Statistical learning and data analysis

Assignment 3 - Norms, Regression and Regularization

NathanP

June 2, 2025

1. Norms

(a) Show that equation (1) defines a norm for $p \geq 1$

To be a norm, the function $\|\cdot\|_p$ must satisfy:

1. **Positive definiteness:** $\|x\|_p \geq 0$ and $\|x\|_p = 0 \iff x = 0$
2. **Homogeneity:** $\|\alpha x\|_p = |\alpha|\|x\|_p$
3. **Triangle inequality:** $\|x + y\|_p \leq \|x\|_p + \|y\|_p$

1. Positive definiteness: All $|x_i|^p \geq 0$, so the sum is non-negative. If $\|x\|_p = 0$, then $|x_i|^p = 0 \Rightarrow x_i = 0 \ \forall i$, so $x = 0$.

2. Homogeneity:

$$\|\alpha x\|_p = \left(\sum_{i=1}^{n} |\alpha x_i|^p\right)^{1/p} = \left(|\alpha|^p \sum_{i=1}^{n} |x_i|^p\right)^{1/p} = |\alpha| \cdot \|x\|_p$$

3. Triangle inequality: This follows from **Minkowski's inequality**:

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p \quad \text{for } p \geq 1$$

Hence, $\|\cdot\|_p$ is a norm for $p \geq 1$.

(b) Show that equation (1) does not define a norm for $p < 1$

For $0 < p < 1$, the triangle inequality does **not** hold.

Let $x = (1, 0)$, $y = (0, 1)$, $p = 0.5$:

$$\|x + y\|_p = \|(1, 1)\|_p = (1^{0.5} + 1^{0.5})^{1/0.5} = (2)^2 = 4$$

$$\|x\|_p + \|y\|_p = 1^2 + 1^2 = 2$$

$$\Rightarrow \|x + y\|_p = 4 > 2 = \|x\|_p + \|y\|_p$$

So, the triangle inequality fails $\Rightarrow$ not a norm.

(c) Does the triangle inequality hold for $p = 0$?

No. The expression:

$$\|x\|_0 := \#\{i : x_i \neq 0\}$$

counts the number of non-zero elements in x. It is not a norm because:

- It is not homogeneous: $\|\alpha x\|_0 = \|x\|_0$ if $\alpha \neq 0$
- It violates the triangle inequality:

$$\|x + y\|_0 \leq \|x\|_0 + \|y\|_0 \quad \text{may fail, e.g. overlapping support}$$

So $\|\cdot\|_0$ is not a norm.

(d) Show that for $p \geq 1$, the unit ball $B_p := \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$ is convex

Let $x_1, x_2 \in B_p$, and $0 < \alpha < 1$. Then:

$$\|\alpha x_1 + (1-\alpha)x_2\|_p \leq \alpha\|x_1\|_p + (1-\alpha)\|x_2\|_p \leq \alpha + (1-\alpha) = 1$$

Hence, $\alpha x_1 + (1-\alpha)x_2 \in B_p$, so B_p is convex.

(e) Show that for $0 < p < 1$, the unit ball B_p is not convex

Let $x_1 = (1,0), x_2 = (0,1) \Rightarrow \|x_1\|_p = \|x_2\|_p = 1 \Rightarrow x_1, x_2 \in B_p$

Now, take the midpoint:

$$z = \frac{1}{2}x_1 + \frac{1}{2}x_2 = \left(\frac{1}{2}, \frac{1}{2}\right)$$

$$\|z\|_p = \left(2 \cdot \left(\frac{1}{2}\right)^p\right)^{1/p} = (2^{1-p})^{1/p} = 2^{\frac{1-p}{p}} > 1 \quad \text{for } p < 1$$

So $z \notin B_p$, and B_p is not convex.

(f) Show that $\|A\|_F^2 = \mathbf{Tr}(A^\top A) = \mathbf{Tr}(AA^\top)$

Let $A \in \mathbb{R}^{m\times n}$, and define the Frobenius norm:

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2} \Rightarrow \|A\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2$$

Now observe:

$$\mathrm{Tr}(A^\top A) = \sum_{i=1}^{n}(A^\top A)_{ii} = \sum_{i=1}^{n}\sum_{k=1}^{m} A_{ki}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2 = \|A\|_F^2$$

Similarly,

$$\mathrm{Tr}(AA^\top) = \|A\|_F^2$$

(g) Provide a closed-form expression for the operator norm of $A \in \mathbb{R}^{m \times n}$ w.r.t. $p = 1$

The operator norm induced by the ℓ_1-norm is:

$$\|A\|_{1 \to 1} = \max_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}|$$

i.e., the maximum absolute column sum.

(h) Provide a closed-form expression for the operator norm of $A \in \mathbb{R}^{m \times n}$ w.r.t. $p = \infty$

The operator norm induced by the ℓ_∞-norm is:

$$\|A\|_{\infty \to \infty} = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|$$

i.e., the maximum absolute row sum.

(i*) Provide a closed-form expression for the operator norm of A w.r.t. $p = 2$

The $\ell_2 \to \ell_2$ operator norm is the largest singular value of A:

$$\|A\|_{2 \to 2} = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\top A)}$$

(j*) Show that the operator norm can equivalently be defined as:

$$\|A\|_{op} = \sup_{v \neq 0, v \in V} \frac{\|Av\|}{\|v\|} = \sup_{\|v\|=1, v \in V} \|Av\|$$

Proof: For any non-zero vector v, define $u = \frac{v}{\|v\|} \Rightarrow \|u\| = 1$, so:

$$\frac{\|Av\|}{\|v\|} = \left\| A\left(\frac{v}{\|v\|}\right) \right\| = \|Au\| \Rightarrow \sup_{v \neq 0} \frac{\|Av\|}{\|v\|} = \sup_{\|v\|=1} \|Av\|$$

(k) Show submultiplicativity for square matrices $A, B \in \mathbb{R}^{n \times n}$:

$$\|AB\| \le \|A\| \cdot \|B\|$$

Let $x \in \mathbb{R}^n$ with $\|x\| = 1$. Then:

$$\|ABx\| \le \|A\| \cdot \|Bx\| \le \|A\| \cdot \|B\| \cdot \|x\| = \|A\| \cdot \|B\| \Rightarrow \sup_{\|x\|=1} \|ABx\| \le \|A\| \cdot \|B\| \Rightarrow \|AB\| \le \|A\| \cdot \|B\|$$

2. Least Squares and Matrix Derivatives

(a) Show that $\nabla_\beta(z^\top \beta) = z$, where $z, \beta \in \mathbb{R}^{n\times 1}$

We apply the identity for gradients of linear functions:

$$\nabla_\beta(z^\top \beta) = \nabla_\beta(\beta^\top z) = z$$

This follows from the fact that $z^\top \beta$ is a scalar and the gradient of a scalar linear form is the coefficient vector.

(b) Show that $\nabla_\beta(\beta^\top H\beta) = 2H\beta$, where $H \in \mathbb{R}^{n\times n}$ is symmetric

We use the identity for the gradient of a quadratic form:

$$\nabla_\beta(\beta^\top H\beta) = (H + H^\top)\beta$$

Since H is symmetric ($H = H^\top$), we get:

$$\nabla_\beta(\beta^\top H\beta) = 2H\beta$$

(c) Given a sample $\{x_i\}_{i=1}^n \subset \mathbb{R}$, find

$$c_1^* = \arg\min_{c\in\mathbb{R}} \sum_{i=1}^n |x_i - c|$$

The solution to minimizing the sum of absolute deviations is the **median**:

$$c_1^* = \text{median}(x_1, x_2, \ldots, x_n)$$

(d) Given a sample $\{x_i\}_{i=1}^n \subset \mathbb{R}$, find

$$c_2^* = \arg\min_{c\in\mathbb{R}} \sum_{i=1}^n (x_i - c)^2$$

This is the classic least squares minimizer. Taking derivative and setting to zero:

$$\frac{d}{dc}\sum_{i=1}^n (x_i - c)^2 = -2\sum_{i=1}^n (x_i - c) = 0 \Rightarrow c_2^* = \frac{1}{n}\sum_{i=1}^n x_i$$

So the minimizer is the **mean** of the data:

$$c_2^* = \bar{x}$$

(e*) Given a sample $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$, find

$$c_1^* = \arg\min_{c\in\mathbb{R}^d} \sum_{i=1}^n \|x_i - c\|$$

There is no closed-form solution for this problem (sum of Euclidean distances). The solution is known as the **geometric median**, which must be computed using iterative algorithms (I read about Weiszfeld's algorithm).

(f) Given a sample $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$, find

$$c_2^* = \arg\min_{c\in\mathbb{R}^d} \sum_{i=1}^n \|x_i - c\|^2$$

This is the multivariate least squares problem. The objective is minimized when c is the **mean** of the vectors:

$$c_2^* = \frac{1}{n}\sum_{i=1}^n x_i$$

3. Regularized Least Squares Regression

We define Ridge and Lasso regression:

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \|X\beta - y\|_2^2 + \lambda_{\text{Ridge}}\|\beta\|_2^2$$

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \|X\beta - y\|_2^2 + \lambda_{\text{Lasso}}\|\beta\|_1$$

(a) Show that the closed-form solution to equation (2) is:

$$\hat{\beta}_{\text{Ridge}} = (X^\top X + \lambda I)^{-1} X^\top y$$

Solution:
We start with the Ridge objective function:

$$J(\beta) = \|X\beta - y\|_2^2 + \lambda\|\beta\|_2^2 = (X\beta - y)^\top (X\beta - y) + \lambda\beta^\top\beta$$

Take the gradient with respect to β:

$$\nabla_\beta J = 2X^\top(X\beta - y) + 2\lambda\beta$$

Set gradient to zero:

$$2X^\top X\beta - 2X^\top y + 2\lambda\beta = 0 \Rightarrow X^\top X\beta + \lambda\beta = X^\top y \Rightarrow (X^\top X + \lambda I)\beta = X^\top y$$

Solving for β, we get:

$$\hat{\beta}_{\text{Ridge}} = (X^\top X + \lambda I)^{-1} X^\top y$$

(b) Provide sufficient conditions for the invertibility of $X^\top X + \lambda I$

Answer:

- $X^\top X$ is a symmetric positive semi-definite matrix (PSD). - Adding λI, where $\lambda > 0$, results in a strictly positive definite matrix:

$$X^\top X + \lambda I \succ 0$$

Sufficient condition:

$$\lambda > 0$$

This ensures that $X^\top X + \lambda I$ is strictly positive definite and hence invertible.

If $X^\top X$ is not full rank (e.g. if X is not full column rank), $X^\top X$ may be singular — but the addition of λI ensures it becomes full rank and invertible as long as $\lambda > 0$.