

## Problem Set 7, Nov 9, 2021 (Theory Questions Part)

### 2. Support Vector Machines using Coordinate Descent

1. The dual objective function that we have to optimise is the following :

$$\begin{aligned} & \underset{\alpha}{\text{maximize}} && f(\alpha) = \alpha^\top \mathbf{1} - \frac{1}{2\lambda} \alpha^\top Q \alpha \\ & \text{subject to} && \alpha \in [0, 1]^N \end{aligned}$$

where  $Q := \text{diag}(\mathbf{y}) X X^\top \text{diag}(\mathbf{y})$ . For computing coordinate update for one coordinate  $n$ , consider the following one variable sub-problem:

$$\begin{aligned} & \underset{\gamma \in \mathbb{R}}{\text{maximize}} && f(\alpha + \gamma e_n) \\ & \text{subject to} && 0 \leq \alpha_n + \gamma \leq 1 \end{aligned}$$

where  $e_n = [0, \dots, 1, \dots, 0]^\top$  (all zero vector except at the  $n^{\text{th}}$  position). We can explicitly develop  $f(\alpha + \gamma e_n)$  as a polynomial in  $\gamma$ , indeed:

$$\begin{aligned} f(\alpha + \gamma e_n) &= (\alpha + \gamma e_n)^\top \mathbf{1} - \frac{1}{2\lambda} (\alpha + \gamma e_n)^\top Q (\alpha + \gamma e_n) \\ &= \alpha^\top \mathbf{1} - \frac{1}{2\lambda} \alpha^\top Q \alpha + \gamma - \frac{1}{2\lambda} (\gamma^2 e_n^\top Q e_n + \gamma \alpha^\top Q e_n + \gamma e_n^\top Q \alpha) \\ &= f(\alpha) - \frac{\gamma^2}{2\lambda} Q_{nn} + \gamma(1 - \frac{1}{\lambda} \alpha^\top Q e_n) \end{aligned}$$

Hence we recognize a concave second degree polynomial, which has a unique maximum over  $\mathbb{R}$ . Differentiating with respect to  $\gamma$  and setting to 0 to obtain its maximum over  $\mathbb{R}$ , we get :

$$\begin{aligned} -\frac{\gamma^*}{\lambda} Q_{nn} + (1 - \frac{1}{\lambda} \alpha^\top Q e_n) &= 0 \\ \gamma^* &= \frac{\lambda}{Q_{nn}} (1 - \frac{1}{\lambda} \alpha^\top Q e_n) \end{aligned}$$

Note that  $Q_{nn} = \mathbf{x}_n^\top \mathbf{x}_n y_n^2 = \mathbf{x}_n^\top \mathbf{x}_n$  and  $\alpha^\top Q e_n = \sum_{i=1}^N \alpha_i Q_{i,n} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i^\top \mathbf{x}_n y_n$ . Using  $\mathbf{w}(\alpha) = \frac{1}{\lambda} \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$ , we get  $\alpha^\top Q e_n = \lambda y_n \mathbf{w}^\top \mathbf{x}_n$  and thus

$$\gamma^* = \frac{\lambda}{\mathbf{x}_n^\top \mathbf{x}_n} (1 - y_n \mathbf{w}^\top \mathbf{x}_n)$$

We conclude

$$\begin{aligned} \alpha_n^{\text{new}} &= \alpha_n + \gamma^* \\ &= \alpha_n + \frac{\lambda}{\mathbf{x}_n^\top \mathbf{x}_n} (1 - y_n \mathbf{w}^\top \mathbf{x}_n) \end{aligned}$$

In the previous equation it *seems* as if  $\alpha_n^{\text{new}}$  depends on  $\alpha_n$ , which should not be the case, you can check that it is indeed not the case. Since we have a constraint  $\alpha \in [0, 1]^N$  and we know that function  $f$  is quadratic with respect to  $\alpha_n$ , the optimal  $\alpha_n$  is the projection of  $\alpha_n^{\text{new}}$  onto the set  $[0, 1]^N$ :

$$\alpha_n^{\text{new}} := \min \left\{ \max \left\{ \alpha_n + \frac{\lambda}{\mathbf{x}_n^\top \mathbf{x}_n} (1 - y_n \mathbf{w}^\top \mathbf{x}_n), 0 \right\}, 1 \right\}$$

2. In lecture we seen the relation between primal and dual variables that is

$$\mathbf{w}(\boldsymbol{\alpha}) = \frac{1}{\lambda} \sum_{n=1}^N \alpha_n y_n \mathbf{x}_n$$

Assume that for current dual variable  $\boldsymbol{\alpha}$  the corresponding  $\mathbf{w}(\boldsymbol{\alpha})$  is known. After a coordinate update over a coordinate  $n_1$  the dual variable changes as  $\boldsymbol{\alpha}^{\text{new}} = \boldsymbol{\alpha} + \gamma^* e_{n_1}$ , and then the primal variable changes as

$$\mathbf{w}^{\text{new}}(\boldsymbol{\alpha}^{\text{new}}) = \frac{1}{\lambda} \sum_{n=1}^N \alpha_n^{\text{new}} y_n \mathbf{x}_n = \frac{1}{\lambda} \sum_{n=1}^N \alpha_n y_n \mathbf{x}_n + \frac{1}{\lambda} \gamma^* y_{n_1} \mathbf{x}_{n_1} = \mathbf{w} + \frac{1}{\lambda} \gamma^* y_{n_1} \mathbf{x}_{n_1}$$

4. We see that for the dataset used in this exercise, optimizing with coordinate descent is faster than using SGD. Coordinate descent also achieves a higher accuracy. We can also see that the duality gap goes to zero.

In general practical performance of SGD for SVM with well tuned stepsize is identical to dual Coordinate Descent, however, the advantage of CD is that no stepsize tuning is needed.

### 3. Kernels

- First we will prove that the sum of two valid kernels  $k_1$  and  $k_2$   $k = k_1 + k_2$  is a valid kernel. We need to construct a feature vector  $\phi(\mathbf{x})$  such that  $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$ , then by definition  $k$  would be a valid kernel.

Because kernels  $k_1$  and  $k_2$  are valid kernels

$$k_1(\mathbf{x}, \mathbf{x}') = \phi_1(\mathbf{x})^\top \phi_1(\mathbf{x}'), \quad k_2(\mathbf{x}, \mathbf{x}') = \phi_2(\mathbf{x})^\top \phi_2(\mathbf{x}'),$$

for some feature vectors  $\phi_1(\mathbf{x})$  and  $\phi_2(\mathbf{x})$ .

Lets take  $\phi(\mathbf{x}) = \begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \end{pmatrix}$ , then

$$\begin{aligned} \phi(\mathbf{x})^\top \phi(\mathbf{x}') &= (\phi_1(\mathbf{x})^\top, \phi_2(\mathbf{x})^\top) \begin{pmatrix} \phi_1(\mathbf{x}') \\ \phi_2(\mathbf{x}') \end{pmatrix} = \phi_1(\mathbf{x})^\top \phi_1(\mathbf{x}') + \phi_2(\mathbf{x})^\top \phi_2(\mathbf{x}') \\ &= k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') \end{aligned}$$

Therefore  $k = k_1 + k_2$  is a valid kernel.

- Second, we will prove that the product  $k = k_1 \cdot k_2$  of two valid kernels is a valid kernel.

Let's denote  $n_1$  and  $n_2$  dimensions of a feature vectors  $\phi_1(\mathbf{x})$  and  $\phi_2(\mathbf{x})$  (i.e.  $\phi_1(\mathbf{x}) \in \mathbb{R}^{n_1}$ ,  $\phi_2(\mathbf{x}) \in \mathbb{R}^{n_2}$ ).

$$k_1(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{n_1-1} \phi_{1,i}(\mathbf{x}) \phi_{1,i}(\mathbf{x}'), \quad k_2(\mathbf{x}, \mathbf{x}') = \sum_{j=0}^{n_2-1} \phi_{2,j}(\mathbf{x}) \phi_{2,j}(\mathbf{x}'),$$

Then the kernel  $k = k_1 \cdot k_2$  is

$$k(\mathbf{x}, \mathbf{x}') = \left( \sum_{i=0}^{n_1-1} \phi_{1,i}(\mathbf{x}) \phi_{1,i}(\mathbf{x}') \right) \left( \sum_{j=0}^{n_2-1} \phi_{2,j}(\mathbf{x}) \phi_{2,j}(\mathbf{x}') \right) = \sum_{i=0}^{n_1-1} \sum_{j=0}^{n_2-1} (\phi_{1,i}(\mathbf{x}) \phi_{2,j}(\mathbf{x})) (\phi_{1,i}(\mathbf{x}') \phi_{2,j}(\mathbf{x}'))$$

Lets introduce a feature vector  $\phi(\mathbf{x}) \in \mathbb{R}^{n_1 n_2}$ , such that  $\phi_{in_2+j}(\mathbf{x}) = \phi_{1,i}(\mathbf{x}) \phi_{2,j}(\mathbf{x})$  for  $i \in [0, \dots, n_1-1], j \in [0, \dots, n_2-1]$ . Note that for such  $i$  and  $j$  the index of the feature vector  $\phi$  is correct:  $in_2+j \in [0, \dots, n_1 n_2-1]$ . Then,

$$\begin{aligned} \phi(\mathbf{x})^\top \phi(\mathbf{x}') &= \sum_{l=0}^{n_1 n_2-1} \phi_l(\mathbf{x}) \phi_l(\mathbf{x}') = \sum_{i=0}^{n_1-1} \sum_{j=0}^{n_2-1} \phi_{in_2+j}(\mathbf{x}) \phi_{in_2+j}(\mathbf{x}') \\ &= \sum_{i=0}^{n_1-1} \sum_{j=0}^{n_2-1} (\phi_{1,i}(\mathbf{x}) \phi_{2,j}(\mathbf{x})) (\phi_{1,i}(\mathbf{x}') \phi_{2,j}(\mathbf{x}')) = k_1(\mathbf{x}, \mathbf{x}') \cdot k_2(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}'). \end{aligned}$$

Therefore  $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') \cdot k_2(\mathbf{x}, \mathbf{x}')$  is a valid kernel.

- Third we need to show that if  $k_1$  is a valid kernel, then  $k = c \times k_1$  with  $c \geq 0$  is also a valid kernel. Since  $k_1$  is valid, we can write for all  $x$  and  $x'$ :  $k_1(x, x') = \phi_1(x)^T \phi_1(x')$ . Now let  $\phi(x) = \sqrt{c} \phi_1(x)$ . Notice that  $k(x, x') = c \cdot k_1(x, x') = (\sqrt{c} \phi_1(x))^T (\sqrt{c} \phi_1(x')) = \phi(x)^T \phi(x')$ . Hence  $c \times k_1$  is also a valid kernel.
  - Since  $f$  is only composed of the three previous operations (sum, product, multiplication by positive scalar), we can conclude that  $(x, x') \mapsto f(k_1(x, x'))$  is a valid kernel.
2. It suffices to apply the hint to the sequence of kernels  $k_n(x, x') = \sum_{i=0}^n \frac{1}{i!} k_1(x, x')^i$ . According to our previous result these are valid kernels. Since  $k_n(x, x') \xrightarrow{n \rightarrow +\infty} \exp(k_1(x, x'))$  we can apply the hint and conclude that  $k$  is a valid kernel.

**Bonus.** For the curious who are familiar with matrices and the trace operator, here is an elegant and more natural way of showing that the product of two valid kernels is a valid kernel. Notice that for  $x, x' \in \mathcal{X}^2$ :

$$\begin{aligned}
k(x, x') &= k_1(x, x') \cdot k_2(x, x') \\
&= \phi_1(x)^T \phi_1(x') \phi_2(x)^T \phi_2(x') \\
&= \phi_1(x)^T \phi_1(x') \phi_2(x')^T \phi_2(x) \\
&= \text{trace}(\phi_1(x)^T \phi_1(x') \phi_2(x')^T \phi_2(x)) \\
&= \text{trace}(\phi_1(x') \phi_2(x')^T \phi_2(x) \phi_1(x)^T) \\
&= \text{trace}((\phi_2(x') \phi_1(x')^T)^T \phi_2(x) \phi_1(x)^T) \\
&= \langle \phi_2(x) \phi_1(x)^T, \phi_2(x') \phi_1(x')^T \rangle_F
\end{aligned}$$

Second equality is the definition of the valid kernels  $k_1$  and  $k_2$ , third is due to  $x^T y = y^T x$ , fourth is noticing that for  $z \in \mathbb{R}$   $\text{trace}(z) = z$ , fifth is that  $\text{trace}(AB) = \text{trace}(BA)$ , seventh is due to  $xy^T = (yx^T)^T$ , and last is the definition of the Frobenius inner product for matrices. Hence by letting  $\phi(x) = \phi_2(x) \phi_1(x)^T \in \mathbb{R}^{n_2 \times n_1}$  we obtain  $k(x, x') = \langle \phi(x), \phi(x') \rangle_F$ .