Homework 12

This homework is due at 11 PM on April 19, 2024.

Submission Format: Your homework submission should consist of a single PDF file that contains all of your answers (any handwritten answers should be scanned), as well as a printout of your completed Jupyter notebook(s).

1. Median Versus Mean

For a given vector $\vec{v} \in \mathbb{R}^n$, the mean can be found as the solution to the optimization problem

$$\min_{x \in \mathbb{R}} \left\| \vec{v} - x\vec{1} \right\|_2^2, \tag{1}$$

where $\vec{1}$ is the vector of ones in $\mathbb{R}^n$. Similarly, the median (any value x such that there is an equal number of values in $\vec{v}$ above or below x) can be found via

$$\min_{x \in \mathbb{R}} \left\| \vec{v} - x\vec{1} \right\|_1. \tag{2}$$

We consider a robust version of the mean problem (1):

$$\min_{x} \max_{\vec{u} \,:\, \|\vec{u}\|_\infty \leq \lambda} \left\| \vec{v} + \vec{u} - x\vec{1} \right\|_2^2, \tag{3}$$

in which we assume that the components of $\vec{v}$ can be independently perturbed by a vector $\vec{u}$ whose magnitude is bounded by a given number $\lambda \geq 0$.

(a) Is the robust problem (3) convex? Justify your answer precisely, based on expression (3), and without further manipulation.

(b) Show that problem (3) can be expressed as

$$\min_{x \in \mathbb{R}} \sum_{i=1}^{n} \left(|v_i - x| + \lambda \right)^2. \tag{4}$$

(c) Express the problem (4) as a QP. State precisely the variables, and constraints if any.

(d) Show that when λ is large, the solution set approaches that of the median problem (2). *HINT: Given variable a, constants b, c, where $c \gg 1$, and the optimization problem $\min_a \frac{1}{c}(b-a)^2 + |b-a|$. The minimizer a^* tends to minimize the second term only.*

(e) It is often said that the median is a more robust notion of "middle" value than the mean, when noise is present in $\vec{v}$. Based on the previous part, justify this statement.

2. Sphere Enclosure

Let B_i, $i = 1, \ldots, m$, be m Euclidean balls in $\mathbb{R}^n$, with centers $\vec{x}_i$, and radii $\rho_i \geq 0$. We wish to find a ball B with center $\vec{c} \in \mathbb{R}^n$ of minimum radius $r \geq 0$ that contains all the B_i, $i = 1, \ldots, m$. Cast this problem as an SOCP.

3. LASSO vs. Ridge

Say that we have the data set $\{(\vec{x}^{(i)}, y^{(i)})\}_{i=1,\dots,n}$ of samples $\vec{x}^{(i)} \in \mathbb{R}^d$ and values $y^{(i)} \in \mathbb{R}$.

Define $X = \begin{bmatrix} \vec{x}^{(1)} & \dots & \vec{x}^{(n)} \end{bmatrix}^\top$ and $y = \begin{bmatrix} y^{(1)} & \dots & y^{(n)} \end{bmatrix}^\top$.

For the sake of simplicity, assume that each feature of the data has mean 0 and variance 1 and the features are uncorrelated, i.e. $X^\top X = nI$. Consider the linear least squares regression with regularization in the ℓ_1-norm, also known as LASSO:

$$\vec{w}^\star = \underset{\vec{w} \in \mathbb{R}^d}{\operatorname{argmin}} \ \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_1 . \tag{5}$$

This problem will compare ℓ_1-regularization with ℓ_2-regularization (ridge regression) to understand their similarities and differences. We will do this by looking at the elements of $\vec{w}^\star$ in the solution to each problem.

(a) First, we decompose this optimization problem into d univariate optimization problems over each element of $\vec{w}$. Let $X = \begin{bmatrix} \vec{x}_1 & \dots & \vec{x}_d \end{bmatrix}$ and recall that $X^\top X = nI$.

(b) If $w_i^\star > 0$, then what is the value of $w_i^\star$? What is the condition on $\vec{y}^\top \vec{x}_i$ for this to be possible? *HINT: Use the first order condition.*

(c) If $w_i^\star < 0$, then what is the value of $w_i^\star$? What is the condition on $\vec{y}^\top \vec{x}_i$ for this to be possible?

(d) What can we conclude about $w_i^\star$ if $\left|\vec{y}^\top \vec{x}_i\right| \leq \dfrac{\lambda}{2}$? How does the value of λ impact the individual entries $w_i^\star$?

(e) Now consider the case of ridge regression, which uses the the ℓ_2 regularization $\lambda \|\vec{w}\|_2^2$.

$$\vec{w}^\star = \underset{\vec{w} \in \mathbb{R}^d}{\operatorname{argmin}} \ \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_2^2 . \tag{6}$$

Write down the new condition for $\vec{w}_i^\star$ to be 0. How does this differ from the condition obtained in part (d) and what does this suggest about LASSO?

4. More Fun with Lasso and Ridge

Complete the Jupyter notebook `ridge_vs_lasso.ipynb` which demonstrates differences between ridge regression and LASSO.

5. Connecting Ridge Regression, LASSO, and Constrained Least Squares

This question aims to help you develop an understanding of how a constraint in an optimization problem has the same effect as a penalty term in the objective, and apply it to the context of regularized least squares. More formally, let $f\colon \mathbb{R}^n \to \mathbb{R}$ be strictly convex and such that $\lim_{t\to\infty} f(\vec{x}_t) = \infty$ for any sequence $(\vec{x}_t)_{t=0}^{\infty}$ such that $\lim_{t\to\infty} \|\vec{x}_t\|_2 = \infty$. Let $g\colon \mathbb{R}^n \to \mathbb{R}_+$ be convex and take non-negative values. Further, suppose that there exists $\vec{x}_0 \in \mathbb{R}^n$ such that $g(\vec{x}_0) = 0$. For $\lambda \geq 0$ and $k \geq 0$, define the "penalty" and "constraint" programs

$$P(\lambda) \doteq \operatorname*{argmin}_{\vec{x}} \{f(\vec{x}) + \lambda g(\vec{x})\} \tag{7}$$

$$C(k) \doteq \operatorname*{argmin}_{\vec{x}\colon\, g(\vec{x}) \leq k} f(\vec{x}). \tag{8}$$

We will show that for every λ there exists k such that $P(\lambda) = C(k)$, and vice versa.

(a) Show that, for $k \geq 0$ and $\lambda \geq 0$, both $P(\lambda)$ and $C(k)$ have exactly one element, i.e., each problem has exactly one optimal solution.

HINT: You may use without proof that $P(\lambda)$ and $C(k)$ have at least one element each (this is true from assumptions but requires some analysis to show). Thus, you just need to show that there are not multiple optimal solutions to each problem. For this, use strict convexity of the objectives.

(b) Prove that for all $\lambda \geq 0$ there exists $k \geq 0$ such that $P(\lambda) = C(k)$.

HINT: Let $\vec{x}^\star \in P(\lambda)$ and show that $\vec{x}^\star \in C(k)$ for $k = g(\vec{x}^\star)$. Use the fact, from part 5(a), that $P(\lambda)$ and $C(k)$ have exactly one element.

(c) Prove that for all $k > 0$ there exists $\lambda \geq 0$ such that $P(\lambda) = C(k)$.

HINT: Prove that strong duality holds for the constraint problem, let $\vec{x}^\star \in C(k)$ and $\mu^\star$ be optimal primal and dual variables for the constraint problem and show that $\vec{x}^\star \in P(\lambda)$ for $\lambda = \mu^\star$.

Now we apply our findings to regularized least squares, in order to understand why LASSO promotes sparsity more than ridge regression. Let $A \in \mathbb{R}^{m\times n}$ have full column rank, and let $\vec{y} \in \mathbb{R}^m$. In the course, we have looked at LASSO:

$$\text{LASSO}(\lambda) \doteq \operatorname*{argmin}_{\vec{x}} \left\{ \|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_1 \right\} \tag{9}$$

and ridge regression:

$$\text{Ridge}(\lambda) \doteq \operatorname*{argmin}_{\vec{w}} \left\{ \|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2 \right\} \tag{10}$$

which add an ℓ^1 and ℓ^2 norm penalty to the least squares objective, respectively. The analogous constrant programs are the ℓ^1- and ℓ^2-constrained least squares problems:

$$\ell^1\text{CLS}(k) \doteq \operatorname*{argmin}_{\vec{x}\colon\, \|\vec{x}\|_1 \leq k} \|A\vec{x} - \vec{y}\|_2^2 \tag{11}$$

$$\ell^2\text{CLS}(k) \doteq \operatorname*{argmin}_{\vec{x}\colon\, \|\vec{x}\|_2^2 \leq k} \|A\vec{x} - \vec{y}\|_2^2 . \tag{12}$$

(d) Show that the result from part 5(b) and part 5(c) can be used to show the equivalence of LASSO with ℓ^1CLS and the equivalence of ridge regression with ℓ^2CLS. Namely, for each pair of equivalent formulations, find f and g, prove that f is strictly convex, prove that g is convex, and prove that there is an $\vec{x}_0$ such that $g(\vec{x}_0) = 0$.

(e) Complete the Jupyter notebook `ridge_lasso_constrained.ipynb`, which will use this equivalence to show geometrically why LASSO solutions tend to be sparse (i.e. have many zeros) while ridge regression doesn't, and attach a PDF printout of your answers.

6. Homework Process

With whom did you work on this homework? List the names and SIDs of your group members.

NOTE: If you didn't work with anyone, you can put "none" as your answer.