HW1 - Deeplearning -

Tina Halimi 400101078

سوال ۱: الف) در SVM، اگر داده ای را حذف کنیم و آن داده support vector نباشد، تاثیری در مرز ندارد

اما اگر داده support vector باشد، تاثیرگذار است و مرز از داده حذف شده دور می شود

چون می خواهد $\|W\|_2^2$ را min کند. در رگرسیون لجستیک همه داده ها تاثیرگذارند و حذف یک داده

قطعا تاثیرگذار است.

ب) i)

$0 \leq \epsilon_i \leq 1$ طبقه بندی درست ولی مرز را رد کرده

طبقه بندی نادرست $\epsilon_i > 1$ ⟸

$\Rightarrow \sum \epsilon_i > \sum 1 <$ تعداد طبقه بندی نادرست

: طبقه بندی نادرست

: طبقه بندی نادرست

ii) اگر مقدار C کم باشد، نمونه های بیشتری اجازه عبور از margin را دارند.

$C \to \infty \to$ Hard margin SVM $\to$ narrow margin

$C \to 0 \to$ شروط مرز راحت تر در نظر گرفته می شود و یک مرز نرم تعیین می شود $\to$ large margin

iii)

SVM ⟵ یک مرز ایجاد می کند و هدفش این است که Margin ماکزیمم باشد.

Logistic Regression ⟵ با توجه به احتمال تعلق هر داده به کلاس مرز تعیین می شود، با توابع sigmoid یا softmax

اگر تک داده پرت وجود داشته باشد تاثیر آن بر SVM زیاد است اما بر LR خیر.

و SVM برای داده هایی که کاملا قابل جداسازی باشد خوب عمل می کند.

iv) Soft SVM اجازه عبور برخی sample ها را می دهد اما همچنان هدف ماکزیمم کردن margin

است و میزان اجازه رد کردن margin، به C بستگی دارد. در صورتی که LR چون با احتمالات

کلاس بندی را انجام می دهد می تواند بهتر عمل کند (در زمانی که جداساز خطی نباشند)

سوال ۲: الف)

i)

$$\|\hat{x}_i - \hat{x}_j\| = \|V_{1:k} z_i - V_{1:k} z_j\| = \underbrace{(V_{1:k} z_i - V_{1:k} z_j)^T}_{(z_i^T V_{1:k}^T - z_j^T V_{1:k}^T)} (V_{1:k} z_i - V_{1:k} z_j)$$

$$= z_i^T \overbrace{V_{1:k}^T V_{1:k}}^{I} z_i - z_i^T V_{1:k}^T V_{1:k} z_j - z_j^T V_{1:k}^T V_{1:k} z_i + z_j^T V_{1:k}^T V_{1:k} z_j$$

$$= z_i^T z_i + z_j^T z_j - z_i^T z_j - z_j^T z_i = (z_i - z_j)^T (z_i - z_j) = \|z_i - z_j\| \checkmark$$

ii)

$$\sum_{i=1}^{n} \|x_i - \hat{x}_i\|_2^2 = (n-1)\sum_{j=k+1}^{p} \lambda_j \qquad \hat{x}_i = V_{1:k} z_i = V_{1:k} V_{1:k}^T x_i$$

$$\Rightarrow \sum_{i=1}^{n} \|x_i - \hat{x}_i\|_2^2 = \sum_{i=1}^{n} \|x_i - V_{1:k} V_{1:k}^T x_i\|_2^2 = \sum_{i=1}^{n} \|(I - V_{1:k} V_{1:k}^T) x_i\|_2^2$$

$$= \sum_{i=1}^{n} \|V_{k+1:p} V_{k+1:p}^T x_i\|_2^2 = \sum_{i=1}^{n} x_i^T V_{k+1:p} V_{k+1:p}^T V_{k+1:p} V_{k+1:p}^T x_i$$

$$= \sum_{i=1}^{n} x_i^T V_{k+1:p} V_{k+1:p}^T x_i = \sum_{i=1}^{n} (V_{k+1:p}^T x_i)^T (V_{k+1:p}^T x_i)$$

$$= \sum_{i=1}^{n} \|V_{k+1:p} x_i\|^2 = \sum_{i=1}^{n} \sum_{j=k+1}^{p} (v_j x_i)^2 = \sum_{i=1}^{n} \sum_{j=k+1}^{p} v_j^T x_i x_i^T v_j$$

$$= \sum_{j=k+1}^{p} \sum_{i=1}^{n} v_j^T x_i x_i^T v_j = \sum_{j=k+1}^{p} v_j^T \underbrace{\sum_{i=1}^{n} x_i x_i^T}_{S} v_j$$

$$= (n-1) \sum_{j=k+1}^{p} \lambda_j v_j^T v_j = (n-1) \sum_{j=k+1}^{p} \lambda_j \checkmark$$

سوال ۳: الف)

$Xw=y$ $\quad X \rightsquigarrow$ tall $\quad X\in\mathbb{R}^{m\times n}$ $\quad m>n$ $\rightsquigarrow$ جواب ندارد

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

least square : y در فضای ستون X نیست و نزدیک‌ترین حالت برای تصویر آن در فضای ستون X به دست می‌آید.

$Xw=y$ $\quad X \rightsquigarrow$ رتبه کامل ستونی $\Rightarrow X^TX \rightsquigarrow$ invertable

$$\Rightarrow X^TXw=X^Ty \Rightarrow w=(X^TX)^{-1}X^Ty$$

ب)

$$X=U\Sigma V^T \Rightarrow w=(V\Sigma^T\overbrace{U^TU}^{I}\Sigma V^T)^{-1}V\Sigma^TU^Ty$$

$$=(V\underbrace{\Sigma^T\Sigma}V^T)^{-1}V\Sigma^TU^Ty$$

$$=V\underbrace{\Sigma^{+}\Sigma^{+T}\underbrace{V^TV}_{I}\Sigma^T}_{I}U^Ty$$

$$=V\Sigma^{+}U^Ty=V\begin{bmatrix}1/\sigma_1 & & 0 & \cdots & 0\\ & \ddots & & & \vdots\\ 0 & \cdots & \frac{1}{\sigma_n} & & 0\end{bmatrix}V^Ty$$

ج)

$A X=(X^TX)^{-1}X^TX=I \rightsquigarrow$ معکوس چپ

د)

$Xw=y$ $\qquad \min w^Tw$ s.t. $Xw=y$ $\qquad \xrightarrow{\text{لاگرانژین}} \mathcal{L}(w,\lambda)=w^Tw+\lambda^T(Xw-y)$

$$\nabla_w\mathcal{L}=2w+X^T\lambda=0$$

$$\nabla_\lambda\mathcal{L}=Xw-y=0$$

$$\Rightarrow w=-\frac{1}{2}X^T\lambda \qquad \Rightarrow -\frac{1}{2}XX^T\lambda=y \Rightarrow \lambda=-2(XX^T)^{-1}y$$

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

ه)

$X = U\Sigma V^T$ $\qquad w = V\Sigma^T U^T (U\Sigma \underbrace{V^T V^T}_{I} \Sigma^T U^T)^{-1} y$

$$= V\Sigma^T U^T U \Sigma^{+T}\Sigma^{+} U^T y$$

$$= V\Sigma^{+} U^T y = V \begin{bmatrix} \frac{1}{\sigma_1} & & \\ & \ddots & \\ & & 1/\sigma_m \\ 0 & \cdots & 0 \end{bmatrix} U^T y$$

و)

$XB = XX^T(XX^T)^{-1} = I \rightsquigarrow$ معکوس راست

سوال ۵ ۴ ۱

$$\ell(w) = \frac{1}{2}\|y - Fw\|_2^2 \qquad \nabla_w \ell(w) = \frac{\partial}{\partial w}\left(\frac{1}{2}(Fw-y)^T(Fw-y)\right) = F^T(Fw-y)$$

$$H = \frac{\partial}{\partial w}\left(F^T(Fw-y)\right) = F^TF$$

تقریب سری تیلور $\rightsquigarrow \hat{L}(w) = L(w^*) + (w-w^*)^T \underbrace{\nabla_w L(w^*)}_{0 \text{ در بهینه}} + \frac{1}{2}(w-w^*)^T H (w-w^*)$

$$\Rightarrow \hat{L}(w) = L(w^*) + \frac{1}{2}(w-w^*)^T H (w-w^*) \Rightarrow \nabla_w \hat{L}(w) = H(w-w^*)$$

$$= F^TF(w-w^*)$$

$$\nabla_w L(w^*) = 0 = F^T(Fw^* - y) = 0 \Rightarrow F^TFw^* = F^Ty$$

$$\Rightarrow \nabla_w \hat{L}(w) = F^TFw - F^Ty \rightsquigarrow \text{gradient descent: } w_t = w_{t-1} - \eta\left(F^T(Fw_{t-1} - y)\right)$$

$$w_t = w_{t-1} - \eta F^TFw_{t-1} + \eta F^Ty$$

$$\Rightarrow w_t = (I - \eta F^TF)w_{t-1} + \eta F^Ty$$

کوشی شوارتز

نامساوی مثلث $\Rightarrow \|w_t\| \leq \|(I - \eta F^TF)w_{t-1}\| + \eta\|F^Ty\|$ $\qquad \|F^Ty\| < \|F^T\|\|y\| = \alpha\|y\|$

$$\|(I - \eta F^TF)w_{t-1}\|_2 \leq \|I - \eta F^TF\|\|w_{t-1}$$

$$F^TF = Q\Lambda Q^T \rightsquigarrow I - \eta F^TF = I - \eta Q\Lambda Q^T = Q(I - \eta\Lambda)Q^T$$

$$\|I-\eta F^TF\| = \|Q(I-\eta\Lambda)Q^T\| \le \underbrace{\|Q\|_2}_{1}\,\|I-\eta\Lambda\|_2\,\underbrace{\|Q^T\|_2}_{1} = \|I-\eta\Lambda\|_2$$

$$= \sqrt{\underbrace{\lambda_{max}\left((I-\eta\Lambda)^T(I-\eta\Lambda)\right)}} \longrightarrow \text{مقدار ویژه } \lambda'_{max} = (1-\eta\sigma_{i\,min}^2)^2, \quad 0 \le \sigma_{min}^2 \le 1 \;\Big\}\Rightarrow \lambda'_{max} < 1$$

$$\Rightarrow \|I-\eta\Lambda\|_2 < 1 \Rightarrow \|I-\eta F^TF\| < 1 \Rightarrow \|I-\eta F^TF\|_2\,\|w_{t-1}\|_2 \le \|w_{t-1}\|_2$$

$$\Rightarrow \|(I-\eta F^TF)\|_2\,\|w_{t-1}\| \le \|w_{t-1}\|_2$$

همچنین داشتیم $\|F^Ty\| < \alpha\|y\|$, $\|w_t\| \le \|(I-\eta F^TF)w_{t-1}\| + \eta\|F^Ty\|$

$$\Rightarrow \|w_t\|_2 \le \|w_{t-1}\|_2 + \eta\alpha\|y\|_2 \quad \checkmark$$

برای اینکه همگرا بشود $\rightarrow \|I-\eta F^TF\| = \|I-\eta\Lambda\|_2 = \|1-\varepsilon\sigma_{min}^2\| < 1$

اگر واگرا بشود می‌توان نتیجه گرفت $\Rightarrow \|1-\varepsilon\sigma_{min}^2\| \ge 1$

سوال ۵ : الف)

$$E_{y\sim p(y|x),D}\left[(y-\hat{f}_{\hat{\theta}(D)}(x))^2\right] = E_{y\sim p(y|x),D}\left[\hat{f}_{\hat{\theta}}(x)-y\right]$$
$$+E_D\left[\left(\hat{f}_{\hat{\theta}}(x)-E_D[\hat{f}_{\hat{\theta}(D)}(x)]\right)^2\right]$$
$$+\sigma^2$$

$$E_{y\sim p(y|x),D}\left[(y-h(x)+h(x)-\hat{f}_{\hat{\theta}(D)}(x))^2\right]$$

$$= E_{y\sim p(y|x),D}\left[(y-h(x))^2+(h(x)-\hat{f}_{\hat{\theta}(D)}(x))^2+2(y-h(x))(h(x)-\hat{f}_{\hat{\theta}(D)}(x))\right]$$

$$= E\left[(y-h(x))^2\right]+E\left[(h(x)-\hat{f}_{\hat{\theta}(D)}(x))^2\right]+\underbrace{2E\left[(y-h(x))(h(x)-\hat{f}_{\hat{\theta}(D)}(x))\right]}$$

$$E\left[(h(x)-\hat{f}_{\hat{\theta}(D)}(x))\underbrace{E(\overbrace{y-h(x)}^{\epsilon\sim N(0,\sigma^2)})}_{0}\right]$$

$$= \underbrace{E\left[(y-h(x))^2\right]}_{\text{noise}}+E\left[(h(x)-\hat{f}_{\hat{\theta}(D)}(x))^2\right] \qquad \bar{f} = E_D\left[\hat{f}_{\hat{\theta}(D)}(x)\right]$$

$$E\left[(h(x)-\bar{f}+\bar{f}-\hat{f}_{\hat{\theta}(D)}(x))^2\right] = E\left[(\bar{f}-\hat{f}_{\hat{\theta}(D)}(x))^2+(h(x)-\bar{f})^2\right.$$
$$\left.+2(h(x)-\bar{f})(\bar{f}-\hat{f}_{\hat{\theta}(D)}(x))\right]$$

$$= \underset{y\sim p(y|x)}{E}\left[\underbrace{E_D((\bar{f}-\hat{f}_{\hat{\theta}(D)}(x))^2)}_{\text{var}(x)}+\underbrace{\overbrace{E_D((h(x)-\bar{f})^2)}^{(\bar{f}(x)-h(x))^2}}_{\text{bias}(x)}\right]$$

$$= \text{var} + \text{bias}$$

$$\Rightarrow E\left[(y-\hat{f}_{\hat{\theta}(D)}(x))^2\right] = E\left[(y-h(x))^2\right]+E\left[(\bar{f}-\hat{f}_{\hat{\theta}(D)}(x))^2\right]$$
$$+E\left((\bar{f}(x)-h(x))^2\right)$$

$$= \text{noise} + \text{var} + \text{bias}$$

$y = X\theta^* + \varepsilon \qquad \hat{\theta} = (X^T X)^{-1} X^T y$

$$\text{bias} = x^T\theta^* - E_D[x\hat{\theta}] = x^T\theta^* - E_D[x^T(X^TX)^{-1}Xy]$$

$$= x^T\theta^* - E_D[x^T(X^TX)^{-1}X^T(X\theta^* + \varepsilon)]$$

$$= x^T\theta^* - E_D[x^T\theta^*] - E_D[x(X^TX)^{-1}X^T\varepsilon]$$

$$= x^T\theta^* - \underbrace{E_D[x^T\theta^*]}_{x^T\theta^*} - \underbrace{x(X^TX)^{-1}X^T E[\varepsilon]}_{0} = 0$$

$$\text{var} = E[(f_D(x) - \bar{f}_D(x))^2] = E[(x^T\hat{\theta} - \underbrace{E_D[f_D(x)]}_{x^T\theta^*})^2]$$

بدون بایاس $\leftarrow x^T\theta^*$

$$= E[(x^T(X^TX)^{-1}X^T \overset{X\theta^* + \varepsilon}{y} - x^T\theta^*)^2] = E((\cancel{x^T\theta^*} + x^T(X^TX)^{-1}X^T\varepsilon - \cancel{x^T\theta^*})^2)$$

$$= E[x^T(X^TX)^{-1}X^T\varepsilon\,\varepsilon^T X(X^TX)^{-1}x] = x^T(X^TX)^{-1}X^T \underbrace{E[\varepsilon\varepsilon^T]}_{\overset{I}{}\ \text{var}[\varepsilon_i] = 1} X(X^TX)^{-1}x$$

مستقلند بقیه صفر

$$= x^T(X^TX)^{-1}X^TX(X^TX)^{-1}x = x^T(X^TX)^{-1}x$$