Complexity / Classification / Regression / Supervised learning

در مسئله‌ی یادگیری با سرپرست با یک جفت Feature/label
X Y

X و Y متغیرهای تصادفی هستند.

به دنبال پیدا کردن یک mapping از X به Y هستیم.

$$T = \{(x_i, y_i)\}_{i=1}^{n} \rightarrow \text{training set.}$$

برای X و Y دو حالت داریم

Y ← { یا به صورت quantitative خواهد بود $Y = [0,1]$
یا به صورت qualitative خواهد بود. مریض/سالم

X ← مجموعه متغیرهایی که Y از آن گرفته

Y ← مجموعه متغیرهایی که Y از آن گرفته

گاهی Y یک چیز بینابینی است ← نزدیک ۱، متوسط، کوچک

سعی می‌کنیم qualitative را به یک سری عدد (وزن‌ها) روکدکنیم.

مریض ← ۱

سالم ← ۰

اگر y به صورت کیفی باشه اما بیش از دو حالت داشته باشه چیه؟

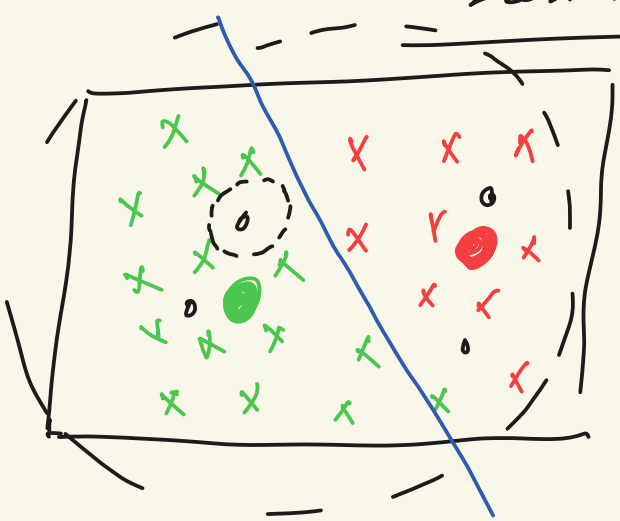
y مثلاً 10 حالت دارد

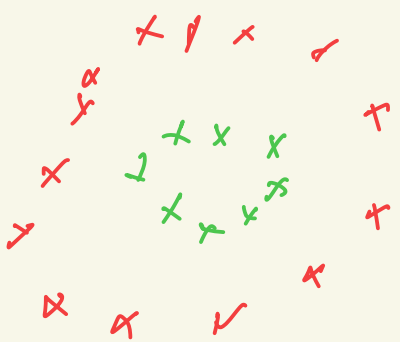
* اعداد صحیح از 1 تا 10
* بردار one hot ← 10 بعدی
* کد کردن با 4 بیت

اگر y به صورت کمی (quantitative) باشه ← Regression

اگر y به صورت کیفی باشه ← classification

Least Squares. — K-NN





به مدل تصمیم را چه یاد داریم؟

دو تا ایده ← ۱) کلی کلیک مدل مثل خط

۲) نزدیک ترین را پیدا کنیم و بدانیم اون علامت میاد یعنی جدید علامت بدیم.

← K Nearest Neighbor ← KNN

$x \rightarrow$ میخواهیم $\hat{y}$ پیش‌بینی کنیم

$$\hat{y} = f(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

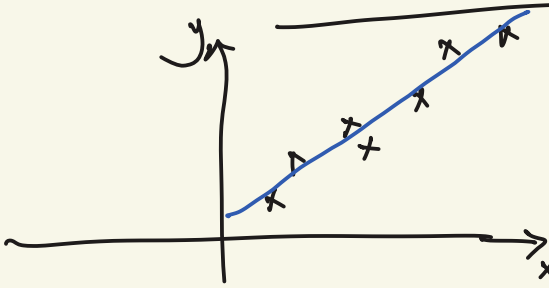
$$\hat{y} = \begin{cases} 1 & f(x) \geq 0.5 \\ 0 & o.w \end{cases}$$

برای classification

وقتی می‌گیم یک خط جدا کننده داریم یعنی

$$\hat{y} = x^T \beta$$

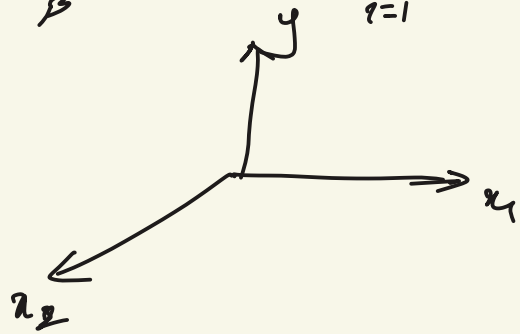
یک سری نمونه داریم و می‌خوایم β را تخمین بزنیم:



$(\hat{y} - y)^2$ می‌گیریم

$$\hat{y} = x^T \beta = \beta^T x$$

$$L_\beta(T) = \sum_{i=1}^{N} (y_i - \beta^T x_i)$$



$$y - \beta^T x = 0$$

$$\begin{bmatrix} x_1 \\ \vdots \\ x_p \\ y \end{bmatrix}^T \beta = 0$$

$$\underset{\beta}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 = \mathcal{L}_\beta(T)$$

$$\frac{d}{d\beta}\mathcal{L}_\beta(T) = 0$$

فرم ماتریسی:

$$\mathcal{L}_\beta(T) = \|Y - X\beta\|_2^2$$

$$\mathcal{L}_\beta(T) = (Y - X\beta)^T (Y - X\beta)$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_1^T \\ \hline 1 \\ \hline \vdots \\ \hline x_n^T \end{bmatrix}_{n \times p}, \quad \beta \in \mathbb{R}^p$$

$$\nabla_\beta \mathcal{L} = 2X^T(Y - X\beta) = 0$$

$$\Rightarrow X^T Y = (X^T X)\beta \Rightarrow \beta = (X^T X)^{-1} X^T Y$$

اگر $(X^T X)$ وارون‌پذیر باشد مثلاً $p \leq n$ باشد وارون‌پذیر است

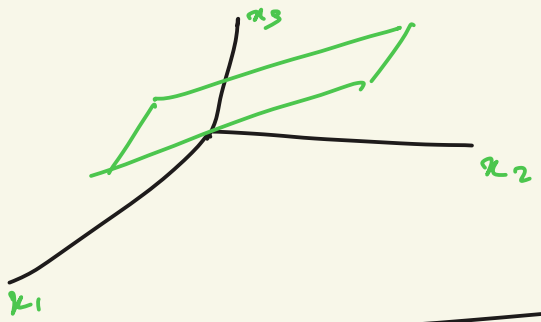
اگر $p \leq n$

$$X^T X = U \Lambda U^T$$

Λ ← ماتریس قطری

$$(X^T X)^+ = U \Lambda^+ U^T = \Lambda^+ = \begin{bmatrix} 1/\lambda_1 & & & \\ & \ddots & & \\ & & 1/\lambda_n & \\ & & & 0 \\ & & & & \ddots \end{bmatrix}_{p \times p}$$

$$\beta = (X^T X)^+ X^T Y$$

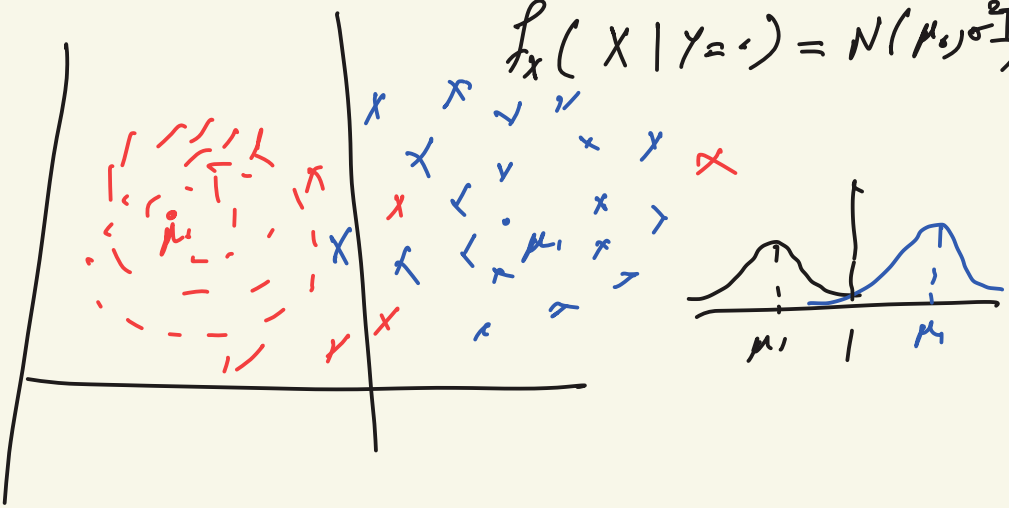


Least Squares

توضیح داده شد.

$P(Y=1) = 1/2 \rightarrow f_X(X \mid Y=1) = N(\mu_1, \sigma^2 I)$

$f_X(X \mid Y=0) = N(\mu_0, \sigma^2 I)$



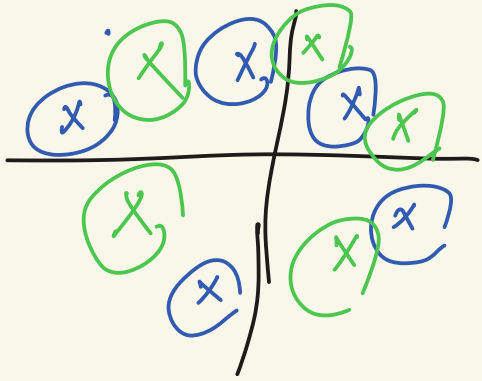
$P(y=1) = 1/2$

$\mu_1, \mu_2, \cdots, \mu_5 \overset{i.i.d.}{\sim} N(0, \sigma^2 I)$

$\mu_6, \cdots, \mu_{10} \sim N(0, \sigma^2 I)$

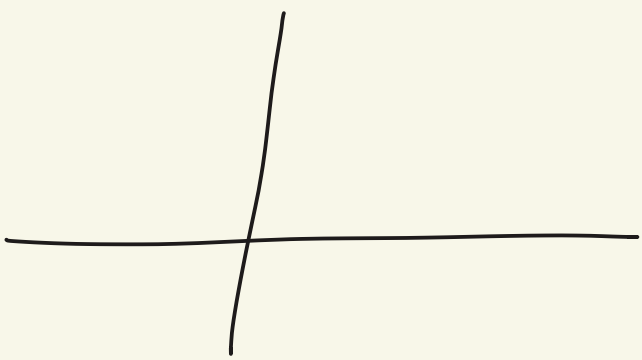
$f_X(x \mid y=1) = \frac{1}{5} \sum_{i=1}^{5} N(\mu_i, \sigma^2 I)$

$f_X(x \mid y=0) = \frac{1}{5} \sum_{i=6}^{10} N(\mu_i, \sigma^2 I)$



برای این توزیع ← KNN جواب بهتریست.

سؤال: خطای training کدام یک از KNN با $K=1$ (استفاده از کل) کمتر است؟ صفر



تابع Loss را تعریف کنیم.

$$\min_f E[L(y, f(x))]$$

این در Statistical learning

مفروضات تحلیلی:

یک مجموعه $T = \{(x_i, y_i)\}_{i=1}^n$ داریم که به صورت مستقل از توزیع $P(X,Y)$ نمونه برداری شده اند

$$\{(x_i, y_i)\}_{i=1}^n \sim D^n(X,Y)$$

هدف: شما مدلی پیدا کنید به صورت $\hat{y} = f(x)$ که این مدل روی نمونه‌های که از توزیع $D(x,y)$ میاید در آینده خوب عمل کند.

سوال: خوب یعنی چی؟

جواب: خوب چیزیه که تابع هزینه‌ای $\rightarrow \ell : Y \times Y \rightarrow R_+$ کمتر باشه.

در صورتی که این تابع و توزیع را داشته باشیم مسئله‌ی زیر را حل خواهیم کرد.

$$\underset{f}{\text{minimize}} \quad \underset{x,y \sim D}{E}\left[\ell(f(x), y)\right] \leftarrow$$

مثلا:

$$\ell(\hat{y}, y) = (\hat{y} - y)^2$$

$$\underset{f}{\text{minimize}} \quad \underset{x,y \sim D}{E}\left[(Y - f(x))^2\right]$$

$$\int (y - f(x))^2 p(x,y)\, dx\, dy \qquad \underset{x}{E}\, \underset{y|x}{E}\left[(y - f(x))^2\right] = 0$$

$$\int (y - f(x))^2 p(x) p(y|x)\, dx\, dy$$

$$\text{minimization} \int p(x) \left[\int (y - f(x))^2 p(y|x)\, dy\right] dx$$

$$\text{minimization} \int (y - c)^2 p(y|x = x)\, dy$$

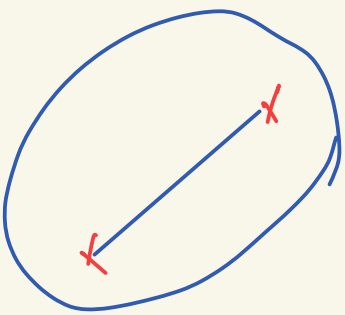
$$\frac{d}{dc} : -\int (y - c)\, p(y|x = x)\, dy = 0$$

$$\int y\, p(y|x=x)\, dy = c = f(x)$$

Regression Function ← $f(x) = E_{y|x=x}[Y|X=x]$ ←

اگر نمونه داشته باشیم به جای توزیع

$$\hat{f}(x) = \frac{1}{k}\sum_{x_i=x} y_i = \frac{1}{k}\sum_{x_i \in N(x)} y_i$$

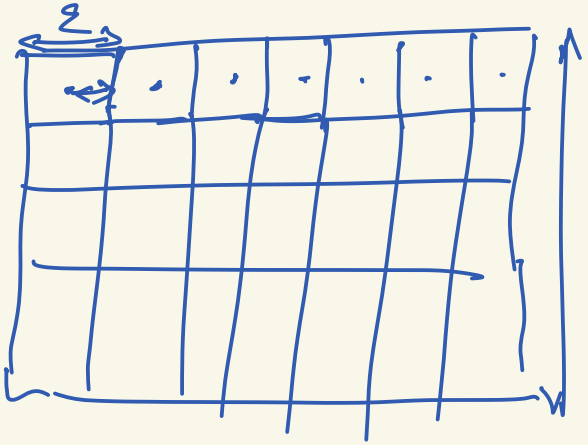


1- به جای اعمال همسایگی روی تعریف ها

2- با جای خود نقطه‌ی x نقاط همسایگی x

} با دوتا Relaxation

$$\left.\begin{array}{l} n \to \infty \\ K \to \infty \\ \frac{K}{n} \to 0 \end{array}\right\} \to KNN \to \text{Regression Function}$$



تعداد مکعبی که می‌شه $\left(\frac{1}{\varepsilon}\right)^d$

2^d

فرض کنید نمونه به صورت یکنواخت از این مکعب برداشته شود.

چند نمونه نیاز داریم که تقریبا مکعب کوچک مدل به نمونه نیفتد؟

به عنوان تمرین این مسئله رو حل کنید ← In Expected coupon collector

Curse of dimensionality

یک راه دیگر برای تابع f را به صورت یک تابع چند جمله‌ای بنویسیم

$$f(x) = \beta^T x$$

$$\min_{\beta} E_{x,y}\left[(x^T\beta - y)^2\right] = R$$

$$\nabla_\beta R = 2E_{x,y}\left[x(x^T\beta - y)\right] = 0$$

$$E[xx^T]\beta = E[xy]$$

$$\Sigma$$

$$\Sigma \text{ وارون پذیر باشد} \Rightarrow \beta = \Sigma^{-1}E[xy]$$

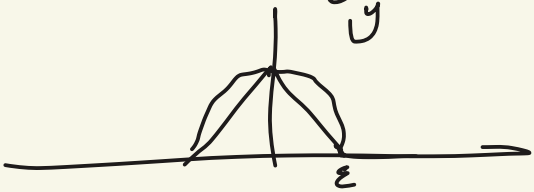
تخمین با نمونه‌ی محدود برای این محاسبه می‌شود جواب Least Square

---

مسئله classification با تابع هزینه‌ی 0-1

$$\min_f E_{x,y}\left[1(y \neq f(x))\right] \Rightarrow$$

$$f(x) = \underset{y}{\operatorname{argmax}}\ P(y \mid X = x)$$



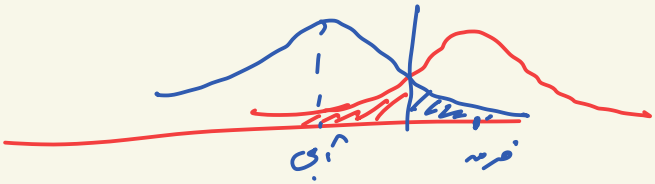
$$E_X\, E_{Y|X}\left[\mathbb{1}(y \neq f(x))\right] =$$

$$= E_X\left[\sum_{y_i} \mathbb{1}(f(x) \neq y_i)\, P(y_i \mid x)\right]$$

Bayes classifier ←

$$f(x) = \underset{y_i}{\operatorname{argmax}}\ P(y_i \mid X = x)$$

$$E_X\left[1 - \max_{y_i} P(y_i \mid X = x)\right]$$



به عنوان تمرین

$$\log \frac{P(Y=1)}{1 - P(Y=1)} = \text{odds} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

($\log 2$ ↑)

$$\frac{P(Y=1)}{P(Y=0)} = \underbrace{e^{\beta_1 X_1}}_{2}\left(e^{\beta_0} e^{\beta_2 X_2}\right)$$

(↑ 1)

Regression تابع هدف ← Loss تابع

$$\min_f E[l(y, f(x))] \xrightarrow{\text{argmin}} f^*$$

$$l(y,\hat{y}) = (y-\hat{y})^2 \Rightarrow f^* = E[Y|X=x]$$

$$l(y,\hat{y}) = |y-\hat{y}| \Rightarrow f^* = ?$$

$$E[|y - f(x)|] =$$

$x, y \sim D \rightarrow$ توزیع مشترک x, y

$$E_x \, E_{y|x} [|y - \underbrace{f(x)}_{\theta}|]$$

$$\min_f \int_x \left(\int_y |y - f(x)| \, p(y|x) \, dy \right) p(x) \, dx$$

$$\geq \int_x \left(\min_f \underbrace{\int |y - f(x)| \, p(y|x) \, dy}_{f^*} \right) p(x) \, dx$$

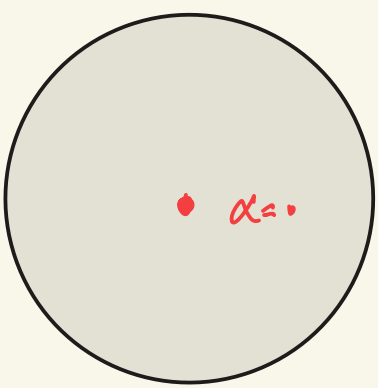
$$\frac{\partial}{\partial f} \int |y - f(x)| \, p(y|x)\, dy$$

$$\int \left(2\,\mathbb{1}(y - f(x) > 0) - 1\right) p(y|x)\, dy$$

$$= P_{Y|x}[y > f(x)] - P_{Y|x}[y < f(x)] = 0$$

$$\Rightarrow f(x) = \text{median}(Y \mid X = x)$$

اگر برای y داشته باشیم $F_{Y|X}(y \mid X = x) = 1/2$



$$y = e^{\alpha \|x\|_2}$$

n نمونه داریم.

فرض کنید n نمونه داریم $T = \{(x_1, y_1), \dots, (x_n, y_n)\}$

می‌خواهیم با استفاده KNN مقدار تابع در $x = 0$ را تخمین بزنیم

نمونه‌ها به صورت تکراری از داخل کره به دست آمده‌اند.

$$\min_i \|x_i\|_2 = z_T$$

$$\text{median } z_T = ?$$

مدین می‌شه این

$$P[z_T \geq t] = 1/2 = 1 - \delta$$

$$P[\min_i \|x_i\|_2 \geq t] = 1/2 = 1 - \delta$$

$$= P[\forall i \ \|x_i\|_2 \geq t] = \prod P[\|x_i\|_2 \geq t]$$

$$\prod_i P[\|x_i\|_2 \le t] = \prod_{i=1}^{n} 1 - \left(\frac{t}{R}\right)^d$$

$$= \gamma_2 = \left(1 - (t/R)^d\right)^n \quad \gamma = 1-\delta$$

$$\to 1 - (t/R)^d = (1/2)^{1/n} = (1-\delta)^{1/n}$$

$$\to \frac{t}{R} = \left(1 - (1/2)^{1/n}\right)^{1/d} = \left(1 - (1-\delta)^{1/n}\right)^{1/d}$$

$$R=1 \to t = \left(1 - (1/2)^{1/n}\right)^{1/d}$$

$E[z] = ?$ تمرین

EPE : Expected Prediction Error

$X \sim \text{Uniform}(S(1,d))$ $\hat{y} =$ تخمین

توزیع یکنواخت کروی بعد d

$y = e^{-8\|x\|_2^2} = f(x)$

$$E_T\left[(f(x)-\hat{y})^2\right] = ?$$

نوی نقطی صفر.

$$\hat{y} = e^{-8z^2}$$

$$E_T\left[(f(x) \pm E_T[\hat{y}] - \hat{y})^2\right] =$$

$$E_T\left[(f(x)-E_T[\hat{y}])^2\right] \Rightarrow \left(f(x)-E_T(\hat{y})\right)^2$$

$$+ E_T\left[(\hat{y}-E[\hat{y}])^2\right] \Rightarrow Var_T(\hat{y})$$

$$+2E_T\left[(f(x)-E_T[\hat{y}])(\hat{y}-E_T[\hat{y}])\right] \to 0$$

$$= \underbrace{\left(f(x)-E_T(\hat{y})\right)^2}_{Bias^2} + Var_T(\hat{y})$$

$e^{-\gamma \|x\|_2^2} \rightarrow (x_1^2 + 1)$

$x \sim D_x$

$y = x^T \beta + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$

ε از x مستقل است.

فرض کنید یک تخمین $\hat{\beta}$ از روی نمونه‌های Brain بزنیم. نمونه‌های آموزشی در T هستند.

$$E_{y|x} E_T [(y - \hat{y})^2] = EPE[x]$$

$(x, y) \sim P_{x,y}$

$x \sim P_x$ ، $y \sim P_{y|x}$

$$X \sim P_X, \quad y = f(X) + \varepsilon$$

$y \in \mathbb{R}$

$x \in \mathbb{R}^d$

$$y = \beta^T x + \varepsilon$$

$$\boxed{\varepsilon \sim N(0, \sigma^2)}$$

یک مجموعه از نمونه‌های آموزشی.

$$\min_{\hat{\beta}} \; E_x \, E_{y|x} \left[(\hat{\beta}^T x - y)^2 \right]$$

$$\min_{\beta} \min \frac{1}{n} \sum_{i \in T} (\beta^T x_i - y_i)^2$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \leftarrow$$

$$X = \begin{bmatrix} \text{---} \; x_1^T \; \text{---} \\ \vdots \\ \text{---} \; x_n^T \; \text{---} \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y_i = \beta^T x_i + \varepsilon_i \Rightarrow Y = X\beta + \underline{\varepsilon}$$

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta + \underline{\varepsilon})$$

$$\hat{\beta} = \beta + (X^T X)^{-1} X^T \underline{\varepsilon}$$

$$\hat{\beta} = \beta + \sum_{i=1}^{n} l_i(x)\, \varepsilon_i$$

$$\underset{n,\Sigma}{E}\left[\hat{\beta} - \beta\right] = 0 \;\Rightarrow\; \underset{T}{E}\left[\hat{\beta}\right] = \beta$$

$$E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T\right] = COV(\hat{\beta})$$

$$E\left[(\Sigma l_i \varepsilon_i)(\Sigma l_i \varepsilon_i)^T\right]$$

$$= E\left[\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} l_i l_j^T \varepsilon_i \varepsilon_j + \sum_{i=1}^{n} l_i l_i^T \varepsilon_i^2\right]$$

$$= \sigma^2 \sum_i l_i l_i^T$$

$$E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T\right] =$$

$$E_{X,\varepsilon}\left[(X^TX)^{-1} X^T \varepsilon \varepsilon^T X (X^TX)^{-1}\right]$$

$$E_X \, E_{\varepsilon}\left[(X^TX)^{-1} X^T \varepsilon \varepsilon^T X (X^TX)^{-1}\right]$$

$$\sigma^2 E_X\left[(X^TX)^{-1} X^T X (X^TX)^{-1}\right]$$

$$\sigma^2 E_X\left[(X^TX)^{-1}\right] \sim \sigma^2 Cov(X)^{-1}$$

$$= Cov(\hat{\beta})$$

برای یک نقطه x_0 و y_0 نسبت به $\hat{y}_0$ چه فرضی داریم.

$$y_0 = \beta^T x_0 + \varepsilon_0 \rightarrow E[y_0] = \beta^T x_0$$

$$Var(y_0) = Var(\varepsilon_0) = \sigma^2$$

$$\hat{y}_0 = \hat{\beta}^T x_0$$

$$E_{\varepsilon, T, x_0}\left[(y_0 - \hat{y}_0)^2\right]$$

$$E_{x_0}\, E_\varepsilon\, E_T\left[(y_0 - \hat{y}_0)^2\right]$$

$$E_\varepsilon E_T\left[\left(y_0 \pm E_T[\hat{y}_0] - \hat{y}_0\right)^2\right]$$

(I)
$$E_\varepsilon E_T\left[(y_0 - E_T[\hat{y}_0])^2\right] \rightarrow ?$$

(II)
$$+ E_\varepsilon E_T\left[(\hat{y}_0 - E_T[\hat{y}_0])^2\right] \rightarrow \mathrm{Var}(\hat{y}_0)$$

$$+ 2 E_\varepsilon E_T\left[(y_0 - E_T[\hat{y}_0])(\hat{y}_0 - E_T[\hat{y}_0])\right] = 0$$

(I)
$$E_\varepsilon E_T\left[(y_0 \pm E_\varepsilon[y_0] - E_T[\hat{y}_0])^2\right]$$

$$E_\Sigma E_T\left[(y_0 - E_\Sigma[y_0])^2\right] \to \sigma^2$$

$$+ E_\Sigma E_T\left[(E_\Sigma[y_0] - E_T[\hat{y}_0])^2\right]$$

$$+ 2E_\Sigma E_T\left[(y_0 - E_\Sigma[y_0])(E_\Sigma[y_0] - E_T[\hat{y}_0])\right]$$

$$= \sigma^2 + \text{bias}^2$$

$$\text{bias} = (E_\Sigma[y_0] - E_T[\hat{y}_0])$$

$$E_{x_0} E_\Sigma E_T\left[(y_0 - \hat{y}_0)^2\right] =$$

$$E_{x_0}\left[\sigma^2 + \text{bias}_T^2(\hat{y}_0|x_0) + \text{var}_T(\hat{y}_0|x_0)\right]$$

$$\underline{\sigma^2} + \underline{\text{bias}_1^2} + \underline{\text{var}}$$

σ^2 — $\sigma^2 d/N$ ...

$$bias = E_{\varepsilon}[y_0] - E_T[\hat{y}_0]$$

$$y_0 = \beta^T x_0 + \varepsilon \rightarrow E_{\varepsilon}[y_0] = \beta^T x_0$$

$$\hat{y}_0 = \hat{\beta}^T x_0 \rightarrow E_T[\hat{y}_0] = E[\hat{\beta}]^T x_0$$

$$= \beta^T x_0$$

$$bias = \beta^T x_0 - \beta^T x_0 = 0$$

$$Var = E_{x_1}\left[E\left[\left(\hat{y}_0 - E[\hat{y}_0]\right)^2\right]\right]$$

$$E_{x_0} E\left[\left(\hat{\beta}^T x_0 - \beta^T x_0\right)^2\right]$$

$$E_{x_0} E\left[\left[\left(\hat{\beta} - \beta\right)^T x_0\right]^2\right]$$

$$E_{x_0} E_T\left[x_0^T (\hat{\beta} - \beta)(\hat{\beta} - \beta)^T x_0\right]$$

$$E_{x_0}\left[x_0^T \underbrace{E\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T\right]}_{\sigma^2 cov(X)^{-1}} x_0\right]$$

$$\sigma^2 E_{x_0}\left[x_0^T cov(X)^{-1} x_0\right]$$

$$\sigma^2 E_{x_0}\left[Tr(x_0^T cov(X)^{-1} x_0)\right]$$

$$= \sigma^2 E_{x_0}\left[Tr(x_0 x_0^T cov(X)^{-1})\right]$$

$$= \sigma^2 Tr(cov(x_0)\, cov(X)^{-1})$$

$$cov(x_0) \to \frac{cov(x_0)}{n}, \quad \frac{1}{n}$$

$$= \sigma^2 d/n$$

$$A = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

$$E_T\left[(A^T A)^{-1}\right] \sim \frac{1}{E_T[A^T A]}$$

$$A^T = \begin{bmatrix} x_1 & | & \cdots & | & x_n \end{bmatrix}_{d \times n}$$

$$A = \begin{bmatrix} \xrightarrow{\quad x_1 \quad} \\ \vdots \\ \xrightarrow{\quad x_n \quad} \end{bmatrix}_{n \times d}$$

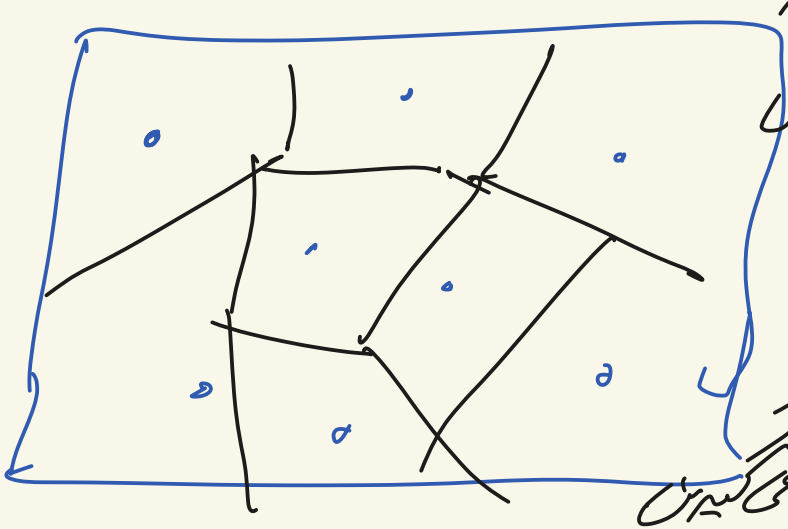
$B = A^T A \rightarrow$ [illegible] B_{ij} [illegible]

$\downarrow$

$(i,k)(k,j)$

$E[B] = n\,cov(x)$

$E[(\hat{y}_i - y)^2] = \sigma^2 + \sigma^2 d/n$



به طریقه تقسیم می‌کنیم و برای هر کدوم پیاده‌ایم در نقاط مختلف

$$f(x) = \sum_{i=1}^{n} f_i(x)\,\beta_i$$

$$\hat{\beta}_2 \underset{\beta}{\text{minimize}} \sum_{i=1}^{n} (\beta^T x_i - y_i)^2 + \lambda \|\beta\|_2^2 \quad (A)$$

تا اینجا بیشتر گذاشتی 23 صفحه‌ای از مورد

regression نیاز مدل‌های degrees of freedom (DF)

و DF را برای kNN حل نهایی loss
برای Squared Error و loss رگرسیون رسم با
د (برای AG) نتیجه.