

## مکانی ۱ تئوری یادگیری ماشین

401110642

۱. افت) هر آنکه این تابع بردارها را از میدان  $\mathbb{R}^D \sim [0,1]^D$  کلاس  $\mathcal{C}$  کند به صورتی که جمع تمام درایهای بردار را با ۱ نمود. این خصیت تابع Softmax آن را برای یادگیری توابع احتمالات Classifications بسیار مناسب می‌سازد. همین فرم این تابع دشمنی آن پیگیرهای است که محاسبات پیوسته به الگوریتم gradient descent را ساده نمایند.

ب) بالا بودن داریانی شناخته این است که مدل به تعدادی داده های train خطا را ندارد اما علیرغم آن دوره داده های test مقابل نمی‌باشد. در این حالت مدل پیش‌بینی داده های train خطا را ندارد اما علیرغم آن دوره داده های test مقابل نمی‌باشد. این مکار را Regularization (روشنی‌کاری) می‌نامند (Ridge یا Lasso).

پ) درین Ridge فرآیند را کمپکت نمایند و دارای اثبات آن هاست این فرضیه که درین Lasso عکیل داریانی شناخته است. (آن جایی که فرض کردن هر ضریب مغایر خوب دیگری را بخواهد نمی‌شود در این مسائل درین Ridge را به معنی آنچه می‌دمم)

ت) در مورد آن که فرآیند بزرگ باشد، وزن های معمولی آن کمپکت باشند این فرضیه به مسائل مثل Bias، Variance کم در مدل نمایند. به طور عالی در مورد آن که آن کمپکت باشد، وزن های کوانته معمولی بزرگ باشند که بفرآیند بزرگ باشند و Variance، Bias با این نسبت

2. انت

$$\hat{Y} = w_j X_j$$

$$\rightarrow \bar{J} = \| Y - \hat{Y} \|_2^2 = (Y - w_j X_j)^T (Y - w_j X_j)$$

$$= Y^T Y - Y^T w_j X_j - w_j X_j^T Y + w_j^2 X_j^T X_j$$

$$\rightarrow \frac{\partial}{\partial w_j} \bar{J} = 0 \Rightarrow -Y^T X_j - X_j^T Y + 2 w_j X_j^T X_j = 0$$

$$\Rightarrow 2 X_j^T Y = 2 w_j X_j^T X_j$$

$$\Rightarrow w_j = \frac{X_j^T Y}{X_j^T X_j}$$

(.)

$$\bar{J}(w) = \| Y - X^T w \|_2^2 = (Y - X^T w)^T (Y - X^T w)$$

$$= (Y^T - w^T X)(Y - X^T w) = Y^T Y - w^T X Y - Y^T X^T w + w^T X X^T w$$

$$= Y^T Y - 2 w^T X Y + w^T X X^T w$$

$$\Rightarrow \frac{\partial \bar{J}}{\partial w} = -2 X Y + 2 X X^T w = 0$$

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

حالاً داریم مسئل حل آن و معادله بهینه شده را در

$$X_k^T X_j = \begin{cases} X_j^T X_j & k=j \\ 0 & k \neq j \end{cases}$$

$$\Rightarrow X X^T = \begin{bmatrix} X_1^T X_1 & & \\ & X_2^T X_2 & \emptyset \\ \emptyset & & \ddots \\ & & X_L^T X_L \end{bmatrix} \Rightarrow (X X^T)^{-1} = \begin{bmatrix} \frac{1}{X_1^T X_1} & & \\ & \frac{1}{X_2^T X_2} & \emptyset \\ \emptyset & & \frac{1}{X_L^T X_L} \end{bmatrix}$$

$$\Rightarrow w_j = \frac{X_j^T y}{X_j^T X_j}$$

$$J = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - w_0 X_j^{(0)} - w_j) ^2$$

$$\frac{\partial J}{\partial w_0} = -2 \sum (y_i - w_0 X_j^{(0)} - w_j) = 0 \rightarrow \sum_{i=1}^N w_0 = \sum y_i - w_j X_j^{(0)}$$

$$\Rightarrow w_0 = \frac{1}{N} \left( \sum y_i - w_j \sum X_j^{(0)} \right)$$

$$\frac{\partial J}{\partial w_j} = -2 \sum X_j^{(0)} (y_i - w_0 X_j^{(0)} - w_j) = 0$$

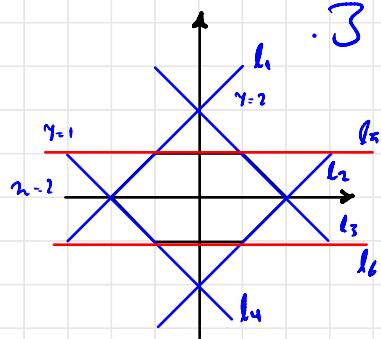
$$\Rightarrow \sum X_j^{(0)} y_i - w_j \sum X_j^{(0)2} - w_0 \sum X_j^{(0)} = 0$$

$$\Rightarrow X_j^T y - X_j^T X_j w_j - (\sum X_j) \left( \frac{\sum y_i - w_j \sum X_j^{(0)}}{N} \right)$$

$$\Rightarrow w_j = \frac{\frac{1}{N} \sum X_j^{(0)} y_i - \left( \frac{1}{N} \sum X_j^{(0)} \right) \left( \frac{1}{N} \sum y_i \right)}{\frac{1}{N} \sum (X_j^{(0)})^2 - \left( \frac{1}{N} \sum X_j^{(0)} \right)^2}$$

$$w_j = \frac{\bar{X_j y} - \bar{X_j} \bar{y}}{\bar{X_j^2} - \bar{X_j}^2}, w_0 = \bar{y} - w_j \bar{X_j}$$

نایمه سه‌ضلعی محدوده در مکان با صورت انتزاع نایمه را مل رفع آبی  
وزیر ترز در مکان کوت اسلام عادل خط حاره ای یا چون رفاهادل و بطب دین رای دین



$$l_1: n_2 - n_1 = 2 \rightarrow n_1 - n_2 + 2 \geq 0$$

$$l_2: n_2 - n_1 = -2 \Rightarrow -n_1 + n_2 + 2 \geq 0$$

$$l_3: n_1 + n_2 = 2 \rightarrow -n_1 - n_2 + 2 \geq 0$$

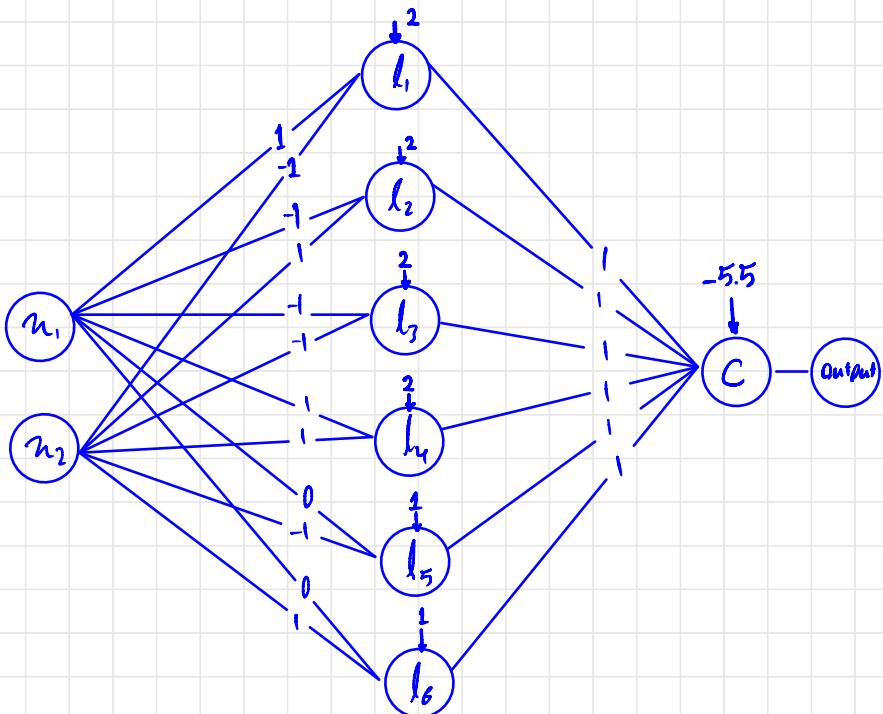
$$l_4: n_1 + n_2 = -2 \rightarrow n_1 + n_2 + 2 \geq 0$$

$$l_5: n_2 = 1 \rightarrow 1 - n_2 \geq 0$$

$$l_6: n_2 = -1 \rightarrow 1 + n_2 \geq 0$$

$$\text{step}(n) = \begin{cases} 1 & n > 0 \\ 0 & n \leq 0 \end{cases}$$

حال با درست کردن شبکه حرکات هر کدام از نایمه‌های بات را پاک می‌کنند و پس یک نایم  
دسته ای از این نایمه‌ها کام نموده ایجاد می‌کنند



$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}}$$

(ال 4)

$$\rightarrow \frac{\partial \hat{y}_i}{\partial z_i} = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}} - \frac{e^{z_i}}{\left(\sum_{j=1}^n e^{z_j}\right)} \cdot e^{z_i} = \hat{y}_i - \hat{y}_i^2 = \hat{y}_i (1 - \hat{y}_i)$$

$$\rightarrow \frac{\partial \hat{y}_i}{\partial z_k} = \frac{-e^{z_i}}{\left(\sum_{j=1}^n e^{z_j}\right)^2} \cdot e^{z_k} = \frac{-e^{z_i}}{\sum e^{z_j}} \cdot \frac{e^{z_k}}{\sum e^{z_k}} = -\hat{y}_i \hat{y}_k$$

(ك)

$$L = -\sum_{i=1}^n y_i \log(\hat{y}_i)$$

$$\begin{aligned} \rightarrow \frac{\partial L}{\partial z_k} &= \frac{\partial}{\partial z_k} \left( -\sum y_i \log \hat{y}_i \right) = -\sum_{i=1}^n y_i \cdot \frac{1}{\hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial z_k} \\ &= -\sum_{\substack{i=1 \\ i \neq k}}^n y_i \cdot \frac{1}{\hat{y}_i} \cdot (-\hat{y}_i \hat{y}_k) + -y_k \cdot \frac{1}{\hat{y}_k} \cdot (\hat{y}_k (1 - \hat{y}_k)) \\ &= \sum_{i=1}^n y_i \hat{y}_k - y_k + y_k \hat{y}_k \end{aligned}$$

$$\rightarrow \frac{\partial L}{\partial z_k} = -y_k + \left(\sum_{i=1}^n y_i\right) \hat{y}_k$$

٥. (الـ) ابره جواب و دوست نامی

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

$$\Rightarrow \frac{\partial \mathcal{J}}{\partial \beta} = -2X^T Y + 2X^T X\beta = 0 \Rightarrow \hat{\beta}_{LS} = (X^T X)^{-1} X^T Y$$

$$\mathcal{J}(\beta, \lambda) = \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 = Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta + \lambda \beta^T \beta$$

$$\Rightarrow \frac{\partial \mathcal{J}}{\partial \beta} = -2X^T Y + 2X^T X\beta + 2\lambda \beta = 0 \Rightarrow \hat{\beta}_{Ridge}^{(\lambda)} = (X^T X + \lambda I)^{-1} X^T Y$$

$$\Rightarrow \hat{\beta}_{Ridge}^{(\lambda)} = (X^T X + \lambda I)^{-1} X^T X \cdot (X^T X)^{-1} X^T Y = W_{\lambda} \hat{\beta}_{LS}$$

$$W_{\lambda} = (X^T X + \lambda I)^{-1} X^T X$$

$$\text{Var}(\hat{\beta}_{LS}) = \text{Var}((X^T X)^{-1} X^T Y)$$

$$= (X^T X)^{-1} X^T \text{Var}(Y) (X^T X)^{-1} X^T$$

$$= (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1}$$

$$\text{Var}(\hat{\beta}_{Ridge}^{(\lambda)}) = \text{Var}(W_{\lambda} \hat{\beta}_{LS})$$

$$= W_{\lambda} \text{Var}(\hat{\beta}_{LS}) W_{\lambda}^T$$

$$= (X^T X + \lambda I)^{-1} X^T X \cdot (X^T X)^{-1} \sigma^2 X^T X (X^T X + \lambda I)^{-1}$$

$$= \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}$$

$$\begin{aligned}\text{Var}(\hat{\beta}_{\text{Ridge}}^{(A)}) &= \sigma^2 \cdot (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} \\ &= \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X)^{-1} X^T X (X^T X + \lambda I)^{-1} \\ &\equiv \sigma^2 S^T (X^T X)^{-1} S \quad ; \quad S = X^T X (X^T X + \lambda I)^{-1}\end{aligned}$$

$$\begin{aligned}\text{Var}(\hat{\beta}_{\text{LS}}) - \text{Var}(\hat{\beta}_{\text{Ridge}}^{(A)}) &= \sigma^2 (X^T X)^{-1} - \sigma^2 S^T (X^T X)^{-1} S \\ &= \sigma^2 [S^T (S^T)^{-1} (X^T X)^{-1} S^{-1} S - S^T (X^T X)^{-1} S] \\ &= \sigma^2 S^T [(S^T)^{-1} (X^T X)^{-1} S^{-1}] S \\ &= \sigma^2 S^T [(X^T X)^{-1} (X^T X + \lambda I) (X^T X)^{-1} (X^T X + \lambda I) (X^T X)^{-1} (X^T X)^{-1}] S \\ &= \sigma^2 S^T [(I + \lambda (X^T X)^{-1}) (X^T X)^{-1} (I + \lambda (X^T X)^{-1}) (X^T X)^{-1} (X^T X)^{-1}] S \\ &= \sigma^2 S^T [(X^T X)^{-1} + 2\lambda (X^T X)^{-2} + \lambda^2 (X^T X)^{-3} (X^T X)^{-1}] S \\ &= \sigma^2 (X^T X + \lambda I)^{-1} (X^T X) [2\lambda (X^T X)^{-2} + \lambda^2 (X^T X)^{-3}] (X^T X) (X^T X + \lambda I)^{-1} \\ &= \sigma^2 (X^T X + \lambda I)^{-1} (2\lambda I + \lambda^2 (X^T X)^{-1}) (X^T X + \lambda I)^{-1}\end{aligned}$$

$\underline{V \neq 0} \rightarrow G \vdash \sim A \vdash V$ .  $\therefore$  Positive definite  $(X^T X + \lambda I)^{-1}$  w/  $\lambda > 0$  (Q.E.D.)

$$Z = (X^T X + \lambda I)^{-1} V \neq 0$$

$$\begin{aligned}\Rightarrow V^T [\text{Var}(\hat{\beta}_{\text{LS}}) - \text{Var}(\hat{\beta}_{\text{Ridge}}^{(A)})] V &= \sigma^2 Z^T [2\lambda I + \lambda^2 (X^T X)^{-1}] Z \\ &= 2\lambda \sigma^2 Z^T Z + \lambda^2 \sigma^2 Z^T (X^T X)^{-1} Z > 0\end{aligned}$$

$$\Rightarrow \underline{\text{Var}(\hat{\beta}_{\text{LS}}) > \text{Var}(\hat{\beta}_{\text{Ridge}}^{(A)})}$$

$$\hat{Y}(\lambda) = X \beta_{\text{Ridge}}^{(1)}$$

$$= X (X^T X + \lambda I)^{-1} X^T y$$

$$\Rightarrow \text{Var}(\hat{Y}(\lambda)) = [X (X^T X + \lambda I)^{-1} X^T]^T \text{Var}(y) [X (X^T X + \lambda I)^{-1} X^T]$$

$$= \sigma^2 X (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} X^T$$

$$X^T X = U D_x U^T, \quad U U^T = I$$

$$\Rightarrow \text{Var}(\hat{Y}) = \sigma^2 U D_x U^T (U (D_x + \lambda I) U^T)^{-1} U D_x U^T U D_x U^T$$

$$(U (D_x + \lambda I) U^T)^{-1} U D_x U$$

$$= \sigma^2 U D_x (D_x + \lambda I)^{-1} D_x^2 (D_x + \lambda I)^{-1} D_x U$$

$$\Rightarrow \text{tr} \{ \text{Var}(\hat{Y}) \} = \text{tr} \left[ \sigma^2 D_x (D_x + \lambda I)^{-1} D_x^2 (D_x + \lambda I)^{-1} D_x \right]$$

$$= \text{tr} \left[ \sigma^2 D_x^4 (D_x + \lambda I)^{-2} \right] ; \quad \square \text{ oder } D_x$$

$$\Rightarrow \text{tr} \{ \text{Var}(\hat{Y}(\lambda)) \} = \sigma^2 \sum_{j=1}^p \frac{(D_{xj})_{\text{adj}}^4}{((D_{xj})_{\text{adj}} + \lambda)^2}$$