第二次作业 胡延伸 PB22050983

1. 正则化的贝叶斯解释

(a) 证:

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|x,y) = \arg\max_{\theta} \frac{p(\theta,x,y)}{p(x,y)}$$

其中, $\frac{p(\theta,x,y)}{p(x,y)} = \frac{p(\theta y|x,\theta)}{p(x,y)} \cdot p(x,\theta)$

由 $p(\theta)=p(\theta|x) \Rightarrow p(x,\theta)=p(x)p(\theta)$ 则:

$$\frac{p(\theta,x,y)}{p(x,y)} = \frac{p(y|x,\theta)}{p(x,y)} \cdot p(x)p(\theta) = \frac{p(y|x,\theta)p(\theta)}{p(y|x)}$$

由 $p(y|x)$ 与 θ 无关得: 当 $p(y|x,\theta)p(\theta)$ 最大, $p(y|x,\theta)p(\theta)/p(y|x)$ 最大

故 $\theta_{MAP} = \arg\max_{\theta} p(y|x,\theta)p(\theta)$

(b) 证:

$$\arg\min_{\theta} -\log p(y|\theta,x) + \lambda\|\theta\|_2^2 = \arg\max_{\theta} \frac{p(y|\theta,x)}{e^{\lambda\|\theta\|_2^2}}$$

由 (a) 得: $\theta_{MAP} = \arg\max_{\theta} p(y|\theta)p(\theta)$

故只需证: $p(\theta) = e^{-\lambda\|\theta\|_2^2} = e^{-\lambda\theta^T\theta}$

由 $\theta \sim N(0,\eta^2 I) \Rightarrow p(\theta) = e^{-\frac{1}{2}\theta^T\theta\cdot\frac{1}{\eta^2}}$

则 $\exists \lambda = \frac{1}{2\eta^2}$, s.t. $\theta_{MAP} = \arg\min_{\theta} -\log p(y|x,\theta) + \lambda\|\theta\|_2^2$

(c) 证:

由题得: $\vec{y} = X^T\theta + E$, 其中, $E = (\epsilon^{(1)}, \epsilon^{(2)}, \cdots)^T$, $\epsilon^{(i)}$ 相互之间独立同分布

则 $E = y - X^T\theta \sim N(0,\sigma^2 I) \Rightarrow p(y|\theta,X) = e^{-\frac{1}{2\sigma^2}\|\vec{y}-X^T\theta\|_2^2}$

故由 (b) 得:

$$\theta_{MAP} = \arg\min_{\theta} -\log p(y|x,\theta) + \lambda\|\theta\|_2^2 = \arg\min_{\theta} \frac{1}{2\sigma^2}\|\vec{y}-X^T\theta\|_2^2 + \frac{1}{2\eta^2}\|\theta\|_2^2$$

(d) 证:

由题得:

$X^T\theta - \vec{y} = E \sim N(0,\sigma^2 I)$ 即 $p(y|\theta,x) = e^{-\frac{1}{2\sigma^2}\|X^T\theta-\vec{y}\|_2^2}$

又由 θ_i 独立同分布得: $p(\theta) = \prod p(\theta_i) = (\frac{1}{2b})^n e^{-\frac{\|\theta\|_1}{b}}$, n 为 θ 的维度

由 (a) 得:

$$\theta_{MAP} = \arg\max_{\theta} p(y|\theta,x)p(\theta) = \arg\max_{\theta} e^{-\frac{1}{2\sigma^2}\|X^T\theta-\vec{y}\|_2^2 - \frac{1}{b}\|\theta\|_1} = \arg\min_{\theta}\left(\frac{1}{2\sigma^2}\|X^T\theta-\vec{y}\|_2^2 + \frac{1}{b}\|\theta\|_1\right)$$

$$\Rightarrow \theta_{MAP} = \arg\min_{\theta} \|X^T\theta - \vec{y}\|_2^2 + \frac{2\sigma^2}{b}\|\theta\|_1$$

即当 $\gamma = \frac{2\sigma^2}{b}$ 时，$\theta_{MAP} = \arg\min_{\theta} J(\theta)$.

2. 逻辑回归：训练稳定性

(a)证：

$$\lim_{C\to+\infty} P(y=1 \mid x^{(i)}) = \lim_{C\to+\infty} \frac{1}{1+e^{-\theta^T x^{(i)}}}$$

对于正例，$\theta^T x^{(i)} > 0$ 则 $\lim_{C\to+\infty} \frac{1}{1+e^{-\theta^T x^{(i)}}} = 1$；

对于正确分类的负例，$\theta^T x^{(i)} < 0$ 则 $\lim_{C\to+\infty} \frac{1}{1+e^{-\theta^T x^{(i)}}} = 0$；

(b)证：

①

对于正确分类的正例，$L(\theta) = -\log P(y=1 \mid x^{(i)}) \to 0 \ (C\to+\infty)$

对于正确分类的负例，$L(\theta) = -\log(1 - P(y=1 \mid x^{(i)})) \to 0 \quad (C\to+\infty)$

则 $C\to+\infty$，$L(\theta)\to 0$ 即无限下降趋于0.

②

在实际情况中，这意味着 $\|\theta\|$ 不断增大，$L(\theta)$ 才能下降.

当参数更新步长太长，会剧烈震荡；当步长太小，会导致收敛速度极慢.

(c)

当存在错误分类：

对于 $y^{(i)}=0$，$\exists\ \theta^T x > 0$ 此时 $L(\theta) = -\log(1 - P(y=1 \mid x^{(i)}) \to +\infty \ (C\to+\infty)$

对于 $y^{(i)}=1$，$\exists\ \theta^T x < 0$ 此时 $L(\theta) = -\log(P(y=1 \mid x^{(i)}) \to +\infty \ (C\to+\infty)$

$L(\theta)$ 显然存在大于0的下界，因为错误分类的点对 $L(\theta)$ 的贡献恒 $\geq -\log 0.2$.

此时 $L(\theta)$ 取最优时，$\|\theta\|$ 为有限的值 则 梯度下降会使得 θ 往最优点靠近.

(d)

(1) 应用不同学习率不会改变数据集线性可分这一性质，故依然不会收敛

(2) 学习率下降会导致收敛速度更慢

(3) 线性缩放依然维持数据集线性可分，故仍不会收敛.

(4) 可以收敛，这样不会使得 $L(\theta)$ 最优时，$\|\theta\|$ 趋于无穷

(5) 可以收敛；增加噪声可能使数据集变为线性不可分.

3. 多分类问题.

(a)

$$\vec{a} = SM(\vec{z}) = \left(\frac{e^{-1}}{e^{-1}+e^{0}+e^{1}}, \frac{e^{0}}{e^{-1}+e^{0}+e^{1}}, \frac{e^{1}}{e^{-1}+e^{0}+e^{1}} \right) = (0.090, 0.245, 0.665)$$

(b)

$$NLL(\vec{a}, \vec{y}) = -\sum_{j=1}^{N_L} y_j \ln a_j^L = -(0,0,1)(\ln 0.3, \ln 0.5, \ln 0.2)^T \approx 1.609$$

(c)

$$NLL = -\sum_{j=1}^{n_L} y_j \ln a_j^L = -\sum_{j=1}^{n_L} y_j \ln \frac{e^{z_j}}{\sum e^{z_i}} = -\sum_{j=1}^{n_L} y_j z_j + \left(\ln \sum e^{z_k}\right) \sum y_j$$

由于 $\vec{y}$ 为 one-hot 编码则 $\sum y_j = 1$

则: $NLL = -\sum_{j=1}^{n_L} y_j z_j + \ln \sum e^{z_k} \Rightarrow \frac{\partial NLL}{\partial W_{kj}^L} = \frac{\partial NLL}{\partial z_j} \cdot \frac{\partial z_j}{\partial W_{k,j}^L} = (-y_j + a_j) x_k.$

对于, $W^L = \begin{bmatrix} 1 & -1 & -2 \\ -1 & 2 & 1 \end{bmatrix} \Rightarrow \vec{z} = xW = (0, 1, -1) \Rightarrow \vec{a} = (0.245, 0.665, 0.090)$

$$\Rightarrow \nabla_{W^L} NLL = \begin{pmatrix} 0.245 & -0.335 & 0.090 \\ 0.245 & -0.335 & 0.090 \end{pmatrix}$$

(d)

由 $\vec{a} = (0.245, 0.665, 0.090) \Rightarrow P_1 = 0.665$

(e)

$$W^L = W^L - 0.5 \times \nabla_{W^L} NLL = \begin{pmatrix} 0.8775 & -1.3325 & -2.045 \\ -1.1225 & 1.6675 & 0.955 \end{pmatrix}$$

(f)

$\vec{z} = xW = (-0.245, 0.335, -1.09) \Rightarrow \vec{a} = (0.311, 0.555, 0.134)$

$\Rightarrow P_1 = 0.555$

4.

(a)

$(z^1)^T = x^T W^1 + (W_0^1)^T = \begin{pmatrix} 3 & 14 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} + (-1\ -1\ -1\ -1) = (2\ \ 13\ -4\ -15)$

$\Rightarrow (a^1)^T = (2\ \ 13\ \ 0\ \ 0) = (f^1(z_1^1), f^1(z_2^1), f^1(z_3^1), f^1(z_4^1))$

$\Rightarrow (z^2)^T = (a^1)^T W^2 + (W_0^2)^T = (2\ 13\ 0\ 0) \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} + (0\ \ 2) = (15\ -13)$

$\Rightarrow (a_1^2, a_2^2) = \left(\frac{e^{15}}{e^{15}+e^{-13}}, \frac{e^{-13}}{e^{15}+e^{-13}} \right) \approx (1.00, 0.00)$

(b)

$(z^1)^T = X^T W^1 + (W_0^1)^T = \begin{pmatrix} 0.5 & 0.5 \\ 0 & 2 \\ -3 & 0.5 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} + (-1\ -1\ -1\ -1) = \begin{pmatrix} -0.5 & -0.5 & -1.5 & -1.5 \\ -1 & 1 & -1 & -3 \\ -4 & -0.5 & 2 & -1.5 \end{pmatrix}$

$\Rightarrow a_1^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$

(c)

(1)

$(z^2)^T = (f^1(z_1^1), f^1(z_2), f^1(z_3^1), f^1(z_4^1)) \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} + (0\ \ 2) = (0\ \ 2)$

$\Rightarrow (a_1^2, a_2^2) = \left(\frac{e^0}{e^0+e^2}, \frac{e^2}{e^0+e^2} \right)$

(2)

同上, $(z^2)^T = (1, -1) + (0, 2) = (1, 1) \Rightarrow (a_1^2, a_2^2) = (0.5, 0.5)$

(3)

$(z^2)^T = (3, -3) + (0, 2) = (3, -1) \Rightarrow (a_1^2, a_2^2) = \left(\frac{e^3}{e^3+e^{-1}}, \frac{e^{-1}}{e^3+e^{-1}} \right) \approx (0.982, 0.018)$

(d)

损失函数 $L = -y_1 \ln a_1^2 - y_2 \ln a_2^2 = -y_1 \ln \frac{e^{z_1^2}}{e^{z_1^2}+e^{z_2^2}} - y_2 \ln \frac{e^{z_2^2}}{e^{z_1^2}+e^{z_2^2}} = -y_1 z_1^2 - y_2 z_2^2 + (y_1+y_2) \ln (e^{z_1^2}+e^{z_2^2})$

由于为 one-hot 编码 $\Rightarrow L = -y_1 z_1^2 - y_2 z_2^2 + \ln (e^{z_1^2}+e^{z_2^2})$

由链式法则:

$$\frac{\partial L}{\partial W^2} = \frac{\partial L}{\partial z_1^2} \cdot \frac{\partial z_1^2}{\partial W^2} + \frac{\partial L}{\partial z_2^2} \cdot \frac{\partial z_2^2}{\partial W^2} = \left(\frac{\partial L}{\partial z_1^2} \ \ \frac{\partial L}{\partial z_2^2} \right) \begin{pmatrix} \frac{\partial z_1^2}{\partial W^2} \\ \frac{\partial z_2^2}{\partial W^2} \end{pmatrix}$$

由 $\frac{\partial L}{\partial z_1^2} = -z_1^2 + a_1^2 = -15+1 = -14$, $\frac{\partial L}{\partial z_2^2} = -z_2^2 + a_2^2 = 13+0 = 13$

$$\frac{\partial z_1^2}{\partial W^2} = \begin{pmatrix} a_1^1 & 0 \\ a_2^1 & 0 \\ a_3^1 & 0 \\ a_4^1 & 0 \end{pmatrix} \qquad \frac{\partial z_2^2}{\partial W^2} = \begin{pmatrix} 0 & a_1^1 \\ 0 & a_2^1 \\ 0 & a_3^1 \\ 0 & a_4^1 \end{pmatrix}$$

$$\Rightarrow \frac{\partial L}{\partial W^2} = -14 \cdot \begin{pmatrix} 2 & 0 \\ 13 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} + 13 \begin{pmatrix} 0 & 2 \\ 0 & 13 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -28 & 26 \\ -182 & 169 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

同理, $\frac{\partial L}{\partial W_0^2} = \begin{pmatrix} \frac{\partial L}{\partial z_1^2} & \frac{\partial L}{\partial z_2^2} \end{pmatrix} \begin{pmatrix} \frac{\partial z_1^2}{\partial W_0^2} \\ \partial z_2^2 / \partial W_0^2 \end{pmatrix} = -14 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 13 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -14 \\ 13 \end{pmatrix}$

$$\frac{\partial L}{\partial W^1} = \frac{\partial L}{\partial \vec{z}^2} \cdot \frac{\partial (z_1^2, z_2^2)}{\partial (a_1^1, \cdots, a_4^1)} \cdot \begin{pmatrix} \partial a_1^1 / \partial z_1^1 \\ \vdots \\ \partial a_4^1 / \partial z_4^1 \end{pmatrix} \cdot \begin{pmatrix} \partial z_1^1 / \partial W^1 \\ \vdots \\ \partial z_4^1 / \partial W^2 \end{pmatrix}$$, 其中最后两项为元素相乘.

$$\frac{\partial L}{\partial W_0^1} = \frac{\partial L}{\partial \vec{z}^2} \cdot \frac{\partial (z_1^2, z_2^2)}{\partial (a_1^1, \cdots, a_4^1)} \cdot \begin{pmatrix} \partial a_1^1 / \partial z_1^1 \\ \vdots \\ \partial a_4^1 / \partial z_4^1 \end{pmatrix} \cdot \begin{pmatrix} \partial z_1^1 / \partial W_0^1 \\ \vdots \\ \partial z_4^1 / \partial W_0^1 \end{pmatrix}$$

由 $\frac{\partial (z_1^2, z_2^2)}{\partial (a_1^1, \cdots, a_4^1)} = (W^2)^T$, $\begin{pmatrix} \partial a_1^1 / \partial z_1^1 \\ \vdots \\ \partial a_4^1 / \partial z_4^1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$,

$$\frac{\partial z_1^1}{\partial W^1} = \begin{pmatrix} x_1 & 0 & 0 & 0 \\ x_2 & 0 & 0 & 0 \end{pmatrix}, \quad \frac{\partial z_2^1}{\partial W^1} = \begin{pmatrix} 0 & x_1 & 0 & 0 \\ 0 & x_2 & 0 & 0 \end{pmatrix}, \cdots, \quad \frac{\partial z_4^1}{\partial W^2} = \begin{pmatrix} 0 & 0 & 0 & x_1 \\ 0 & 0 & 0 & x_2 \end{pmatrix}$$

$$\frac{\partial z_1^1}{\partial W_0^1} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \cdots, \frac{\partial z_4^1}{\partial W_0^1} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

得:

$$\frac{\partial L}{\partial W^1} = (-14 \quad 13) \cdot \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \partial z_1^1 / \partial W_0^1 \\ \vdots \\ \partial z_4^1 / \partial W_0^1 \end{pmatrix} = -27 \begin{pmatrix} x_1 & x_2 & 0 & 0 \\ x_2 & x_2 & 0 & 0 \end{pmatrix} = \begin{pmatrix} -81 & -81 & 0 & 0 \\ -378 & -378 & 0 & 0 \end{pmatrix}$$

$$\frac{\partial L}{\partial W_0^1} = -27 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -27 \\ -27 \\ 0 \\ 0 \end{pmatrix}$$

则更新后: $W^2 = \begin{pmatrix} 3.8 & -3.6 \\ 19.2 & -17.9 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$, $W_0^2 = \begin{pmatrix} 1.4 \\ 0.7 \end{pmatrix}$

$$W^1 = \begin{pmatrix} 9.2 & 8.1 & -1 & 0 \\ 37.8 & 38.8 & 0 & -1 \end{pmatrix}, \quad W_0^1 = \begin{pmatrix} 1.7 \\ 1.7 \\ -1 \\ -1 \end{pmatrix}$$