SMAI-S25-07: Linear Regression

C. V. Jawahar

IIIT Hyderabad

January 28, 2025

Recap:

- Two problems of interest:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
  - Learn Feature Transformation as a step to find useful representations.
- Three Classification Schemes:
  - Nearest Neighbour Algorithm
  - Linear Classification
  - Decide as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2.
- Performance Metrics:
  - Classification: Accuracy, TP/FP etc., Confusion Matrix; Ranking: Precision, Recall, F-Score, AP
- Supervised Learning:
  - Notion of Training, Validation and Testing
  - Linear Regression
  - Notion of Loss Function, MSE (today)
  - Role of Optimization, Gradient Descent(today)

Regression

Regression: Regression is a statistical technique that relates a dependent variable to one or more independent variables.

In the context of machine learning, a regression problem starts with many $(\mathbf{x}_i, y_i)$ with $y_i \in R$ examples, we would like to learn a function

$$y_i = f(\mathbf{w}, \mathbf{x}_i)$$

such that it works well on unseen data.

Note: y_i could also be a real vector. Also note the difference (and similarity) with the classification problem.

Linear Regression: Linear regression algorithm provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events

Examples

- Given height, weight of 1000 students of a school (Age 4-17), develop an algorithm to predict weight of a student, given the height.
- Given the processor speed, memory, hard disk space, predict the price of a laptop.

Linear Regression

Given N examples $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$. Our goal is to find a solution of the form: $y_i = mx_i + c$; or $y_i = w_1 x_i + w_2$ or

$$\hat{y}_i = \mathbf{w}^T \mathbf{x}_i$$

If $\hat{y}_i$ is the predicted value and y_i is the actual value, our goal is to minimize a Loss function (which capture the discrepancy between the them):

$$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

or

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

We need to solve this optimization problem to obtain $\mathbf{w}$

Solution - I: Closed Form Solution

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{w}$$

$$Min_{\mathbf{w}} \frac{1}{N}[\mathbf{Y} - \hat{\mathbf{Y}}]^T[\mathbf{Y} - \hat{\mathbf{Y}}]$$

$$Min_{\mathbf{w}} \frac{1}{N}[\mathbf{Y} - \mathbf{X}\mathbf{w}]^T[\mathbf{Y} - \mathbf{X}\mathbf{w}]$$

$$Min_{\mathbf{w}} \frac{1}{N}(\mathbf{Y}^T\mathbf{Y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2(\mathbf{X}\mathbf{w})^T\mathbf{Y})$$

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

1

[1] Tom Minka, Old and New Matrix Algebra Useful for Statistics
https://tminka.github.io/papers/matrix/minka-matrix.pdf

Solution-II: Gradient Descent Solution

Gradient Descent update for Minimizing J

- Start with $\mathbf{w}^0$
- Iterate

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k - \nabla J$$

$$\min_{\mathbf{w}} J = \min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{w^T x_i})^2$$

$$\nabla J = \frac{2}{N} \sum_{i=1}^{N} (y_i - \mathbf{w^T x_i})(-\mathbf{x_i})$$

Q: Assuming $y_i = mx_i + c$ form, derive update rule for m and c

Discussion Points

- What are the relative advantages and disadvantageous of these two solutions. When to use which?
- Does the optimization problem have multiple solutions or only one solution? (global vs local optima; convex vs non-convex problems)
- Why do you think, this is a good loss function? Any alternate loss functions possible?
- What happens when the data is "noisy"? or "erroneous"? (say 10% samples are corrupted)

Discussions Point

We are interested in fitting a solution in presence of outliers. A student came up with an original, yet simple idea. (i) Detect outliers from the data set $\mathcal{D}$ and remove them (ii) Build models on what is left out $\mathcal{D}'$. Here are the formal steps:

1. Find $\mathbf{w}$, the solution for linear regression problem with data set as $\mathcal{D}$.
2. Remove all samples where $|y_i - \mathbf{w}^T \mathbf{x}_i| > \theta$ from $\mathcal{D}$ to create an inlier set $\mathcal{D}'$.
3. Now find a new $\mathbf{w}$ with $\mathcal{D}'$.

- Student tested it with θ as very small (say zero) and very large (say infinity). In both cases, the algorithm came out to be ineffective. What could be the reason?
- Another student helped to solve the problem by coming with a nice intuitive way of finding θ. It did reasonably well. Suggest a simple and effective way to address outliers (or finding θ).

Problem I

Consider the problem of finding the minima of $f(x) = x^2 - 2x + 4$

$$\min_x x^2 - 2x + 4$$

The optimum value of the function f^* is:

- (a) 1
- (b) 2
- (c) 3
- (d) 4
- (e) -2
- (f) none of the above

Problem - 2

Consider a simple regression problem with $x_i \in R$ and $y_i \in R$. We would like to learn a model of $y_i = w_1 x_i + w_2$

Given the data $\mathcal{D}$ of $N = 4$ samples, as $\{(1,2)(2,3),(3,4),(4,5)\}$

Find $\mathbf{w} = [w_1, w_2]^T$ using a closed form expression. Write the associated matrices and then used the closed form solution

Problem - 3

Consider a simple regression problem with $x_i \in R$ and $y_i \in R$. We would like to learn a model of $y_i = w_1 x_i + w_2$

Given the data $\mathcal{D}$ of $N = 4$ samples, as $\{(1,2)(2,3),(3,4),(4,5)\}$

Find $\mathbf{w} = [w_1, w_2]^T$ using gradient descent, starting from (a) $\mathbf{w} = [1,0]^T$ and (b) $[1,1]^T$. Show only one iteration for both.

Problem - 4

There are problems where we also have certain confidence in each of the data samples (eg. how reliable they are). Say there are α_i and they are in $[0, 1]$ with 0 being non-confident and 1 being confident.
Now Consider the problem of linear regression where we minimize the loss

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

Derive the corresponding closed form solution

Problem - 5

There are problems where we also have certain confidence in each of the data samples (eg. how reliable they are). Say there are α_i and they are in $[0, 1]$ with 0 being non-confident and 1 being confident.
Now Consider the problem of linear regression where we minimize the loss

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \alpha_i (y_i - (mx_i + c))^2$$

Derive the corresponding gradient descent solution