Assignment 3

CS6230: Optimization Methods in Machine Learning

Due: 21st Nov 2016 01:00 pm
Max Marks: 100

INSTRUCTIONS

- Please submit your solutions to theory questions on an A4 sheet and hand it over to the TA (Adepu Ravi Sankar) in Room No 611, 6^{th} floor, Academic Block A by the due date (and time). Solutions to programming questions should be uploaded as a single ZIP file named '⟨YourRollNo⟩_assign2.zip' through the Google Classroom submission link.
- Your solution to coding problems should include plots and whatever explanation necessary to answer the questions asked.
- 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 5 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.
- 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.

SECTION A (38 POINTS)

1. [$3*5=15$ **points**] For each of the following functions on $\mathbb{R}^n$, explain how to calculate the sub-gradient at a given x.

   (a) $f(x) = \sup_{0\leq t\leq 1} p(t)$, where $p(t) = x_1 + x_2 t + \cdots + x_n t^{n-1}$.

   (b) $f(x) = x_{[1]} + x_{[2]} + \cdots + x_{[k]}$ where $x_{[i]}$ denotes the ith largest element of x.

   (c) $f(x) = ||Ax - b||_2 + ||x||_2$ where $A \in \mathbb{R}^{m\times n}$

   (d) $f(x) = \lambda_{\max}(W + \mathrm{diag}(x))$ where $W \in \mathbf{S}^n$

   (e) $f(x) = \sup_{Ay\leq b} x^T y$, where $A \in \mathbb{R}^{m\times n}$ and the polyhedron defined by $Ay \leq b$ is nonempty and bounded.

2. [$3*4=12$ **points**] For each subproblem, give a formula or simple algorithm for evaluating the proximal mapping

$$\mathrm{prox}_f = \mathrm{argmin}_u \left(f(u) + \frac{1}{2}||u - x||_2^2 \right)$$

   (a) $f(x) = \|x\|_1$ with dom $f = \{x \in \mathbb{R}^n, |\|x\|_\infty \leq 1\}$

   (b) $f(x) = \|Ax - b\|_1$ where $AA^T = D$ with D positive diagonal

(c) $f(x) = \max_k x_k$

(d) $f(x) = \|x\|_2$ with domain $\mathbb{R}^n_+$

3. **[3 points]** Show that the dual norm of a dual norm is the original norm itself.

4. **[4 points]** Given f is closed and convex, show why given any x, y:

$$x \in \partial f^*(y) \iff y \in \partial f(x) \iff f(x) + f^*(y) = x^T y$$

where f^* is the conjugate function of f.

5. **[4 points]** If $f(x) = ||x||_p$, where $p \geq 1$, show why its conjugate is:

$$f^*(y) = \mathbb{I}_{\{z:||z||_* \leq 1\}}(y)$$

Section B (62 points)

1. **[8 points]** Take the least squares regression problem (for $X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$):

$$\min_{\beta\in\mathbb{R}^p} \left(\|y - X\beta\|_2\right)^2 \tag{0.1}$$

Prove that an equivalent dual of this problem is

$$\min_{v\in\mathbb{R}^n} \|y - v\|_2^2 \text{ subject to } X^T v = 0 \tag{0.2}$$

(Hint: in deriving the dual, you may start by introducing the auxiliary variable $z = X\beta$.) What is the relationship between the primal and the dual solutions, implied by the KKT conditions? Explain why this relationship makes sense, given what you know about projections onto linear subspaces.

2. **Invariance Under Affine Transformation: [4+4 = 8 points]**

(a) Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and twice differentiable, $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ be invertible. Define g as $g(x) = f(Ax + b)$ for all x and let $u_0 \in \mathbb{R}^n$ be arbitrary but fixed. A step of Newton's method applied to f at u_0 results in

$$u_1 = u_0 - \left(\nabla^2 f(u_0)\right)^{-1} \nabla f(u_0). \tag{0.3}$$

Show that a step of the Newton's method applied to g at $x_0 = A^{-1}(u_0 - b)$ results in $x_1 = A^{-1}(u_1 - b)$. This will imply that $g(x_1) = f(u_1)$, that is, the criterion values match after a Newton step. This will continue to be true at all iterations, and thus we say that Newton's method is affine invariant.

(b) Show that gradient descent is not invariant under affine transformation by providing a concise counterexample. Be specific, that is, define a function f, an affine transformation A, b and an initial u_0 at least.

3. **SVMs and Duality: [46 points]** In binary classification, we are interested in finding a hyperplane that separates two clouds of points living in, say, $\mathbb{R}^p$. 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:

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

Part (a) (18 points)

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

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

where $\beta \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}$, $\xi = (\xi_1,\dots,\xi_n) \in \mathbb{R}^n$ are our variables, and C_1, C_2 are positive costs, chosen by the implementer. (Just to remind you of some of the intuition here: when $C_1 = C_2$, the above problem can be viewed as another way of writing a squared ℓ_2-norm penalized hinge loss minimization problem.)

(i) *[2 points]* Does strong duality hold for the above problem? Why or why not?

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

(iii) *[6 points]* Show that the cost-sensitive SVM dual problem can be written as

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

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

(iv) *[3 points]* Give an expression for the optimal β in terms of the optimal α variables. Explain why the optimal β can be thought of as "cost-sensitive".

(v) *[2 points]* What kind of problem class are the primal and the dual, and why? You may choose none, one, or more than one of the following:

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

Part (b) (28 points)

Please submit your code for this part:.

(a) *[6 points]* Implement the above primal SVM using a standard QP solver. 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: (i) $C_1 = C_2 = 1$; (ii) $C_1 = 1, C_2 = 10$; and (iii) $C_1 = 10, C_2 = 1$. For each pair of penalty parameters, report the objective value of the optimal solution.

(b) *[4 points]* 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?

(c) *[3 points]* Looking back at the KKT conditions derived in part (a, ii) and the form of the primal solution in part (a, iv), what can be said about the influence of the data points that lie strictly on the right side of the margin (points i with $y_i(x_i^\top \beta + \beta_0) > 1$)? How would the decision boundary change if we removed these data points from the dataset and recomputed the optimal solution? (Give a qualitative answer, no need to actually implement that).

(d) *[6 points]* Implement now the dual SVM in the above problem using again a standard QP solver and report the optimal objective value of the dual for the same penalty parameters as in (a).

(e) *[6 points]* What can in general be said about the location of a data point $i \in \mathscr{S}_k$ 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 α_i obtained from the dual SVM.

(f) *[3 points]* 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.