Homework 1, CPSC 8420, Fall 2020

Sadeghi Tabas, Sadegh

September 28, 2020

Solution to Problem 1

we need to find μ and U_q for our objective then we start to get derivative based on μ

$\frac{\delta}{\delta\mu}(\sum_1^N (x_i - \mu - U_q v_i)^T (x_i - \mu - U_q v_i)) = \sum_{i=1}^N -2(x_i - \mu - U_q v_i) = 0 \hookrightarrow$
$\mu = \bar{x} - U_q(\frac{1}{N}\sum_{i=1}^N v_i)$
where $\bar{x}$ is mean of x_i.
Then, we get derivative wrt ν_i. We change the summation as following

$\sum_{i=1}^N [(x_i - \mu)^T - 2(x_i - \mu)^T U_q v_i + v_i^T U_q^T U_q^T v_i] \hookrightarrow$

$-2((x_i - \mu)^T U_q)^T + (U_q^T U_q^T + U_q^T U_q)v_i = 0 \hookrightarrow derivative: \quad U_q^T(x_i - \mu) = U_q^T U_q v_i = v_i$

since $U_q^T U_q = I_p$ so we use the μ_i in the equation

$v = \bar{x} - U_q U_q^T(\bar{x} - v) \hookrightarrow \mathrm{U}_q U_q^T(\bar{x} - \mu) = \bar{x} - \mu$
then μ can satisfy:

$(I - U_q U_q^T)(\bar{x} - \mu) = 0$
$\hookrightarrow \mathrm{v}_i = U_q^T(x_i - \bar{x})$

Solution to Problem 2

Since $(\mathbf{X}\phi)^T\mathbf{X}\phi = ||\mathbf{X}\phi||_2^2 = ||\mathbf{X}\phi|| \quad ||\mathbf{X}\phi||$ the above unconstrained problem is equivalent to the constrained problem

$$\begin{aligned} \underset{\phi}{\text{maximize}} \quad & (\mathbf{X}\phi)^T\mathbf{X}\phi \\ \text{subject to} \quad & ||\phi||_2^2 = 1 \end{aligned} \tag{1}$$

$\longmapsto$

$$\begin{aligned} \underset{\phi}{\text{maximize}} \quad & \phi^T\mathbf{X}^T\mathbf{X}\phi \\ \text{subject to} \quad & ||\phi||_2^2 = 1 \end{aligned} \tag{2}$$

we know that $\mathbf{X}^T\mathbf{X}$ is symmetic positive semi definite matrix. Therefore, eigenvalue decomposition is $\mathbf{X}^T\mathbf{X} = Q\Lambda Q$ where Q is an diagonal matrix, $QQ^T = I$ and Λ is a diagonal matrix with non-negative eigenvalues λ_i such that $\lambda_1 \geq \lambda_2 ... \geq \lambda_n \geq 0$. Therefore, $\phi^T\mathbf{X}^T\mathbf{X}\phi = \phi^T Q\Lambda Q\phi = w^T\Lambda w = \sum_{i=1}^{p} \lambda_i {w_i}^2$.

Since ϕ is constrained to have a norm of one, then we have for w as $|||w||_2 = ||Q^T\phi||_2 = ||\phi||_2 = 1$ w.r.t Q is orthogonal.

But, if we want to maximize the $\sum_{i=1}^{p} \lambda_i {w_i}^2$ under constraint $\sum_{i=1}^{p} {w_i}^2 = 1$ then the best choice is $w = e_1$ where $w_1 = 1$ and $w_i = 0$ for any $i1$. Then, back to main problem for ϕ, we get $\phi^* = Qe_1 = q_1$. q_1 is the first column of Q.

Therefore, the eigenvector corresponding to the largest eigenvalue of X^TX is q_1. Moreover, the numerical value of objective function is λ_1

Solution to Problem 3

Vanilla Reg loss function: $J(\beta) = ||y - A\beta||_2^2$
$\nabla_\beta J(\beta) = \nabla_\beta(||y - A\beta||_2^2) = 2A^T(y - A\beta) = 0 \hookrightarrow \hat{\beta} = (A^TA)^{-1}A^Ty$

Ridge Reg loss function: $J(\beta) = ||y - A\beta||_2^2 + \lambda||\beta||_2^2$
$\nabla_\beta J(\beta) = \nabla_\beta(||y - A\beta||_2^2 + \lambda||\beta||_2^2) = 2A^T(y - A\beta) + 2\lambda\beta = 0 \hookrightarrow \hat{\beta} = (A^TA + \lambda I)^{-1}A^Ty$

Lasso Reg loss function: $J(\beta) = \frac{1}{2}||y - A\beta||_2^2 + \lambda||\beta||_1$
for a simple problem of $A = I$ we have:
$J(\beta) = \frac{1}{2}||y - A\beta||_2^2 + \lambda||\beta||_1 = \sum_{j=1}^{d}\{\frac{1}{2}(y_j - \beta_j)^2 + \lambda|\beta_j|_1\}$
Since the loss is separable so we need to minimize individual terms, so we have:
$\hat{\beta} = argmin\{\frac{1}{2}(y - \beta)^2 + \lambda|\beta|\} = max(|y| - \lambda, 0)sign(y) = S_\lambda(y)$

Subset selection model loss function: $J(\beta) = \frac{1}{2}||y - A\beta||_2^2 + \lambda||\beta||_0$

In the subset problem with subset size k, the $||\beta||_0$ counts the number of non zeroes in β and is given by $||\beta||_0 = \sum_{i=1}^{p} 1(\beta_i \neq 0)$ where 1(.) denotes the indicator function. The cardinality constraint makes the problem NP-hard. So A straightforward approach to choosing a subset of variables for linear regression is to try all possible combinations and pick one that minimizes some criterion. For the relationship of different $\hat{\beta}$ please see figure 1 at the end of homework.

Solution to Problem 4

By using minimization of sum of absolute residuals, it will have equal emphasis to all the data and so it is resistant to the outliers (large residuals) while the least squares, by squaring the residuals, gives more weight to large residuals.

The K-medians takes the median value of the datapoints belonging to the same cluster to calculate its centroid.

k-means minimizes within-cluster variance, which equals squared Euclidean distances (squared deviations from the mean) while k-medians minimizes absolute deviations, which equals Manhattan distance. So it is more robust to outliers. So in the figure 2 the circle a is a outlayer data. based on each method we have: the the K-means clustring assumes the the circle a belongs to claster A while the K-medians assuming it belongs to cluster B.

Solution to Problem 5

5.1. $A = A^T$ is a symmetric matrix, so $Z = A^T A = A^2$. we have $Av = \lambda v$, then $Zv = A^2 v = \lambda^2 v$. So the A's eigenvectors are also the Z's eigenvectors with eigenvalue λ^2. So we can say that the A's eigenvector basis is also an eigenvector basis for the Z and also forms a complete system of singular vectors for A. So we have $\sigma^2 = \lambda^2$

5.2. We know that any symmetric matrix A is positive definite if and only if all of its eigenvalues are strictly positive, so $\lambda_j > 0$ (i), and on the other hand we know that a singular value cannot be a negative value (ii). As a result, based on the proof of 5.1. and (i) and (ii) we can conclude that $\sigma = \lambda$

proof of (i): assuming $A > 0$

so $u_j^T A u_j = u_j^T(\lambda_j u_j) = \lambda_j u_j^T u_j = \lambda_j ||u_j||^2 = \lambda_j > 0$

since $u_1, ..., u_n$ is an arthonormal basis for R^n, for any x, $x = c_1 u_1 + ... + c_n u_n$ so, $x^T A x = (c_1 u_1 + ... + c_n u_n)^T A(c_1 u_1 + ... + c_n u_n) = (c_1 u_1 + ... + c_n u_n)^T (c_1 \lambda_1 u_1 + ... + c_n \lambda_n u_n) = \lambda_1 c_1^2 + ... + \lambda_n c_n^2 > 0$

5.3. if $u_i \perp u_j$ then it's axiomatic that $u_i \perp -u_j$ then orthonormality constraint is still met. We know that $|Xu_i|^2 = |-Xu_i|^2$. Moreover, we know that basis vectors of a space need to be independent and orthogonal to each other whereas $u_i - u_i$. So, minus of basis vector couldn't expand the space.

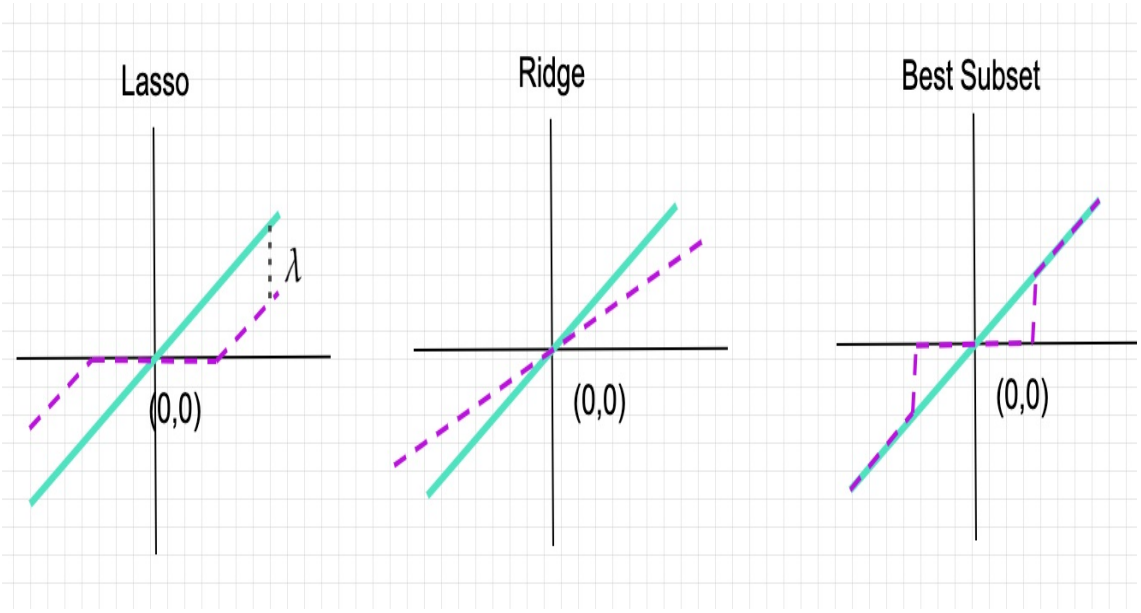

Figure 1: the relation of different $\hat{\beta}$

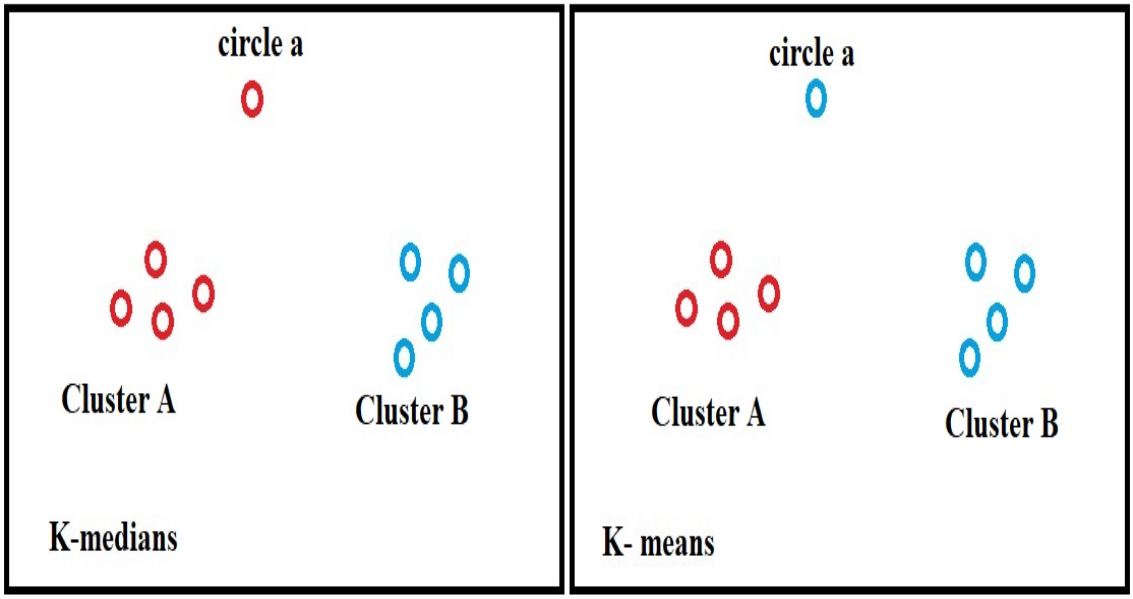

Figure 2: K-means vs. K-medians clustering methods