Homework Set 1, CPSC 8420, Fall 2020

Reference Solution

Due 09/28/2020, Monday, 11:59PM EST

Problem 1

For PCA, from the perspective of minimizing reconstruction error, please derive the solution to $\underset{\boldsymbol{\mu},\{\mathbf{v}_i\},\mathbf{U}_q}{\text{minimize}} \sum_{i=1}^{N} \|\mathbf{X}_i - \boldsymbol{\mu} - \mathbf{U}_q\mathbf{v}_i\|_2^2, s.t.\ \mathbf{U}_q^T\mathbf{U}_q = \mathbf{I}_q$, where $\mathbf{X} \in \mathbb{R}^{p\times N}, \boldsymbol{\mu} \in \mathbb{R}^p, \mathbf{U} \in \mathbb{R}^{p\times q}, \mathbf{v}_i \in \mathbb{R}^q$.
Denote $\mathbf{J} = \sum_{i=1}^{N} \|\mathbf{X}_i - \boldsymbol{\mu} - \mathbf{U}_q\mathbf{v}_i\|_2^2$*, taking the derivative w.r.t.* $\boldsymbol{\mu}$ *and set it to be 0:*

$$\begin{aligned}\frac{\partial \mathbf{J}}{\partial \boldsymbol{\mu}} &= -\sum_{i=1}^{N} 2 * (\mathbf{X}_i - \boldsymbol{\mu} - \mathbf{U}_q\mathbf{v}_i) = 0 \\ \implies \boldsymbol{\mu} &= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_i - \mathbf{U}_q\mathbf{v}_i) = \bar{\mathbf{X}} - \mathbf{U}_q\frac{\sum_{i=1}^{N}\mathbf{v}_i}{N}\end{aligned} \tag{1}$$

In addition, by taking the derivative w.r.t. $\mathbf{v}_{\boldsymbol{i}}$ *and set it to be 0:*

$$\begin{aligned}\frac{\partial \mathbf{J}}{\partial \mathbf{v}_{\boldsymbol{i}}} &= -2\mathbf{U}_q^T * (\mathbf{X}_i - \boldsymbol{\mu} - \mathbf{U}_q\mathbf{v}_i) = 0 \\ \implies \mathbf{v}_{\boldsymbol{i}} &= \mathbf{U}_q^T * (\mathbf{X}_i - \boldsymbol{\mu})\end{aligned} \tag{2}$$

Now plug in Eq. (2) to Eq. (1), we will have:

$$(\mathbf{I} - \mathbf{U}_q\mathbf{U}_q^T)(\bar{\mathbf{X}} - \boldsymbol{\mu}) = 0 \tag{3}$$

where we come the conclusion that $\boldsymbol{\mu} = \bar{\mathbf{X}}$ *optimizes the objective, then* $\mathbf{v}_{\boldsymbol{i}} = \mathbf{U}_q^T * (\mathbf{X}_i - \bar{\mathbf{X}})$*. Therefore, the we can reformulate the objective as* $\sum_{i=1}^{N} \|\mathbf{X}_i - \bar{\mathbf{X}} - \mathbf{U}_q\mathbf{U}_q^T * (\mathbf{X}_i - \bar{\mathbf{X}})\|_2^2$*, now denote* $\tilde{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}}$*, we have* $\sum_{i=1}^{N} \|\mathbf{X}_i - \bar{\mathbf{X}} - \mathbf{U}_q\mathbf{U}_q^T * (\mathbf{X}_i - \bar{\mathbf{X}})\|_2^2 = \sum_{i=1}^{N} \|\tilde{\mathbf{X}}_i - \mathbf{U}_q\mathbf{U}_q^T\tilde{\mathbf{X}}_i\|_2^2 = \|\tilde{\mathbf{X}} - \mathbf{U}_q\mathbf{U}_q^T\tilde{\mathbf{X}}\|_F^2 = \boldsymbol{tr}\left(\mathbf{U}_q^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{U}_q\right)$*, assume* $[\mathbf{U}, \boldsymbol{\Lambda}, \mathbf{U}] = svd(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)$*, then* $\mathbf{U}_q = \mathbf{U}[:, 1:q]$*.*

Problem 2

For PCA, from the perspective of maximizing variance, please show that the sotution of $\boldsymbol{\phi}$ to maximize $\|\mathbf{X}\boldsymbol{\phi}\|_2^2, s.t.\ \|\boldsymbol{\phi}\|_2 = 1$ is exactly the first column of $\mathbf{U}$, where $[\mathbf{U}, \mathbf{S}, \mathbf{U}] = svd(\mathbf{X}^T\mathbf{X})$. (Note: you need prove why it is optimal than any other reasonable combinations of $\mathbf{U}_i$, say $\hat{\boldsymbol{\phi}} = 0.8 * \mathbf{U}(:, 1) + 0.6 * \mathbf{U}(:, 2)$ which also satisfies $\|\hat{\boldsymbol{\phi}}\|_2 = 1$.)
Since columns of U *span the space, thus we assume* $\boldsymbol{\phi}=\sum_{i=1}^{p} \lambda_i\mathbf{U}(:, i)$*, and the constraint* $\|\boldsymbol{\phi}\|_2 = 1$ *is equivalent to* $\sum_{i=1}^{p} \lambda_i^2 = 1$*. Also we can rewrite* $\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{p} \mathbf{S}(i, i)\mathbf{U}(:, i)\mathbf{U}(:, i)^T$*, therefore* $\|\mathbf{X}\boldsymbol{\phi}\|_2^2 = \sum_{i=1}^{p} \mathbf{S}(i, i)\lambda_i^2 \leq \mathbf{S}(1, 1)\sum_{i=1}^{p} \lambda_i^2 = S(1, 1)$*, the equation holds iff* $\lambda_1 = 1$*.*

Problem 3

For vanilla linear regression model: $\min \|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|_2^2$, we denote the solution as $\hat{\boldsymbol{\beta}}_{LS}$; for ridge regression model: $\min \|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|_2^2 + \lambda * \|\boldsymbol{\beta}\|_2^2$, we denote the solution as $\hat{\boldsymbol{\beta}}_\lambda^{Ridge}$; for Lasso model: $\min \frac{1}{2}\|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|_2^2 + \lambda * \|\boldsymbol{\beta}\|_1$, we denote the solution as $\hat{\boldsymbol{\beta}}_\lambda^{Lasso}$; for Subset Selection model: $\min \frac{1}{2}\|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|_2^2 + \lambda * \|\boldsymbol{\beta}\|_0$, we denote the solution as $\hat{\boldsymbol{\beta}}_\lambda^{Subset}$, now please derive each $\hat{\boldsymbol{\beta}}$ given $\mathbf{y}, \mathbf{A}$($s.t.$ $\mathbf{A}^T\mathbf{A} = \mathbf{I}$), λ. Also, show the relationship of (each element in) $\hat{\boldsymbol{\beta}}_\lambda^{Ridge}, \hat{\boldsymbol{\beta}}_\lambda^{Lasso}, \hat{\boldsymbol{\beta}}_\lambda^{Subset}$ with (that in) $\hat{\boldsymbol{\beta}}_{LS}$ respectively. (up to 5 bonus points will be given if you illustrate the relationship with figures appropriately.)

1. ***Least Squares**: Denote $\mathbf{J} = \|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|_2^2$, then taking the derivative of $\mathbf{J}$ w.r.t. $\boldsymbol{\beta}$ and set it to be 0, we will get $\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} = \mathbf{A}^T\mathbf{y}$.*

2. ***Ridge Regression**: Following **Least Squares** by taking the derivative, we have: $\hat{\boldsymbol{\beta}}_\lambda^{Ridge} = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{y} = \frac{\mathbf{A}^T\mathbf{y}}{\lambda+1} = \frac{\hat{\boldsymbol{\beta}}_{LS}}{\lambda+1}$.*

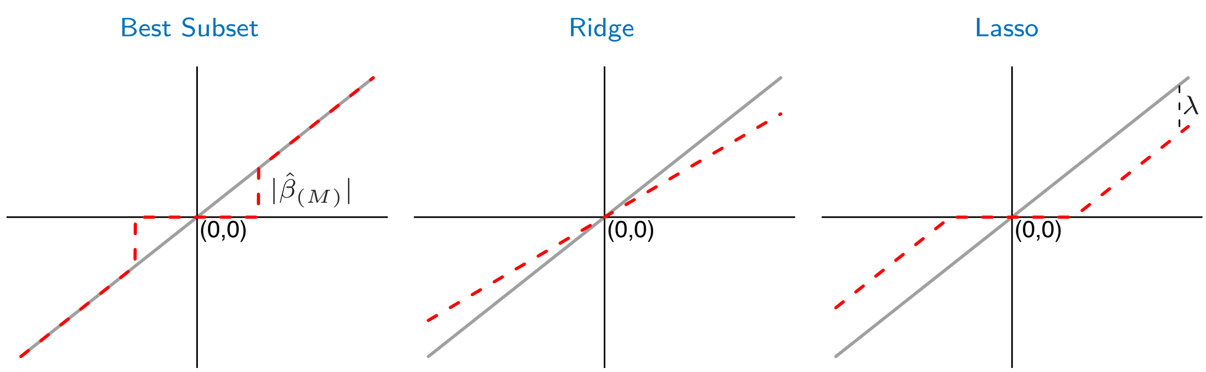
3. ***Lasso**: Since $\underset{\boldsymbol{\beta}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|_2^2 + \lambda * \|\boldsymbol{\beta}\|_1$ is equivalent to $\underset{\boldsymbol{\beta}}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{\beta} - \mathbf{A}^T\mathbf{y}\|_2^2 + \lambda * \|\boldsymbol{\beta}\|_1$, now for element in $\hat{\boldsymbol{\beta}}_\lambda^{Lasso}$, we divide into cases whether it is positive or negative. After some reformulation we have:*

$$\hat{\beta}^{Lasso} = \begin{cases} \mathbf{A}^T\mathbf{y} - \lambda = \hat{\boldsymbol{\beta}}_{LS} - \lambda & \textit{if} \quad \hat{\boldsymbol{\beta}}_{LS} > \lambda, \\ \mathbf{A}^T\mathbf{y} + \lambda = \hat{\boldsymbol{\beta}}_{LS} + \lambda & \textit{if} \quad \hat{\boldsymbol{\beta}}_{LS} \leq -\lambda, \\ 0 & \textit{else.} \end{cases} \tag{4}$$

4. ***Best Subset**: Since $\underset{\boldsymbol{\beta}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|_2^2 + \lambda * \|\boldsymbol{\beta}\|_0$ is equivalent to $\underset{\boldsymbol{\beta}}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{\beta} - \mathbf{A}^T\mathbf{y}\|_2^2 + \lambda * \|\boldsymbol{\beta}\|_0$, now for element in $\hat{\boldsymbol{\beta}}_\lambda^{Subset}$, we divide into cases whether it is 0 or not. After some reformulation we have:*

$$\hat{\beta}^{Subset} = \begin{cases} \hat{\boldsymbol{\beta}}_{LS} & \textit{if} \quad |\hat{\boldsymbol{\beta}}_{LS}| > \sqrt{2\lambda} := \hat{\boldsymbol{\beta}}_{(M)}, \\ 0 & \textit{else.} \end{cases} \tag{5}$$

In the figure below, the grey line represents Least Squares while red dash line denotes its variation in each section respectively.



Problem 4

Why might we prefer to minimize the sum of absolute residuals instead of the residual sum of squares for some data sets? Recall clustering method K-means when calculating the controid, it is to take the mean value of the datapoints belonging to the same cluster, so what about K-medians? What is its advantage over of K-means? Please use a synthetic (toy) experiment to illustrate your conclusion.

Since in real-world datasets, some data is noise or even outliers. By calculating the mean value may lead the centroid close to the outlier to avoid huge loss cost. However, the K-medians will take the median instead of mean. Assume we have data $\{-3, -2, 5, 6, 7, 20\}$, where point 20 is an outlier and positive or negative sign data should be in the same cluster respectively. Assume $0, 11$ are the initial centroids, then $\{-3, -2, 5\}$ will be in the same cluster, while the rest in another cluster via K-means. However, the clustering result is not correct. But if by making use of K-meadians, after several iterations $\{-3, -2\}$ will be in the same cluster, while the rest in another. This is the correct clustering in accordance to our preassumption.

Problem 5

Please show that:

1. if a matrix is symmetric, denote its eigenvalue and singular value as $\boldsymbol{\lambda}, \boldsymbol{\sigma}$ respectively (descending order in magnitude), then we have: $\boldsymbol{\lambda}^2 = \boldsymbol{\sigma}^2$.

   Assume $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, then $\mathbf{X}^T = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T$, then $\mathbf{X}^T\mathbf{X} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T$, which implies $\mathbf{X}^2\mathbf{V} = \mathbf{V}\boldsymbol{\Sigma}^2$, apparently each column of $\mathbf{V}$ is an eigenvector of $\mathbf{X}^2$, thus should be also eigenvector of $\mathbf{X}$, assume $\mathbf{X}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}$, then $\mathbf{X}^2\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}^2$, then $\boldsymbol{\lambda}^2 = \boldsymbol{\sigma}^2$.

2. if the matrix is symmetric and positive definite, then $\boldsymbol{\lambda} = \boldsymbol{\sigma}$.

   Consider $\mathbf{v}^T\mathbf{X}\mathbf{v} = \mathbf{v} * \lambda\mathbf{v} = \lambda\|\mathbf{v}\|^2$. Since $\mathbf{X}$ is positive definite, then $\lambda\|\mathbf{v}\|^2 = \mathbf{v}^T\mathbf{X}\mathbf{v} > 0$, therefore $\boldsymbol{\lambda} > 0$, as singuar value is always positive, and $\boldsymbol{\lambda}^2 = \boldsymbol{\sigma}^2$, then $\boldsymbol{\lambda} = \boldsymbol{\sigma}$.

3. for PCA, the loading vectors can be directly computed from the q columns of $\mathbf{U}$ where $[\mathbf{U}, \mathbf{S}, \mathbf{U}] = svd(\mathbf{X}^T\mathbf{X})$, please show that any $[\pm\mathbf{u}_1, \pm\mathbf{u}_2, \ldots, \pm\mathbf{u}_q]$ will be equivalent to $[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q]$ in terms of the same variance while satisying the orthonormality constraint.

   Obviously $\|\pm\mathbf{u}_i\|^2 = \|\mathbf{u}_i\|^2 = 1$. For $i \neq j, \langle\pm\mathbf{u}_i, \pm\mathbf{u}_j\rangle = \pm\langle\mathbf{u}_i, \mathbf{u}_j\rangle = 0$, that is the orthonormality constraint still holds. For variance $\|\mathbf{X}\mathbf{u}\|^2$, one can verify that $\|\mathbf{X}\mathbf{u}\|^2 = \boldsymbol{tr}\left(\mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u}\right) = \boldsymbol{tr}\left((-\mathbf{u})^T\mathbf{X}^T\mathbf{X}(-\mathbf{u})\right) = \|\mathbf{X}(-\mathbf{u})\|^2$, thus variance remains the same.