HW 07

(a) I work on this homework with [redacted]. We orally discussed the question. A bit flared for the last question

(b) I certify that all solution are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up. Wang Zijian 8/7

HW 07

2. (a) Since $\phi_i > 0$, multiply by ϕ_i just change ξ_i's weight.

Intuitively, the corresponding unconstrained optimization problem for the SVM with custom margins is:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \phi_i \max\left(1 - y_i(w^T x_i - b), 0\right)$$

$$\text{s.t.} \quad (1-\xi_i) - y_i(w^T x_i - b) \le 0 \quad \forall i$$

$$-\xi_i \le 0 \quad \forall i$$

(b) $\therefore$ Lagrangian for the problem:

$$\mathcal{L}(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\phi_i\xi_i + \sum_{i=1}^{n}\alpha_i\left((1-\xi_i) - y_i(w^T x_i - b)\right) + \sum_{i=1}^{n}\beta_i(-\xi_i)$$

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i y_i(w^T x_i - b) + \sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}(C\phi_i - \alpha_i - \beta_i)\xi_i$$

$\therefore$ The dual is $\max_{\alpha \ge 0, \beta \ge 0} g(\alpha, \beta)$

where $g(\alpha, \beta) = \min_{w,b,\xi} \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i y_i(w^T x_i - b) + \sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}(C\phi_i - \alpha_i - \beta_i)\xi_i$

Use KKT conditions to find the optimal dual variables. Verify that the primal problem is convex in the primal variables as stated in notes 22. We know that from the stationary conditions, evaluated at the optimal dual values α^* and β^*, and the optimal primal values w^*, b^*, ξ_i^*: $\frac{\partial \mathcal{L}}{\partial w_i} = \frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial \xi_i} = 0$

$\nabla_w \mathcal{L} = w^* - \sum_{i=1}^{n}\alpha_i^* y_i x_i = 0 \Rightarrow w^* = \sum_{i=1}^{n}\alpha_i^* y_i x_i$

$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n}\alpha_i^* y_i = 0.$

$\frac{\partial \mathcal{L}}{\partial \xi_i} = C\phi_i - \alpha_i^* - \beta_i^* = 0 \Rightarrow 0 \le \alpha_i^* \le C\phi_i$, α_i^* are restricted to being less than the hyperparameter C.

Verify that other KKT also hold, indicating strong duality.

$\Rightarrow \mathcal{L}(w, b, \xi, \alpha^*, \beta^*) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i^* y_i(w^T x_i) + b\sum_{i=1}^{n}\alpha_i^* y_i + \sum_{i=1}^{n}\alpha_i^* + \sum_{i=1}^{n}(C\phi_i - \alpha_i - \beta_i)\xi_i$

$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i^* y_i(w^T x_i) + \sum_{i=1}^{n}\alpha_i^*.$

Since the primal problem is convex, from the KKT conditions we have that the optimal primal variables w^*, b^*, ξ^* minimize $\mathcal{L}(w, b, \xi, \alpha^*, \beta^*)$

$\therefore g(\alpha^*, \beta^*) = \min_{w,b,\xi} \mathcal{L}(w, b, \xi, \alpha^*, \beta^*)$

$= \mathcal{L}(w^*, b^*, \xi^*, \alpha^*, \beta^*)$

(same process in notes 22) $= \frac{1}{2}\|\sum_{i=1}^{n}\alpha_i^* y_i x_i\|^2 - \sum_{i=1}^{n}\alpha_i^* y_i\left(\left(\sum_{j=1}^{n}\alpha_j^* y_j x_j\right)^T x_i\right) + \sum_{i=1}^{n}\alpha_i^*$

$= \frac{1}{2}\left\|\sum_{i=1}^{n}\alpha_i^* y_i x_i\right\|^2 - \sum_{i=1}^{n}\left(\alpha_i^* y_i x_i^T\left(\sum_{j=1}^{n}\alpha_j^* y_j x_j\right)\right) + \sum_{i=1}^{n}\alpha_i^*$

$= \alpha^{*T}\mathbf{1} - \frac{1}{2}\alpha^{*T} Q \alpha^*.$ where $Q_{ij} = y_i(x_i^T x_j) y_j$

and $Q = (\text{diag}\, y) X X^T (\text{diag}\, y)$

∴ The dual is only in terms of α and X and y.

$$\max_{\alpha} \alpha^T \mathbf{1} - \frac{1}{2}\alpha^T Q \alpha$$

s.t. $\sum_{i=1}^{n} \alpha_i y_i = 0$

$0 \le \alpha_i \le C\phi_i \quad i = 1, \cdots n$

(c) We have $\max_{\alpha} \alpha^T \mathbf{1} \ge \frac{1}{2}\alpha^T Q \alpha$, where $Q = (\text{diag } y)\, XX^T (\text{diag } y)$

Notice that there is our desired term XX^T in Q, so we just Kernalize it - by replacing XX^T by gram matrix K, $K = XX^T$.

We found that custom margin SVM has only one difference compared to normal (the dual problem of) soft margin SVM, which is $0 \le \alpha_i \le C\phi_i$.

∴ ϕ_i here limit the value of α_i.

Adjusting ϕ_i could therefore adjust the range of α_i.

3. (a) $P(y|\theta) = P(y_1, y_2, y_3, y_4, y_5 | \theta)$

$= \frac{8!}{y_1! \cdots y_5!} \cdot p_1^{y_1} \cdots p_5^{y_5}$

$\therefore \log P(y|\theta) = \log\left(\frac{8!}{1 \cdot 1 \cdot 2! \cdot 2! \cdot 3!}\right) + y_1 \log p_1 + y_2 \log(p_2) + y_3 \log(p_3) + y_4 \log p_4 + y_5 \log p_5$

$= \log\left(\frac{8!}{24}\right) + y_1 \log\left(\frac{1}{2}\right) + y_2 \log\left(\frac{\theta}{4}\right) + 4 \log\left(\frac{1}{4}(1-\theta)\right) + 3 \log\frac{\theta}{4}$

where $y_1 + y_2 = 1$

$P(y|\theta) = \frac{8!}{24} \cdot \left(\frac{1}{2}\right)^{y_1} \cdot \left(\frac{\theta}{4}\right)^{3+y_2} \cdot \left(\frac{1-\theta}{4}\right)^4$

(b)

$q_0^{t+1} = P_\theta(Y_1 = 0, Y_2 = 1 \mid y_{sum})$

$= \frac{P_\theta(y_{sum} \mid Y_1 = 0, Y_2 = 1)\, P_\theta(Y_1 = 0, Y_2 = 1)}{P_\theta(y_{sum})}$ (Bayes' Rule)

$= \frac{1 \cdot P_\theta(Y_1 = 0, Y_2 = 1)}{P_\theta(Y_1 = 0, Y_2 = 1) + P_\theta(Y_1 = 1, Y_2 = 0)}$

$= \frac{\frac{8!}{24} \cdot \left(\frac{1}{2}\right)^0 \cdot \left(\frac{\theta}{4}\right)^4 \cdot \left(\frac{1-\theta}{4}\right)^4}{\frac{8!}{24}\left(\left(\frac{1}{2}\right)^0 \cdot \left(\frac{\theta}{4}\right)^4 \cdot \left(\frac{1-\theta}{4}\right)^4 + \left(\frac{1}{2}\right)^1 \cdot \left(\frac{\theta}{4}\right)^3 \cdot \left(\frac{1-\theta}{4}\right)^4\right)}$

$= \frac{\left(\frac{\theta^t}{4}\right)^4}{\left(\frac{\theta^t}{4}\right)^4 + \frac{1}{2} \cdot \left(\frac{\theta^t}{4}\right)^3} = \frac{\frac{\theta^t}{4}}{\frac{\theta^t}{4} + \frac{1}{2}} = \frac{\theta^t}{2 + \theta^t}$

Similarly, $q_1^{t+1} = P_\theta(Y_1 = 1, Y_2 = 0 \mid y_{sum})$

$= \frac{\left(\frac{1}{2}\right)^1 \cdot \left(\frac{\theta^t}{4}\right)^3}{\left(\frac{\theta^t}{4}\right)^4 + \frac{1}{2}\left(\frac{\theta^t}{4}\right)^3} = \frac{2}{2 + \theta^t}$

(c) $\theta^{t+1} = \arg\max_\theta E_{q^t}[\log(P_\theta(Y \mid y_{sum}, y_3, y_4, y_5))]$, denote $y_0! y_1! \cdots y_5!$ as k $= y_3! y_4! y_5! = k$

$= \arg\max_\theta \left[q_0^{t+1} \cdot \log(P_\theta(Y_0 \mid y_{sum}, y_3, y_4, y_5)) + q_1^{t+1} \log(P_\theta(Y_1 \mid y_{sum}, y_3, y_4, y_5)) \right]$

$= q_0^{t+1} \cdot \log\left(\frac{n!}{k} \cdot \left(\frac{1}{2}\right)^0 \cdot \left(\frac{\theta}{4}\right)^{y_5+1} \cdot \left(\frac{1-\theta}{4}\right)^{y_3+y_4}\right) + q_1^{t+1} \log\left(\frac{n!}{k}\left(\frac{1}{2}\right)^1 \cdot \left(\frac{\theta}{4}\right)^{y_5} \cdot \left(\frac{1-\theta}{4}\right)^{y_3+y_4}\right)$

$= q_0^{t+1}\left(\log\left(\frac{n!}{k}\right) + (y_5+1)\log\left(\frac{\theta}{4}\right) + (y_3+y_4)\log\left(\frac{1-\theta}{4}\right)\right) + q_1^{t+1}\left[\log\frac{n!}{k} + \log\frac{1}{2} + y_5 \log\left(\frac{\theta}{4}\right) + (y_3+y_4)\log\frac{1-\theta}{4}\right]$

where $k = y_3! y_4! y_5!$

(d) As derived in (c), $\theta^{t+1} = \arg\max_{\theta} \left(q_0^{t+1} + q_1^{t+1}\right)\left[\log\left(\frac{h!}{y_1!y_2!y_3!y_4!y_5!}\right) + y_5 \log\frac{\theta}{4} + (y_3+y_4)\log\left(\frac{1-\theta}{4}\right)\right] + q_0^{t+1}\log\frac{\theta}{4} + q_1^{t+1}\cdot\log\frac{1}{4}$

The function is concave, thus we can find the optimal value by setting its gradient to 0 with respect to θ.

$$\therefore \frac{\partial f}{\partial \theta} = \left(q_0^{t+1} + q_1^{t+1}\right)\left[\frac{y_5}{\theta} + \frac{y_3+y_4}{\theta-1}\right] + \frac{q_0^{t+1}}{\theta} = 0 \qquad (*)$$

Since $0<\theta<1$, mutiply both sides by $\theta\cdot(\theta-1)$

We have: $(y_5(\theta-1) + (y_3+y_4)\theta)(q_0^{t+1}+q_1^{t+1}) + q_0^{t+1}(\theta-1) = 0$

$$(\theta(y_3+y_4+y_5) - y_5)(q_0^{t+1}+q_1^{t+1}) + \theta q_0^{t+1} - q_0^{t+1} = 0$$

Simplify it: $\theta q_0^{t+1} + \theta(y_3+y_4+y_5)(q_0^{t+1}+q_1^{t+1}) - y_5(q_0^{t+1}+q_1^{t+1}) - q_0^{t+1} = 0$

$$\Rightarrow \theta^{t+1} = \frac{q_0^{t+1} + y_5(q_0^{t+1}+q_1^{t+1})}{q_0^{t+1} + (y_3+y_4+y_5)(q_0^{t+1}+q_1^{t+1})}$$

(e) $q_0^{t+1} = \frac{\theta^t}{2+\theta^t}$

$q_1^{t+1} = \frac{2}{2+\theta^t}$, since $0<\theta<1$, $\theta < 2$ is always true.

$\therefore q_1^{t+1} > q_0^{t+1}$ for all θ

i.e. $y_1=1, y_2=0$ (The probability is always higher)

(f) $\theta = \arg\max_{\theta} P(y_1, \cdots, y_5 \mid \theta, Y_1=0, Y_2=1) \cdot P(y_1, \cdots, y_5 \mid \theta, Y_1=1, Y_2=0)$

$= \arg\max_{\theta} \log P(y \mid \theta, Y_1=0, Y_2=1) + \log P(y \mid \theta, Y_1=1, Y_2=0)$

$= \arg\min_{\theta} \log \frac{1}{P(y \mid \theta, Y_1=0, Y_2=1)} + \log \frac{1}{P(y \mid \theta, Y_1=1, Y_2=0)}$

where $P(y \mid \theta, Y_1=0, Y_2=1)$ can be calculated by using the result of part (a).

It's an convex function, so we can find the optimal (minimum) θ.

After obtaining θ, we substitute it back to $P(y \mid \theta, Y_1=0, Y_2=1)$
and $P(y \mid \theta, Y_1=1, Y_2=0)$

Pick the one with higher probability and the corresponding Y_1, Y_2 will be selected.

4. (a) Step 1: Initialize p_a, p_b with arbitrary value in range (0,1).

Step 2: In t stage, for $\forall i$, compare $\begin{cases} \binom{5}{h_i}(p_a^t)^{h_i}(1-p_a^t)^{5-h_i} & \cdots P(1) \\ \binom{5}{h_i}(p_b^t)^{h_i}(1-p_b^t)^{5-h_i} & \cdots P(2) \end{cases}$

Compare these two ~~possibility~~ probability. Assign h_i to class a if $P(1)$ is higher, assign the data to class b if $P(2)$ is higher.
(Therefore, data are classfied into 2 groups)

Step 3: calculate p_a^{t+1} and p_b^{t+1}

$$p_a^{t+1} = \frac{\sum \frac{h_i}{5} \cdot (\text{average of probability of points in a})}{\text{number of data in class a}}$$

$$p_b^{t+1} = \frac{\sum \frac{h_j}{5}}{\text{number of data in class b}} \quad (h_j \in \text{class b})$$

Repeat step 2, 3 until p_a, p_b converge.

$\therefore \hat{p}_a = p_a^{t+1}$, $\hat{p}_b = p_b^{t+1}$, $\alpha_a = \frac{\text{number of data in class a}}{\text{number of all data}}$

(b) No.

Try When $n \to \infty$, the number of heads will be their expected value

$P(H=0) = \alpha_a \cdot \binom{5}{0} \cdot p_a^0 \cdot (1-p_a)^5 + \alpha_b \cdot \binom{5}{0} \cdot p_b^0 \cdot (1-p_b)^5$

$= 0.5 [\ 0.6^5 + 0.4^5\ ]$

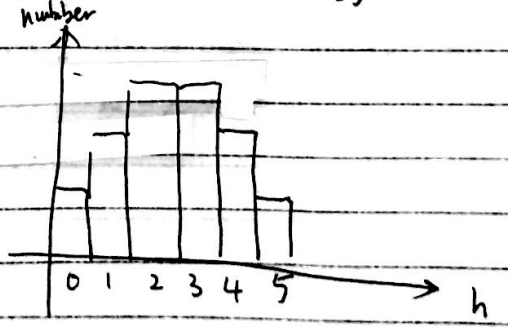
$= \frac{11}{250}$

$P(H=5) = 0.5\ [0.4^5 + 0.6^5] = \frac{11}{250}$

$P(H=1) = P(H=4) = 0.5 \cdot [\ 5 \cdot (0.4 \times 0.6^4 + 0.6 \times 0.4^4)\ ] = \frac{21}{125}$

$P(H=2) = P(H=3) = 0.5 \cdot [\ 10\ (0.4^2 \times 0.6^3 + 0.6^2 \times 0.4^3)\ ] = \frac{36}{125}$

Thus,
- $h=0 \to \frac{11}{250} n$ data
- $h=1 \to \frac{21}{125} n$ data
- $h=2 \to \frac{36}{125} n$ data
- $h=3 \to \frac{36}{125} n$ data
- $h=4 \to \frac{21}{125} n$ data
- $h=5 \to \frac{11}{250} n$ data.



(a) Initial $P_a^0 = 0.1$, $P_b^0 = 0.9$

$\therefore$ $h = 0, 1, 2$ will be assigned to class a (less head)

$h = 3, 4, 5$ will be assigned to class b (more head)

$$P_a^1 = \frac{\frac{11}{250}\cdot\frac{5}{5} + \frac{21}{125}\cdot\frac{1}{5} + \frac{36}{125}\cdot\frac{2}{5}}{\frac{1}{2}} = 0.2976$$

$$P_b^1 = \frac{\frac{11}{250}\cdot\frac{5}{5} + \frac{21}{125}\cdot\frac{4}{5} + \frac{36}{125}\cdot\frac{3}{5}}{\frac{1}{2}} = 0.7024$$

In second iteration, $h = 0, 1, 2$ will still be assigned to class a

$h = 3, 4, 5$ will still be assigned to class b

$\therefore$ $P_a^2 = 0.2976$

$P_b^2 = 0.7024$, they converge but not equal to 0.4 and 0.6.

$\therefore$ "k-means" based estimates will not give the correct estimates, No matter what initial values we assign to P_a and P_b.

(b)

$$P(H=0) = \alpha_a \cdot \binom{5}{0} \cdot P_a^0 \cdot (1-P_a)^5 + \alpha_b \cdot \binom{5}{0} P_b^0 \cdot (1-P_b)^5$$

$$= 0.5\left[0.6^5 + 0.4^5\right]$$

$$= \frac{11}{250}$$

$$P(H=5) = 0.5\left[0.4^5 + 0.6^5\right] = \frac{11}{250}$$

$$P(H=1) = P(H=4) = 0.5\cdot\left[5\cdot(0.4\times0.6^4 + 0.6\times0.4^4)\right] = \frac{132}{625}$$

$$P(H=2) = P(H=3) = 0.5\cdot\left[10\cdot(0.6^3\times0.4^2 + 0.4^3\times0.6^2)\right] = \frac{132}{625}$$

Thus, there are $\frac{11}{250}\cdot N$ data that $N = 0$ or 5

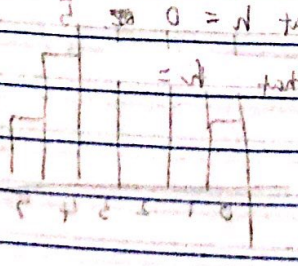
$N = 1 \rightarrow \frac{132}{625} N$ data

$N = 2 \rightarrow \frac{132}{625} N$ data

$N = 3 \rightarrow \frac{132}{625} N$ data

$N = 4 \rightarrow \frac{132}{625} N$ data

$N = 5 \rightarrow \frac{11}{250} N$ data.



(c) (0) initialize p_a^0, p_b^0, α_a^0

(i) E-step.

$$p^{t+1}(z_i = k \mid h_i) \qquad (z_i \text{ is the label for } h_i,\ k = a, b)$$

$$= \frac{\alpha_k^t\, p(h_i \mid z_i = k)}{\alpha_a^t p(h_i \mid z_i = a) + \alpha_b^t p(h_i \mid z_i = b)} \qquad \text{(using the result from notes)}$$

$$\therefore\ p^{t+1}(z_i = a \mid h_i) = \frac{\alpha_a^t\, p(h_i \mid z_i = a)}{\alpha_a^t p(h_i \mid z_i = a) + \alpha_b^t p(h_i \mid z_i = b)}, \quad \alpha_b^t = 1 - \alpha_a^t$$

$$= \frac{\alpha_a^t \binom{L}{h_i} (p_a^t)^{h_i} (1 - p_a^t)^{L - h_i}}{\alpha_a^t \binom{L}{h_i} (p_a^t)^{h_i} (1 - p_a^t)^{L-h_i} + \alpha_b^t \binom{L}{h_i} (p_b^t)^{h_i} (1 - p_b^t)^{L-h_i}}$$

$$p^{t+1}(z_i = b \mid h_i) = \frac{\alpha_b^t \binom{L}{h_i} (p_b^t)^{h_i} (1 - p_b^t)^{L - h_i}}{\alpha_a^t \binom{L}{h_i} (p_a^t)^{h_i} (1 - p_a^t)^{L-h_i} + \alpha_b^t \binom{L}{h_i} (p_b^t)^{h_i} (1 - p_b^t)^{L-h_i}}$$

(ii) M step. - denote $p^{t+1}(z_i = a \mid h_i) = q_{i|a}^{t+1}$

$$p^{t+1}(z_i = b \mid h_i) = q_{i|b}^{t+1}$$

~~The expected complete log likelihood is~~

$$\mathbb{E}_{q^{t+1}}[\ell_c(h, z)] = \sum_{i=1}^{n} \sum_{k=a,b} q_{i|k}^{t+1} \log$$

Let θ denote $\{\alpha_a, p_a, p_b\}$ (all data).

Recall the log likelihood $\ell(h) = \sum_{i=1}^{n} \log \left(\sum_{k=a,b} p(h_i \mid z_i = k)\, p(z_i = k) \right)$

The expected complete log likelihood is:

$$\arg\max_{\hat{p}_a, \hat{p}_b, \hat{\alpha}_a} \left(\qquad \right)$$

E step: give the probabilities of a coin belongs to class a and class b.

M step: Given observation data, find the optimal p_a, p_b, α_a that maximize the probability of a coin belongs to class a / b.