subject :

date :

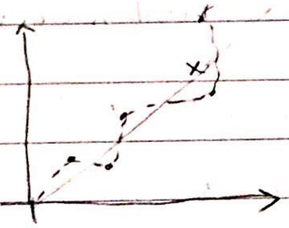
سوال اول :

الف) طبق فرمول برای به دست آوردن هر y باید w ها و x ها با هم ضرب شوند و در نهایت جمع شوند در نتیجه هر چه یک w بزرگتر باشد تاثیر آن فیچر را نشان می دهد اما در صورتی که فیچر گفته شده با یکی دیگر از فیچرها وابستگی بالا داشته باشد و ضریب آن مخالف فیچر قبلی باشد این دو یکدیگر را خنثی می کنند در نتیجه برای دانستن میزان اهمیت این ضریب نیاز به اطلاعات بیشتر داریم ← گزینه ی سوم صحیح است.

ب)

(a) طبق فرمول می دانیم بایاس $[h(x) - E_D(y(x;D))]$ در نتیجه هر چه تعداد دیتاست ها را بیشتر کنیم مقدار آن به $h(x)$ نزدیک تر می شود در نتیجه بایاس ما کمتر خواهد شد ← جمله درست است

(b) با کم کردن خطای training در واقع داریم مدل را به داده های ورودی فیت می کنیم در واقع در trade off بین بایاس و واریانس داریم بایاس را تا حد ممکن کم کرده و واریانس را زیاد می کنیم یعنی با اضافه شدن یک داده ی جدید مدل به آن داده فیت شده و شکل خود را تغییر می دهد تا ارور را کمینه کند و این باعث خراب شدن مدل و over fit شدن آن می شود ← گزاره غلط است



(c) در این مثال اگر بخواهیم خطای مدل را کمینه کنیم باید به گونه ای عمل کنیم که از نقطه هایی بگذرد در این صورت خطای x که داده ی ترین هاست بیشتر می شود ولی اگر سعی کنیم بایاس را کمی زیاد کنیم مدل خطی شده و خطای x کمتر خواهد شد ← گزاره غلط است

(d) همیشه خطای test زیاد به علت پیچیدگی بیش از اندازه نیست. گاهی ممکن است مدل بیش از اندازه ساده باشد و نتوانسته باشد به خوبی یاد بگیرد و نیاز است که مدل را پیچیده تر کنیم ← گزاره غلط است

سوال ۲:

(a) برای اینکه w_{opt} را به دست آوریم باید خطا را کمینه کنیم ← مشتق می گیریم تا کمینه را به دست آوریم

$$\min E = \sum_{n=1}^{N} (y_n - w^T X_n)^2$$

$$\rightarrow \nabla E = 2\sum_{n=1}^{N} (y_n - X_n^T w) X_n = 0 \rightarrow \sum_{n=1}^{N} y_n X_n = w^T \sum_{n=1}^{N} X_n X_n^T$$

X_n برداری است که مقادیر تابع linear Reg به ازای هر x در آن قرار دارد در نتیجه X یک ماتریس است که در هر یک از N سطر آن یکی از X_n ها قرار دارد در نتیجه Σ ها را با ضرب ماتریسی جایگزین می کنیم:

$$X^T = \begin{bmatrix} X_1 & X_2 & \cdots & X_N \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \xrightarrow{\text{ضرب این دو}} \begin{bmatrix} \sum_{i=1}^{N} t_i x_i^1 \\ \vdots \\ \sum_{i=1}^{N} t_i x_i^M \end{bmatrix}$$

(M طول بردار X است)

می‌توان نمونه کرد که ماتریس ضرب سطر های $x_n y_n$ $\sum_{n=1}^{N}$ است در واقع می‌توان جمله‌ی سمت چپ معادله را با X^TY جایگزین کرد

حال سمت راست را معادلات ماتریسی جایگزین می‌کنیم و با توجه به سمت قبل می‌توان دید که آن معادله به صورت $(X^TX)w$ بدست می‌آید:

$$X^Ty = (X^TX)w \rightarrow w = (X^TX)^{-1}X^TY$$

در نتیجه داریم:

(b) اگر ما را اضافه کنیم جمله‌ی ارور را اضافه می‌شود در نتیجه دوباره مشتق را بر حسب w با احتساب جمله‌ی اضافه حساب می‌کنیم:

$$\min E = \sum_{n=1}^{N}(y_n - w^Tx_n)^2 + \sum_{n=1}^{N}\lambda w_i^2$$

$$\nabla E \rightarrow \cancel{2}\sum_{n=1}^{N}(y_n - x_n^Tw)x_n + \cancel{2}\lambda\sum_{n=1}^{N} w_i = 0 \rightarrow \sum_{N=1} y_n x_n = -\sum \lambda w_i + w\sum x_n x_n^T$$

دوباره معادله‌ها را به ماتریس تبدیل می‌کنیم

$$X^TY = (X^TX)w + \lambda I w \rightarrow w = (X^TX+\lambda I)^{-1}X^TY$$

(c) در این مسئله باید عبارت را رد و یا ثابت کنیم

$$(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y = (X^TX)X^TY \Longleftrightarrow \Sigma X = XF$$

می‌دانیم که طرف اول این عبارت غلط است زیرا اگر Y ها صفر باشند بدون توجه به مقدار F می‌توان w را نیز صفر در نظر گرفت.

حال طرف دوم را بررسی می‌کنیم:

$$(X^TX)X^TY = \underline{(X^T\Sigma^{-1}X)}^{-1}X^T\underline{\Sigma^{-1}}Y$$

ماتریس F nonsingular است در نتیجه عملیات های زیر قابل انجام است

$$\Sigma X = XF \rightarrow \Sigma^{-1}X = XF^{-1} \rightarrow X^T\Sigma^{-1} = (F^{-1})^TX^T$$

$$\rightarrow ((F^{-1})^TX^TX)^{-1}(F^{-1})^TX^TY = (X^TX)^{-1}\underbrace{F^T(F^{-1})^T}_{I}X^TY$$

$$= (X^TX)^{-1}X^TY$$

طرف دوم شرط برقرار است

(d) L_1 را بررسی می‌کنیم:

$$L_1: \sum_{n=1}^{N}(y_n - w^Tx_n)^2 + \sum\lambda w$$

برای اینکه معادلش کنه شکل صورت سوال به شکل بالا اسکالر باید جمله‌ی $\lambda\|w\|^2$ حذف بشود. برای اینکار این جمله را حذف کنیم و با data point اضافه کنیم. جمله‌ی اول نگاه می‌کنیم:

$$\|y - Xw\|^2$$

در این جمله باید λ ها را اضافه کنیم. چون در حال حاضر این مقدار به سمت صفر میل می‌کند می‌توانیم به قطر اصلی X، λ^2 را اضافه کنیم یعنی $(X' = \sqrt{\lambda} I + X)$ در این صورت جمله‌ی بالا به شکل Reg ما را در فرا کند:

$$\|y - (\sqrt{\lambda} I + X) w\|^2 + \lambda_2 \|w\|_1$$

سوال ۳:

$$q(x) = N(x|\mu, \Sigma) \rightarrow Q = (2\pi)^{-k/2} \det(\Sigma)^{-1/2} e^{-1/2 (x-\mu)^T \Sigma^{-1} (x-\mu)}$$

طبق تعریف Q داریم:

$$\rightarrow KL(P\|Q) = \int P_x \log \frac{P_x}{Q_x} dx = \int P_x \log P_x dx - \int P_x \log Q_x dx$$

$$E_p[x] = \mu = \int x p(x) dx \quad , \quad \Sigma = Var_p(x)$$

می‌خواهیم تابع را تخمین بزنیم یعنی باید سعی کنیم KL این دو تابع را مینیمم کنیم در نتیجه از تابع بدست آمده برحسب μ مشتق می‌گیریم:

$$\int P_x \log P_x dx - \int P_x \log (2\pi)^{-k/2} dx - \int P_x \log \det(\Sigma^{-1/2}) -$$

$$\int P_x \log e^{-1/2 (x-\mu)^T \Sigma^{-1} (x-\mu)} dx$$

$$\rightarrow \int P_x \log P_x dx + \frac{k}{2} \int P_x (\log 2\pi) dx + \frac{1}{2} \int P_x \log (\det \Sigma) dx$$

$$+ \int P_x \frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) dx = A$$

برای بدست آوردن پارامترهای برحسب Σ دیفرانسیل برحسب μ مشتق می‌گیریم:

$$\rightarrow \frac{dA}{d\Sigma} = \frac{1}{2} \left(\int P_x \left[\frac{d}{d\Sigma} \log \det(\Sigma) + ((x-\mu)^T \Sigma^{-1} (x-\mu)) + 0 \right] dx \right.$$

$$= \frac{1}{2} \int P_x \left[\frac{1}{\Sigma} - \frac{1}{\Sigma^2} (x-\mu)^T (x-\mu) \right] dx = 0 \rightarrow \frac{1}{2\Sigma} = \frac{1}{2\Sigma^2} \int P_x$$

$$(x-\mu)^T (x-\mu)] \rightarrow \Sigma = \int P_x (x-\mu)^T (x-\mu) dx \rightarrow \Sigma = Var_p[x]$$

با توجه به تعریف واریانس

حال از عبارت بالا نسبت به μ مشتق می‌گیریم:

$$\frac{dA}{d\mu} = \frac{1}{2}\int p_x \frac{d}{d\mu}\left[(x-\mu)^T\Sigma(x-\mu)\right]dx = \frac{1}{2}\int p_x\left[-\Sigma^{-1}(x-\mu) - \Sigma^{-1}(x-\mu)^T\right]dx = 0 \longrightarrow \frac{1}{2}\int p_x\left[-x+\mu-x^T+\mu\right]dx$$

$$\frac{2\mu}{2} = \frac{1}{2}\int p_x(-x-x^T)\,dx \longrightarrow \mu = \int p_x\, x\, dx$$

که طبق تعریف برابر با $E(x)$ است.

سوال ۴:

a) اگر بخواهیم فقط فیچر j را نگه داریم باید ضرایب w را برای تمام فیچرهای دیگر صفر کنیم. حال طبق فرمول می‌دانیم:

$$\min E = \sum_{n=1}^{N}(y_n - w_n^T x_n)^2 \Rightarrow w$$ مشتق می‌گیریم تا w را optimal محاسبه کنیم

$$2\sum_{n=1}^{N}(y_n - w_n^T x_n)\,x_n \longrightarrow \sum_{n=1}^{N} y_n x_n = w^T\sum_{n=1}^{N} x_n x_n^T$$

از آنجایی که ضرایب برای بقیه فیچرها صفر است فقط برای فیچر j ام w داریم. همانطور که در سوال ۲ توضیح داده شد محاسبات را به فرم ماتریسی می‌نویسیم.

$$X^T y = (X^T X)\,w$$ درایه‌های ماتریس به جز j ام همه صفر هستند ← بردار j ام ماتریس میانه

$$\longrightarrow x_j^T y = x_j^T x_j\, w \longrightarrow w = \frac{x_j^T y}{x_j^T x_j}$$

b) ستون‌های X متعامد هستند $x_i^T x_j = 0$

$$X^T y = (X^T X)\,w$$

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

$$\begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^n \\ x_2^1 & x_2^2 & \cdots & x_2^n \\ \vdots & & & \vdots \\ x_n^1 & \cdots & & x_n^n \end{bmatrix}\begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_n^1 \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & & \vdots \\ x_1^n & x_2^n & \cdots & x_n^n \end{bmatrix} = \begin{bmatrix} x_1^T x_1 + 0 \\ x_2^T x_2 + 0 \\ \vdots \\ x_n^T x_n + 0 \end{bmatrix} = X^T X$$

معکوس آن ← ۱ تقسیم بر آنها می‌شود

$$\longrightarrow w = (X^T X)^{-1} X^T y = \frac{x_1 y}{x_1^T x_1} + \frac{x_2 y}{x_2^T x_2} + \cdots + \frac{x_n y}{x_n^T x_n}$$

که همان حاصل جمع فیچرهای تک‌تک است.

Scanned with CamScanner

(c) دو تا متغیر در نظر می‌گیریم. اگر مقادیر متغیرها به صفر است چنین یک مدلی با این هم داریم

$$E = \sum_{n=0}^{N} (y - w^T x_n)^2$$

می‌خواهیم w ها را حساب کنیم که دیگر رسید
۲ تا w مشتق می‌گیریم

$$\frac{dE}{dw_0} \rightarrow 2 \sum (y - w^T x_n)\, x_0 = 0$$

$$\rightarrow \sum y x_0 - \underbrace{w^T}_{\text{ثابت ها یک مقدار دارند}} \sum x_n x_0^T = 0 \longrightarrow x_0^T y = w_0 n + w_j x_0^T x_j$$

$$\rightarrow w_0 = \frac{x_0^T y - w_j x_0^T x_j}{n} \xrightarrow{\text{طبق تعریف}} E(y) - w_j E(x_j)$$

$$\frac{dE}{dw_j} = 0 \Rightarrow \sum y x_j - w^T \sum x_n^T x_j \longrightarrow x_j^T y = w_0 x_0^T x_j + w_j x_j^T x_j$$

جای‌گذاری معادله اول در دومی

$$\rightarrow n x^T y = [x_0^T y - w_j x_0^T x_j]\, x_0^T x_j + n w_j x_j^T x_j$$

$$= x_0^T y\, x_0^T x_j - w_j x_0^T x_j\, x_0^T x_j + n w_j x_j^T x_j \rightarrow$$

$$n x^T y = (x_0^T y)^2 - w_j [(x_0^T x_j)^2 - n x_j^T x_j] \rightarrow$$

$$w_j = \frac{(x_0^T y)^2 - n(x^T y)}{(x_0^T x_j)^2 - n x_j^T x_j} \longrightarrow w_j = \frac{cov(x_j, y)}{var(x_j)}$$