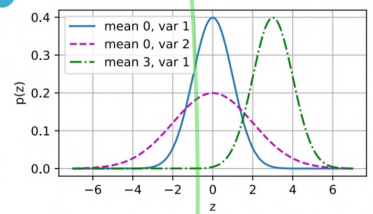
April 2 /2020

4. 概率派和ML Round 2

Normal distribution



One way to motivate linear regression with the mean squared error loss function is to formally assume that observations arise from noisy observations, where the noise is normally distributed as follows

$$y = \mathbf{w}^\top \mathbf{x} + b + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

we can now write out the likelihood of seeing a particular y for a given x via

$$p(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \mathbf{w}^\top \mathbf{x} - b)^2\right).$$

MLE 极大似然估计

Now, according to the maximum likelihood principle, the best values of b and w are those that maximize the likelihood of the entire dataset:

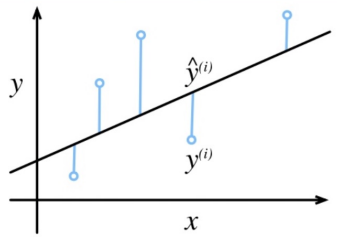
$$P(Y \mid X) = \prod_{i=1}^{n} p(y^{(i)}|\mathbf{x}^{(i)}).$$

σ is some fixed constant

$$-\log p(\mathbf{y}|\mathbf{X}) = \sum_{i=1}^{n} \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)} - b\right)^2$$

It follows that minimizing squared error is equivalent to maximum likelihood estimation of a linear model under the assumption of additive Gaussian noise.

Linear Regression



Loss Function

Loss function

$$L(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} l^{(i)}(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\left(\mathbf{w}^\top \mathbf{x}^{(i)} + b - y^{(i)}\right)^2.$$

Analytic Solution

$$\mathbf{w}^*, b^* = \underset{\mathbf{w},b}{\operatorname{argmin}}\ L(\mathbf{w}, b).$$

Not analytically solvable

Gradient descent

gradient descent

Even in cases where we cannot solve the models analytically, and even when the loss surfaces are high- dimensional and nonconvex, it turns out that we can still train models effectively in practice.

解方程

MLE

Loss function

最大似然 ⟺ 最小化 $y-\hat{y}$

Of course!

logistic regression 再转!!!

Reference : Dive into deep learning