Solution to CS229 Problem Set 2

Son Nguyen

6/1/2020

Problem 1

(a)

The algorithm converges on data set A but does not seem to converge on data set B

(b)

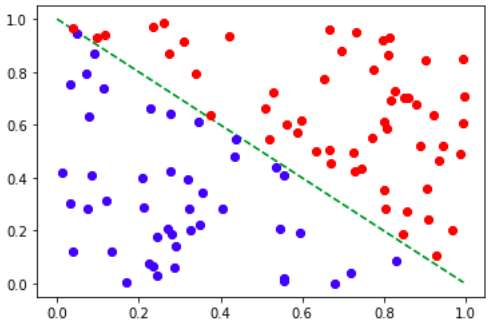

Figure 1: Data set B after 10000 iterations

From the above plot, the data set B is linearly separable. Hence, the actual maximum likelihood $L(\theta) = \prod p(Y = y|x;\theta)$ is equal to 1. However, since we choose to model the data set using the formula

$$p(Y = y|x;\theta) = \frac{1}{1 + e^{-yx^T\theta}}$$

our algorithm will try to maximize the likelihood by keep increasing θ in order to bring the probability closer to 1. However, this can be achieved only when θ approaches ∞. Hence, the algorithm never converges.

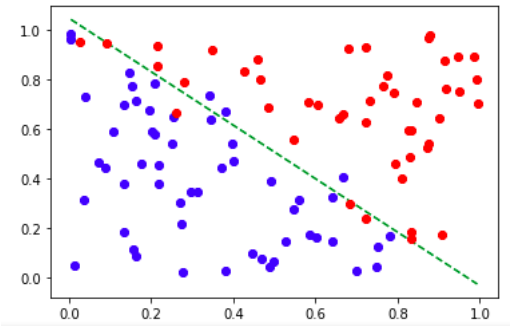

Figure 2: Data set A after 30000 iterations

In contrast, the data set A is not linearly separable, hence the possible maximum likelihood is smaller than 1. Hence, the algorithm will converge to some real values of θ.

(c)

(i) Using a different constant learning rate would not lead the logistic regression converging on linearly separable data set. Since the solution of θ is ∞, the algorithm will try to increase magnitude of θ. Changing the constant learning rate would only increase or decrease the speed of the algorithm but do not change the "direction" of the algorithm towards larger magnitude of θ.

(ii) Decreasing the learning rate over time will make the algorithm converges on data set B, as when the learning rate is small enough, the change in θ by a gradient descent step is infinitesimal.

Below are the figures of data set A and B final decision boundary, using the learning rate α scaling rule with decay constant $\gamma = 0.001$ at i-th iteration:

$$\alpha = \alpha * \frac{1}{1 + \gamma i}$$

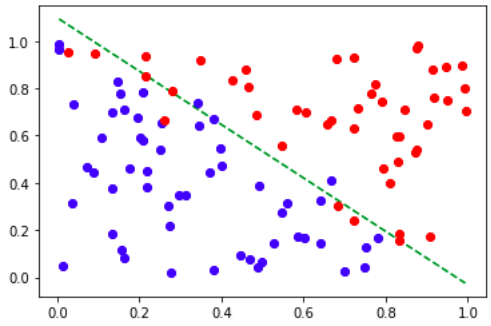

Figure 3: Data set A after 266 iterations

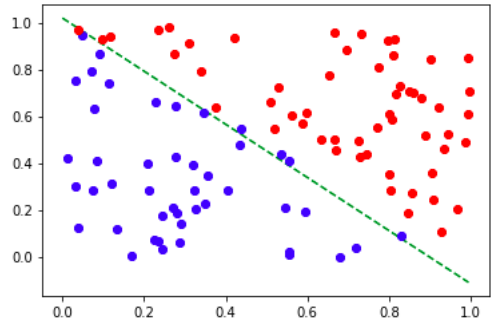

Figure 4: Data set B after 266 iterations

(iii)

Adding a regularization term $||\theta||_2^2$ to the loss function does make the algorithm converges on linearly separable data set such as B. The regularized loss with regularizing parameter λ would be come

$$J_{reg}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + exp(-y^{(i)} \theta^T x^{(i)})\right) + \frac{\lambda}{2} ||\theta||_2^2$$

The gradient of $J_{reg}(\theta)$ with respect to θ is

$$\nabla_\theta J_{reg}(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{1+exp(-y^{(i)}\theta^T x^{(i)})}\left(-(x^{(i)})^T y^{(i)}\right)\right) + \lambda\theta$$

Hence, the new gradient descent update rule with learning rate α is

$$\begin{aligned}\theta^{(t+1)} &= \theta^{(t)} - \alpha\nabla_\theta J_{reg}(\theta) \\ &= \theta^{(t)} - \alpha\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{1+exp(-y^{(i)}\theta^T x^{(i)})}\left(-(x^{(i)})^T y^{(i)}\right) + \lambda\theta\right) \\ &= \theta^{(t)} + \alpha\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{1+exp(-y^{(i)}\theta^T x^{(i)})}\left((x^{(i)})^T y^{(i)}\right)\right) - \alpha\lambda\theta\end{aligned}$$

The new gradient descent rule tends to reduce the magnitude of θ when it becomes to large. Below are the figures of the logistic regression converges on data set A and B with $\lambda = 1$:

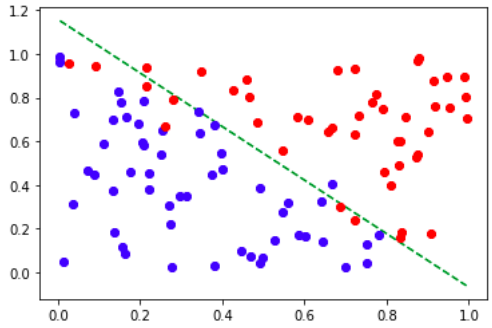

Figure 5: Data set A after 16042 iterations

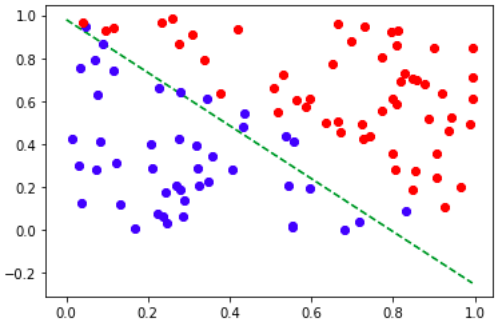

Figure 6: Data set B after 15791 iterations

(iv)

Linear scaling of the input features would not lead to the logistic regression converges on linearly separable data set. Feature scaling only results in the relative "balance" in magnitude between θ's, but the magnitude of the θ's can still be large.

(v)

Adding a little noise to the linearly separable data set would result in the logistic regression being able to converge. The zero-mean Gaussian noise added directly to the training data will make the initial data set no longer linearly separable, as the original data is shifted by a small distance randomly. Also, one should only add a little noise, i.e. limit the variance of the Gaussian noise to around 0.1, to prevent the data set being deviated too far from the original.

Below are the decision boundary of data set A and B after adding zero-mean Gaussian noise with variance 0.1:

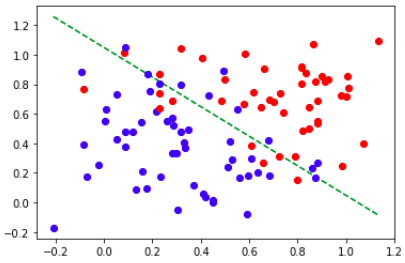

Figure 7: Data set A after 8056 iterations

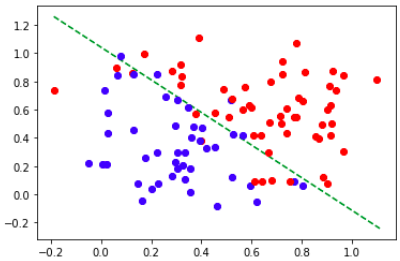

Figure 8: Data set B after 25245 iterations

(d)

Since support vector machines uses hinge loss

$$\begin{aligned}\phi_{hinge}(z) &= max(1-z,0)\\ &= max(1-yx^T\theta,0)\end{aligned}$$

when $yx^T\theta$ becomes larger than 1, the loss will become 0 and gradient descent will not update θ as $\dfrac{\partial}{\partial z}\phi_{hinge}(z)$ is 0.

Problem 2

(a)

The hypothesis function of logistic regression is

$$h_\theta(x^{(i)}) = \frac{1}{1+e^{\theta^T x^{(i)}}}$$

From above equation, $h_\theta(x^{(i)}) \in (0,1)$ for all $x^{(i)} \in \mathbb{R}^{n+1}$, $\theta \in \mathbb{R}^{n+1}$. Therefore,

$$\begin{aligned}I_{0,1} &= \{i | i \in \{1,...,m\}, h_\theta(x^{(i)}) \in (0,1)\}\\ &= \{1,...,m\}\end{aligned}$$

So, $|I_{0,1}| = m$.
Denote $l(\theta)$ as the log-likelihood

$$l(\theta) = \sum_{i=1}^{m}\left(y^{(i)}\log\left(h_\theta(x)\right) + (1-y^{(i)})(1-\log\left(h_\theta(x)\right))\right)$$

Let $j \in \{0,2,...,n\}$. Note that:

$$\begin{aligned}\frac{\partial}{\partial\theta_j}l(\theta) &= \sum_{i=1}^{m}\left[y^{(i)}\frac{\partial}{\partial\theta_j}\left(-\log\left(1+e^{-\theta^T x^{(i)}}\right)\right) + (1-y^{(i)})\frac{\partial}{\partial\theta_j}\left(-\theta^T x^{(i)} - \log\left(1+e^{-\theta^t x^{(i)}}\right)\right)\right]\\ &= \sum_{i=1}^{m}\left[y^{(i)}x_j^{(i)} - x_j^{(i)} - \frac{\partial}{\partial\theta_j}\log\left(1+e^{\theta^T x^{(i)}}\right)\right]\\ &= \sum_{i=1}^{m}\left[y^{(i)}x_j^{(i)} - x_j^{(i)} - \frac{x_j^{(i)}e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right]\\ &= \sum_{i=1}^{m}\left[(y^{(i)}-1)x_j^{(i)} - \frac{x_j^{(i)}e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right]\end{aligned}$$

Since θ is the maximum likelihood parameter, $\dfrac{\partial}{\partial\theta_j}l(\theta)$ must be 0 for all j. Consider the bias term, i.e. $j=0$:

$$\frac{\partial}{\partial\theta_0} = \sum_{i=1}^{m}\left((y^{(i)}-1) + \frac{e^{\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right) = 0 \qquad (1)$$

Hence,

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{1+e^{-\theta^T x^{(i)}}} = \frac{1}{m}\sum_{i=1}^{m} y^{(i)}$$

which is equivalent to

$$\frac{\sum_{i\in I_{0,1}} P(y^{(i)}=1|x^{(i)};\theta)}{|\{i\in I_{0,1}\}|} = \frac{\sum_{i\in I_{0,1}} \mathbf{1}\{y^{(i)}=1\}}{|\{i\in I_{0,1}\}|}(Q.E.D)$$

(may need to prove convexity of $l(\theta)$)

(b)

A perfectly calibrated binary classification model is not necessarily perfectly accurate, since being well-calibrated only means that the average of $h_\theta(x^{(i)})$ is equal to the percentage of positive examples but does not mean that the predicted number of positives is exactly equal to the number of positives from ground truth.

For example, consider a set of 5 examples, in which 3 examples are positive. One possible set of corresponding hypotheses given by the model is $(0.6, 0.6, 0.6, 0.3, 0.9)$. The average of all hypotheses is

$$h_\theta(x^{(i)}) = \frac{0.6+0.6+0.6+0.3+0.9}{5} = \frac{3}{5}$$

but the model is not perfectly accurate since it gives out four 4 labels instead of 3 (an example is positive if $h_\theta(x^{(i)}) \geq 0.5$).

The converse is also not necessarily true. Even when the average of all $h_\theta(x^{(i)}$ is equal to the percentage of positives, it does not mean that the model will label the individual examples accurately.

Continuing from the above examples, one possible set of corresponding hypotheses is $(0.7, 0.2, 0.9, 0.7, 0.1, 0.1)$. The average of all hypotheses is

$$h_\theta(x^{(i)}) = \frac{0.7+0.2+0.9+0.7+0.1}{5} = 0.52$$

which is different from $\frac{3}{5}$.

(c)

When including L_2 regularization with parameter $\lambda \geq 0$ in logistic regression, the objective to maximize with respect to θ is

$$l_r(\theta) = l(\theta) - \frac{\lambda}{2}||\theta||^2$$

The partial derivative w.r.t θ_0 of the above function is

$$\frac{\partial}{\partial \theta_0} l_r(\theta) = \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right) - \lambda \theta_0$$

Since θ is the maximum likelihood parameter,

$$\frac{\partial}{\partial \theta_0} l_r(\theta) = \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right) - \lambda \theta_0 = 0$$

Hence,

$$\frac{\left(\sum_{i \in I_{0,1}} h_\theta(x^{(i)}) \right) + \lambda \theta_0}{|\{i \in I_{0,1}\}|} = \frac{\sum_{i \in I_{0,1}} \mathbf{1}\{y^{(i)} = 1\}}{|\{i \in I_{0,1}\}|}$$

From the above equation, the logistic regression is only perfectly calibrated if $\lambda = 0$ (i.e. the L_2 regularization term is not included in objective), or the parameter of bias term θ_0 is equal to 0.

Problem 3

The maximum likelihood estimate θ_{ML} of θ is

$$\begin{aligned}
\theta_{ML} &= \arg\max_\theta \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) \\
&= \arg\max_\theta \log \left(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) \right) \\
&= \arg\max_\theta \sum_{i=1}^{m} y^{(i)} \log \left(h_\theta(x^{(i)}) \right) + (1 - y^{(i)}) \log \left(1 - h_\theta(x^{(i)}) \right) \\
&= \arg\max_\theta (l(\theta))
\end{aligned}$$

Since $\theta \sim N(0, \tau^2 I)$,

$$\begin{aligned}
\log \left(p(\theta) \right) &= \log \left(\frac{1}{(2\pi)^{(n+1)/2} |\tau^2 I|^{1/2}} \exp\left(-\frac{1}{2} \theta^T (\tau^2 I) \theta \right) \right) \\
&= -\log \left(2\pi \right)^{(n+1)/2} - \log \left(\tau^{2(n+1)1/2} \right) - \frac{\tau^2}{2} \sum_{j=0}^{n} \theta_j^2 \\
&= -\frac{n+1}{2} \log \left(2\pi \right) - (n+1) \log \tau - \frac{\tau^2}{2} \sum_{j=0}^{n} \theta_j^2
\end{aligned}$$

(note that if A is a n-by-n matrix and $k \in \mathbb{R}$, $|kA| = k^n |A|$)

The maximum-a-posteriori estimate θ_{MAP} of θ is

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} p(\theta) \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta) \\ &= \arg\max_{\theta} \log\left(p(\theta)\right) \sum_{i=1}^{m} y^{(i)} \log\left(h_\theta(x^{(i)})\right) + (1 - y^{(i)}) \log\left(1 - h_\theta(x^{(i)})\right)\end{aligned}$$

Since n and τ are constant that are independent of θ,

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} -\frac{1}{2}\sum_{j=0}^{n} \theta_j^2 + \sum_{i=1}^{m} y^{(i)} \log\left(h_\theta(x^{(i)})\right) + (1 - y^{(i)}) \log\left(1 - h_\theta(x^{(i)})\right) \\ &= \arg\max_{\theta}(l(\theta) - \frac{1}{2}\sum_{j=0}^{n} \theta_j^2) \\ &= \arg\max_{\theta}(l(\theta) - \frac{1}{2}||\theta||_2^2)\end{aligned}$$

Assume that $||\theta_{MAP}||_2 > ||\theta_{ML}||_2$ (1).

Since $\theta_{ML} = \arg\max_{\theta}\left(l(\theta)\right)$, $l(\theta_{ML}) \geq l(\theta_{MAP})$ (2).

From (1) and (2)

$$l(\theta_{MAP}) - \frac{1}{2}||\theta_{MAP}||_2^2 < l(\theta_{ML}) - \frac{1}{2}||\theta_{ML}||_2^2$$

which raises a contradiction because $\theta_{MAP} = \arg\max_{\theta}\left(l(\theta) - \frac{1}{2}||\theta||_2^2\right)$.

Therefore, $||\theta_{MAP}||_2 \leq ||\theta_{ML}||_2$.

Problem 4

(a) $K(x, z)$ is a valid kernel

Note that since K_1 and K_2 are symmetric matrices,

$$\begin{aligned}K_{ij} &= (K_1 + K_2)_{ij} \\ &= (K_1)_{ij} + (K_2)_{ij} \\ &= (K_1)_{ji} + (K_2)_{ji} \\ &= K_{ji}\end{aligned}$$

Hence K is a symmetric matrix.

Also, since $K(x,z) = K_1(x,z) + K_2(x,z),\ K = K_1 + K_2$. Therefore, for all $p \in \mathbb{R}^m$

$$\begin{aligned} p^T K p &= p^T (K_1 + K_2) p \\ &= p^T (K_1) p + p^T (K_2) p \\ &\geq 0 \end{aligned}$$

since K_1, K_2 are positive semi-definite matrices.

(b)

...

(c) $K(x,z)$ is a valid kernel

Since K_1 is a symmetric matrix,

$$\begin{aligned} K_{ij} &= a(K_1)_{ij} \\ &= a(K_1)_{ji} \\ &= K_{ji} \end{aligned}$$

Hence K is a symmetric matrix.

Also, since $K(x,z) = aK_1(x,z)$, $K = a * K_1$. Therefore, for all $p \in \mathbb{R}^m$

$$\begin{aligned} p^T K p &= p^T (aK_1) p \\ &= a * (p^T K_1 p) \\ &\geq 0 \end{aligned}$$

because K_1 is a positive semi-definite matrix.

(d)

...

(e) $K(x,z)$ is a valid kernel

Since $K(x,z) = K_1(x,z)K_2(x,z)$, $K_{ij} = (K_1)_{ij}(K_2)_{ij}$.

Since $K_1(x,z)$ and $K_2(x,z)$ are kernels, denote $(K_1)_{ij} = K_1(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$ and $(K_2)_{ij} = K_2(x^{(i)}, x^{(j)}) = \omega(x^{(i)})^T \omega(x^{(j)})$. Also, let ϕ_{k1} and ω_{k2} be the $k1$-th and $k2$-th entry of corresponding vectors.

For all $p \in \mathbb{R}^m$,

$$\begin{aligned}
p^T K p &= \sum_i \sum_j p_i K_{ij} p_j \\
&= \sum_i \sum_j p_i p_j (K_1)_{ij} (K_2)_{ij} \\
&= \sum_i \sum_j p_i p_j \left[\phi(x^{(i)})^T \phi(x^{(j)})\right] \left[\omega(x^{(i)})^T \omega(x^{(j)})\right] \\
&= \sum_i \sum_j p_i p_j \left(\sum_{k1} \phi_{k1}(x^{(i)}) \phi_{k1}(x^{(j)})\right) \left(\sum_{k2} \omega_{k2}(x^{(i)}) \omega_{k2}(x^{(j)})\right) \\
&= \sum_i \sum_j p_i p_j \left(\sum_{k1} \sum_{k2} \phi_{k1}(x^{(i)}) \phi_{k1}(x^{(j)}) \omega_{k2}(x^{(i)}) \omega_{k2}(x^{(j)})\right) \\
&= \sum_{k1} \sum_{k2} \sum_i \sum_j p_i p_j \phi_{k1}(x^{(i)}) \phi_{k1}(x^{(j)}) \omega_{k2}(x^{(i)}) \omega_{k2}(x^{(j)}) \\
&= \sum_{k1} \sum_{k2} \sum_i \sum_j \left(p_i \phi_{k1}(x^{(i)}) \omega_{k2}(x^{(i)})\right) \left(p_j \phi_{k1}(x^{(j)}) \omega_{k2}(x^{(j)})\right) \\
&= \sum_{k1} \sum_{k2} \sum_i \left(p_i \phi_{k1}(x^{(i)}) \omega_{k2}(x^{(i)})\right)^2 \\
&\geq 0
\end{aligned}$$

Hence K is a positive semi-definite matrix.

Also, since K_1 and K_2 are symmetric matrices,

$$\begin{aligned}
K_{ij} &= (K_1)_{ij} (K_2)_{ij} \\
&= (K_1)_{ji} (K_2)_{ji} \\
&= K_{ji}
\end{aligned}$$

Therefore, K is a symmetric matrix.

(f) $K(x, z)$ is a valid kernel

Note that

$$\begin{aligned}
K_{ij} &= f(x^{(i)}) f(x^{(j)}) \\
&= f(x^{(j)}) f(x^{(i)}) \\
&= K_{ji}
\end{aligned}$$

Hence K is a symmetric matrix.

Furthermore, for all $p \in \mathbb{R}^m$,

$$\begin{aligned} p^T K p &= \sum_i \sum_j p_i \left(f(x^{(i)}) f(x^{(j)}) \right) p_j \\ &= \sum_i \sum_j \left(f(x^{(i)}) p_i \right) \left(f(x^{(j)}) p_j \right) \\ &= \sum_i \left(f(x^{(i)}) p_i \right)^2 \\ &\geq 0 \end{aligned}$$

Therefore, K is a positive semi-definite matrix.

(g) $K(x, z)$ is a valid kernel

Since $K_3(x, z)$ is a kernel, K_3 is a symmetric matrix. Hence K is also symmetric.

Since $K_3(x, z)$ is a valid kernel, $K_3(x, z) = \omega(x)^T \omega(z)$. Note that for all $p \in \mathbb{R}^n$,

$$\begin{aligned} p^T K_{ij} p &= p^T (K_3)_{ij} p \\ &= p^T (\omega(\phi(x^{(i)}))^T \omega(\phi(x^{(j)}))) p \\ &= \sum_i \sum_j p_i p_j \left(\sum_k \omega_k(\phi(x^{(i)})) \omega_k(\phi(x^{(j)})) \right) \\ &= \sum_k \sum_i \sum_j p_i \omega_k(\phi(x^{(i)})) p_j \omega_k(\phi(x^{(j)})) \\ &= \sum_k \sum_i \left(p_i \omega_k(\phi(x^{(i)})) \right)^2 \\ &\geq 0 \end{aligned}$$

Hence K is a positive semi-definite matrix.

(h)

...

Problem 5

(a)

Denote m as the number of training examples. An important observation (which is inspired by the representer theorem) is that $\theta^{(i)}$ can be written as a linear

combination of the feature mappings of all training examples:

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

where $\beta_j \in \mathbb{R}$, $\forall j \in \{1, 2, ..., m\}$. Therefore, we can implicitly represent $\theta^{(i)}$ by a set of real numbers $\beta_1, \beta_2, ..., \beta_m$.

Initially, all $\beta_1, ..., \beta_m$ are set to 0, hence the initial value $\theta^{(0)} = \vec{0}$ is represented by all the $\beta_1, ..., \beta_m$ being 0.

(b)

Denote $K(x^{(j)}, x^{(i)}) = \phi(x^{(j)})^T \phi(x^{(i)})$. We can compute $h_{\theta^{(i)}}(x^{(i+1)})$ using the above representation without explicitly calculating the θ:

$$\begin{aligned} h_{\theta^{(i)}}(x^{(i+1)}) &= y^{(i+1)} \frac{{\theta^{(i)}}^T \phi(x^{(i+1)})}{|{\theta^{(i)}}^T \phi(x^{(i+1)})|} \\ &= y^{(i+1)} \frac{\sum_{j=1}^{m} \beta_j \phi(x^{(j)})^T \phi(x^{(i+1)})}{|\sum_{j=1}^{m} \beta_j \phi(x^{(j)})^T \phi(x^{(i+1)})|} \\ &= y^{(i+1)} \frac{\sum_{j=1}^{m} \beta_j K(x^{(j)}, x^{(i+1)})}{|\sum_{j=1}^{m} \beta_j K(x^{(j)}, x^{(i+1)})|} \end{aligned}$$

(c)

Firstly, note that:

$$\mathbf{1}\{g({\theta^{(i)}}^T \phi(x^{(i+1)}))y^{(i+1)}\} = \frac{y^{(i+1)}}{2} \left(y^{(i+1)} + \frac{{\theta^{(i)}}^T \phi(x^{(i+1)})}{|{\theta^{(i)}}^T \phi(x^{(i+1)})|} \right)$$

Hence we can transform the update rule:

$$\begin{aligned} \theta^{(i+1)} &:= \theta^{(i)} + \alpha \mathbf{1}\{g({\theta^{(i)}}^T \phi(x^{(i+1)}))y^{(i+1)}\} y^{(i+1)} \phi(x^{(i+1)}) \\ &= \theta^{(i)} + \alpha \frac{y^{(i+1)}}{2} \left(y^{(i+1)} + \frac{{\theta^{(i)}}^T \phi(x^{(i+1)})}{|{\theta^{(i)}}^T \phi(x^{(i+1)})|} \right) y^{(i+1)} \phi(x^{(i+1)}) \end{aligned}$$

From the newly derived update rule, we have

$$\begin{aligned} \theta^{(i+1)} &:= \theta^{(i)} + \alpha \frac{y^{(i+1)}}{2} \left(y^{(i+1)} + \frac{{\theta^{(i)}}^T \phi(x^{(i+1)})}{|{\theta^{(i)}}^T \phi(x^{(i+1)})|} \right) y^{(i+1)} \phi(x^{(i+1)}) \\ &= \sum_{j=1}^{m} \beta_j \phi(x^{(j)}) + \alpha \frac{y^{(i+1)}}{2} \left(y^{(i+1)} + \frac{{\theta^{(i)}}^T \phi(x^{(i+1)})}{|{\theta^{(i)}}^T \phi(x^{(i+1)})|} \right) y^{(i+1)} \phi(x^{(i+1)}) \\ &= \sum_{j \neq i+1} \beta_j \phi(x^{(j)}) + \left(\beta_{i+1} + \alpha \frac{(y^{(i+1)})^2}{2} \left(y^{(i+1)} + \frac{{\theta^{(i)}}^T \phi(x^{(i+1)})}{|{\theta^{(i)}}^T \phi(x^{(i+1)})|} \right) \right) \phi(x^{(i+1)}) \end{aligned}$$

Therefore, the update rule for new β_{i+1} is

$$\begin{aligned}\beta_{i+1} &:= \beta_{i+1} + \alpha\frac{(y^{(i+1)})^2}{2}\left(y^{(i+1)} + \frac{\theta^{(i)^T}\phi(x^{(i+1)})}{|\theta^{(i)^T}\phi(x^{(i+1)})|}\right) \\ &= \beta_{i+1} + \alpha\frac{(y^{(i+1)})^2}{2}\left(y^{(i+1)} + \frac{\sum_{j=1}^{m}\beta_j\phi(x^{(j)})^T\phi(x^{(i+1)})}{|\sum_{j=1}^{m}\beta_j\phi(x^{(j)})^T\phi(x^{(i+1)})|}\right)\end{aligned}$$

Hence, we can derive at an update rule of $\beta_1, ...\beta_m$, which implicitly represents θ:

$$\forall i \in \{1, 2, ..., m\}, \beta_i := \beta_i + \alpha\frac{(y^{(i)})^2}{2}\left(y^{(i)} + \frac{\sum_{j=1}^{m}\beta_j K(x^{(j)}, x^{(i)})}{|\sum_{j=1}^{m}\beta_j K(x^{(j)}, x^{(i)})|}\right)$$