Homework 12

Self grades are due at 11 PM on April 26, 2024.

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.

Solution: The robust problem is convex, since the objective function is the pointwise maximum (over $\vec{u}$) of convex functions, $x \to \left\| \vec{v} + \vec{u} - x\vec{1} \right\|_2^2$.

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

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

Solution: For any vector $\vec{z} \in \mathbb{R}^n$, we have

$$\max_{\vec{u}:\|\vec{u}\|_\infty \leq \lambda} \|\vec{z} + \vec{u}\|_2^2 = \max_{\vec{u}:\|\vec{u}\|_\infty \leq \lambda} \sum_{i=1}^{n} (z_i + u_i)^2 \tag{5}$$

$$= \max_{|u_1| \leq \lambda, |u_2| \leq \lambda, \ldots, |u_n| \leq \lambda} \sum_{i=1}^{n} (z_i + u_i)^2 \tag{6}$$

$$= \sum_{i=1}^{n} \max_{u_i \,:\, |u_i| \leq \lambda} (z_i + u_i)^2. \tag{7}$$

The last equality follows since the problem is decomposable and it is optimal to maximize each of these terms. Next we have,

$$\sum_{i=1}^{n} \max_{u_i \,:\, |u_i| \leq \lambda} (z_i + u_i)^2 = \sum_{i=1}^{n} \max_{\eta \,:\, |\eta| \leq \lambda} (z_i + \eta)^2 \tag{8}$$

$$= \sum_{i=1}^{n} (|z_i| + \lambda)^2, \tag{9}$$

the last line resulting from

$$\forall\, \eta,\ \ |\eta| \le \lambda\ :\ \ |z_i + \eta| \le |z_i| + \lambda, \tag{10}$$

with upper bound attained with $\eta = \lambda\,\mathrm{sign}(z_i)$. Taking $\vec{z} = \vec{v} - x\vec{1}$ we get the desired form.

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

Solution: A QP formulation is

$$\min_{x,t} \sum_{i=1}^{n} (t_i + \lambda)^2\ :\ t_i \ge \pm(v_i - x),\ \ i = 1, \ldots, n. \tag{11}$$

(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.*

Solution: The objective function takes the form

$$\sum_{i=1}^{n} (|v_i - x| + \lambda)^2 = n\lambda^2 + 2\lambda \left\|\vec{v} - x\vec{1}\right\|_1 + \left\|\vec{v} - x\vec{1}\right\|_2^2, \tag{12}$$

The corresponding optimization problem has the same minimizers as the problem

$$\min_{x}\ \left\|\vec{v} - x\vec{1}\right\|_1 + \frac{1}{2\lambda}\left\|\vec{v} - x\vec{1}\right\|_2^2, \tag{13}$$

When λ is large, the minimizer will tend to minimize the first term only, which implies the desired result.

(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.

Solution: The median problem can be interpreted as a robust version of the mean problem, when the uncertainty is large.

© UCB EECS 127/227AT, Spring 2024. All Rights Reserved. This may not be publicly shared without explicit permission.

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.

Solution: Let $\vec{c} \in \mathbb{R}^n$ and $r \geq 0$ denote the center and radius of the enclosing ball B, respectively. We express the given balls B_i as

$$B_i = \{\vec{x} : \vec{x} = \vec{x}_i + \vec{\delta}_i, \ \left\|\vec{\delta}_i\right\|_2 \leq \rho_i\}, \quad i = 1, \ldots, m. \tag{14}$$

We have that $B_i \subseteq B$ if and only if

$$\max_{\vec{x} \in B_i} \|\vec{x} - \vec{c}\|_2 \leq r. \tag{15}$$

Note that

$$\max_{\vec{x} \in B_i} \|\vec{x} - \vec{c}\|_2 = \max_{\left\|\vec{\delta}_i\right\|_2 \leq \rho_i} \left\|\vec{x}_i - \vec{c} + \vec{\delta}_i\right\|_2 = \|\vec{x}_i - \vec{c}\|_2 + \rho_i. \tag{16}$$

The last step follows by choosing $\vec{\delta}_i$ in the direction of $\vec{x}_i - \vec{c}$. The problem is then cast as the following SOCP

$$\min_{\vec{c}, r} \quad r \tag{17}$$

$$\text{s.t.} \quad \|\vec{x}_i - \vec{c}\|_2 + \rho_i \leq r, \qquad i = 1, \ldots, m. \tag{18}$$

© UCB EECS 127/227AT, Spring 2024. All Rights Reserved. This may not be publicly shared without explicit permission.

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:

$$\vec{w}^\star = \underset{\vec{w}\in\mathbb{R}^d}{\operatorname{argmin}} \ \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_1 . \tag{19}$$

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 $X^\top X = nI$.

Solution:

$$\|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_1 = \sum_{i=1}^{d} \left[n w_i^2 - 2\vec{y}^\top \vec{x}_i w_i + \lambda |w_i| \right] + \vec{y}^\top \vec{y}. \tag{20}$$

Hence the original problem becomes

$$\min_{\vec{w}\in\mathbb{R}^d} \sum_{i=1}^{d} \left[n w_i^2 - 2\vec{y}^\top \vec{x}_i w_i + \lambda |w_i| \right] + \vec{y}^\top \vec{y}. \tag{21}$$

Since the objective is separable in w_i the problem decomposes into d univariate optimization problems and hence we have

$$\sum_{i=1}^{d} \min_{w_i\in\mathbb{R}} \left[n w_i^2 - 2\vec{y}^\top \vec{x}_i w_i + \lambda |w_i| \right] + \vec{y}^\top \vec{y}. \tag{22}$$

(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.*

Solution: If $w_i^\star > 0$, then the first order optimality conditions for $w_i^\star$ write

$$2nw_i^\star - 2\vec{y}^\top \vec{x}_i + \lambda = 0, \tag{23}$$

from which we obtain

$$w_i^\star = \frac{2\vec{y}^\top \vec{x}_i - \lambda}{2n}, \tag{24}$$

which is positive when

$$\vec{y}^\top \vec{x}_i > \frac{\lambda}{2}. \tag{25}$$

(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?

Solution: If $w_i^\star < 0$, then the first order optimality conditions for $w_i^\star$ write

$$2nw_i^\star - 2\vec{y}^\top \vec{x}_i - \lambda = 0, \tag{26}$$

from which we obtain

$$w_i^\star = \frac{2\vec{y}^\top \vec{x}_i + \lambda}{2n}, \tag{27}$$

© UCB EECS 127/227AT, Spring 2024. All Rights Reserved. This may not be publicly shared without explicit permission.

which is negative when

$$\vec{y}^\top \vec{x}_i < -\frac{\lambda}{2}. \tag{28}$$

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

Solution: From the previous parts we have $w_i^\star \neq 0 \Rightarrow \left|\vec{y}^\top \vec{x}_i\right| > \frac{\lambda}{2}$. Hence if $\left|\vec{y}^\top \vec{x}_i\right| \leq \frac{\lambda}{2}$ then we must have $w_i^\star = 0$. This means that a larger value of λ will force more entries of $\vec{w}$ to be zero — i.e. larger λ will imply higher sparsity.

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

$$\vec{w}^\star = \underset{\vec{w}\in\mathbb{R}^d}{\operatorname{argmin}} \ \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_2^2 . \tag{29}$$

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?

Solution: In the case of ridge regression the optimal weight vector $\vec{w}$ is given by

$$w_i^\star = \frac{\vec{y}^\top \vec{x}_i}{n + \lambda}, \ i = 1, \ldots, d. \tag{30}$$

So $w_i^\star$ is only zero when $\vec{y}^\top \vec{x}_i = 0$, in contrast to LASSO where $w_i^\star$ is zero when $\vec{y}^\top \vec{x}_i \in \left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right]$. This suggest that LASSO forces a lot of coordinates to be zero, i.e. induces sparsity to the optimal weight vector.

© UCB EECS 127/227AT, Spring 2024. All Rights Reserved. This may not be publicly shared without explicit permission.

4. More Fun with Lasso and Ridge

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

© UCB EECS 127/227AT, Spring 2024. All Rights Reserved. This may not be publicly shared without explicit permission.

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

$$P(\lambda) \doteq \operatorname*{argmin}_{\vec{x}} \{f(\vec{x}) + \lambda g(\vec{x})\} \tag{31}$$

$$C(k) \doteq \operatorname*{argmin}_{\vec{x}\colon\, g(\vec{x}) \leq k} f(\vec{x}). \tag{32}$$

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.

Solution: Since f is strictly convex and g is convex, both problems are convex with strictly convex objective. Thus both problems have at most one solution. Since we know from the assumptions that both problems have at least one solution, $P(\lambda)$ and $C(k)$ have exactly one element each.

We give some clarity on the fact that $P(\lambda)$ and $C(k)$ have at least one element each. This requires some analysis to prove, and so **it is all out of scope of the course**. We know that convex functions (hence also strictly convex functions) are continuous (see this MathStackExchange link). The assumption that $f(\vec{x}) \to \infty$ as $\|\vec{x}\|_2 \to \infty$ is called "coercivity". We know that continuous coercive functions on closed sets (such as $\mathbb{R}^n$ and $\{\vec{x} \in \mathbb{R}^n \mid g(\vec{x}) \leq k\}$) attain their global minima, i.e., their problems have at least one solution (see this MathStackExchange link). This is how we are able to say that $P(\lambda)$ and $C(k)$ have at least one element each.

(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.

Solution: Let $\lambda \geq 0$ and let $\vec{x}^\star \in P(\lambda)$, so that $P(\lambda) = \{\vec{x}^\star\}$. Set $k = g(\vec{x}^\star)$. Certainly $\vec{x}^\star$ is feasible for the constraint problem with this k. We claim that $\vec{x}^\star \in C(k)$, so that $C(k) = \{\vec{x}^\star\}$. Suppose for the sake of contradiction that $C(k) = \{\vec{z}\}$ where $\vec{z} \neq \vec{x}^*$. Then $\vec{z}$ must be feasible for the constraint problem, so $g(\vec{z}) \leq k = g(\vec{x}^\star)$, and it must be better than $\vec{x}^\star$, so $f(\vec{z}) < f(\vec{x}^\star)$. Then we have

$$f(\vec{z}) + \lambda g(\vec{z}) < f(\vec{x}^\star) + \lambda g(\vec{z}) \tag{33}$$

$$\leq f(\vec{x}^\star) + \lambda g(\vec{x}^\star) \tag{34}$$

so $\vec{x}^\star \notin P(\lambda)$, a contradiction. Thus $\vec{x}^\star \in C(k)$, i.e., $C(k) = \{\vec{x}^\star\}$.

(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 $\lambda = \mu^\star$.

© UCB EECS 127/227AT, Spring 2024. All Rights Reserved. This may not be publicly shared without explicit permission.

Solution: We know that there is a point $\vec{x}_0$ such that $g(\vec{x}_0) = 0 < k$. Thus the constraint problem has a strictly feasible point. Since the constraint problem is a convex problem with a strictly feasible point, strong duality holds for it. The Lagrangian of the constraint problem is given by

$$L_k(\vec{x}, \mu) = f(\vec{x}) + \mu(g(\vec{x}) - k). \tag{35}$$

Since strong duality holds, let $(\vec{x}^\star, \mu^\star)$ be optimal primal and dual variables for the constraint problem (so that $\vec{x}^\star \in C(k)$, i.e., $C(k) = \{\vec{x}^\star\}$) such that

$$p^\star = L_k(\vec{x}^\star, \mu^\star) = d^\star. \tag{36}$$

Since the constraint problem is convex, and strong duality holds, we have

$$\begin{aligned}
\vec{x}^\star &\in \operatorname*{argmin}_{\vec{x}} L_k(\vec{x}, \mu^\star) && (37)\\
&= \operatorname*{argmin}_{\vec{x}} \{f(\vec{x}) + \mu^\star(g(\vec{x}) - k)\} && (38)\\
&= \operatorname*{argmin}_{\vec{x}} \{f(\vec{x}) + \mu^\star g(\vec{x}) - \mu^\star k\} && (39)\\
&= \operatorname*{argmin}_{\vec{x}} \{f(\vec{x}) + \mu^\star g(\vec{x})\} && (40)\\
&= P(\mu^\star), && (41)
\end{aligned}$$

so $\vec{x}^\star \in P(\lambda)$, i.e., $P(\lambda) = \{\vec{x}^\star\}$, with $\lambda = \mu^\star$.

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:

$$\text{LASSO}(\lambda) \doteq \operatorname*{argmin}_{\vec{x}} \left\{ \|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_1 \right\} \tag{42}$$

and ridge regression:

$$\text{Ridge}(\lambda) \doteq \operatorname*{argmin}_{\vec{w}} \left\{ \|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2 \right\} \tag{43}$$

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:

$$\ell^1\text{CLS}(k) \doteq \operatorname*{argmin}_{\vec{x}\colon \|\vec{x}\|_1 \le k} \|A\vec{x} - \vec{y}\|_2^2 \tag{44}$$

$$\ell^2\text{CLS}(k) \doteq \operatorname*{argmin}_{\vec{x}\colon \|\vec{x}\|_2^2 \le k} \|A\vec{x} - \vec{y}\|_2^2 . \tag{45}$$

(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 $g(\vec{x}_0) = 0$.

Solution: For all formulations,

$$f(\vec{x}) \doteq \|A\vec{x} - \vec{y}\|_2^2 \tag{46}$$

has Hessian

$$\nabla_{\vec{x}}^2 f(\vec{x}) = 2A^\top A \tag{47}$$

© UCB EECS 127/227AT, Spring 2024. All Rights Reserved. This may not be publicly shared without explicit permission.

which is PD since A has full column rank. Thus f is strictly convex (and in fact is strongly convex), and in particular $f(\vec{x}_t) \to \infty$ as $t \to \infty$ for any $(\vec{x}_t)_{t=0}^{\infty}$ such that $\|\vec{x}_t\|_2 \to \infty$.

In the LASSO-ℓ^1CLS formulations, $g(\vec{x}) = \|\vec{x}\|_1$, which is convex since it is a norm. In the LASSO-ℓ^2CLS formulations, $g(\vec{x}) = \|\vec{x}\|_2^2$, which is (strictly) convex since it has Hessian $2I$ which is PD.

For both g, $g(\vec{0}) = 0$.

(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.

Solution: See the Jupyter notebook solutions.

© UCB EECS 127/227AT, Spring 2024. All Rights Reserved. This may not be publicly shared without explicit permission.

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.

© UCB EECS 127/227AT, Spring 2024. All Rights Reserved. This may not be publicly shared without explicit permission.