Machine Learning
Problem Set 3

Ozaner Hansha

November 3, 2020

Question 1

Problem: Give the dual of the soft-margin SVM optimization problem as a QP in canonical form. That is define $\mathbf{H}, \mathbf{f}, \mathbf{A}, \mathbf{a}, \mathbf{B}, \mathbf{b}$ such that:

$$\begin{array}{lll}
\underset{\boldsymbol{\alpha}}{\arg\min} & \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} + \mathbf{f}^\top \boldsymbol{\alpha} & \\
\text{subject to} & \mathbf{A}\boldsymbol{\alpha} \leq \mathbf{a} & (\leq \text{ is pointwise}) \\
 & \mathbf{B}\boldsymbol{\alpha} = \mathbf{b} & (\mathbf{b} \text{ is unrelated to the bias } b)
\end{array}$$

Solution: Before we start, let us define the following matrix $\mathbf{M}$ for convenience:

$$\mathbf{M} = \begin{bmatrix} y_1\boldsymbol{\phi}(\mathbf{x}_1) \\ \vdots \\ y_i\boldsymbol{\phi}(\mathbf{x}_i) \\ \vdots \\ y_n\boldsymbol{\phi}(\mathbf{x}_n) \end{bmatrix}$$

Now, recall that primal optimization problem for an SVM with soft margin is given by:

$$\underset{\mathbf{w}}{\arg\min} \left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}[1 - y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b)]_+ \right)$$

Introducing a slack variable $\xi_i = [1 - y_i(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i) + b)]_+$ for each training example $(\boldsymbol{\phi}(\mathbf{x}_i), y_i)$, we can reformulate the primal problem as one with constraints:

$$\begin{array}{lll}
\underset{\mathbf{w},b}{\arg\min} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i & \\
\text{subject to} & y_i(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i) + b) \geq 1 - \xi_i & \text{for } i = 1, \dots, n \\
 & \xi_i \geq 0 & \text{for } i = 1, \dots, n
\end{array}$$

Now let us take the Lagrangian of this primal problem:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) &= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i(y_i(\boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w} + b) - 1 + \xi_i) - \sum_{i=1}^{m} r_i\xi_i \\
&= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i y_i\boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w} - b\sum_{i=1}^{m}\alpha_i y_i + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m} r_i\xi_i \\
&= \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}_n \cdot \boldsymbol{\xi} - (\mathbf{M}^\top \boldsymbol{\alpha}) \cdot \mathbf{w} - b\mathbf{y} \cdot \boldsymbol{\alpha} + \mathbf{1}_n \cdot \boldsymbol{\alpha} - \boldsymbol{\alpha} \cdot \boldsymbol{\xi} - \mathbf{r} \cdot \boldsymbol{\xi} \\
&= \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}_n^\top \boldsymbol{\xi} - (\mathbf{M}^\top \boldsymbol{\alpha})^\top \mathbf{w} - b\mathbf{y}^\top \boldsymbol{\alpha} + \mathbf{1}_n^\top \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \boldsymbol{\xi} - \mathbf{r}^\top \boldsymbol{\xi}
\end{aligned}$$

To find the dual problem, we must first minimize the Lagrangian w.r.t. our parameters $\mathbf{w}, b, \boldsymbol{\xi}$. Since $\mathcal{L}$ is the sum of convex functions, it too is convex and thus has a single minimum. We can find that minimum by settings its partial derivatives to 0:

$$\begin{aligned}
\mathbf{0} &= \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) \\
&= \frac{1}{2}\nabla_{\mathbf{w}}\|\mathbf{w}\|^2 + C\nabla_{\mathbf{w}}\mathbf{1}_n^\top\boldsymbol{\xi} - \nabla_{\mathbf{w}}(\mathbf{M}^\top\boldsymbol{\alpha})^\top\mathbf{w} - b\nabla_{\mathbf{w}}\mathbf{y}^\top\boldsymbol{\alpha} + \nabla_{\mathbf{w}}\mathbf{1}_n^\top\boldsymbol{\alpha} - \nabla_{\mathbf{w}}\boldsymbol{\alpha}^\top\boldsymbol{\xi} - \nabla_{\mathbf{w}}\mathbf{r}^\top\boldsymbol{\xi} \\
&= \mathbf{w} + \mathbf{0} - (\mathbf{M}^\top\boldsymbol{\alpha}) - \mathbf{0} + \mathbf{0} - \mathbf{0} - \mathbf{0} \\
\mathbf{w} &= \mathbf{M}^\top\boldsymbol{\alpha} \qquad (1)
\end{aligned}$$

$$\begin{aligned}
0 &= \frac{\partial}{\partial b}\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) \\
&= \frac{1}{2}\frac{\partial}{\partial b}\|\mathbf{w}\|^2 + C\frac{\partial}{\partial b}\mathbf{1}_n^\top\boldsymbol{\xi} - \frac{\partial}{\partial b}\left(\mathbf{M}^\top\boldsymbol{\alpha}\right)^\top\mathbf{w} - \frac{\partial}{\partial b}b\mathbf{y}^\top\boldsymbol{\alpha} + \frac{\partial}{\partial b}\mathbf{1}_n^\top\boldsymbol{\alpha} - \frac{\partial}{\partial b}\boldsymbol{\alpha}^\top\boldsymbol{\xi} - \frac{\partial}{\partial b}\mathbf{r}^\top\boldsymbol{\xi} \\
&= 0 + 0 - 0 - \mathbf{y}^\top\boldsymbol{\alpha} + 0 - 0 - 0 \\
\mathbf{y}^\top\boldsymbol{\alpha} &= 0 \qquad (2)
\end{aligned}$$

$$\begin{aligned}
\mathbf{0} &= \nabla_{\boldsymbol{\xi}}\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) \\
&= \frac{1}{2}\nabla_{\boldsymbol{\xi}}\|\mathbf{w}\|^2 + C\nabla_{\boldsymbol{\xi}}\mathbf{1}_n^\top\boldsymbol{\xi} - \nabla_{\boldsymbol{\xi}}(\mathbf{M}^\top\boldsymbol{\alpha})^\top\mathbf{w} - b\nabla_{\boldsymbol{\xi}}\mathbf{y}^\top\boldsymbol{\alpha} + \nabla_{\boldsymbol{\xi}}\mathbf{1}_n^\top\boldsymbol{\alpha} - \nabla_{\boldsymbol{\xi}}\boldsymbol{\alpha}^\top\boldsymbol{\xi} - \nabla_{\boldsymbol{\xi}}\mathbf{r}^\top\boldsymbol{\xi} \\
&= \mathbf{0} + C\mathbf{1}_n - \mathbf{0} - \mathbf{0} + \mathbf{0} - \boldsymbol{\alpha} - \mathbf{r} \\
\mathbf{r} &= C\mathbf{1}_n - \boldsymbol{\alpha} \qquad (3)
\end{aligned}$$

Plugging these equations, which hold for optimal $\mathbf{w}, b, \boldsymbol{\xi}$, into $\mathcal{L}$ we arrive at:

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}_n^\top\boldsymbol{\xi} - (\mathbf{M}^\top\boldsymbol{\alpha})^\top\mathbf{w} - b\mathbf{y}^\top\boldsymbol{\alpha} + \mathbf{1}_n^\top\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top\boldsymbol{\xi} - \mathbf{r}^\top\boldsymbol{\xi} \\
&= \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}_n^\top\boldsymbol{\xi} - \mathbf{w}^\top\mathbf{w} - b\mathbf{y}^\top\boldsymbol{\alpha} + \mathbf{1}_n^\top\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top\boldsymbol{\xi} - \mathbf{r}^\top\boldsymbol{\xi} && \text{(eq. 1)} \\
&= \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}_n^\top\boldsymbol{\xi} - \|\mathbf{w}\|^2 - b\mathbf{y}^\top\boldsymbol{\alpha} + \mathbf{1}_n^\top\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top\boldsymbol{\xi} - \mathbf{r}^\top\boldsymbol{\xi} \\
&= \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}_n^\top\boldsymbol{\xi} - \|\mathbf{w}\|^2 + \mathbf{1}_n^\top\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top\boldsymbol{\xi} - \mathbf{r}^\top\boldsymbol{\xi} && \text{(eq. 2)} \\
&= \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}_n^\top\boldsymbol{\xi} - \|\mathbf{w}\|^2 + \mathbf{1}_n^\top\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top\boldsymbol{\xi} - (C\mathbf{1}_n - \boldsymbol{\alpha})^\top\boldsymbol{\xi} && \text{(eq. 3)} \\
&= \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}_n^\top\boldsymbol{\xi} - \|\mathbf{w}\|^2 + \mathbf{1}_n^\top\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top\boldsymbol{\xi} - C\mathbf{1}_n^\top\boldsymbol{\xi} + \boldsymbol{\alpha}^\top\boldsymbol{\xi} \\
&= \frac{1}{2}\|\mathbf{w}\|^2 - \|\mathbf{w}\|^2 + \mathbf{1}_n^\top\boldsymbol{\alpha} \\
&= -\frac{1}{2}\|\mathbf{w}\|^2 + \mathbf{1}_n^\top\boldsymbol{\alpha} \\
&= \min_{\mathbf{w}, b, \boldsymbol{\xi}}\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) = \theta_{\mathcal{D}}(\boldsymbol{\alpha})
\end{aligned}$$

And with this we can finally give the dual problem of our soft-margin SVM:

$$\begin{aligned}
\arg\max_{\boldsymbol{\alpha}} \quad & \theta_{\mathcal{D}}(\boldsymbol{\alpha}) \\
\text{subject to} \quad & 0 \le \boldsymbol{\alpha} \le C\mathbf{1}_n \quad (\le \text{ is pointwise}) \\
& \mathbf{y}^\top\boldsymbol{\alpha} = 0
\end{aligned}$$

Recall that each Lagragian multiplier that relates to an inequality must be nonnegative. So we have that $0 \leq \alpha_i, r_i$ for all i. Further consider the following for all i:

$$
\begin{aligned}
0 &\leq r_i && \text{(see above)} \\
0 &\leq C - \alpha_i && \text{(eq. 3)} \\
\alpha_i &\leq C
\end{aligned}
$$

Putting these together we get our first condition $0 \leq \alpha_i \leq C$. The second condition is simply eq. 2.

We will now transform our dual problem into a canonical QP problem:

$$
\begin{array}{ll}
\arg\max\limits_{\boldsymbol{\alpha}} & \theta_{\mathcal{D}}(\boldsymbol{\alpha}) \\
\text{subject to} & 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}_n \\
& \mathbf{y}^\top \boldsymbol{\alpha} = 0
\end{array}
=
\begin{array}{ll}
\arg\max\limits_{\boldsymbol{\alpha}} & -\frac{1}{2}\|\mathbf{w}\|^2 + \mathbf{1}_n^\top \boldsymbol{\alpha} \\
\text{subject to} & 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}_n \\
& \mathbf{y}^\top \boldsymbol{\alpha} = 0
\end{array}
\qquad \text{(def. of } \theta_D\text{)}
$$

$$
=
\begin{array}{ll}
\arg\min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{1}_n^\top \boldsymbol{\alpha} \\
\text{subject to} & 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}_n \\
& \mathbf{y}^\top \boldsymbol{\alpha} = 0
\end{array}
\qquad \text{(negation of max = min)}
$$

$$
=
\begin{array}{ll}
\arg\min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}\mathbf{w}^\top \mathbf{w} - \mathbf{1}_n^\top \boldsymbol{\alpha} \\
\text{subject to} & 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}_n \\
& \mathbf{y}^\top \boldsymbol{\alpha} = 0
\end{array}
\qquad \text{(def. of } L_2 \text{ norm)}
$$

$$
=
\begin{array}{ll}
\arg\min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}(\mathbf{M}^\top \boldsymbol{\alpha})^\top (\mathbf{M}^\top \boldsymbol{\alpha}) - \mathbf{1}_n^\top \boldsymbol{\alpha} \\
\text{subject to} & 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}_n \\
& \mathbf{y}^\top \boldsymbol{\alpha} = 0
\end{array}
\qquad \text{(eq. 1)}
$$

$$
=
\begin{array}{ll}
\arg\min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{M}\mathbf{M}^\top \boldsymbol{\alpha} - \mathbf{1}_n^\top \boldsymbol{\alpha} \\
\text{subject to} & 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}_n \\
& \mathbf{y}^\top \boldsymbol{\alpha} = 0
\end{array}
$$

$$
=
\begin{array}{ll}
\arg\min\limits_{\boldsymbol{\alpha}} & \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{M}\mathbf{M}^\top \boldsymbol{\alpha} - \mathbf{1}_n^\top \boldsymbol{\alpha} \\
\text{subject to} & \left[\begin{smallmatrix} -I_n \\ I_n \end{smallmatrix}\right] \boldsymbol{\alpha} \leq \left[\begin{smallmatrix} \mathbf{0}_n \\ C\mathbf{1}_n \end{smallmatrix}\right] \\
& \mathbf{y}^\top \boldsymbol{\alpha} = 0
\end{array}
$$

You'll notice in the last equality, in order to represent both $-\boldsymbol{\alpha} \leq 0$ and $\boldsymbol{\alpha} \leq C$ in a single matrix equation, we stack some matrices and vectors on top of each other to satisfy all $2n$ inequality conditions.

At this point it should be clear what $\mathbf{H}, \mathbf{f}, \mathbf{A}, \mathbf{a}, \mathbf{B}, \mathbf{b}$ should be, but we give them below for good measure:

$$
\begin{aligned}
\mathbf{H} &= \mathbf{M}\mathbf{M}^\top & \mathbf{f} &= -\mathbf{1}_n \\
\mathbf{A} &= \begin{bmatrix} -I_n \\ I_n \end{bmatrix} & \mathbf{a} &= \begin{bmatrix} \mathbf{0}_n \\ C\mathbf{1}_n \end{bmatrix} \\
\mathbf{B} &= \mathbf{y}^\top & \mathbf{b} &= [0] = 0
\end{aligned}
$$

Recall that $\mathbf{M}$ was defined at the start of our solution to be $\mathbf{M}_i = y_i \mathbf{x}_i$ for each row i.

Question 2

Problem: Given the Lagragian multipliers α of a soft-margin kernel SVM, how would you calculate the bias term b? (Assume there exits at least one support vector i such that $0 < a_i < C$).

Solution: Recall that all support vectors (at least one of which was guaranteed to exist) lie on the margin, that is to say for any support vector x_i:

$$y_i(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i) + b) = 1$$

And so we have the following:

$$\begin{aligned}
1 &= y_i(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i) + b) && (x_i \text{ is a support vector}) \\
y_i &= \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i) + b && (y_i^2 = (\pm 1)^2 = 1) \\
b &= y_i - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i) \\
&= y_i - \sum_{j=1}^{n} \alpha_j y_j \phi(\mathbf{x}_j)^\top \boldsymbol{\phi}(\mathbf{x}_i) && (\text{eq. } 1) \\
&= y_i - \sum_{j;\alpha_j > 0}^{n} \alpha_j y_j \phi(\mathbf{x}_j)^\top \boldsymbol{\phi}(\mathbf{x}_i) && (\text{only support vectors contribute to } \mathbf{w})
\end{aligned}$$

However the above corresponds to the kernel $\langle \cdot, \cdot \rangle$. For a general kernel $K(\cdot, \cdot)$ we apply the kernel trick to arrive at:

$$b = y_i - \sum_{j;\alpha_j > 0}^{n} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)$$