Statistical Machine Learning

Linus Falk

November 4, 2022

Contents

Lecture 2: Linear regreesion

wednesday 02 nov 10:15

1 Linear regression discussion points

Different reasons we look at the square error, will be discussed.

Cost landscape in parameter space.

minimizes can be hard to find

- In what type of problems is the squared prediction error an unsuitable performance measure:

It symmetric, we maybe don't want to under overestimate the outcome/prediction. Think of medication or battery. Another way yo do it pinball loss. two alternatives

$$L(x, y, \theta) = \left[Pinballloss\right](y - f(x, \theta))\alpha, y \geq f(x, \theta) or \\ (y - f(x, \theta))(1 - \alpha), y < f(x, \theta)) \tag{1}$$

- How would you visualize a regression model with two inputs.

A plane for a linear regression model. A surface lives in a subspace of the d parameters space.

- **When is there a unique linear regression model that minimizes the average squared-error loss**

IF we get to little data we can generate a unique model for an example. Think of just having one data point and fitting a line to it. All interpolate the training data

Necessary for uniqueness that $n \geq (d + 1)$, (not sufficient)

Linear regression model

$$\overline{X} = \begin{bmatrix} -x_1^T - \\ \vdots \\ -x_n^T - \end{bmatrix} \tag{2}$$

$X^T X$ is invertible. (d+1) when they are linearly independent we can find a unique solution. (rand($\overline{X}$) = d+1)

- 4

Parameter space Numerical search e.g. gridding. Evaluate J at every gridpoint. Might not find the actual minimum. Might work when d i small ≤ 3. Another way is gradient search.

- Consider alternative ways to regularize the least-square method.

$J(\theta) = 1/nnorm(\overline{y} - \overline{X\theta})$

Positive quadratic function in θ (convex) local minimum $\rightarrow$ global minimum of J(θ)

Local min (*) $\nabla_0 J(\theta) = 0 < - > \overline{X^T}\overline{X} = X^T\overline{y}$

**Notes if small amount of data, choose fewer d, but which one to choose?

Different norms for example norm 2 for ridge regression affects the cost. Lasso is another example using norm 1.

Regularization reduces the sensitivity.

- How does one interpret the 'best' regression function?

Finding the balancing point of the distribution for each point.

- In what types of problem can it still produce poor performance

Balancing point between two "hills" give a prediction between them. Data will never come there. Multi modal distribution. The population splits into two parts. In the blood pressure it could be difference between male and female there could therefore be necessary add a new feature.

1.1 Family of Gaussian distributions

Why not model the best regression function