Note Title 4/26/2021

Linear Regression

Given data $\{(x_1, y_1), \ldots (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, $i = 1 \ldots n$

$$L(\theta) = \frac{1}{2} \sum_{i=1}^{n} (x_i^T \theta - y_i)^2$$

Minimize L: Define $A = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$ ← Feature matrix, $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

$$A\theta = \begin{bmatrix} x_1^T \theta \\ \vdots \\ x_n^T \theta \end{bmatrix} \qquad A\theta - Y = \begin{bmatrix} x_1^T \theta - y_1 \\ \vdots \\ x_n^T \theta - y_n \end{bmatrix} \qquad (A\theta - Y)^T (A\theta - Y) = \sum_{i=1}^{n} (x_i^T \theta - y_i)^2$$

So, $$L(\theta) = \frac{1}{2} (A\theta - Y)^T (A\theta - Y)$$

Minimizing L: $\nabla_\theta L(\theta) = 0$

$$\nabla_\theta \left((A\theta - Y)^T (A\theta - Y) \right) = 0$$

$$\Leftrightarrow \quad A^T (A\theta - Y) = 0$$

$$\Leftrightarrow \quad \boxed{(A^T A)\theta = A^T Y} \longrightarrow$$ Least square solution

Note! Can we write $\theta = (A^T A)^{-1} A^T Y$?

Yes, if A is full rank [why? full rank $A \Rightarrow A^T A$ is invertible.

How to argue this? Show that $\text{rank}(A) = \text{rank}(A^T A)$

by showing $\underbrace{N(A) = N(A^T A)}_{\text{H.W.}}$.

[Take an $x \in N(A)$ & show $x \in N(A^T A)$ and the other way]

Maximum likelihood and least squares:

Suppose the data is generated according to the following model:

$$x \longrightarrow \boxed{\begin{array}{c}\text{Linear}\\ \text{Model}\end{array}} \longrightarrow y = \theta^T x + \epsilon$$

ϵ → zero-mean Noise, e.g. $\epsilon \sim$ Gaussian with mean 0 and variance $\left(\frac{1}{\beta}\right)$

Dataset $\mathcal{D} = \{ (x_i, y_i),\ i = 1 \ldots n \}$ generated in an i.i.d fashion

i.i.d → independent & identically distributed

Maximum likelihood (ML) approach:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(-\frac{\beta}{2} \left(y_i - \theta^T x_i \right)^2 \right)$$

→ Why this function? (Module 6 on probability)

ML approach: Find a θ that maximizes $\mathcal{L}(\theta)$

Instead of $\max_\theta \mathcal{L}(\theta)$, we could consider $\max_\theta \log \mathcal{L}(\theta)$

Since log is increasing, the optima of these two problems are the same.

$$\log \mathcal{L}(\theta) = \frac{n}{2}\log\beta - \frac{n}{2}\log 2\pi - \beta\left[\frac{1}{2}\sum(y_i - \theta^T x_i)^2\right]$$

is the least squares objective that we minimized wrt θ

Message: "Maximizing $\log \mathcal{L}(\theta)$" is the same as "$\min_\theta \frac{1}{2}\sum(y_i - \theta^T x_i)^2$"

(or) Least squares regression solves a maximum likelihood estimation problem under a linear model.

Polynomial regression:

Consider one-dimensional data

$$\mathcal{D} = \{ (x_i, y_i), \ i = 1 \ldots n \} \qquad x_i, y_i \in \mathbb{R}, \ \forall i$$

Previously, we tried fitting a line through this data.

Here, we generalize to any polynomial, say of degree m.

Transformed features:

$$\hat{y}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_m x^m$$

$$= \sum_{j=0}^{m} \theta_j \, \phi_j(x), \qquad \text{where } \phi_j(x) = x^j$$

For a given x, the transformed feature vector $\phi(x) = (1, x, x^2, \ldots, x^m)$

$$\hat{y}(x) = \theta^T \phi(x), \qquad \theta = (\theta_0, \ldots, \theta_m)$$

Using these transformed features, perform linear regression.

$$A = \begin{bmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix}, \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$(A^T A)\theta = A^T Y$$

Different wrt regular linear regression is that we are using transformed features & then performing regression.

Regularized version of linear regression (a.k.a.) Ridge regression

Instead of solving $\min_\theta \frac{1}{2} \sum_{i=1}^{n} (x_i^T \theta - y_i)^2$, we solve the following regularized version:

$$\min_{\theta} \left\{ J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (x_i^T\theta - y_i)^2 + \underbrace{\lambda \|\theta\|^2}_{\text{regularization term}} \right\}$$

Repeat the calculation leading to least squares solution, we obtain

$$(A^T A + \lambda I)\, \theta_{reg} = A^T Y$$

$$\theta_{reg} = (A^T A + \lambda I)^{-1} A^T Y$$

H.W.: Show $(A^T A + \lambda I)$ is invertible even if A is not full rank.

Note: regularization Parameter λ controls overfitting.

Too small a $\lambda \Rightarrow$ overfitting

Too large a $\lambda \Rightarrow$ underfitting