	**Label** (y)
100 cm	round	30cm	yellow	dog
40 cm	triangle	20cm	white	cat
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- We need to establish the mapping beteween (x_1, x_2, x_3, x_4) and the discrite lable $y \in \{0, 1\}$ in which 1 indicates dog while 0 indicates cat

An intuitive approach

- We associate each feature item x_i with a weight w_i
- An intuivie hard classification approach is

$$(x_1, x_2, \cdots, x_d) \quad \longrightarrow \quad y = \begin{cases} 1 & \text{if } \sum_{i=1}^{d} x_i w_i > c \\ 0 & \text{otherwise} \end{cases}$$

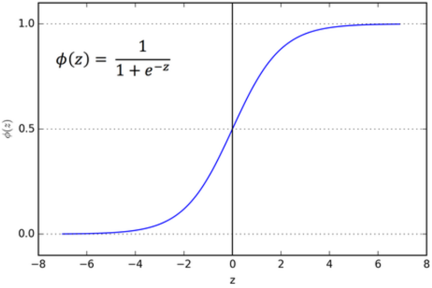
  where c is a pre-defined threshold

- While intuitive, it is hard to construct smooth loss functions that facilitate to learn the weights (parameters) $\{w_i\}_{i=1}^{d}$

Sigmoid function

- Now we consider a different approach
- Sigmoid function maps $[-\infty, +\infty]$ to $[0, 1]$

$$\phi(z) = \frac{1}{1 + e^{-z}}$$



Predicted probability

- With sigmoid function, we can map $(x_1, \cdots, x_d)$ to a probability

$$p(z) = \frac{1}{1 + e^{-z}} \in (0, 1) \quad \text{where} \quad z = \sum_{i=1}^{d} w_i x_i \tag{3}$$

- With (3), we map $(x_1, \cdots, x_d)$ to a probability distribution

$$(x_1, \cdots, x_d) \quad \longrightarrow \quad \begin{bmatrix} p(z) \\ 1 - p(z) \end{bmatrix} \in \mathbb{R}^2$$

where $p(z)$ is the probability that $(x_1, \cdots, x_d)$ belongs to class 1

Real probability

- Given the label y, the real probability distribution is

$$\begin{bmatrix} y \\ 1-y \end{bmatrix} \in \mathbb{R}^2$$

  where label $y \in \{0, 1\}$ can be regarded as the probability of class 1

- We need to measure the difference between

$$\text{(Predicted prob.)} \quad \begin{bmatrix} p(z) \\ 1-p(z) \end{bmatrix} \quad \text{and} \quad \text{(Real prob.)} \quad \begin{bmatrix} y \\ 1-y \end{bmatrix}$$

Cross entropy

- Cross entropy can measure the difference between two probability distributions $\boldsymbol{p} \in \mathbb{R}^d$ and $\boldsymbol{q} \in \mathbb{R}^d$

$$H(\boldsymbol{p}, \boldsymbol{q}) = -\sum_{i=1}^{d} p_i \log(q_i)$$

  Smaller cross entropy indicates smaller difference between $\boldsymbol{p}$ and $\boldsymbol{q}$.

- Examples:

$$\boldsymbol{p} = (1, 0, 0, 0) \quad \boldsymbol{q} = (0.25, 0.25, 0.25, 0.25) \quad \longrightarrow \quad H(\boldsymbol{p}, \boldsymbol{q}) = 2$$

$$\boldsymbol{p} = (1, 0, 0, 0) \quad \boldsymbol{q} = (0.91, 0.03, 0.03, 0.03) \quad \longrightarrow \quad H(\boldsymbol{p}, \boldsymbol{q}) = 0.136$$

Loss function

- Given a data pair $(\boldsymbol{x}, y)$ where $\boldsymbol{x} \in \mathbb{R}^d$ is the feature vector and $y \in \{0, 1\}$ is the label. Using the sigmoid function, we can predict the probability:

$$\begin{bmatrix} \frac{1}{1+\exp(-\boldsymbol{x}^\top \boldsymbol{w})} \\ \frac{\exp(-\boldsymbol{x}^\top \boldsymbol{w})}{1+\exp(-\boldsymbol{x}^\top \boldsymbol{w})} \end{bmatrix} \in \mathbb{R}^2$$

- The difference between the predicted and real probability is given by

$$\ell(\boldsymbol{x}, y; \boldsymbol{w}) = -y \log\left(\frac{1}{1 + \exp(-\boldsymbol{x}^\top \boldsymbol{w})}\right) - (1-y) \log\left(\frac{\exp(-\boldsymbol{x}^\top \boldsymbol{w})}{1 + \exp(-\boldsymbol{x}^\top \boldsymbol{w})}\right) \quad (4)$$

- Given the dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, the loss function is to measure the averaged difference

$$L(\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N; \boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w})$$

where $\ell(\boldsymbol{x}_i, y_i; \boldsymbol{w})$ is in (4).

Logisic regression

- By solving the following optimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}) \tag{5}$$

  where $\ell(\boldsymbol{x}_i, y_i; \boldsymbol{w})$ is defined as

$$\ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}) = -y_i \log\left(\frac{1}{1+\exp(-\boldsymbol{x}_i^\top \boldsymbol{w})}\right) - (1-y_i)\log\left(\frac{\exp(-\boldsymbol{x}_i^\top \boldsymbol{w})}{1+\exp(-\boldsymbol{x}_i^\top \boldsymbol{w})}\right),$$

  we can achieve the model parameters $\boldsymbol{w}^\star$.

- Given $\boldsymbol{w}^\star$ and a new feature vector $\boldsymbol{x}$, we can decide its lable by

$$y = \begin{cases} 1 & \text{if } p \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad \text{where} \quad p = \frac{1}{1+\exp(-\boldsymbol{x}^\top \boldsymbol{w})}$$

Logisic regression: simplified loss

- The loss in (4) can be written as

$$\ell(\boldsymbol{x}, y; \boldsymbol{w}) = \begin{cases} \log(1 + \exp(-\boldsymbol{x}^\top \boldsymbol{w})) & \text{if } y = 1 \\ \log(1 + \exp(\boldsymbol{x}^\top \boldsymbol{w})) & \text{if } y = 0 \end{cases} \tag{6}$$

- If we modify the label as follows:

$$y \leftarrow \begin{cases} 1 & \text{if } y = 1 \\ -1 & \text{if } y = 0 \end{cases}$$

the loss in (6) becomes

$$\ell(\boldsymbol{x}, y; \boldsymbol{w}) = \log\left(1 + \exp(-y\boldsymbol{x}^\top \boldsymbol{w})\right) \tag{7}$$

Logisic regression: simplified loss

- Substituting (7) to (5), logistic regression becomes

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \quad \frac{1}{N} \sum_{i=1}^{N} \ln(1 + \exp(-y_i \boldsymbol{x}_i^\top \boldsymbol{w}))$$

where $y \in \{+1, -1\}$ is the modified label

- Exercise: the gradient descent recursion to solve the above problem

A code example

Multi-class classification

To be added

Summary

- Linear regression

$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{2N} \sum_{i=1}^{N} (\boldsymbol{x}_i^\top \boldsymbol{w} - y_i)^2$$

- Logistic regression

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \quad \frac{1}{N} \sum_{i=1}^{N} \ln(1 + \exp(-y_i \boldsymbol{x}_i^\top \boldsymbol{w}))$$