تمرین سری ۱

اثبات : $Var(x+y) = Var(x) + Var(y) + 2cov(x,y)$

$\rightarrow Var(x) = E(x^2) - E^2(x) \quad , \quad cov(x,y) = E(xy) - E(x)E(y)$

$\Rightarrow Var(x) + Var(y) + 2cov(x,y) = E(x^2) - E^2(x) + E(y^2) - E^2(y) + 2(E(xy) - E(x)E(y)) = E(x^2) + E(y^2) + 2E(xy) - (E^2(x) + E^2(y) + 2E(x)E(y)) \rightarrow E(A) + E(B) = E(A+B) \Rightarrow$

$E(x^2 + y^2 + 2xy) - (E(x) + E(y))^2 = E((x+y)^2) - E^2(x+y)$

$\rightarrow x+y = z \rightarrow E(z^2) - E^2(z) = Var(z) = Var(x+y)$

$\rightarrow Var(x+y) = Var(x) + Var(y) + 2cov(x,y)$ ✓

تمرین تئوری ۲

$J_1(w) = \|y - Xw\|_2^2 + \lambda_2 \|w\|_2^2 + \lambda_1 \|w\|_1$

$J_2(w) = \|\tilde{y} - \tilde{X}\tilde{w}\|_2^2 + c\lambda_1 \|w\|_1$

$c = (1+\lambda_2)^{-1/2}$, $\tilde{X} = c\begin{pmatrix} X \\ \sqrt{\lambda_2}\, I_d \end{pmatrix}$, $\tilde{y} = \begin{pmatrix} y \\ 0_{d\times 1} \end{pmatrix}$

اثبات : $\arg\min J_1(w) = c\,(\arg\min J_2(w))$ or $J_1(cw) = J_2(w)$

$J_1(w) = \|y - Xw\|_2^2 + \lambda_2 \|w\|_2^2 + \lambda_1 \|w\|_1 \longrightarrow J_1(cw) \Rightarrow$

$J_1(cw) = \|y - Xcw\|_2^2 + \lambda_2 \|cw\|_2^2 + \lambda_1 \|cw\|_1 = \|y - cXw\|_2^2 + \|c\sqrt{\lambda_2}\, w\|_2^2$

$+ \|\lambda_1 c w\| \rightarrow \underline{\underline{*}} \Rightarrow J_1(cw) = \left\|\begin{pmatrix} y - cXw \\ -c\sqrt{2}\, w \end{pmatrix}\right\|_2^2 + \|\lambda_1 c w\|_1$

$\Rightarrow J_1(cw) = \left\|\begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} cXw \\ -c\sqrt{2}\, w \end{pmatrix}\right\|_2^2 + \|\lambda_1 c w\|_1 = \left\|\tilde{y} - c\begin{pmatrix} X \\ -\sqrt{2}\, I \end{pmatrix} w\right\|_2^2$

$+ \|\lambda_1 c w\| = \|\tilde{y} - \tilde{X}w\|_2^2 + c\lambda_1 \|w\| = J_2(w)$

$\Rightarrow \underline{\underline{J_1(cw) = J_2(w)}}$ ✓

$\underline{\underline{*}}$: $\left\|\begin{bmatrix} A \\ B \end{bmatrix}\right\|_2^2 = \|A\|_2^2 + \|B\|_2^2$ (A, B Block) $\rightarrow \|y - cXw\|_2^2 +$

$\|c\sqrt{\lambda_2}\, w\|_2^2 = \left\|\begin{pmatrix} y - cXw \\ +c\sqrt{\lambda_2}\, w \end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix} y - cXw \\ -c\sqrt{\lambda_2}\, w \end{pmatrix}\right\|_2^2$

تمرین تئوری ۳

در مدل برنولی ـ گوسی فرض می‌کنیم هر وزن w_i به این صورت تعریف می‌شود که احتمال π وزن غیرصفر است و مقدار آن از گوسی محاسبه می‌شود و احتمال $1-\pi$ صفر خواهد بود.

حال احتمال مشترک تمامی همه‌ی وزن‌ها ضرب احتمال آن‌هاست. چرا که مستقل هستند

$$P(w) = \prod \Big(\underbrace{(1-\pi)}_{\text{احتمال صفر}} \delta(w_i) + \underbrace{\pi}_{\text{غیرصفر}} \underbrace{N(w_i|0,\sigma^2)}_{\text{گوسی}} \Big)$$

حال برای هزینه‌سازی داریم (MAP):

$$-\log P(w)$$

$$\overset{w_i \neq 0}{\Longrightarrow} -\log P(w) = -\log(\pi N(w_i)) = -\log(\pi) + \frac{w_i^2}{2\sigma^2} + \underbrace{C}_{\text{ثابت}}$$

$$\overset{w_i = 0}{\Longrightarrow} -\log(1-\pi)$$

\# مقدار از تابع گوسی = تعداد غیرصفرهای w

$$\Rightarrow \text{هزینه کل} = -\log\left(\frac{\pi}{1-\pi}\right) \times \|w\|_0 + \underbrace{C}_{\text{ثابت}}$$

* برای همین می‌گویند از منظم‌ساز نرم صفر استفاده می‌شود، چرا که اگر وزن غیرصفر به وجود آید هزینه‌ای برای ما خواهد داشت.

تمرین تئوری ۴ :

اول از همه مدل‌های لاسو و ریج را خواهیم نوشت

Lasso : $\min L(w) + \lambda \|w\|_1$, Ridge : $\min L(w) + \lambda \|w\|_2^2$

در Lasso قیود $\|w\|_1 < t$ و در Ridge قیود $\|w\|_2 < t$ داریم.
برای همین از دید هندسی، در Lasso یک لوزی خواهد شد (الماس)، در Ridge نیز دایره (کره)
خواهد شد. احتمال برخورد سطح‌های MSE با Lasso بیشتر خواهد بود (روی محورها که
ضرایب صفر خواهند بود). حال اگر از دید آماری نگاه کنیم باید تفاوت‌های لاپلاسی و گوسی را بدانیم -
در Lasso که لاپلاسی است داده‌ها اکثراً میل به وسط و صفر دارند و در Ridge که
گوسی است این اتفاق نمی‌افتد. و در Ridge مقادیر نزدیک به صفر تولید می‌شود ولی
دقیقاً صفر کم پیش می‌آید. پس Lasso ← Sparse.
از ویژگی‌های Ridge می‌توان به کاهش واریانس و جلوگیری از overfitting اشاره کرد.