Understanding ML with perspective of analy

Regularization

Problem

When N <= d, X^TX is not existing!

Solution: regulariza

L2 norm, ridge regression → $\frac{\lambda}{2}||\theta||^2$

adding

Loss function → $J(\theta) = \frac{1}{2}(y - X\theta)^T(y - X\theta) + \frac{\lambda}{2}||\theta||^2$

Gradient → $\nabla J(\theta) = (X^TX + \lambda I)\theta - X^Ty$

$= 0$

Analytical Solution → $\theta = (X^TX + \lambda I)^{-1}X^Ty$

With L2 norm, the traverse of matrix $X^TX + \lambda I$ are **always be able to solve**

Similarity

Question 1

Why and where should we get similarity?

Why

Similarity means correlation between samples or datasets

So ML is actually predicting similarity between new data and trained data

Where

$$f_\theta(X) = X\theta = X(X^TX + \lambda I)^{-1}X^Ty$$

Push-through identity: $(\lambda I + PQ)^{-1}P = P(\lambda I + PQ)^{-1}$

$$f_\theta(X) = X(X^TX + \lambda I)^{-1}X^Ty = XX^T(XX^T + \lambda I)^{-1}y$$

$$f_\theta(X) = K(K + \lambda I)^{-1}y, \quad K = XX^T$$

$K_{ij} = x_i^Tx_j$

so we get dot production of vectors!

$$f_\theta(x) = (x^Tx_1, ..., x^Tx_n)(XX^T + \lambda I)^{-1}y$$

Similarity | weights | Labels

Concepts 1: ML actually compute samples' similarity, then using statistical idea to predict on new samples.

ytical solution

Analytical solution

$\theta = (X^T X)^{-1} X^T y$

tion

L1 norm, LASSO → $\lambda||\theta||_1$

L2 norm from perspective of Analytical Solution:
Limiting whole dataset scale

adding

Loss function → $J(\theta) = \frac{1}{2}(y - X\theta)^T(y - X\theta) + \lambda||\theta||_1$

Gradient → $\nabla J(\theta) = -X^T y + X^T X\theta + \lambda sgn(\theta)$ ---------- $sgn(x) = \begin{cases} 1, x > 0 \\ 0, x = 0 \\ -1, x < 0 \end{cases}$

L2 norm from perspective of Analytical Solution:
Limiting whole dataset scale

Question 2

How dot production could always reflecting similarities?

Kernel Matrix

Kernel function transform data into **linear**

Kernel function speeds up **computation**

$\phi : \mathbb{R}^d \rightarrow \mathbb{R}^h$ → After transformation, data are able to compute similarity

$\cos\alpha = \frac{a \cdot b}{||a||||b||} = \frac{a}{||a||} \cdot \frac{b}{||b||}$

Here **dot production is reflecting similarity between vectors!** (when a and b in unit length)

$\theta = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T y$

$f_\theta(X) = K(K + \lambda I)^{-1} y$, $K = \phi(x_i)^T \phi(x_j)$ → **Kernel Matrix**

We introduce Kernel Matrix in order to indicate that we **can simplify computation by considerably select or design kernels**

Concepts 2: To better compute similarity, it is useful to mapping data into different form or dimension and so reducing the difficulty in computation as well.