LINEAR REGRESSION

Linear regression is a common statistical tool for modeling the relationship between some "explanatory" variables and some real-valued outcome.

Definition (Linear Regression)

Given domain set $X \in \mathbb{R}^d$ and label set $Y \in \mathbb{R}$. We would like to learn a linear function $h: \mathbb{R}^d \to \mathbb{R}$ that best approximates the relationship between our variables.

The hypothesis class of linear regression predictors is the set of linear functions.

$$\mathcal{H} = L_d = \{ x \mapsto \langle w, x \rangle + b : w \in \mathbb{R}^d, b \in \mathbb{R} \}$$

We need to define a loss function for regression. A common choice is squared-loss function

Squared-loss function

$$\ell(h, (x,y)) = \underbrace{h(x)}_{\text{prediction}} - \underbrace{y}_{\text{label}})^2$$

Why use squared-loss function?

Unlike classification problem, where the loss function is $\ell(h, (x,y)) = 1_{h(x) \neq y}$

Regression problem won't always give the "perfect" number, i.e $1, -1$

Empirical Risk Function (Mean Squared Error)

For that loss function, the impirical risk is defined as:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} (h(x_i) - y_i)^2$$

This is called Mean Squared Error

Least Squares Algorithm

Least Squares is the algorithm that solves ERM problem for the hypothesis class of linear regression predictors with respect to squared loss. Given training set S:

$$\underset{w}{\arg\min}\, L_S(h_w) = \underset{w}{\arg\min} \frac{1}{m} \sum_{i=1}^{m} (\langle w, x_i \rangle - y_i)^2$$

Let $\left\{ \begin{array}{l} X \text{ be the matrix where columns are made of examples from } S. \\ y \text{ be the vector of labels.} \\ w \text{ be the vector of coefficients.} \end{array} \right.$

Such that $X.w = y$

Visually: $X = (x_1 | \dots | x_m)$ $\quad w = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$ $\quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$

Recall from ALAFF, to solve Least Squares problem, we try to find $\hat{w}$ that is:

$$\| y - X.\hat{w} \| = \min \| y - X.w \|$$

Also mentioned in LAFF are some ways to solve this, some are:

- If X is invertable: $\hat{w} = X^{-1} y$
- Normal equations: $X^H X . \hat{w} = X^H y$
- SVD: Decompose $X = U \Sigma V^H$
  Solve $\hat{w} = V \Sigma^{-1} U^H . y$
- Eigenvalue decomposition (since X is symmetric):
  - Decompose $C = Q D Q^T$ where $\left\{ \begin{array}{l} Q \text{ is orthonormal matrix} \\ D \text{ is diagonal matrix} \end{array} \right.$
  - Solve $C \hat{w} = Q Q^T b$ $\quad C = X^T X \;;\; b = X^T y$

Linear Regression for Polynomial Regression Tasks

Take for instance, a one dimensional polynomial function of degree n:

$$p(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$$

$$= \begin{pmatrix} 1 \\ x \\ \vdots \\ x^n \end{pmatrix}^T \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}}_{\text{vector of coefficients}}$$

- For simplicity, we focus on one dimensional, n-degree, polynomial regression predictors, namely: $\mathcal{H}^n_{poly} = \{ x \mapsto p(x) \}$

Note that: domain set $\mathcal{X} \in \mathbb{R}$ and label set $\mathcal{Y} \in \mathbb{R}$

- One way to learn this class is to reduce it to linear regression in the form of vectors (as shown above). Given mapping $\psi(x) = (1 | x | \dots | x^n)$, we can rewrite $p(x)$ as:

$$p(x) = \langle \psi(x), a \rangle$$

We solve this by finding vector a using Least Squares algorithm.

Maximum Likelihood Function

Given a dataset S, if we assume the dataset follows some distribution, for example, normal distribution $S \sim N(\mu, \sigma^2)$, how can we determine the values of μ and σ^2?

We can answer this question using Principle of Maximum Likelihood, which states that the best estimate to these parameters is the one that maximize likelihood function.

Likelihood function:

Probability of observing data x given the parameters

$$L(\mu, \sigma^2 \mid S = \{x_1, \dots x_m\}) = \prod_{i=1}^{m} f(x_i \mid \mu, \sigma^2)$$

where $f(x_i \mid \mu, \sigma^2)$ is the PDF of the i^{th} data point given μ, σ^2;
here we assume all $x_i \in S$ are i.i.d.

In practice, if is more convinient to maximise the <u>log-likelihood</u>, which is

Log-likelihood function

$$\log L(\mu, \sigma^2 \mid S = \{x_1, \dots, x_m\}) = \sum_{i=1}^{m} \log f(x_i \mid \mu, \sigma^2)$$

To find the parameters, set the derivative of log-likelihood w.r.t. to the parameter to zero.

For example, finding parameter μ:

- Likelihood Function:

$$L(\mu, \sigma^2 \mid S) = \prod_{i=1}^{m} \underbrace{\frac{1}{\sqrt{2\pi}\,\sigma} e^{\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)}}_{\text{PDF of Normal Distribution}}$$

- Convert to Log-Likelihood Function:

$$\log L(\mu, \sigma^2 \mid S) = \sum_{i=1}^{m} \log\left[\frac{1}{\sqrt{2\pi}\,\sigma} e^{\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)}\right]$$

$$= \sum_{i=1}^{m}\left[\log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) - \left(\frac{x_i-\mu}{2\sigma^2}\right)^2\right]$$

- Derivative w.r.t μ:

$$\nabla_\mu \log L(\mu, \sigma^2 \mid S) = \nabla_\mu \sum_{i=1}^{m}\left[-(x_i-\mu)^2\right]$$

$$= 2\left[\sum_{i=1}^{m}(x_i-\mu)\right]$$ derivative w.r.t. μ

- Find best estimate $\hat{\mu}_{ML}$ by setting derivative equals 0:

$$\hat{\mu}_{ML} = \arg\max \log L(\mu, \sigma^2 \mid S)$$

Solve by: $\nabla_\mu \log L(\mu, \sigma^2 \mid S) = 0$

$$\Rightarrow 2\left[\sum_{i=1}^{n} (x_i - \hat{\mu}_{ML})\right] = 0$$

$$\Rightarrow \hat{\mu}_{ML} = \underbrace{\frac{1}{m} \sum_{i=1}^{m} x_i}_{\text{mean value of } S}$$

Notes:

This heavily depends on what distribution you assume the data is distributed on.

If you assume $S \sim N(\mu, \sigma^2)$, then you estimate μ and σ^2

If you assume $S \sim \text{Expo}(\lambda)$, then you estimate λ

Linear Regression - Maximum Liklihood

Consider a simple linear regression:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$= \underbrace{\begin{pmatrix} 1 \\ x \end{pmatrix}^T}_{\text{vector instance } x} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\text{weight vector}} + \underbrace{\varepsilon}_{\text{error term } \varepsilon \sim N(0, \sigma^2)}$$

What is the probability observing dataset $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ given β_0, β_1?

We can answer this question by finding the log-likelihood function.

- First, find the likelihood function:

$$L(\beta_0, \beta_1 \mid S = \{(x_1, y_1), \dots, (x_m, y_m)\})$$

Since $y = \beta_0 + \beta_1 x + \varepsilon$, we can rewrite this as:

$$L(\beta_0, \beta_1, x_i \mid y_i) = \prod_{i=1}^{m} f(y_i \mid \beta_0, \beta_1, x_i)$$

Since $\varepsilon \sim N(0, \sigma^2) \Rightarrow (y - \beta_0 - \beta_1 x) \sim N(0, \sigma^2)$

$$\Rightarrow y \sim N(\beta_0 + \beta_1 x, \sigma^2):$$

$$L(\beta_0, \beta_1, x_i \mid y_i) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2}}$$

- Second, convert to log-likelihood

$$\log L(\beta_0, \beta_1, x_i \mid y_i) = -\frac{m}{2} \log 2\pi - m \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Lastly, we maximize the likelihood / find best estimate for β_0 and β_1 by setting derivative w.r.t. β_0, β_1 to 0:

$$\underset{\beta_0, \beta_1}{\arg\max} \underbrace{\sum_{i=1}^{m} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2}_{\text{Linear Least Squares}}$$

Coefficients / weights in Linear regression:

- Geometric: coefficients of the line that minimizes squared distances from line to labels

$$\arg\min \| y - X^T b \|$$

- Statistic: coefficients give the maximum likelihood estimator for a training set generated by $y \sim N(\beta_0 + \beta_1 x, \varepsilon)$

$$\arg\max \sum_{i=1}^{m} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2$$