DANIEL WU
1895-2460-22
ddwu@usc.edu

Written Assignment 3
Summer 2020: CSCI 567

2020-07-26

1 Problem 1

1.1

$$
\begin{aligned}
&\min_{\mathbf{w}} -\sum_{i=1}^{N} y_i \mathbf{w}^T \mathbf{x}_i \qquad \text{subject to } \ \|\mathbf{w}\|^2 - 1 \leq 0 \\
\implies &\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} -\sum_{i=1}^{N} y_i \mathbf{w}^T \mathbf{x}_i \qquad \text{subject to } \ \|\mathbf{w}\|^2 - 1 \leq 0 \\
&\mathcal{L} = -\sum_{i=1}^{N} y_i \mathbf{w}^T \mathbf{x}_i + \lambda(\|\mathbf{w}\|^2 - 1) \\
&\quad = -\Bigg[\sum_{i:\mathbf{x}_i \in \mathcal{C}_1} \mathbf{w}^T \mathbf{x}_i - \sum_{j:\mathbf{x}_i \in \mathcal{C}_{-1}} \mathbf{w}^T \mathbf{x}_j\Bigg] + \lambda(\|\mathbf{w}\|^2 - 1) \\
&\quad = \sum_{j:\mathbf{x}_j \in \mathcal{C}_{-1}} \mathbf{w}^T \mathbf{x}_j - \sum_{i:\mathbf{x}_i \in \mathcal{C}_1} \mathbf{w}^T \mathbf{x}_i + \lambda(\|\mathbf{w}\|^2 - 1)
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\mathbf{w}} \mathcal{L} &= \sum_{j:\mathbf{x}_j \in \mathcal{C}_{-1}} \mathbf{x}_j - \sum_{i:\mathbf{x}_i \in \mathcal{C}_1} \mathbf{x}_i + 2\lambda \mathbf{w}^* = 0 \\
\implies \mathbf{w}^* &= \frac{1}{2\lambda}\Bigg[\sum_{i:\mathbf{x}_i \in \mathcal{C}_1} \mathbf{x}_i - \sum_{j:\mathbf{x}_j \in \mathcal{C}_{-1}} \mathbf{x}_j\Bigg] \\
\mathbf{w}^* &\propto \sum_{i:\mathbf{x}_i \in \mathcal{C}_1} \mathbf{x}_i - \sum_{j:\mathbf{x}_j \in \mathcal{C}_{-1}} \mathbf{x}_j, \quad \text{where } \lambda \geq 0
\end{aligned}
$$

1.2

Let $\mathbf{x}$ be a test sample classified as $+1$.

$$
\begin{aligned}
\mathbf{x} \in \hat{\mathcal{C}}_1 \iff \ & \mathbf{w}^T \mathbf{x} \geq 0 \\
\implies \ & \frac{1}{2\lambda}\Bigg[\sum_{i:\mathbf{x}_i \in \mathcal{C}_1} \mathbf{x}_i - \sum_{j:\mathbf{x}_j \in \mathcal{C}_{-1}} \mathbf{x}_j\Bigg]^T \mathbf{x} \geq 0 \\
\implies \ & \Bigg[\sum_{i:\mathbf{x}_i \in \mathcal{C}_1} \mathbf{x}_i - \sum_{j:\mathbf{x}_j \in \mathcal{C}_{-1}} \mathbf{x}_j\Bigg]^T \mathbf{x} \geq 0
\end{aligned}
$$

WLOG, this can also be shown for any test sample $\mathbf{x}' \in \hat{\mathcal{C}}_{-1}$

1.3

Primal:

$$\min_{\mathbf{w}} -\sum_{i=1}^{N} y_i \mathbf{w}^T \phi(\mathbf{x}_i) \qquad \text{subject to } \ \|\mathbf{w}\|^2 - 1 \leq 0$$

1.4

To obtain the dual formulation, I solve the primal lagrangian ($\mathcal{L}_p$) and substitute for the primal variable ($\mathbf{w}$).

$$\begin{aligned}
\mathcal{L}_p &= -\sum_{i=1}^{N} y_i \mathbf{w}^T \phi(\mathbf{x}_i) + \alpha(\|\mathbf{w}\|^2 - 1) \\
\nabla_w \mathcal{L}_p &= -\sum_{i=1}^{N} y_i \phi(\mathbf{x}_i) + 2\alpha \mathbf{w} = 0 \\
\implies \ \mathbf{w}^* &= \frac{1}{2\alpha} \sum_{i=1}^{N} y_i \phi(\mathbf{x}_i)
\end{aligned}$$

Dual:

$$\begin{aligned}
&\max_{\alpha} \left\{ -\sum_{i=1}^{N} y_i \left(\frac{1}{2\alpha} \sum_{i=1}^{N} y_i \phi(\mathbf{x}_i) \right)^T \phi(\mathbf{x}_i) + \alpha \left(\left\| \frac{1}{2\alpha} \sum_{i=1}^{N} y_i \phi(\mathbf{x}_i) \right\|^2 - 1 \right) \right\} \quad \text{s.t.} \quad \alpha \geq 0 \\
\iff &\max_{\alpha} \left\{ -\frac{1}{2\alpha} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \phi(\mathbf{x}_j)^T \phi(\mathbf{x}_i) + \frac{1}{4\alpha} \left\| \sum_{i=1}^{N} y_i \phi(\mathbf{x}_i) \right\|^2 - \alpha \right\} \quad \text{s.t.} \quad \alpha \geq 0 \\
\iff &\max_{\alpha} \left\{ -\frac{1}{4\alpha} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \phi(\mathbf{x}_j)^T \phi(\mathbf{x}_i) - \alpha \right\} \quad \text{s.t.} \quad \alpha \geq 0 \\
\iff &\max_{\alpha} \left\{ -\frac{1}{4\alpha} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j k(\mathbf{x}_j, \mathbf{x}_i) - \alpha \right\} \quad \text{s.t.} \quad \alpha \geq 0
\end{aligned}$$

1.5

The primal problem in 1.3 is not kernelizable because we cannot substitute in the value for optimal weights $\mathbf{w}^*$ to complete the kernel, since it is what the primal problem is searching for (but the dual in 1.4 is kernelizable). Thus, in order to solve the primal problem outright, one must compute $\phi(\mathbf{x})$. However, once you solve for $\mathbf{w}^*$, the prediction rule is kernelizable, shown as follows:

$$
\begin{aligned}
\mathbf{x} \in \hat{\mathcal{C}}_1 &\iff \mathbf{w}^{*T}\phi(\mathbf{x}) \geq 0 \\
&\implies \frac{1}{2\alpha}\sum_{i=1}^{N} y_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) \geq 0 \\
&\implies \frac{1}{2\alpha}\sum_{i=1}^{N} y_i k(\mathbf{x}_i, \mathbf{x}) \geq 0
\end{aligned}
$$

2 Problem 2

2.1

$$
\text{Linearly Seperable} \iff \exists\, w \ \ni \ \begin{aligned}&(1)\ wx_i \geq t,\ \forall i : y_i = 1, \\ &(2)\ wx_j < t,\ \forall j : y_j = -1\end{aligned}
$$

$$
\begin{aligned}
&(a)\ \{(x_3, y_3)\} = (1, 1) \implies w \geq t \\
&(b)\ \{(x_2, y_2)\} = (-1, -1) \implies -w < t \\
&(c)\ \{(x_4, y_4)\} = (3, -1) \implies 3w < t
\end{aligned}
$$

$$
(b) \text{ and } (c) \implies 2w < 2t \implies w < t \not\Longrightarrow w \geq t
$$

There is a contradiction in general form, so this data cannot be linearly separated by any hyperplane.

2.2

$$
k(x, x') = \phi(x)^T \phi(x') = xx' + \sin\left(\frac{\pi}{2}x\right)\sin\left(\frac{\pi}{2}x'\right)
$$

Consider $\mathbf{w} = [0\ \ 1],\ b = 0$

$$
\begin{aligned}
\text{sign}(\mathbf{w}^T\phi(x_1)) &= \text{sign}\left[\sin\left(\frac{-3\pi}{2}\right)\right] = 1 = y_1 \\
\text{sign}(\mathbf{w}^T\phi(x_2)) &= \text{sign}\left[\sin\left(\frac{-\pi}{2}\right)\right] = -1 = y_2 \\
\text{sign}(\mathbf{w}^T\phi(x_3)) &= \text{sign}\left[\sin\left(\frac{\pi}{2}\right)\right] = 1 = y_3 \\
\text{sign}(\mathbf{w}^T\phi(x_4)) &= \text{sign}\left[\sin\left(\frac{3\pi}{2}\right)\right] = -1 = y_4
\end{aligned}
$$

The data is separable after the transformation for this particular hyperplane.

2.3

Since the transformed data is linearly separable, we can use the hard margin SVM.

General Form Primal:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b) \geq 1, \forall n \end{aligned} \tag{1}$$

Plugging in dataset, Primal is:

$$\begin{aligned} \min_{w_1,w_2,b} \quad & \frac{1}{2}(w_1^2+w_2^2) \\ \text{s.t.} \quad & -3w_1+w_2+b \geq 1, \\ & w_1+w_2-b \geq 1, \\ & w_1+w_2+b \geq 1, \\ & -3w_1+w_2-b \geq 1 \end{aligned} \tag{2}$$

Dual:

$$\begin{aligned} \max_{\{\alpha_n\}} \quad & \sum_{n=1}^{4}\alpha_n - \frac{1}{2}\underbrace{\sum_{n=1}^{4}\sum_{m=1}^{4} y_m y_n \alpha_m \alpha_n k(\mathbf{x}_m,\mathbf{x}_n)}_{:= G \text{ (for subproblem 2.4)}} \\ \text{s.t.} \quad & \sum_{n=1}^{4} y_n\alpha_n = 0, \\ & \alpha_n \geq 0, \forall\, n \in \{1,2,3,4\} \end{aligned} \tag{3}$$

2.4

$$\begin{aligned} G = \; & y_1^2\alpha_1^2\left[x_1^2+\sin^2\left(\frac{\pi}{2}x_1\right)\right] && + && (4) \\ & y_1y_2\alpha_1\alpha_2\left[x_1x_2+\sin\left(\frac{\pi}{2}x_1\right)\sin\left(\frac{\pi}{2}x_2\right)\right] && + && (5) \\ & \vdots && \vdots && (6) \\ & y_4y_3\alpha_4\alpha_3\left[x_4x_3+\sin\left(\frac{\pi}{2}x_4\right)\sin\left(\frac{\pi}{2}x_3\right)\right] && + && (7) \\ & y_4^2\alpha_4^2\left[x_4^2+\sin^2\left(\frac{\pi}{2}x_4\right)\right] && && (8) \end{aligned}$$

Notice between line (4) and (8), there are 16 terms. The off-diagonal terms appear twice, and the diagonal terms appear once. This is a weighted sum of elements in a covariance/-symmetric matrix, thus:

$$\begin{aligned} G &= \sum_{k=1}^{4} y_k^2\alpha_k^2\left[x_k^2+\sin^2\left(\frac{\pi}{2}x_k\right)\right] + \sum_{\substack{n,m,\\ n\neq m}}^{4} y_m y_n \alpha_m \alpha_n\left[x_n x_m + \sin\left(\frac{\pi}{2}x_n\right)\sin\left(\frac{\pi}{2}x_m\right)\right] \\ &= 10\alpha_1^2+2\alpha_2^2+2\alpha_3^2+10\alpha_4^2+2\big(-2\alpha_1\alpha_2-2\alpha_1\alpha_3+10\alpha_1\alpha_4+2\alpha_2\alpha_3-2\alpha_2\alpha_4-2\alpha_3\alpha_4\big) \end{aligned}$$

$$\alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 = 0 \qquad \text{(from dual constraint } \sum_n y_n \alpha_n = 0\text{)}$$

$$\alpha_1 = \alpha_4 \implies \alpha_2 = \alpha_3$$

$$\begin{aligned} \implies G &= 20\alpha_1^2 + 4\alpha_2^2 + 2(10\alpha_1^2 + 4\alpha_2^2 - 8\alpha_1\alpha_2) \\ &= 40\alpha_1^2 + 8\alpha_2^2 - 16\alpha_1\alpha_2 \end{aligned}$$

Therefore the dual formulation for the dataset and after substituting $\alpha_4 = \alpha_1$ and $\alpha_2 = \alpha_3$ is:

$$\begin{aligned} \max_{\alpha_1, \alpha_2} \quad & 2\alpha_1 + 2\alpha_2 - 20\alpha_1^2 - 4\alpha_2^2 + 8\alpha_1\alpha_2 \\ \text{s.t.} \quad & \alpha_1 \geq 0, \\ & \alpha_2 \geq 0 \end{aligned} \tag{9}$$

$$\mathcal{L}_d = 2\alpha_1 + 2\alpha_2 - 20\alpha_1^2 - 4\alpha_2^2 + 8\alpha_1\alpha_2$$

$$\begin{aligned} \frac{\partial \mathcal{L}_d}{\partial \alpha_1} &= 2 - 40\alpha_1 + 8\alpha_2 = 0 \\ \frac{\partial \mathcal{L}_d}{\partial \alpha_2} &= 2 + 8\alpha_1 - 8\alpha_2 = 0 \end{aligned}$$

2 eqs, 2 unknowns $\implies$

$$\boxed{\begin{gathered} \alpha_1 = 1/8, \\ \alpha_4 = 3/8, \\ \alpha_2 = 3/8, \\ \alpha_3 = 1/8 \end{gathered}}$$

All data points are support vectors.

3 Problem 3

Note: In this question, my notation is as follows: Bold $\mathbf{x}_n$ refers to the nth data point, unbold x_d refers to the dth feature in the feature space, and $x_{n,d}$ refers to the dth feature for the nth datapoint. For other variables apart from y_n, the subscript refers to the iteration number in the adaboost algorithm.

3.1

With this weak classifier, we can only partition the space with a vertical or horizontal line (along only 1 feature). Evidently, we will always mis-classify at least two points with the best line in the first iteration, so I just choose $x_1 = 0$ as my first hyperplane.

$$f_1 = \text{argmin}_{h\in\mathcal{H}} \sum_{n=1}^{4} \frac{1}{4}\mathbb{I}[y_n \neq h(\mathbf{x}_n)] = h_{\{-1,1\},0,1} = \text{sign}(x_1 - 0) = \boxed{\text{sign}(x_1)}$$

$$\epsilon_1 = \sum_{n=1}^{4} \frac{1}{4}\mathbb{I}[y_n \neq \text{sign}(x_1)] = \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 1 = \boxed{\frac{1}{2}}$$

$$\beta_1 = \frac{1}{2}\ln\left(\frac{1-\epsilon_1}{\epsilon_1}\right) = \frac{1}{2}\ln 1 = \boxed{0}$$

$$w_2 = w_1 e^{-\beta_1 y_n h_1} \implies$$

$$w_2(1) = \frac{1}{4}e^{-0} = \frac{1}{4},\ w_2(2) = \frac{1}{4}e^{0} = \frac{1}{4},\ w_2(3) = \frac{1}{4}e^{0} = \frac{1}{4},\ w_2(4) = \frac{1}{4}e^{-0} = \frac{1}{4}$$

Normalize:

$$\sum_n w_2(n) = 1 \implies \boxed{w_2(1) = \frac{1}{4},\ w_2(2) = \frac{1}{4},\ w_2(3) = \frac{1}{4},\ w_2(4) = \frac{1}{4}}$$

3.2

Again, with uniform weighting and no way to misclassify less than 2 data points with the weak classifier, I just choose $x_2 = 0$ as my second hyperplane.

$$f_2 = \text{argmin}_{h\in\mathcal{H}} \sum_{n=1}^{4} \frac{1}{4}\mathbb{I}[y_n \neq h(\mathbf{x}_n)] = h_{\{-1,1\},0,2} = \text{sign}(x_2 - 0) = \boxed{\text{sign}(x_2)}$$

$$\epsilon_2 = \sum_{n=1}^{4} \frac{1}{4}\mathbb{I}[y_n \neq \text{sign}(x_2)] = \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 0 = \boxed{\frac{1}{2}}$$

$$\beta_2 = \frac{1}{2}\ln\left(\frac{1-\epsilon_2}{\epsilon_2}\right) = \frac{1}{2}\ln 1 = \boxed{0}$$

$$w_3(1) = \frac{1}{4}e^{0} = \frac{1}{4},\ w_3(2) = \frac{1}{4}e^{-0} = \frac{1}{4},\ w_3(3) = \frac{1}{4}e^{0} = \frac{1}{4},\ w_3(4) = \frac{1}{4}e^{-0} = \frac{1}{4}$$

Normalize:

$$\sum_n w_3(n) = 1 \implies \qquad \boxed{w_3(1) = \frac{1}{4},\ w_3(2) = \frac{1}{4},\ w_3(3) = \frac{1}{4},\ w_3(4) = \frac{1}{4}}$$

No values were updated in the second iteration, so the third iteration will not be updated either. This is because the algorithm fails to assign nonzero importance to any classifier. Thus, the outcome of this adaboost algorithm will be $F(x) = \text{sign}(0)$. It will just classify every point as $+1$.

3.3

Now the XOR takes diamond formation. We can find a best weak classifier such that one is always misclassified. I just arbitrarily choose $x_1 = 1$.

$$f_1 = \text{argmin}_{h \in \mathcal{H}} \sum_{n=1}^{4} \frac{1}{4} \mathbb{I}[y_n \neq h(\mathbf{x}_n)] = h_{\{-1,1\},1,1} = \boxed{\text{sign}(x_1 - 1)}$$

$$\epsilon_1 = \sum_{n=1}^{4} \frac{1}{4} \mathbb{I}[y_n \neq \text{sign}(x_1 - 1)] = \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 = \boxed{\frac{1}{4}}$$

$$\beta_1 = \frac{1}{2} \ln\left(\frac{1 - \epsilon_1}{\epsilon_1}\right) = \boxed{\frac{1}{2} \ln 3}$$

$$w_2(1) = w_2(2) = w_2(3) = \frac{1}{4} e^{-\frac{1}{2}\ln(3)} = \frac{1}{4\sqrt{3}}$$

$$w_2(4) = \frac{1}{4} e^{\frac{1}{2}\ln(3)} = \frac{\sqrt{3}}{4}$$

Normalize:

$$w_2(1) = w_2(2) = w_2(3) = \frac{w_2(3)}{\sum_{n=1}^{4} w_2(n)} = \frac{1/4\sqrt{3}}{3/4\sqrt{3} + \sqrt{3}/4} = \boxed{\frac{1}{6}}$$

$$w_2(4) = \frac{w_2(4)}{\sum_{n=1}^{4} w_2(n)} = \frac{\sqrt{3}/4}{3/4\sqrt{3} + \sqrt{3}/4} = \boxed{\frac{1}{2}}$$

3.4

As long as I choose a classifier such that I don't misclassify data point 4 (which has the highest weight) and only misclassify one other point, I have the best weak classifier, so I choose $x_2 = -1$

$$
\begin{aligned}
f_2 &= \boxed{\text{sign}(x_2+1)} \\
\epsilon_1 &= \sum_{n=1}^{4} w_2(n)\mathbb{I}[y_n \neq \text{sign}(x_2+1)] = \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 0 + \frac{1}{6}\cdot 0 + \frac{1}{6}\cdot 0 = \boxed{\frac{1}{6}} \\
\beta_2 &= \frac{1}{2}\ln\left(\frac{1-\epsilon_2}{\epsilon_2}\right) = \boxed{\frac{1}{2}\ln 5}
\end{aligned}
$$

$$
\begin{aligned}
w_3(1) = w_2(1)e^{\beta_2} &= \frac{1}{6}e^{\ln(\sqrt{5})} = \frac{\sqrt{5}}{6} \\
w_3(2) = w_3(3) = w_2(3)e^{-\beta_2} &= \frac{1}{6}e^{\ln\frac{1}{\sqrt{5}}} = \frac{1}{6\sqrt{5}} \\
w_3(4) = w_2(4)e^{-\beta_2} &= \frac{1}{2}e^{\ln\frac{1}{\sqrt{5}}} = \frac{1}{2\sqrt{5}}
\end{aligned}
$$

Normalize:

$$
\begin{aligned}
w_3(1) &= \frac{w_3(1)}{\sum_{n=1}^{4} w_3(n)} = \frac{\sqrt{5}/6}{\sqrt{5}/3} = \boxed{\frac{1}{2}} \\
w_3(2) = w_3(3) &= \frac{w_3(3)}{\sum_{n=1}^{4} w_3(n)} = \frac{1/6\sqrt{5}}{\sqrt{5}/3} = \boxed{\frac{1}{10}} \\
w_3(4) &= \frac{w_3(4)}{\sum_{n=1}^{4} w_3(n)} = \frac{1/2\sqrt{5}}{\sqrt{5}/3} = \boxed{\frac{3}{10}}
\end{aligned}
$$

3.5

$$
\begin{aligned}
F(\mathbf{x}_n) &= \text{sign}\big(\beta_1 f_1(\mathbf{x}_n) + \beta_2 f_2(\mathbf{x}_n)\big) \\
&= \text{sign}\left(\frac{1}{2}\ln(3)\cdot\text{sign}(x_{n,1}-1) + \frac{1}{2}\ln(5)\cdot\text{sign}(x_{n,2}+1)\right)
\end{aligned}
$$