Labs
Machine Learning Course
Fall 2019

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Rüdiger Urbanke
www.epfl.ch/labs/mlo/machine-learning-cs-433

Problem Set 7, Oct 31, 2019 (Theory Questions Part)

2. Support Vector Machines using Coordinate Descent

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

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

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

$$\begin{array}{ll} \underset{\gamma\in\mathbb{R}}{\text{maximize}} & f(\boldsymbol{\alpha}+\gamma\boldsymbol{e}_n) \\ \text{subject to} & 0 \le \alpha_n + \gamma \le 1 \end{array}$$

where $\boldsymbol{e}_n = [0,\cdots,1,\cdots,0]^\top$ (all zero vector except at the n^{th} position). Note that, $\nabla f(\boldsymbol{\alpha}) = \mathbf{1} - \frac{1}{2\lambda}(\boldsymbol{Q}+\boldsymbol{Q}^\top)\boldsymbol{\alpha} = \mathbf{1} - \frac{1}{\lambda}\boldsymbol{Q}\boldsymbol{\alpha}$ ($\boldsymbol{Q}$ is symmetric) and thus $\nabla_n f(\boldsymbol{\alpha}) = 1 - \frac{1}{\lambda}\boldsymbol{e}_n^\top\boldsymbol{Q}\boldsymbol{\alpha}$. Simplifying $f(\boldsymbol{\alpha}+\gamma\boldsymbol{e}_n)$, we get

$$\begin{aligned} f(\boldsymbol{\alpha}+\gamma\boldsymbol{e}_n) &= (\boldsymbol{\alpha}+\gamma\boldsymbol{e}_n)^\top\mathbf{1} - \tfrac{1}{2\lambda}(\boldsymbol{\alpha}+\gamma\boldsymbol{e}_n)^\top\boldsymbol{Q}(\boldsymbol{\alpha}+\gamma\boldsymbol{e}_n) \\ &= \boldsymbol{\alpha}^\top\mathbf{1} - \tfrac{1}{2\lambda}\boldsymbol{\alpha}^\top\boldsymbol{Q}\boldsymbol{\alpha} + \gamma - \tfrac{1}{2\lambda}(\gamma^2\boldsymbol{e}_n^\top\boldsymbol{Q}\boldsymbol{e}_n + \gamma\boldsymbol{\alpha}^\top\boldsymbol{Q}\boldsymbol{e}_n + \gamma\boldsymbol{e}_n^\top\boldsymbol{Q}\boldsymbol{\alpha}) \\ &= f(\boldsymbol{\alpha}) - \tfrac{\gamma^2}{2\lambda}\boldsymbol{Q}_{nn} + \gamma(1 - \tfrac{1}{\lambda}\boldsymbol{\alpha}^\top\boldsymbol{Q}\boldsymbol{e}_n) \end{aligned}$$

Differentiating with respect to γ and equating to 0, we get :

$$\begin{aligned} -\tfrac{\gamma^\star}{\lambda}\boldsymbol{Q}_{nn} + (1 - \tfrac{1}{\lambda}\boldsymbol{\alpha}^\top\boldsymbol{Q}\boldsymbol{e}_n) &= 0 \\ \gamma^\star &= \frac{\lambda}{\boldsymbol{Q}_{nn}}(1 - \tfrac{1}{\lambda}\boldsymbol{\alpha}^\top\boldsymbol{Q}\boldsymbol{e}_n) \end{aligned}$$

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

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

We conclude

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

Since we have a constraint $\boldsymbol{\alpha} \in [0,1]^N$ and we know that function f is quadratic with respect to α_n, the optimal α_n is the projection of α_n^{new} onto the set $[0,1]^N$:

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

3. Kernels

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

     Because kernels k_1 and k_2 are valid kernels

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

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

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

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

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

   - Second, we will prove that a 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 $\phi_1(\boldsymbol{x})$ and $\phi_2(\boldsymbol{x})$ (i.e. $\phi_1(\boldsymbol{x}) \in \mathbb{R}^{n_1}$, $\phi_2(\boldsymbol{x}) \in \mathbb{R}^{n_1}$).

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

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

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

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

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

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

   - Note that every term in the resulting polynomial is a product of kernels and that all these terms have positive coefficients. Hence the result follows by the two statements proved above: that the positive sum of valid kernels is a valid kernel and that the product of valid kernels is a valid kernel.

2. We have $\exp(x) = \sum_{i \geq 0} \frac{x^i}{i!}$. We can hence apply the previous result concerning polynomials with positive coefficients and apply the limit.