به نام خدا

علی قاسم‌زاده ۴۰۱۱۰۶۳۳۹

۱- داریم که ،

$$f(y \mid x, \theta) = \mathcal{N}(2x, \theta x)$$

$$\Rightarrow \quad y_i \mid x_i, \theta \sim \mathcal{N}(2x_i, \theta x_i)$$

$$\Rightarrow \quad L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta x_i}} \exp\left(-\frac{(y_i - 2x_i)^2}{2\theta x_i}\right) \xrightarrow{\ Ln\ }$$

$$Ln\, L(\theta) = -\frac{1}{2}\left(\sum_{i=1}^{n} Ln(2\pi\theta x_i) + \frac{(y_i - 2x_i)^2}{\theta x_i}\right) \xrightarrow{\ \partial/\partial\theta\ }$$

$$-\frac{1}{2}\left(\sum_{i=1}^{n} \frac{1}{\theta} - \frac{(y_i - 2x_i)^2}{\theta^2 x_i}\right) = 0 \longrightarrow$$

$$\frac{n}{\theta} - \frac{1}{\theta^2}\sum_{i=1}^{n} \frac{(y_i - 2x_i)^2}{x_i} = 0 \longrightarrow n\theta = \sum_{i=1}^{n} \frac{(y_i - 2x_i)^2}{x_i}$$

$$\longrightarrow \boxed{\hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - 2x_i)^2}{x_i}}$$

۲-۱. داریم که ،

$$y = f(x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$Err(x) = \mathbb{E}_{D,\varepsilon}\{(y - \hat{f}(x))^2\} = \mathbb{E}_{D,\varepsilon}\{(f(x) + \varepsilon - \hat{f}(x))^2\} =$$

$$\underbrace{\mathbb{E}_{D,\varepsilon}\{\varepsilon^2\}}_{\sigma^2} + \underbrace{\mathbb{E}_{D,\varepsilon}\{(f(x) - \mathbb{E}\{\hat{f}(x)\})^2\}}_{Bias^2(\hat{f}(x))} + \underbrace{\mathbb{E}_{D,\varepsilon}\{(\hat{f}(x) - \mathbb{E}\{\hat{f}(x)\})^2\}}_{Var_D(\hat{f}(x))}$$

همچنین عبارت بالا یعنی ترم ها ساده تر نیز می‌شوند، برای مثال $\mathbb{E}_{D,\varepsilon}\{\varepsilon^2\}$ همان $\mathbb{E}_{\varepsilon}\{\varepsilon^2\}$ است و ــــــ .

$$\Rightarrow Err(x) = Bias(\hat{f}(x))^2 + Var(\hat{f}(x)) + \sigma^2$$

۲-۲. حالا داریم که،

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$D = \{(X_i, Y_i),\quad i=1,\dots,n\}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 0 \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = S_X^2$$

$$\hat{f}_1(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \hat{f}_2(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\mathbb{E}\{\hat{f}_1(x)\} = \mathbb{E}_{\varepsilon}\{\frac{1}{n}\sum_{i=1}^{n} Y_i\} = \beta_0 + \beta_1 \bar{X} = \beta_0$$

$$\Rightarrow Bias(\hat{f}_1(x)) = \beta_0 + \beta_1 x - \beta_0 = \beta_1 x$$

$$Var(\hat{f}_1(x)) = Var(\bar{Y}) = \frac{\sigma^2}{n} \longrightarrow Err(X) = \beta_1^2 x^2 + \frac{\sigma^2}{n} + \sigma^2$$

برای $\hat{f}_2(x)$ باز داریم که،

$$\hat{f}_2(x) = \hat{\beta}_0 + \hat{\beta}_1 x, \quad (\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\arg\min} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2 \qquad \bar{x} = 0$$

$$\rightsquigarrow \mathbb{E}\{\hat{\beta}_0\} = \beta_0, \quad \mathbb{E}\{\hat{\beta}_1\} = \beta_1, \quad Cov(\hat{\beta}_0, \hat{\beta}_1) = 0$$

$$Var(\hat{\beta}_0) = \frac{\sigma^2}{n}, \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{n S_X^2}$$

$\Rightarrow Err(x) = ?$ $\qquad \mathbb{E}\{\hat{f}_2(x)\} = \beta_0 + \beta_1 x = f(x) \longrightarrow \text{Bias}=0$

$$\overset{\hat{f}_2}{\Rightarrow} \; Err(x) = 0 + \left(\frac{\sigma^2}{n} + x^2 \frac{\sigma^2}{n s_x^2}\right) + \sigma^2$$

← $\hat{f}_1$ مدل ساده‌تری است ولی بایاس زیادی دارد ولی درعوض واریانس کمتری $\left(\frac{\sigma^2}{n}\right)$ دارد.

ولی مدل $\hat{f}_2$ پیچیده‌تر است و بایاس ندارد ولی درعوض واریانس بیشتری دارد که با $\hat{f}_1$

نسبت دارد.

۳-۲ Ridge Regression :

$$(\hat{\beta}_0^\lambda, \hat{\beta}_1^\lambda) = \underset{\beta_0, \beta_1}{\arg\min} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2 + \lambda \beta_1^2 \longrightarrow$$

با مشتق‌گیری و برابر صفر گذاشتن داریم که :

$$\hat{\beta}_0^\lambda = \bar{y} \;,\; \hat{\beta}_1^\lambda = \frac{n s_x^2}{n s_x^2 + \lambda} \hat{\beta}_1^{ols}$$

که $\hat{\beta}_1^{ols}$ همان $\hat{\beta}_1$ در سری دوم بخش دوم این سوال است.

$$\Rightarrow \mathbb{E}\{\hat{f}_\lambda(x)\} = \beta_0 + \frac{n s_x^2}{n s_x^2 + \lambda} \beta_1 x \longrightarrow \text{bias} = \beta_1^2 x^2 \frac{\lambda^2}{(n s_x^2 + \lambda)^2}$$

$$\text{var}(\hat{f}_\lambda(x)) = \frac{\sigma^2}{n} + x^2 \left(\frac{\sigma^2 n s_x^2}{(n s_x^2 + \lambda)^2}\right)$$

حالا اگه $\lambda \to 0$ ، بایاس صفر میشه ، $\text{ridge} \to \text{ols}$

اگه $\lambda \to \infty$ ، بایاس یه مدل ثابت میشه ، و واریانس میشه $\frac{\sigma^2}{n}$.

← زیاد کردن λ بایاس را بیشتر میکنه ولی واریانس را کمتر می‌کنه .

۴۲. طراحی که،

$$y_i = \sin(\pi x_i) + \varepsilon_i \quad , \quad x_i \sim unif([0,1]) \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} y_i \Rightarrow \mathbb{E}\{\hat{f}(x)\} = \int_0^1 \sin(\pi x_i)\,dx = \frac{2}{\pi}$$

$$\Rightarrow \mu = \frac{2}{\pi} \quad , \quad V_f = Var(\sin(\pi x)) = \int_0^1 \sin^2(\pi x)\,dx - \mu^2 = \frac{1}{2} - \frac{4}{\pi^2}$$

$$Var(\hat{y}) = \frac{1}{n}(V_f + \sigma^2)$$

$$Gen\,Err = \mathbb{E}_{x,D,\varepsilon}\{(y-\bar{y})^2\} = \mathbb{E}_x\{(\sin(\pi x)-\mu)^2\} + Var(\bar{y}) + \sigma^2$$

$$\Rightarrow Gen\,Err = \underbrace{\left(\frac{1}{2} - \frac{4}{\pi^2}\right)}_{\substack{\text{Misspecification} \\ \text{bias}}} + \frac{1}{n}\underbrace{\left(\frac{1}{2} - \frac{4}{\pi^2} + \sigma^2\right)}_{\text{est Var}} + \underbrace{\sigma^2}_{\text{irreducible noise}}$$

وقتی $n \to \infty$ ترم $\frac{1}{n}$دار ناپدید می‌شود و ارور ما برابر خواهد بود با $\frac{1}{2} - \frac{4}{\pi^2} + \sigma^2$

۳-

$$\frac{\partial \ell}{\partial c} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial c} = -(y-\hat{y}) \cdot 1 = v^T h + c - y$$

$$\frac{\partial \ell}{\partial v} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial v} = -(y-\hat{y})\, h \in \mathbb{R}^m$$

حالا باید ابتدا به h مشتق بگیریم، خواهیم داشت که،

$$\frac{\partial \ell}{\partial h} = \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial h} = -(y-\hat{y})\, v \in \mathbb{R}^m$$

برای ReLU داریم که،

$$\cancel{\frac{\partial \ell}{\partial h} = \frac{\partial \ell}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h} = -(y-\hat{y})\, v \in R^m}$$

$$\sigma'(z_j) = \begin{cases} 1 & z_j > 0 \\ 0 & z_j < 0 \end{cases}$$

$$\Rightarrow \frac{\partial \ell}{\partial z} = \frac{\partial \ell}{\partial h} \odot \sigma'(z) = -(y-\hat{y})\, v \odot \mathbb{1}\{Ux+b>0\}$$

$$\boxed{\frac{\partial \ell}{\partial b} = \frac{\partial \ell}{\partial z} = -(y-\hat{y})\, v \odot \mathbb{1}\{Ux+b>0\}}$$

$$\frac{\partial \ell}{\partial U} = \frac{\partial \ell}{\partial z} \cdot \frac{\partial z}{\partial U} = -(y-\hat{y})\, v \odot \mathbb{1}\{Ux+b>0\}\, X^T$$

$m \times d$ matrix

تو مقادیر پیوسته، احتمال اینکه z_j دقیقاً صفر باشه صفره.

ازطرفی می‌توانیم هر subgradient نیز استفاده کنیم.

ازطرفی ReLU جاهای negative را دیگه آپدیت نمی‌کنه و dead می‌شوند، و باعث sparsity بیشتر و generalization بهتر در یکسری از کیس ها می‌شود.

۴- منظم‌سازی پنهان در روش کاهش گرادیان،

۴.۱- جواب mle عبارت است از،

$$\nabla_\beta (y - X\beta)^T (y - X\beta) = 0 \rightarrow X^T X \beta = X^T y \rightarrow$$

$$\beta_{min} = (X^T X)^{-1} X^T y \longrightarrow$$ با فرض معکوس‌پذیر بودن $X^T X$

حالا در روش کاهش گرادیان داریم که

$$\beta^{(t+1)} = \beta^{(t)} - \eta \nabla \| y - X\beta^{(t)} \|^2 = \beta^{(t)} - 2\eta X^T (X\beta^{(t)} - y) = (I - 2\eta X^T X)\beta^{(t)} + 2\eta X^T y$$

$$\beta^{(t)} = (I-2\eta \underbrace{A}_{X^TX})^t \underbrace{\beta^{(0)}}_{0} + 2\eta \sum_{k=0}^{t-1} (I-2\eta \underbrace{A}_{X^TX})^k X^T y$$

اگر η را به اندازه کافی کوچک بگیریم که تمامی مقادیر ویژه $I-2\eta A$ بین $(-1,1)$ بیفتند.

$$\Rightarrow \underbrace{(I-2\eta A)}_{V(I-2\eta\Lambda)V^T}{}^t \to 0 \quad \text{when } t\to\infty \Rightarrow \qquad M^t, \qquad (1-2\eta\lambda_j)^t \xrightarrow[t\to\infty]{} 0$$

$$\Rightarrow M^t = (I-2\eta A)^t = V(I-2\eta\Lambda)^t V^T \to 0$$

$$\sum_{k=0}^{\infty} M^k = (I-M)^{-1} = (I-(I-2\eta A))^{-1} = (2\eta A)^{-1}$$

$$\Rightarrow \lim_{t\to\infty} \beta^{(t)} = 2\eta(2\eta A)^{-1}X^Ty = A^{-1}X^Ty = (X^TX)^{-1}X^Ty$$

← به همان جواب argmin می‌رسیم.

۴-۲. داریم که

$$\beta^{(t)} = (I-(I-2\eta A)^t)A^{-1}X^Ty$$

λها

$$\hat\beta_{ridge}(\lambda) = (A+\lambda I)^{-1}X^Ty, \quad A = V\Lambda V^T$$

$$\beta_j^{(t)} = (1-(1-2\eta\lambda_j)^t)\frac{v_j^T X^T y}{\lambda_j} \Longrightarrow$$

$$\beta_j^{ridge} = \frac{1}{\lambda+\lambda_j}(v_j^T X^T y) \Rightarrow$$

رگولاریزیشن ←

حال برای اینکه این دو معادل شوند، باید داشته باشیم که،

$$\frac{1-(1-2\eta\lambda_j)^t}{\lambda_j} = \frac{1}{\lambda+\lambda_j} \longrightarrow \boxed{\lambda = \frac{\lambda_j\,(1-2\eta\lambda_j)^t}{1-(1-2\eta\lambda_j)^t}}$$

به ازای هر λ_j ای.

⇐ متوقف شدن در گام t ام معادل است با ridge reg که λ آن مقدار بالا را داشته باشد

۳-۴. مثل مدل خطی روش‌های گرادیان بیس روی شبکه‌های عصبی علاوه بر کمینه کردن خطا با تنظیماتی مثل تعداد گام بروز رسانی، نرخ یادگیری، batch-size به صورت غیر مستقیم پارامترها را به سمتی می‌برند که توابع ساده‌تر بیابند. ~~weight~~

از طرفی اگر early stopping داشته باشیم و آموزش را قبل همگرایی متوقف کنیم، از بروز پدیده overfit جلوگیری می‌شود. کار ridge regression را برای ما می‌کند.

۵- داریم که،

$$\hat{\theta} = \left(1-\frac{c}{\|y\|^2}\right)y \longrightarrow \hat{\theta}-\theta = (y-\theta) - \frac{c}{\|y\|^2}y \Longrightarrow$$

$$\|\hat{\theta}-\theta\|^2 = \|y-\theta\|^2 + c^2\frac{\|y\|^2}{\|y\|^4} - 2c\,\frac{(y-\theta)^T y}{\|y\|^2} \xrightarrow{E}$$

$$E\{\|\hat{\theta}-\theta\|^2\} = E\{\|\varepsilon\|^2\} = d$$

stein identity ←

$$E\{(y-\theta)^T f(y)\} = E\{\operatorname{div} f(y)\}$$

$$f(y) = \frac{y}{\|y\|^2} \longrightarrow \operatorname{div}\frac{y}{\|y\|^2} = \sum_{i=1}^{d}\frac{\partial}{\partial y_i}\left(\frac{y_i}{\|y\|^2}\right) = \frac{d-2}{\|y\|^2}$$

$$\Rightarrow \mathbb{E}\left\{\frac{(y-\theta)^T y}{\|y\|^2}\right\} = \mathbb{E}\left\{\frac{d-2}{\|y\|^2}\right\}$$

$$\Rightarrow \mathbb{E}\{\|\hat{\theta}-\theta\|^2\} = d - 2c\, \mathbb{E}\left\{\frac{d-2}{\|y\|^2}\right\} + c^2\, \mathbb{E}\left\{\frac{1}{\|y\|^2}\right\} =$$

$$d + \left(c^2 - 2c(d-2)\right) \mathbb{E}\left\{\frac{1}{\|y\|^2}\right\}$$

۵-۲. حال مقدار بهینه c را بخواهیم پیدا کنیم میشه:

$$2c - 2(d-2) = 0 \longrightarrow c^* = d-2$$

$$\Rightarrow d - (d-2)^2\, \mathbb{E}\left\{\frac{1}{\|y\|^2}\right\}$$

برای mle اگه بخواهیم بنویسیم درمیاد که $\hat{\theta}_{mle} = y$ چرا که برابر یه گوسی نقطه‌ای اتفاق افتاده که مرکز گوسی باشه بیشترین احتمال را داره ✓

$$\Rightarrow \text{risk}: \mathbb{E}\{\|y-\theta\|^2\} = d$$

ما داشتیم که $c = d-2$

$$\rightsquigarrow d - (d-2)^2 \underbrace{\mathbb{E}\left\{\frac{1}{\|y\|^2}\right\}}_{+} < d$$

۵-۳. برای $d \geq 3$، مقدار James-stein risk ~~estimator~~ تخمین‌گر را کمینه می‌کند.

$$\hat{\theta}_{JS} = \left(1 - \frac{d-2}{\|y\|^2}\right) y \longrightarrow$$

risk کمتری داره.

این از mle در mean-squared error بهتره.

Scanned with CamScanner