Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning

Instructor: Purushottam Kar
Date: August 20, 2019
Total: 80 marks

ASSIGNMENT 1

Problem 1.1 (The Squared SVM Solver)**.** Consider the following problem formulation for n binary labelled data points $(\mathbf{x}^i, y^i)_{i=1,\ldots,n}$ where $\mathbf{x}^i \in \mathbb{R}^d$ and $y^i \in \{-1, +1\}$ that uses the squared hinge loss instead of the hinge loss

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \left(\left[1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \right]_+ \right)^2 \qquad (P1)$$

The above optimization problem (P1) can be rewritten as a new optimization problem (P2)

$$\begin{aligned} \arg\min_{\mathbf{w}\in\mathbb{R}^d, \boldsymbol{\xi}\in\mathbb{R}^n} \ & \frac{1}{2}\|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \xi_i^2 \\ \text{s.t. } & y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1 - \xi_i, \text{ for all } i \in [n] \end{aligned} \qquad (P2)$$

1. Introduce a dual variable α_i for each of the n constraints in (P2) and write the expression for the Lagrangian. Remember, the Lagrangian has no max, min terms and is simply a function of the form $\mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha})$ where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_n] \in \mathbb{R}^n$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n] \in \mathbb{R}^n$. Also take care that in this case, the primal problem (P2) has not only a $\mathbf{w}$ variable that represents the model, but also a $\boldsymbol{\xi}$ variable that encodes the slacks.

   Solution. $\mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \xi_i^2 + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)$

2. Find out the dual problem by eliminating the primal variables $\mathbf{w}, \boldsymbol{\xi}$ using the first order optimality trick i.e. by setting $\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0}$ and $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}} = \mathbf{0}$. Write down the dual optimization problem that you get as a result – let us call this problem (D2). Note that (P1), (P2), (D2) are all the same problem written differently.

   Solution. The dual problem is

   $$\max_{\boldsymbol{\alpha}\geq 0} \left\{ \min_{\mathbf{w}, \boldsymbol{\xi}} \left\{ \frac{1}{2}\|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \xi_i^2 + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle) \right\} \right\}$$

   Using first order optimality gives us $\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0}$ which implies $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y^i \cdot \mathbf{x}^i$, and $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}} = \mathbf{0}$ which implies $\xi_i = \frac{\alpha_i}{2C}$. Substituting these in the above expression gives us the simplified dual problem

   $$\max_{\boldsymbol{\alpha}\geq 0} \left\{ \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^\top \left(Q + \frac{1}{2C} \cdot I_n \right) \boldsymbol{\alpha} \right\},$$

   where I_n denotes the $n \times n$ identity matrix, $\mathbf{1}$ denotes the all ones vector and Q denotes the $n \times n$ matrix where $Q_{ij} = \alpha_i \alpha_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$.

3. *Bonus*: Notice that (P2) does not have the positivity constraints $\xi_i \geq 0$ that the CSVM formulation had. Can you show that the inserting these additional constraints into (P2) does not change the optimization problem at all (i.e. the solution remains the same)?

Solution. Suppose (P2) has an optimal solution $\{\mathbf{w}^*, \boldsymbol{\xi}^*\}$. Then consider the new solution $\left\{\mathbf{w}^*, \tilde{\boldsymbol{\xi}}\right\}$ such that $\tilde{\xi}_i = \max\{\xi_i^*, 0\}$. Now, for every $i \in [n]$, we have

$$y^i \left\langle \mathbf{w}^*, \mathbf{x}^i \right\rangle \geq 1 - \xi_i^* \geq 1 - \tilde{\xi}_i,$$

where the first inequality holds since $\{\mathbf{w}^*, \boldsymbol{\xi}^*\}$ is a solution to (P2) (and hence must satisfy all its constraints) and the second inequality holds since by construction, $\tilde{\xi}_i \geq \xi_i^*$ for all $i \in [n]$. It also holds, due to the same construction, that whenever $\tilde{\xi}_i \neq \xi_i^*$ (which happens exactly when $\xi_i^* < 0$), we have $(\tilde{\xi}_i)^2 < (\xi_i^*)^2$. However, this means that such that if for one or more $i \in [n]$, the solution $\{\mathbf{w}^*, \boldsymbol{\xi}^*\}$ has $\xi_i^* < 0$, then we must have

$$\sum_{i=1}^{n} (\xi_i^*)^2 < \sum_{i=1}^{n} (\tilde{\xi}_i)^2$$

The above means that the solution $\left\{\mathbf{w}^*, \tilde{\boldsymbol{\xi}}\right\}$ satisfies all the constraints of (P2) but offers a better objective value, since

$$\frac{1}{2} \|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^{n} (\xi_i^*)^2 < \frac{1}{2} \|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^{n} (\tilde{\xi}_i)^2,$$

whenever $\xi_i^* < 0$ for one or more $i \in [n]$. This contradicts the fact that $\{\mathbf{w}^*, \boldsymbol{\xi}^*\}$ is an optimal solution to (P2). This means that any optimal solution to (P2) must automatically satisfy $\xi_i^* \geq 0$ and thus, putting the constraint in does not change the solution.

(80 + bonus marks)