Linear Regression

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$
$$= (y_1 - \hat{y}_1)^2 + \cdots + (y_n - \hat{y}_n)^2$$

single Linear Regression : $y = \hat{\beta}_0 + \hat{\beta}_1 x$ model에서

최적의 $\hat{\beta}_0$와 $\hat{\beta}_1$을 구하기 위해서 RSS를 minimize한다.

$$RSS = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

이제 RSS를 minimize하는 $\hat{\beta}_0$와 $\hat{\beta}_1$은 각각으로 RSS를 미분한다

$\dfrac{\partial RSS}{\partial \hat{\beta}_0} = 0$, $\dfrac{\partial RSS}{\partial \hat{\beta}_1} = 0$ 이 되도록 해보자. (이를 만족하면 max or min이다)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

* Single Linear Regression에서 최적의 $\hat{\beta}_0$와 $\hat{\beta}_1$을 얻기 위해서 RSS를 각각 $\hat{\beta}_0$와 $\hat{\beta}_1$로 미분해서 구해준다.

<평가방식>

RSS ⇒ e^2의 sum # data↑ error↑

RSE, RMSE → data수로 나눈다.

$\sqrt{\dfrac{RSS}{n-2}}$ $\sqrt{\dfrac{RSS}{n}}$

R-square : $1 - \dfrac{RSS}{TSS}$ (1이면 Best)

RSS → 우리 model의 오차²의 합

TSS → 실제 분산²의 합, $\sum_{i=1}^{n} (y_i - \bar{y})^2$

→ model이 미처 설명하지 못한 변동.

Maximum liklihood

Why 굳이 $(\ )^2$을 사용하지? 그냥 abs()를 사용해도 되는거 아닌가?

IID : Independent and Identically distribution

⇒ 모든 사건이 독립이고, 동일한 확률분포를 따른다.

주로 data들이 어느 분포에서 얻어진지 모르기에 명확한 판단 하기가 힘든만
IID를 따른다고 가정한다.

Liklihood : 이미 관측된 data가 있는 경우, 그 data를 가장 잘 설명해 주는 분포 parameter θ를 찾기 위한 가운.

일반확률은 θ가 주어진 경우 data가 나올 확률 $p(\text{data} \mid \theta)$ 라면
가능도 함수는 θ를 확률 변수로 두고 $p(\theta \mid \text{data})$를 의미한다.

그래서 $L(\theta ; \underbrace{x_1, x_2 \cdots x_n}_{\text{n개의 data가 관측된 경우의}})$이 가능도이다.

θ 값

⇓

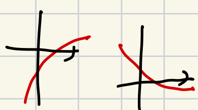
IID 가정이기에

$L(\theta ; x_1, x_2, \cdots, x_n) = p_\theta(x_1) \cdot p_\theta(x_2) \cdots \cdot p_\theta(x_n)$ 이 된다.

결국 모든 확률값의 곱을 maximize하는 θ를 구하는게 가능도 함수의 목적이다.

* Distribution의 모형은 미리 가정 (예를 들어 Gaussian, 라플라시안 ---)

그래서 Gaussian인 경우 μ와 γ를 찾게 되는것이다.

Log-likilihood : 여기의 maximize, minimize를 구할거기에  를 곱해서 해당 value는 변화 X.
monotonus increasing 하기에.

$$\log \prod_{i=1}^{n} p_\theta(x_i) = \sum_{i=1}^{n} \log p_\theta(x_i)$$

$$\underset{x}{\arg\max}\ f(x) = \underset{x}{\arg\max}\ \log f(x)$$

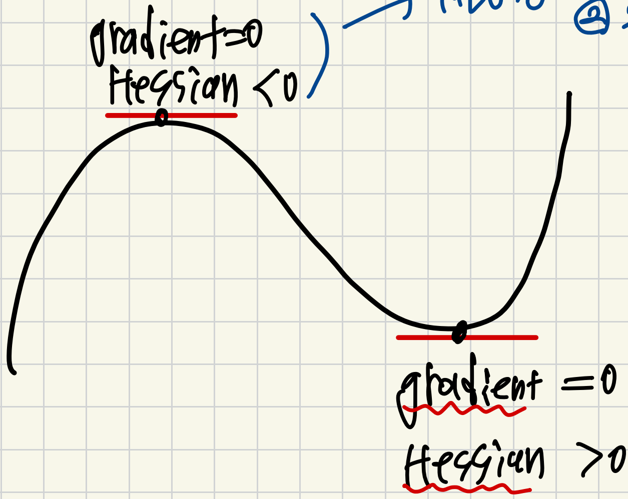
작을시 gradient 구하기에 용이

MLE : Maximum Liklihood Estimation

$$\hat{\theta}(x_1, x_2, \cdots, x_n) = \underset{\theta}{\operatorname{argmax}}\, L(\theta; x_1, x_2, \cdots, x_n)$$

$$= \underset{\theta}{\operatorname{argmax}}\, l(\theta; x_1, x_2, \cdots, x_n)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P_\theta(x_i)$$



→ MLE는 ① θ에 대해 1번 미분해서 0
② 2번 미분한 matrix에 어떠한 v를 곱해도 <0 이면 된다.
(Negative definite)

〈MLE 예시 : Bernoulli distribution〉

$$\underset{\theta}{\operatorname{arg\,max}} \sum_{i=1}^{n} \log P_\theta(x_i) = \underset{\theta}{\operatorname{arg\,max}} \sum_{i=1}^{n} \log\left(\theta^{x_i}(1-\theta)^{(1-x_i)}\right)$$

$$= \underset{\theta}{\operatorname{argmax}}\ \log\theta \sum_{i=1}^{n} x_i + \log(1-\theta)\sum_{i=1}^{n}(1-x_i)$$

$$\frac{\partial\left(\log\theta \sum_{i=1}^{n} x_i + \log(1-\theta)\sum_{i=1}^{n}(1-x_i)\right)}{\partial\theta} = 0$$ 을 풀면

$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i$ 가 나온다. 자연스럽게 평균값이 나오게 된다.

2번 미분해서 Hessian을 구하면

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} = -\frac{1}{\theta^2}\sum_{i=1}^{n} x_i - \frac{1}{(1-\theta)^2}\sum_{i=1}^{n}(1-x_i) \quad \therefore \text{0보다 작다.}$$

$+ \quad + \quad + \quad +$

〈 MLE 예시 : Gaussian distribution 〉

$\theta = \{\mu, \sigma\}$ 이다. // μ와 σ에 대해 모두 미분한다.

$$\hat{\theta}(x_1, x_2, \cdots x_n) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} + \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

$\therefore$ μ에 대해 1차 미분시 $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ 가 된다.

σ에 " $\sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$ 가 된다.

* 정규분포는 가우시안 가정시 MLE를 통해서 나온 결과이다.

MLE for Linear Regression

기본적인 Linear Regression는 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ 표현가능

이를 $y = \beta^T x + \epsilon$ 로 표현이 가능하다.

↳ 에러가 $\sim N(0, \sigma^2)$을 따른다고 가정하자.

⇒ 결국 Linear Regress는 $p(y|x) \sim N(\beta^T x, \sigma^2)$

↳ 우리가 구한 선형을 따르되 특정 σ만큼 분산이 존재한다.

이는 MLE를 구하기 위해 가능도를 구하게 되면

$$\arg\max_{\beta} \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{y-(\beta^T x_i)}{\sigma}\right)^2}$$

$$\doteq \arg\max_{\beta} -\sum_{i=1}^{n} (y_i - \beta^T x)^2$$

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2$$

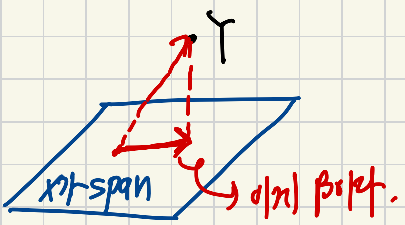
(오차제곱의 합이다)

e를 제곱하는 이유이다.

행렬에 대해서 $\frac{\partial RSS(\beta)}{\partial \beta}$를 구하게 되면 $\beta = (X^T X)^{-1} X^T Y$ 가 된다.

이게 정사영 Matrix이다.

∴ Y 가장 잘 근사하는 β



Bias, Variance and MSE

$$\mathrm{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

$$\mathrm{Var}(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

$$\mathrm{MSE} = E\left(\sum_{i=1}^{d} (\hat{\theta}_i - \theta_i)^2\right)$$

$$\mathrm{MSE} = \mathrm{tr}(\mathrm{Var}) + \|\mathrm{Bias}\|^2$$

$$\mathbb{E}\left[\sum_{j=1}^{d}(\hat{\theta}_j - \theta_j)^2\right] \stackrel{?}{=} \sum_{j=1}^{d}\mathbb{E}\left[(\hat{\theta}_j - \mathbb{E}(\hat{\theta}_j))^2\right] + \sum_{j=1}^{d}(\mathbb{E}(\hat{\theta}_j) - \theta_j)^2$$

For simplicity, let's suppose $d = 1$ without loss of generality.

$$\mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \stackrel{?}{=} \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\right] + (\mathbb{E}(\hat{\theta}) - \theta)^2$$

- Using the Bias-Variance decomposition, we can compute the MSE of linear regression:

$$\mathrm{MSE}(\hat{\theta}|\mathbf{X}) = \underbrace{\boxed{\mathrm{Var}(\hat{\theta}|\mathbf{X})}}_{\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}} + \underbrace{\boxed{\mathrm{Bias}(\hat{\theta}|\mathbf{X})^2}}_{0} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

  - Linear regression is unbiased (bias = 0).
  - With infinite data, variance also converges to 0.

- It can be also proved that the MLE is the best unbiased estimator.
  - That is, no other θ has lower variance than the one found by MLE. (Gauss-Markov Theorem)