Moed A 2016/7 - Introduction to Machine Learning

February 9, 2017

1 Question 1

(a)

1. Possible. Perceptron returns a hyperplane classifier.
2. Possible. SVM returns a hyperplane classifier with maximal margin, and the one in the picture looks like it has a maximal margin.
3. Not possible. Decision tree with decision stumps only returns a classifier whose separating line is piecewise parallel to the axes, which is not the case here.

(b)

1. Possible. Perceptron returns a hyperplane classifier, not necessarily with maximal margin.
2. Not possible. SVM returns a hyperplane classifier with maximal margin, and the one in the picture definitely doesn't have maximal margin (the one (a) has a larger margin).
3. Not possible. Decision tree with decision stumps only returns a classifier whose separating line is piecewise parallel to the axes, which is not the case here.

(c)

1. Not possible. Perceptron returns a hyperplane classifier, but the separating line here is not a hyperplane.
2. Not possible. SVM returns a hyperplane classifier, but the separating line here is not a hyperplane.
3. Possible. Decision tree with decision stumps returns a classifier whose separating line is piecewise parallel to the axes, like the one here.

2 Question 2

(1)

The VC-dimension is **1**. To show this, we first note that if we denote by $\mathbf{v}^P$ the indicator variable for a subset P (that is, $\mathbf{v}_i^P = \delta_{i \in P}$), then,

$$T_P(\mathbf{x}) = 1 \text{ iff } \mathbf{x} = \mathbf{v}^P$$

This is true, since if $\mathbf{x} = \mathbf{v}^P$, then $T_P(\mathbf{x}) = 1 \cdot \ldots \cdot 1 = 1$. Conversely, if $\mathbf{x} \neq \mathbf{v}^P$, then at least one of the terms of $\Pi_{i \in P} x_i \Pi_{i \notin P}(1 - x_i)$ will be zero, so $T_P(\mathbf{x}) = 0$. Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a dataset. Its projection under a single hypothesis T_P, i.e. the vector $\langle T_P(\mathbf{x}_1), \ldots, T_P(\mathbf{x}_m) \rangle$, can have 1 only in one entry, by the reasoning above.

We now prove the claim. Let $S = \{\mathbf{0}\}$. Then, $T_\emptyset(\mathbf{0}) = 1$ and $T_P(\mathbf{0}) = 0$ for any other subset $P \neq \emptyset$, so $\Pi_{C_1}(S) = \{\langle 0 \rangle, \langle 1 \rangle\}$. This shows VC-dim$(C_1) \geq 1$.

Now, let S be a dataset $|S| > 1$. Then, S has distinct two points, which we will denote $\mathbf{x}_1$ and $\mathbf{x}_2$. By the reasoning above, there is no P for which $T_P(\mathbf{x}_1) = T_P(\mathbf{x}_2) = 1$. Thus, any vector of labels which assigns the same label for $\mathbf{x}_1$ and $\mathbf{x}_2$ will not exist in $\Pi_{C_1}(S)$; therefore, S cannot be shatter by C_1 whenever $|S| > 1$, showing VC-dim$(C_1) = 1$ exactly.

(2)

The VC-dimension is **k**. Based on the reasoning in (1), since $T_{P_1,\ldots,P_m}(\mathbf{x}) = \vee_{i=1,\ldots,k} T_{P_i}(\mathbf{x})$, we have

$$T_{P_1,\ldots,P_m}(\mathbf{x}) \text{ iff } \mathbf{x} \in \{\mathbf{v}^{P_1}, \ldots, \mathbf{v}^{P_m}\}$$

Therefore, the vector $\langle T_P(\mathbf{x}_1), \ldots, T_P(\mathbf{x}_m) \rangle$ can have 1 in up to m entries only. Let S by a dataset of size $k < d$. We will show that S can be shattered by C_k. To show that, let $y_1, \ldots, y_k$ be some labeling of S. Define the hypothesis

$$h = \vee_{\{i | y_i = 1\}} T_{P_{\mathbf{x}_i}}$$

where $P_{\mathbf{x}_i} = \{j | (\mathbf{x}_i)_j = 1\}$. The above hypothesis h is in C_k, since there are at most k such subsets. By construction, $h(\mathbf{x}_i) = y_i$. We have shown that there is an hypothesis labeling S by every possible label; thus, C_k shatters S and thus VC-dim$(C_k) \geq k$.

Conversely, let S be a dataset $|S| > k$. Then, S has $k + 1$ distinct points, which we will denote $\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}$. By the reasoning above, there is no P for which $T_P(\mathbf{x}_1) = \ldots = T_P(\mathbf{x}_{k+1}) = 1$. Thus, any vector of labels which assigns the same label for all $\mathbf{x}_i \in S$ will not exist in $\Pi_{C_1}(S)$; therefore, S cannot be shattered by C_k whenever $|S| > k$, showing VC-dim$(C_k) = k$ exactly.

3 Question 3

(1)

$$\begin{aligned}\|\mathbf{w}_{t+1}\|^2 &= \langle \mathbf{w}_{t+1}, \mathbf{w}_{t+1} \rangle \\ &= \langle \mathbf{w}_t + \eta_t y_t \mathbf{x}_t, \mathbf{w}_t + \eta_t y_t \mathbf{x}_t \rangle \\ &= \|\mathbf{w}_t\|^2 + 2\eta_t y_t \langle \mathbf{x}_t, \mathbf{w}_t \rangle + \eta_t^2 y_t^2 \|\mathbf{x}_t\|^2 \\ &\leq \|\mathbf{w}_t\|^2 + \eta_t^2\end{aligned}$$

where we used $y_t^2 = 1$, $\|\mathbf{x}_t\|^2 = 1$ (by assumption), and also, since we know there was a mistake, $y_t \langle \mathbf{x}_t, \mathbf{w}_t \rangle \leq 0$. By induction, and by $\|\mathbf{w}_1\|^2 = 0$, we have

$$\|\mathbf{w}_{t+1}\|^2 \leq \sum_{i=1}^{t} \eta_i^2$$

Note: Some students used the *false* argument $\|\mathbf{w}_{t+1}\|^2 = \|\mathbf{w}_t + \eta_t y_t \mathbf{x}_t\|^2 \leq \|\mathbf{w}_t\|^2 + \|\eta_t y_t \mathbf{x}_t\|^2 = \|\mathbf{w}_t\|^2 + \eta_t^2$. The inequality is wrong; this is *not* the triangle inequality (the triangle inequality is without squares), nor is it the Pythagoras theorem (which is only for perpendicular vectors).

(2)

$$\begin{aligned}\mathbf{w}_{t+1} \cdot \mathbf{w}^* &= \langle \mathbf{w}_t + \eta_t y_t \mathbf{x}_t, \mathbf{w}^* \rangle \\ &= \langle \mathbf{w}_t, \mathbf{w}^* \rangle + \eta_t y_t \langle \mathbf{x}_t, \mathbf{w}^* \rangle \\ &\geq \langle \mathbf{w}_t, \mathbf{w}^* \rangle + \eta_t \gamma\end{aligned}$$

where we used $y_t \langle \mathbf{x}_t, \mathbf{w}^* \rangle \geq \gamma$. By induction, and by $\|\mathbf{w}_1\|^2 = 0$, we have

$$\mathbf{w}_{t+1} \cdot \mathbf{w}^* \geq \gamma \sum_{i=1}^{t} \eta_i$$

(3)

Let M be the number of mistakes. By Cauchy-Schwarz, we have $\mathbf{w}_{M+1} \cdot \mathbf{w}^* \leq \|\mathbf{w}_{M+1}\| \cdot \|\mathbf{w}^*\| = \|\mathbf{w}_{M+1}\|$. Using the bounds from (1) and (2), and the inequalities given, we get

$$\begin{aligned}\frac{\gamma}{2} \log_2(M) \leq \gamma \sum_{i=1}^{M} \eta_i \leq \mathbf{w}_{M+1} \cdot \mathbf{w}^* \leq \|\mathbf{w}_{M+1}\| \leq \left(\sum_{i=1}^{t} \eta_i^2 \right)^{1/2} \leq \left(\frac{\pi^2}{6} \right)^{1/2} \Rightarrow \\ \frac{\gamma}{2} \log_2(M) \leq \frac{\pi}{\sqrt{6}} \Rightarrow \\ \log_2(M) \leq \frac{2\pi}{\sqrt{6}\gamma} \Rightarrow \\ M \leq 2^{\frac{2\pi}{\sqrt{6}\gamma}}\end{aligned}$$

4 Question 4

(a)

We will show that there exists a solution to the hard-SVM dual problem, showing that there exists a separator with zero training error. The SVM solution is given by $h(\mathbf{x}) = \text{sign}\left(\sum_i \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i)\right)$. Set $\alpha_i = 1$ and we get $h(\mathbf{x}_j) = \text{sign}\left(\alpha_j y_j\right) = \text{sign}(y_j) = y_j$. Note that this solution may not be the optimum for SVM; but it shows there is a separator with error zero and therefore, hard SVM will also find one.

(b)

Denote the dimension of $\boldsymbol{\phi}(\mathbf{x})$ by d. Then, since we can separate any classification of the points, it follows that the number of points must be at most the VC dimension of linear classifiers in d dimensions (without bias), which is d. So we have $n \leq d$ and therefore $d \geq n$.

Alternatively, we can say that $K = \phi(X)\phi(X)^T = I$. Therefore $\phi(X)$ consists of n independent rows and therefore its rank is at least n. Therefore it must have at least n columns.

(c)

The classifier for the training data is the sign of the vector $K_S\bar{\boldsymbol{\alpha}}$, where we define $\bar{\alpha}_i = \alpha_i y_i$. Now if we had $\boldsymbol{y} = K_S\bar{\boldsymbol{\alpha}}$, we would have perfect separation. And indeed we can solve $\bar{\boldsymbol{\alpha}} = K_S^{-1}\boldsymbol{y}$, from which we can recover an α_i. Note that this α_i might not be feasible. But still $\boldsymbol{w} = \sum_i \alpha_i y_i \phi(\mathbf{x}_i)$ is a valid linear classifier. So we have linear separation.

Alternatively, since K is invertible and PSD it can be written as $K = VDV^T$ for orthogonal V and diagonal D with strictly positive diagonal. So we can just view it as the kernel matrix of the data $X = V\sqrt{D}$, which is full rank (it is invertible, since its inverse is $\sqrt{D}^{-1}V^T$, and therefore can be linearly separated as we showed in class.

A third argument - the reasoning for (b) still holds here. That is, since $K_S = \phi(X)\phi(X)^T$ is invertible, it is full rank, and therefore $\phi(X)$ must be at least of rank n. But that means $d \leq n$, so we can solve a system of equations $\phi(X)\mathbf{w} = \boldsymbol{y}$ to get an exact separator.

5 Question 5

5.1 (a)

Probability of $X = 1$ is $0.5 \cdot 0.5 + 0.5 \cdot \theta = 0.5(0.5 + \theta)$ and probability of $X = 0$ is $1 - 0.25 - 0.5\theta = 0.75 - 0.5\theta = 0.5(1.5 - \theta)$. The likelihood is:

$$\ell(\theta) = n_0 \log 0.5(1.5 - \theta) + n_1 \log 0.5(0.5 + \theta) \tag{1}$$

Dropping constant terms we get:

$$\ell(\theta) \propto n_0 \log(1.5 - \theta) + n_1 \log(0.5 + \theta) \tag{2}$$

Take derivative wrt θ to get:

$$-\frac{n_0}{1.5 - \theta} + \frac{n_1}{0.5 + \theta} = 0 \tag{3}$$

So:

$$-0.5n_0 - n_0\theta + 1.5n_1 - n_1\theta = 0 \tag{4}$$

Or:

$$-n_0 - 2n_0\theta + 3n_1 - 2n_1\theta = 0 \tag{5}$$

Define $\hat{\theta}$ as the θ that sets this to zero.

$$\hat{\theta} = \frac{3n_1 - n_0}{2n} \tag{6}$$

If $\hat{\theta} \leq 0$ the optimum is $\theta = 0$ which can be seen since the derivative of the function is always negative in this case. Similarly if $\hat{\theta} \geq 1$ the optimum is $\theta = 1$ since the derivative is always positive in this case.

5.2 (b)

The EM function is:

$$Q(\theta) = \sum_{i=1}^{n} \sum_{z_i=0,1} p(Z_i = z_i | x_i; \theta_t) \log p(Z_i = z_i, x_i; \theta) \tag{7}$$

The parameter θ is only involved in the case where $z_i = 1$ so we can restrict attention to that case:

$$Q(\theta) = \sum_{i=1}^{n} p(Z_i = 1 | x_i; \theta_t) \log p(Z_i = 1, x_i; \theta) + (\text{constant}) \tag{8}$$

Denote:

$$r_0 = n_0 p(Z = 1 | X = 0) \quad , \quad r_1 = n_1 p(Z = 1 | X = 1) \tag{9}$$

And:

$$p(Z = 1 | X = 1) = \frac{0.5\theta}{0.5(0.5 + \theta)} = \frac{\theta}{0.5 + \theta} \tag{10}$$

and:

$$p(Z = 1 | X = 0) = \frac{0.5(1 - \theta)}{0.5(1.5 - \theta)} = \frac{1 - \theta}{1.5 - \theta} \tag{11}$$

So the objective is:

$$Q(\theta) = r_0 \log 0.5(1 - \theta) + r_1 \log 0.5\theta + (\text{constant}) \tag{12}$$

Up to constants:

$$Q(\theta) = r_0 \log(1 - \theta) + r_1 \log \theta \tag{13}$$

Which is solved at:

$$\theta_{t+1} \quad = \quad \frac{r_1}{r_0 + r_1}$$

6 Question 6

Denote by X the matrix whose rows are $\mathbf{x}_i$, and by $\mathbf{y}$ the vector of y_i. The objective function we need to minimize is

$$f(\mathbf{a}) = \|\mathbf{y} - X\mathbf{a}\|^2 + \frac{1}{2}\mathbf{a}^T M\mathbf{a}$$

Derive by $\mathbf{a}$ and equate to zero:

$$\frac{\partial f}{\partial \mathbf{a}} = -2X^T\mathbf{y} + 2X^T X\mathbf{a} + M\mathbf{a} = \mathbf{0} \Rightarrow$$
$$(2X^T X + M)\mathbf{a} = 2X^T\mathbf{y}$$

where we used the identity described, and $M = M^T$. Let $\boldsymbol{v} \in \mathbb{R}^d$. M is PD, so $\boldsymbol{v}^T M\boldsymbol{v} > 0$. Similarly, $\boldsymbol{v}^T X^T X\boldsymbol{v} = \|X^T\boldsymbol{v}\|^2 \geq 0$. Thus, $\boldsymbol{v}^T(2X^T X + M)\boldsymbol{v} = 2\boldsymbol{v}^T X^T X\boldsymbol{v} + \boldsymbol{v}^T M\boldsymbol{v} > 0$, showing that $2X^T X + M$ is PD, and therefore invertible. Therefore, we can solve

$$\mathbf{a} = 2(2X^T X + M)^{-1}X^T\mathbf{y} = (X^T X + M/2)^{-1}X^T\mathbf{y}$$

To show that this is a minimum, we derive again to compute the Hessian:

$$\frac{\partial^2 f}{\partial \mathbf{a}^2} = 2X^T X + M$$

We have already shown this matrix is PD, and therefore, this is a minimum.