# Programming Assignment 2

CS6230: Optimization Methods in Machine Learning
IIT-Hyderabad
Aug-Nov 2017

**Max Marks:** 40
**Due:** 10th Nov 2017 11:59 pm

## Instructions

- Please use Google Classroom to upload your submission by the deadline mentioned above. Your submission should comprise of a single ZIP file, named `<Your_Roll_No>_PA2`, with all your solutions.
- For late submissions, 10% is deducted for each day (including weekend) late after an assignment is due. Note that each student begins the course with 6 grace days for late submission of assignments. Late submissions will automatically use your grace days balance, if you have any left. You can see your balance on the CS6230 Marks and Grace Days document under the course Google drive.
- You can use either PYTHON or MATLAB for this assignment.
- Please read the department plagiarism policy. Do not engage in any form of cheating - strict penalties will be imposed for both givers and takers. Please talk to instructor or TA if you have concerns.

## 1 Support vector machines and duality

In binary classification, we are, roughly speaking, interested in finding a hyperplane that separates two clouds of points living in, say, $\mathbb{R}^p$. The support vector machine (SVM), is a pretty popular method for doing binary classification; to this day, it's (still) used in a number of fields outside of just machine learning and statistics.

One issue with the standard SVM, though, is that it doesn't work well in situations where we pay a higher "price" for misclassifications of one of the two point-clouds. For example, a bank will probably want to be quite certain that a customer won't default on their loan before deciding to give them one (here, the "price" that we pay is monetary). In this problem, you will develop a variant of the standard SVM that addresses these issues, called the *cost-sensitive* SVM. You will implement your own cost-sensitive SVM solver (in part (b) of this question), but as a starting point, we will first investigate the cost-sensitive SVM dual problem (in part (a) of this question).

Throughout, we assume that we are given $n$ data samples, each one taking the form $(x_i, y_i)$, where $x_i \in \mathbb{R}^p$ is a feature vector and $y_i \in \{-1, +1\}$ is a class. In order to make our notation more concise, we can transpose and stack the $x_i$ vertically, collecting these feature vectors into the matrix $X \in \mathbb{R}^{n \times p}$; doing the same thing with the $y_i$ lets us write $y \in \{-1, +1\}^n$. It will also be useful for us to define the following sets, containing the indices of the positive (i.e., those with $y_i = +1$) and negative (i.e., those with $y_i = -1$) samples, respectively:

$$\mathcal{S}_1 = \{i \in \{1, \ldots, n\} : y_i = +1\}, \quad \mathcal{S}_2 = \{i \in \{1, \ldots, n\} : y_i = -1\}.$$

## Part (a) (Submit either pdf / scanned copies) 15 Points

One simple way to incorporate misclassification costs into the standard SVM formulation, is to pose the following (primal) cost-sensitive SVM optimization problem:

$$\begin{array}{ll} \underset{\beta \in \mathbb{R}^p,\, \beta_0 \in \mathbb{R},\, \xi \in \mathbb{R}^n}{\text{minimize}} & (1/2)\|\beta\|_2^2 + C_1 \sum_{i \in \mathcal{S}_1} \xi_i + C_2 \sum_{i \in \mathcal{S}_2} \xi_i \\ \text{subject to} & \xi_i \geq 0, \quad y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \quad i = 1, \ldots, n, \end{array} \tag{1}$$

where $\beta \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}$, $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$ are our variables, and $C_1, C_2$ are positive costs, chosen by the implementer.

1. Does strong duality hold for problem (1)? Why or why not? [3 Points]

2. Derive the Karush-Kuhn-Tucker (KKT) conditions for problem (1). Please use $\alpha \in \mathbb{R}^n$ for the dual variables (i.e., Lagrange multipliers) associated with the constraints "$y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i,\ i = 1, \ldots, n$", and $\mu \in \mathbb{R}^n$ for the dual variables associated with the constraints "$\xi_i \geq 0,\ i = 1, \ldots, n$". [5 Points]

3. Show that the cost-sensitive SVM dual problem can be written as

$$\begin{array}{ll} \underset{\alpha \in \mathbb{R}^n}{\text{maximize}} & -(1/2)\alpha \tilde{X} \tilde{X}^T \alpha + 1^T \alpha \\ \text{subject to} & y^T \alpha = 0, \quad 0 \leq \alpha_{\mathcal{S}_1} \leq C_1 1, \quad 0 \leq \alpha_{\mathcal{S}_2} \leq C_2 1, \end{array} \tag{2}$$

   where $\tilde{X} \in \mathbb{R}^{n \times p} = \mathrm{diag}(y)X$, $\alpha_{\mathcal{S}}$ means selecting only the indices of $\alpha$ that are in the set $\mathcal{S}$, and the 1's here are vectors (of the appropriate and possibly different sizes) containing only ones. [5 Points]

4. What kind of problem class are both (1) and (2)? You may choose none, one, or more than one of the following: [2 Points]

   - linear program
   - quadratic program
   - second-order cone program
   - semidefinite program
   - cone program

## Part (b)(Please submit your code to this problem) 25 Points

1. Implement the primal SVM in problem (1) using a standard QP solver, typically available as "quadprog" function (for example in Matlab). Load a small synthetic toy problem with inputs $X \in \mathbb{R}^{100\times 2}$ and labels $y \in \{-1, 1\}^{100}$ from `toy.hdf5` (HDF5 file format) and solve the primal SVM with (1) $C_1 = C_2 = 1$, (2) $C_1 = 1, C_2 = 10$ and (3) $C_1 = 10, C_2 = 1$. For each pair of penalty parameters report the objective value of the optimal solution. [8 Points]

2. For each parameter pair, show a scatter plot of the data and plot the decision border (where the predicted class label changes) as well as the boundaries of the margin (the area in which there is a nonzero penalty for predicting any label) on top. Also highlight the data points $i$ that lie on the wrong side of the margin, that is, points with $\xi_i > 0$. How and why does the decision boundary change with different penalty parameters? [5 Points]

3. Implement now the dual SVM in problem (2) using again a standard QP solver and report the optimal objective value of the dual for the same penalty parameters as in (i). What can in general be said about the location of a data point $i \in \mathcal{S}_k$ with respect of the boundary of the margin if

   - $\alpha_i = 0$;
   - $\alpha_i \in (0, C_k)$;
   - $\alpha_i = C_k$?

   For each pair of penalty parameters, plot the signed distance to the decision boundary of each datapoint $i$ obtained from the primal SVM $y_i(x_i^\top \beta + \beta_0)$ against dual variables $\alpha_i$ obtained from the dual SVM. [8 Points]

4. Cost-sensitive SVMs minimize the (regularized) cost-sensitive hinge-loss, a convex upper bound on the weighted classification error. Predict the class labels for each data point (of the same set that the SVM was trained on) and report the total weighted classification error. A datapoint incurs a loss of $C_1$ if the true label is $+1$ and $-1$ is predicted and $C_2$ if $+1$ is predicted for a data point with true label $-1$. [4 Points]