Recitation 5

DS-GA 1013 Mathematical Tools for Data Science

1. Let $X \in \mathbb{R}^{m \times n}$ have rank n. Define $P \in \mathbb{R}^{m \times m}$ so that $P\vec{v}$ is the orthogonal projection of $\vec{v} \in \mathbb{R}^m$ onto the column space of X.
    1. Give a formula for P using the SVD $X = USV^T$.
    2. Give a formula for P just using X and algebraic operations (i.e., inverses, multiplications, transposes, etc.).

> **Solution:**
>
> 1. $P = UU^T$
> 2. $P = X(X^T X)^{-1} X^T$. To see this, recall that the least squares solution to $\min_{\vec{\beta}} \|X\vec{\beta} - \vec{y}\|_2^2$ is given by $\vec{\beta} = (X^T X)^{-1} X^T \vec{y}$, where $X\vec{\beta}$ is the projection we require. Alternatively, this can be seen to match the SVD based formula above.

2. You are given data matrices $X \in \mathbb{R}^{n \times d}$ and $\vec{y} \in \mathbb{R}^n$ and want to fit a least squares regression model $X\vec{\beta} = \vec{y}$ to find $\vec{\beta}$. You accidently duplicate the data so that each row of X and each entry of $\vec{y}$ occurs exactly twice. How will this effect the estimate of $\vec{\beta}$?

> **Solution:** It won't change, as we obtain an equivalent objective.

3. Under what conditions will training error increase if you add a feature to your regression problem? How does the answer change if you are using ridge regression?

> **Solution:** It never increases for both.

4. Suppose you fit a linear regression model model again but have scaled a feature by a factor of 10.
    1. Under what conditions will this change the new forecast $X\hat{\beta}$?
    2. What impact will this have on ridge regression?

> **Solution:**
>
> 1. Same forecast for standard regression.
> 2. It will have the effect of reducing the penalty on the corresponding coefficient.

5. The ridge regression estimator is given by

$$\vec{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T \vec{y}.$$

Under what conditions on X is this formula valid (i.e., does the inverse exist)?

Solution: It always exists since $X^T X$ is positive semidefinite and λI is positive definite.

6. Let $X \in \mathbb{R}^{n\times p}$, $\vec{\beta} \in \mathbb{R}^p$, $\vec{y} \in \mathbb{R}^n$, $\lambda > 0$, and $M \in \mathbb{R}^{m\times p}$ has full column rank. What is the solution to

$$\arg\min_{\vec{\beta}} \|X\beta - \vec{y}\|_2^2 + \lambda \|M\vec{\beta}\|_2^2?$$

Solution:

$$(X^T X + \lambda M^T M)^{-1} X^T y,$$

where the matrix is invertible since $M^T M$ is positive definite.

7. Suppose you are given data $\vec{y} = X\vec{\beta} + \vec{z}$ (all variables deterministic; $\vec{\beta}, \vec{z}$ unknown) and compute the least squares estimator $\hat{\vec{\beta}}$ for $\vec{\beta}$. Assuming $\|\vec{z}\|_2 = \eta$ is fixed, and X has full column rank, what direction for $\vec{z}$ produces the largest error $\|\hat{\vec{\beta}} - \vec{\beta}\|_2$, and how much is that error?

Solution: If $X \in \mathbb{R}^{n\times p}$ has SVD $X = USV^T$ (with $S \in \mathbb{R}^{p\times p}$) then the error $\hat{\vec{\beta}} - \vec{\beta} = VS^{-1}U^T\vec{z}$. This has maximum norm when the noise $\vec{z}$ points in the direction $U[:, p]$ giving a norm of η/σ_p. The resulting error will point in the direciton $V[:, p]$.

8. Let $\vec{\beta}_{\text{ridge}}$ denote the ridge regression estimator which minimizes

$$\arg\min_{\beta} \|X\vec{\beta} - \vec{y}\|_2^2 + \lambda \|\vec{\beta}\|_2^2.$$

Show that $\vec{\beta}_{\text{ridge}}$ is in the row space of X.

Solution: Write $\vec{\beta} = \vec{\beta}_r + \vec{\beta}_{r^\perp}$ where $\vec{\beta}_r$ is the orthogonal projection of $\vec{\beta}$ onto the row space of X, and $\vec{\beta}_{r^\perp}$ is the orthogonal projection of $\vec{\beta}$ onto the orthogonal complement of the row space of X. Then $X\vec{\beta} = X\vec{\beta}_r$ but $\|\vec{\beta}\|_2^2 = \|\vec{\beta}_r\|_2^2 + \|\vec{\beta}_{r^\perp}\|_2^2$.

9. Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$ be a given dataset. And consider the objective

$$L(h) = \frac{1}{n}\sum_{i=1}^{n} (h(x_i) - y_i)^2,$$

where $h : \mathbb{R} \to \mathbb{R}$. In each of the following, show how to solve the problem using linear least squares.

1. Minimize $L(h)$ for $h \in \mathcal{F}$ where

$$\mathcal{F} = \{h(x) = a + bx + cx^2 + dx^3 : a, b, c, d \in \mathbb{R}\}.$$

2. Minimize $L(h)$ for $h \in \mathcal{F}$ where $\mathcal{F}$ is the set of piecewise cubic polynomials with knot locations $\tau_1 < \cdots < \tau_k$. That is, there are a_j, b_j, c_j, d_j for $i = 1, \ldots, k+1$ such that

$$h(t) = a_j + b_j t + c_j t^2 + d_j t^3$$

for $\tau_{i-1} \leq h(t) < \tau_j$ where it is assumed that all of the data points lie in the interval (τ_0, τ_{k+1}).

Solution:

1. Construct a matrix $A \in \mathbb{R}^{n\times 4}$ where the ith row has the form $(1, x_i, x_i^2, x_i^3)$ and apply least squares.

2. Construct a matrix $A \in \mathbb{R}^{n\times 4(k+1)}$ where the ith row has $k+1$ blocks of 4 entries of the form

$$\begin{aligned}&\mathbf{1}(\tau_{j-1} \leq x_i < \tau_j), (x_i - \tau_{j-1})\,\mathbf{1}(\tau_{j-1} \leq x_i < \tau_j),\\ &(x_i - \tau_{j-1})^2\,\mathbf{1}(\tau_{j-1} \leq x_i < \tau_j), (x_i - \tau_{j-1})^3\,\mathbf{1}(\tau_{j-1} \leq x_i < \tau_j).\end{aligned}$$

   This corresponds to a basis for $\mathcal{F}$ containing functions of the form

$$f_{j,p}(x) = (x - \tau_{j-1})^p\,\mathbf{1}(\tau_{j-1} \leq x < \tau_j).$$

Here we list some useful facts about complex numbers. Below $z \in \mathbb{C}$ and $a, b \in \mathbb{R}$.

- $z = a + bi = \operatorname{Re}(z) + i\operatorname{Im}(z)$
- $(a + bi)(c + di) = ac - bd + (ad + bc)i$
- $|a + bi|^2 = a^2 + b^2 = (a + bi)(a - bi) = (a + bi)(\overline{a + bi})$
- $|zw| = |z||w|$, $|z + w| \leq |z| + |w|$
- $e^{a+bi} = e^a(cos(b) + i\sin(b))$, $e^z e^w = e^{z+w}$
- $|e^{a+bi}| = e^a$
- $z = \overline{z}$ if and only if $z \in \mathbb{R}$
- $z + \overline{z} = 2\operatorname{Re}(z)$ and $z - \overline{z} = 2i\operatorname{Im}(z)$
- $\langle \vec{x}, \vec{y} \rangle = \overline{\langle \vec{y}, \vec{x} \rangle}$
- $\langle c\vec{x}, \vec{y} \rangle = c\langle \vec{x}, \vec{y} \rangle$ and $\langle \vec{x}, c\vec{y} \rangle = \overline{c}\langle \vec{x}, \vec{y} \rangle$
- $\|\vec{x}\|^2 = \langle \vec{x}, \vec{x} \rangle$
- For $\vec{x} \in \mathbb{C}^n$, $\vec{x}^* := \overline{(\vec{x})}^T$
- For $A \in \mathbb{C}^{m\times n}$, $A^* = \overline{A}^T$
- For $\vec{x}, \vec{y} \in \mathbb{C}^n$, $\langle \vec{x}, \vec{y} \rangle = \vec{y}^*\vec{x} = \sum_{i=1}^n \overline{\vec{y}[i]}\vec{x}[i]$

10. Compute $z = 1 + e^{2\pi i/n} + e^{4\pi i/n} + \cdots + e^{2(n-1)\pi i/n}$ where $n \geq 1$.

Solution: If $n = 1$ then $z = n$. Otherwise $e^{2\pi i/n}z = z$ with $e^{2\pi i/n} \neq 1$ so $z = 0$.

11. Show that $|r_1e^{it} - r_2e^{is}| \geq |r_1 - r_2|$ for all $r_1, r_2 > 0$ and $t, s \in \mathbb{R}$.

Solution: By the triangle inequality

$$|r_1e^{it} - r_2e^{is}| + |r_2e^{is}| \geq |r_1e^{it}| \implies |r_1e^{it} - r_2e^{is}| \geq |r_1| - |r_2| = r_1 - r_2.$$

The same argument holds with the two terms flipped giving the result.

12. Prove that Cauchy-Schwarz holds in a complex inner product space:

$$|\langle \vec{x}, \vec{y} \rangle| \leq \|\vec{x}\|\|\vec{y}\|.$$

Solution: Since

$$\langle a\vec{x}, \vec{y} \rangle = a\langle \vec{x}, \vec{y} \rangle \quad \text{and} \quad \langle \vec{x}, a\vec{y} \rangle = \bar{a}\langle \vec{x}, \vec{y} \rangle$$

we see that proving the inequality with $\|\vec{x}\| = \|\vec{y}\| = 1$ is sufficient. Let $c \in \mathbb{C}$ with $|c| = 1$ and observe that

$$\begin{aligned} 0 &\leq \|c\vec{x} - \vec{y}\|^2 \\ &= \langle c\vec{x} - \vec{y}, c\vec{x} - \vec{y} \rangle \\ &= \|c\vec{x}\|^2 + \|\vec{y}\|^2 - \langle c\vec{x}, \vec{y} \rangle - \langle \vec{y}, c\vec{x} \rangle \\ &= 2 - 2\,\mathrm{Re}(c\langle \vec{x}, \vec{y} \rangle). \end{aligned}$$

Let c be given by

$$c = \frac{\overline{\langle \vec{x}, \vec{y} \rangle}}{|\langle \vec{x}, \vec{y} \rangle|}$$

so that the above gives

$$|\langle \vec{x}, \vec{y} \rangle| \leq 1$$

completing the proof.