CS229 - Problem Set 2

Or Haifler

November 2023

Exercise 2

(a)

$$\frac{\partial}{\partial\theta_0}\ell(\theta) = \sum_{i=1}^{m}\frac{\partial}{\partial\theta_0}\log h_\theta(x^{(i)})^{y^{(i)}}(1-h_\theta(x^{(i)}))^{1-y^{(i)}} = y^{(i)}\sum_{i=1}^{m}\frac{\partial}{\partial\theta_0}\log h_\theta(x^{(i)}) + (1-y^{(i)})\sum_{i=1}^{m}\frac{\partial}{\partial\theta_0}\log(1-h_\theta(x^{(i)}))$$

$$= y^{(i)}\sum_{i=1}^{m}(1-h_\theta(x^{(i)}))x_0^{(i)} - (1-y^{(i)})\sum_{i=1}^{m}h_\theta(x^{(i)})x_0^{(i)} = \sum_{i=1}^{m}(y^{(i)}-h_\theta(x^{(i)}))x_0^{(i)} = \sum_{i=1}^{m}(y^{(i)}-h_\theta(x^{(i)}))$$

$$\frac{\partial}{\partial\theta_0}\ell(\theta) = 0 \implies \sum_{i=1}^{m}(y^{(i)}-h_\theta(x^{(i)})) = 0 \implies \sum_{i=1}^{m}\mathbb{I}y^{(i)} = 1 = \sum_{i=1}^{m}p(y^{(i)}=1|x^{(i)};\theta)$$

$$\implies \frac{\sum_{i=1}^{m}\mathbb{I}y^{(i)}=1}{|i\in I_{a,b}|} = \frac{\sum_{i=1}^{m}p(y^{(i)}=1|x^{(i)};\theta)}{|i\in I_{a,b}|}$$

(b)

Both statements aren't true, suppose $(a,b) = (0.5,1)$, then if the model achieves perfect score we get $\sum_{i=1}^{m}\mathbb{I}\{y^{(i)}=1\} = |\{i\in I_{a,b}\}|$, but $0.5 < p(y^{(i)}|x^{(i)};\theta) < 1$, and thus $\frac{\sum_{i=1}^{m}p(y^{(i)}=1|x^{(i)};\theta)}{|i\in I_{a,b}|} < \frac{\sum_{i=1}^{m}\mathbb{I}y^{(i)}=1}{|i\in I_{a,b}|}$, so the model isn't perfectly calibrated.

(c)

If we'll use L_2 regularization(with $\lambda\neq 0$), the loss function will be

$$J(\theta) = \sum_{i=1}^{m}y^{(i)}\log h_\theta(x^{(i)}) + \sum_{i=1}^{m}(1-y^{(i)})\log(1-h_\theta(x^{(i)})) + \frac{\lambda}{2}\|\theta\|_2^2$$

And the learned parameter will be different, because

$$\frac{\partial}{\partial\theta_0}\ell(\theta) = 0 \implies \sum_{i=1}^{m}(y^{(i)}-h_\theta(x^{(i)})) + \lambda\theta_0 = 0 \implies \sum_{i=1}^{m}1y^{(i)} = 1 = \sum_{i=1}^{m}p(y^{(i)}=1|x^{(i)};\theta) + \lambda\theta_0$$

And we got $\sum_{i=1}^{m}1y^{(i)} = 1 \neq \sum_{i=1}^{m}p(y^{(i)}=1|x^{(i)};\theta)$

Exercise 3

(a)

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

(b)

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

$$= \arg\min_{\theta}(-\log p(y|x,\theta) - \log p(\theta))$$

$$p(\theta) = \frac{1}{(2\pi)^{n/2}\eta^n} \exp\Big(-\frac{1}{2\eta^2}\theta^T\theta\Big) = \frac{1}{(2\pi)^{n/2}\eta^n} \exp\Big(-\frac{1}{2\eta^2}\|\theta\|_2^2\Big)$$

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

(c)

$$p(y^{(i)}|x^{(i)},\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2}(y^{(i)} - \theta^T x^{(i)})^2\Big)$$

$$p(Y|X,\theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2}(y^{(i)} - \theta^T x^{(i)})^2\Big)$$

$$\frac{1}{(2\pi)^{m/2}\sigma^m} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2\Big) = \frac{1}{(2\pi)^{m/2}\sigma^m} \exp\Big(-\frac{1}{2\sigma^2}\|Y - X\theta\|_2^2\Big)$$

$$\log p(Y|X,\theta) = \lambda - \frac{1}{2\sigma^2}\|Y - X\theta\|_2^2, \lambda = -\frac{m}{2}\log 2\pi - m\log\sigma$$

$$\theta_{MAP} = \arg\min_{\theta} \frac{1}{2\sigma^2}\|Y - X\theta\|_2^2 + \frac{1}{2\eta^2}\|\theta\|_2^2$$

$$J(\theta) = \frac{1}{2\sigma^2}\|Y - X\theta\|_2^2 + \frac{1}{2\eta^2}\|\theta\|_2^2, \nabla_\theta J(\theta) = \frac{1}{\sigma^2}(\|X\|_2^2\theta - X^T Y) + \frac{1}{\eta^2}\theta = 0$$

$$\theta_{MAP} = \arg\min_{\theta} J(\theta) = (\|X\|_2^2 + \frac{\sigma^2}{\eta^2}I)^{-1}X^T Y$$

(d)

$$p(\theta) = \frac{1}{(2b)^n} \exp\Big(-\frac{1}{b}\|\theta\|_1\Big), \log p(\theta) = -n\log 2b - \frac{1}{b}\|\theta\|_1$$

$$\theta_{MAP} = \arg\min_{\theta} \frac{1}{2\sigma^2}\|Y - X\theta\|_2^2 - \log p(\theta) = \arg\min_{\theta} \frac{1}{2\sigma^2}\|Y - X\theta\|_2^2 + \frac{1}{b}\|\theta\|_1$$

$$J(\theta) = \|Y - X\theta\|_2^2 + \lambda\|\theta\|_1, \theta_{MAP} = \arg\min_{\theta} J(\theta), \lambda = \frac{2\sigma^2}{b}$$

Exercise 4

(a)

K is positive semidefinite, because $\langle Kz, z\rangle = \langle (K_1 + K_2)z, z\rangle = \langle K_1 z, z\rangle + \langle K_2 z, z\rangle \geq 0$

(b)

K doesn't have to be a kernel, for $K_1 \succeq 0, K_2 = 2K_1 \succeq 0$ we got $K = -K_1 \preceq 0$

(c)

K is necessarily a kernel, $\langle Kz, z\rangle = \langle aK_1 z, z\rangle = a\langle K_1 z, z\rangle \geq 0$

(d)

K isn't a kernel, because $\langle Kz, z\rangle = \langle aK_1 z, z\rangle = a\langle K_1 z, z\rangle \leq 0$

(e)

We know that there are ϕ, ψ such that $K_1(x, z) = \phi(x)^T \phi(z), K_2(x, z) = \psi(x)^T \psi(z)$, so

$$\langle Kz, z\rangle = z^T K z = \sum_i \sum_j z_i K_{ij} z_j = \sum_i \sum_j z_i K_1(x^{(i)}, x^{(j)}) K_2(x^{(i)}, x^{(j)}) z_j$$

$$\sum_i \sum_j z_i \phi(x^{(i)})^T \phi(x^{(j)}) \psi(x^{(i)})^T \psi(x^{(j)}) z_j = \sum_i \sum_j z_i z_j \sum_k \phi(x^{(i)})_k \phi(x^{(j)})_k \sum_l \psi(x^{(i)})_l \psi(x^{(j)})_l$$

$$= \sum_i \sum_j \sum_k \sum_l z_i z_j \phi(x^{(i)})_k \phi(x^{(j)})_k \psi(x^{(i)})_l \psi(x^{(j)})_l = \sum_k \sum_l \sum_i [z_i \phi(x^{(i)})_k \psi(x^{(i)})_l]^2 \geq 0$$

(f)

$\langle Kz, z\rangle = \sum_{i=1}^n \sum_{j=1}^n z_i z_j f(x^(i)) f(x^(j)) = [z^T f_x]^2 \geq 0, f_x = \begin{bmatrix} f(x^{(1)}) \\ f(x^{(2)}) \\ \vdots \\ f(x^{(n)}) \end{bmatrix}$

(g)

K is a kernel, we know that $K_{3_{ij}} = K_3(x^{(i)}, x^{(j)})$ is positive semidefinite for $x^{(i)} \in \mathbb{R}^n$, and thus $K_{ij} = K_3(\phi(x^{(i)}), \phi(x^{(j)})) \succeq 0$

(h)

Using (a) and (c) and (e) we conclude that for $a, b \geq 0$, $aK_1^i + bK_1^j \succeq 0$, and hence if we generalize for multiple summands we'll get $K = \sum_{i=1}^n \alpha_i K_1^i \succeq 0$

Exercise 5

(a)

i.

$$\theta^{(0)} = \vec{0}, \theta^{(i)} = \sum_{j=1}^{i} \beta_j \phi(x^{(j)})$$

ii.

$$h_{\theta^{(i)}}(\phi(x^{(i+1)})) = g({\theta^{(i)}}^T \phi(x^{(i+1)})) = \text{sign}({\theta^{(i)}}^T \phi(x^{(i+1)})) = \text{sign}\Big(\sum_{j=1}^{i} \beta_j \phi(x^{(j)})^T \phi(x^{(i+1)})\Big)$$

$$= \text{sign}\Big(\sum_{j=1}^{i} \beta_j \langle \phi(x^{(j)}), \phi(x^{(i+1)}) \rangle\Big) = \text{sign}\Big(\sum_{j=1}^{i} \beta_j K(x^{(j)}, x^{(i+1)})\Big)$$

iii.

$$\theta^{(i+1)} := \theta^{(i)} + \alpha\Big(y^{(i+1)} - h_{\theta^{(i)}}(\phi(x^{(i+1)}))\Big)\phi(x^{(i+1)}) = \sum_{j=1}^{i} \beta_j \phi(x^{(j)}) + \alpha\Big(y^{(i+1)} - \text{sing}\Big(\sum_{j=1}^{i} \beta_j K(x^{(j)}, x^{(i+1)})\Big)\Big)\phi(x^{(i+1)})$$

$$= \sum_{j=1}^{i+1} \beta_j \phi(x^{(j)}), \beta_{i+1} = \alpha\Big(y^{(i+1)} - \text{sing}\Big(\sum_{j=1}^{i} \beta_j K(x^{(j)}, x^{(i+1)})\Big)\Big)$$

(c)

Dot kernel achieves bad accuracy relative to RBF kernel, because it doesn't perform any feature mapping(and thus the model is a linear classifier), and the dataset isn't linearly separable

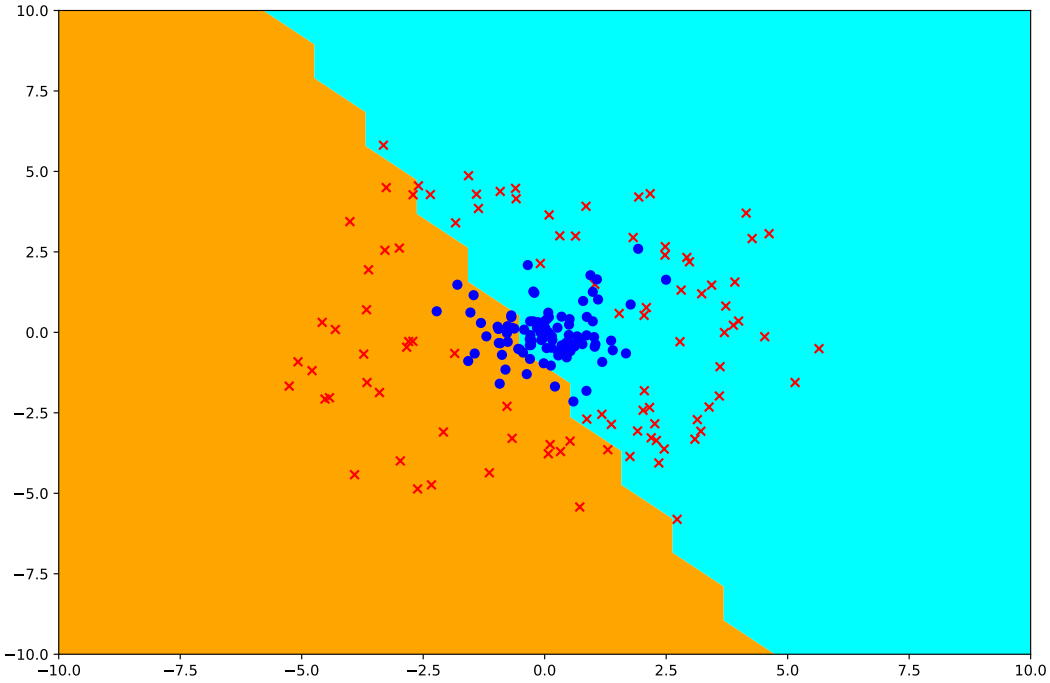

Figure 1: Dot kernel

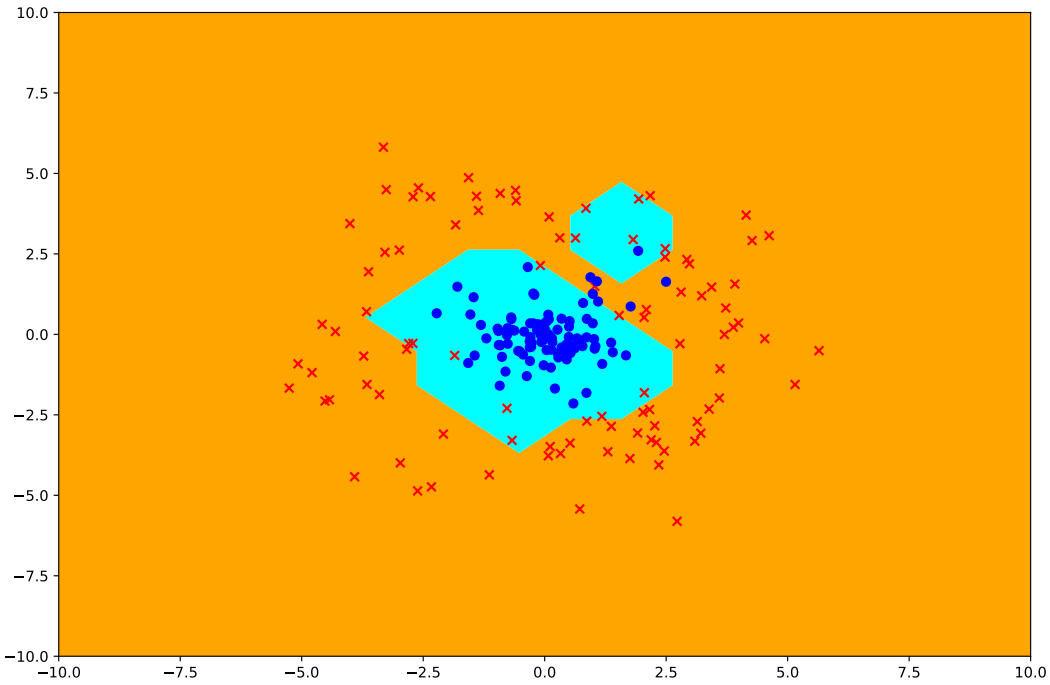

Figure 2: RBF kernel