Moed B 2016/7 - Introduction to Machine Learning

March 11, 2017

1 Question 1

(a)

It is possible. We show a stationary point of the algorithm, with $S_1 = \{0\}$ and $S_2 = \{10, 13, 16, 21\}$. In the next iteration, each of the points in S_2 are closer to the mean $(10 + 13 + 16 + 21)/4 = 15$ than to 0.

(b)

Impossible. It is not possible to find a subset of the points with mean 1. If we exclude 0, then since all the other points are > 1, the mean will also be > 1. If we include just 0, the mean is 0, and if we include any other point, the mean is at least $(0 + 10)/2 = 5 > 1$.

(c)

It is possible. We show a stationary point of the algorithm, $S_1 = \{0, 10\}$ and $S_2 = \{13, 16, 21\}$, with means 5 and $16.66\ldots$ respectively. In the next iteration, each point will remain in its cluster.

2 Question 2

(1)

We show VC-dim$(C) < d + 1$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_{d+1} \in \mathbb{R}^d$. Therefore, there exists a point which is a linear combination of the others; w.l.o.g, let it be $\mathbf{x}_{d+1} = \sum_{i=1}^{d} a_i \mathbf{x}_i$. We will show that no hypothesis $h_{\mathbf{w}}$ can assign the labels $y_1 = \ldots = y_d = 1$ and $y_{d+1} = 0$: If $h_{\mathbf{w}}(\mathbf{x}_1) = y_1, \ldots, h_{\mathbf{w}}(\mathbf{x}_d) = y_d$, then $\mathbf{w}^T\mathbf{x}_1 = \ldots \mathbf{w}^T\mathbf{x}_d = 0$. But then, $h_{\mathbf{w}}(\mathbf{x}_{d+1}) = \mathbf{w}^T(\sum_{i=1}^{d} a_i \mathbf{x}_i) = \sum_{i=1}^{d} a_i \mathbf{w}^T\mathbf{x}_i = 0 \Rightarrow\neq$ $y_{d+1} \neq 0$, as required.

(2)

We show VC-dim$(C) \geq d$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_d$ be the standard basis, and $y_1, \ldots, y_d$ any labeling. Then, the hypothesis $h_{\mathbf{w}}$ with $w_i = 1 - y_i$ satisfies $h_{\mathbf{w}}(\mathbf{x}_i) = y_i$

3 Question 3

(1)

Yes. Assume $\mathbf{x}_t = \mathbf{x}_{t+1} = (1, 0)$ and $y_t = y_{t+1} = 1$. If $\mathbf{w}_t = (-2, 0)$, then $\mathbf{w}_t \cdot \mathbf{x}_t = -2 < 0$ is a mistake, so $\mathbf{w}_{t+1} = \mathbf{w}_t + \mathbf{x}_t = (-1, 0)$. Then, $\mathbf{w}_{t+1} \cdot \mathbf{x}_{t+1} = -1 < 0$ is still a mistake.

(2)

Yes. Assume the same as in (1), but $\mathbf{w}_t = (-1/2, 0)$. Then, $\mathbf{w}_t \cdot \mathbf{x}_t = -1/2 < 0$ is a mistake, but $\mathbf{w}_{t+1} = \mathbf{w}_t + \mathbf{x}_t = (1/2, 0)$ so $\mathbf{w}_{t+1} \cdot \mathbf{x}_{t+1} = 1/2 > 0$ is not a mistake.

(3)

No. If it was not wrong on $\mathbf{x}_t$, then $\mathbf{w}_{t+1} = \mathbf{w}_t$ and will give the same classification.

4 Question 4

The mean of the dot products is

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{v}^T \mathbf{x}_i = \frac{\boldsymbol{v}^T}{n} \sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}$$

So the variance is the mean of squares, and is:

$$\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{v}^T \mathbf{x}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{v}^T \mathbf{x}_i)(\mathbf{x}_i^T \boldsymbol{v}) = \frac{1}{n} \cdot \boldsymbol{v}^T \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right) \boldsymbol{v} = \boldsymbol{v}^T \Sigma \boldsymbol{v} =$$

where we define the empirical covariance matrix $\Sigma = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$. Define $C = \text{diag}(c_1, \ldots, c_d)$. The problem we wish to solve is therefore

$$\begin{aligned} \arg\max_{\boldsymbol{v}} \quad & \boldsymbol{v}^T \Sigma \boldsymbol{v} \\ \text{s.t.} \quad & \boldsymbol{v}^T C \boldsymbol{v} = 1 \end{aligned}$$

To solve this, we define a vector $\mathbf{w}$ with $\mathbf{w}_i = \boldsymbol{v}_i \cdot c_i^{1/2}$; then, $\boldsymbol{v} = C^{-1/2} \mathbf{w}$. Substituting, we get

$$\begin{aligned} \arg\max_{\mathbf{w}} \quad & \mathbf{w}^T C^{-1/2} \Sigma C^{-1/2} \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^T \mathbf{w} = 1 \end{aligned}$$

As we saw in class, the solution for this problem is the eigenvector of $C^{-1/2}\Sigma C^{-1/2}$ corresponding to the largest eigenvalue.

5 Question 5

Denote $q = (q_{0,0}, q_{0,1}, q_{1,0}, q_{1,1})$. The likelihood function of S is:

$$\begin{aligned} L(q;S) &= \prod_{i=1}^{m} Pr(\mathbf{x}^i, y^i; q) \\ &= \prod_{i=1}^{m}\prod_{j=1}^{d} Pr(\mathbf{x}^i_j | y^i; q) \cdot Pr(y^i; q) \\ &= \prod_{i=1}^{m}\prod_{j=1}^{d} q_{\mathbf{x}^i_j, y^i} \cdot 0.5 \end{aligned}$$

Taking a logarithm, and ignoring constants, we get

$$\ell(q;S) = \sum_{i=1}^{m}\sum_{j=1}^{d} \log q_{\mathbf{x}^i_j, y^i} (+C)$$

which we wish to maximize under the constraints $q_{0,0} + q_{1,0} = 1, q_{0,1} + q_{1,1} = 1$. We will ignore the positivity constraints for all q, as they will be achieved by the solution. To solve this optimiziation problem, we use Lagrange multipliers. Our Lagrangian is:

$$\mathcal{L} = \sum_{i=1}^{m}\sum_{j=1}^{d} \log q_{\mathbf{x}^i_j, y^i} - \lambda_1(q_{0,0} + q_{1,0} - 1) - \lambda_2(q_{0,1} + q_{1,1} - 1)$$

Deriving by $q_{1,1}$, denote by $\delta_{i,j}$ an indicator variable which is 1 if and only if $y^i = 1$ and $x^i_j = 1$. Then we get

$$\frac{\partial \mathcal{L}}{\partial q_{1,1}} = \sum_{i=1}^{m}\sum_{j=1}^{d} \delta_{i,j} \frac{\partial \log q_{\mathbf{x}^i_j, y^i}}{\partial q_{1,1}} - \lambda_2 = \sum_{i=1}^{m}\sum_{j=1}^{d} \delta_{i,j} \frac{1}{q_{1,1}} - \lambda_2 = n_{1,1}/q_{1,1} - \lambda_2 = 0$$

$$\Rightarrow q_{1,1} = n_{1,1}/\lambda_2$$

where $n_{1,1} = \sum_{i=1}^{m}\sum_{j=1}^{d} \delta_{i,j}$, the number of times the j-th coordinate was 1 and y was 1. Similarly, we get $q_{0,1} = n_{0,1}/\lambda_2$. Together, from the constraint $q_{0,1} + q_{1,1} = 1$, we get that $\lambda_2 = n_{0,1} + n_{1,1}$, so that the MLE is:

$$q_{1,1} = n_{1,1}/(n_{0,1} + n_{1,1})$$

The derivation for $q_{1,0}$ is identical with

$$q_{1,0} = n_{1,0}/(n_{0,0} + n_{1,0})$$

Note: An alternative derivation is possible substituting $q_{0,1} = 1 - q_{1,1}$, $q_{0,0} = 1 - q_{1,0}$ in the original optimization problem and solving for only these two parameters.

6 Question 6

(1)

Yes. If it is linearly separable, there exists $\mathbf{w}^*$ such that for all i, $y_i \mathbf{w}^* \mathbf{x}_i > 0$. Denote the minimal distance from the separating hyperplane by $a = \min_i(y_i \mathbf{w}^* \mathbf{x}_i)$. If $a < 1$, we can choose $\mathbf{w}^* \leftarrow \mathbf{w}^*/a$, so that at any case $a \geq 1$.

Now, setting $c_1 = \ldots = c_d = 0$ would give simply $f(\mathbf{w}) = \sum_{i=1}^{n} \max[0, 1 - y_i \mathbf{w} \mathbf{x}_i]$. By construction above, $f(\mathbf{w}^*) = 0$; so the SVM problem will either find $\mathbf{w}^*$ or any other hyperplane for which $f(\mathbf{w}) = 0$, which means in particular that the training error is 0.

(2)

Substitute $\xi_i = 1 - y_i \mathbf{w} \mathbf{x}_i$; the problem is now

$$\begin{aligned} \arg\max_{\mathbf{w}} \quad & \frac{1}{2} \sum_{j=1}^{d} c_j w_j^2 + \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & y_i \mathbf{w} \mathbf{x}_i \geq 1 - \xi_i \qquad \forall i \\ \text{s.t.} \quad & \xi_i \geq 0 \qquad \forall i \end{aligned}$$

We will solve using KKT. Denote $C = \text{diag}(c_1, \ldots, c_d)$. The Lagrangian is

$$\mathcal{L} = \frac{1}{2} \mathbf{w}^T C \mathbf{w} + \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i \mathbf{w} \mathbf{x}_i) + \sum_{i=1}^{n} (-r_i) \xi_i$$

Deriving:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} &= C\mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w} = C^{-1} \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \frac{\partial \mathcal{L}}{\partial \xi_i} &= 1 - \alpha_i - r_i = 0 \end{aligned}$$

We note that C is invertible since $c_j > 0$ for all j. Substituting back into $\mathcal{L}$, we get

$$\begin{aligned} \mathcal{L} &= \frac{1}{2} (C^{-1} \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i)^T C (C^{-1} \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i) - (C^{-1} \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i)^T \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i + \sum_{i=1}^{n} \alpha_i \\ &= -\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T C^{-1} \mathbf{x}_j + \sum_{i=1}^{n} \alpha_i \end{aligned}$$

As we have seen in class, the constraints $\alpha_i, r_i \geq 0$ with $1 - \alpha_i - r_i = 0$ can be converted to $0 \leq \alpha_i \leq 1$, giving the dual problem:

$$\begin{aligned} \arg\max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T C^{-1} \mathbf{x}_j + \sum_{i=1}^{n} \alpha_i \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq 1 \qquad \forall i \end{aligned}$$