CS/ECE/ME532 Activity 12

Estimated time: 5 mins for Q1, 15 mins for Q2 (review), 20 mins for Q3, and 20 mins for Q4.

1. Let the n-by-p rank-r ($n > p > r$) matrix $\boldsymbol{X}$ have SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ where $\boldsymbol{U}$ is n-by-r, $\boldsymbol{\Sigma}$ is r-by-r, and $\boldsymbol{V}$ is p-by-r.

a) Find the SVD of $\boldsymbol{Z} = \boldsymbol{X}^T$ in terms of $\boldsymbol{U}, \boldsymbol{\Sigma}$, and $\boldsymbol{V}$.

b) Find the orthonormal basis for the best rank-1 subspace to approximate the rows of $\boldsymbol{Z}$ in terms of $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{\Sigma}$.

2. Uniqueness of solutions and Tikhonov regularization (ridge regression). The least-squares problem is $\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2$. Assume $\boldsymbol{X}$ is n-by-p with $p < n$.

a) Under what conditions is the solution to the least-squares problem not unique?

b) The Tikhonov-regularized least-squares problem is

$$\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_2^2$$

Show that this can be written as an ordinary least-squares problem $\min_{\boldsymbol{w}} ||\hat{\boldsymbol{y}} - \hat{\boldsymbol{X}}\boldsymbol{w}||_2^2$ and find $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{X}}$.

c) Use the results from the previous part to determine the conditions for which the Tikhonov-regularized least-squares problem has a unique solution.

3. Psuedoinverse and truncated SVD. The solution to the ridge regression problem

$$\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_2^2$$

is given by $\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. The *psuedoinverse* of $\boldsymbol{X}$, denoted $\boldsymbol{X}^\dagger$, can be defined by looking at the limit of the ridge regression solution as $\lambda \to 0$ (from above):

$$\boldsymbol{X}^\dagger = \lim_{\lambda \downarrow 0} (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T.$$

a) Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $p \leq n$, have SVD $\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^T = \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$. Show that

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T = \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i^2 + \lambda} \boldsymbol{v}_i \boldsymbol{u}_i^T.$$

Hint: Note that $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\Sigma^2\boldsymbol{V}^T$ *and* $\lambda\boldsymbol{I} = \boldsymbol{V}\lambda\boldsymbol{I}\boldsymbol{V}^T$.

b) Using the limit definition of the psuedoinverse above, show that when $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, then $\boldsymbol{X}^\dagger = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

c) Argue that when $\boldsymbol{X}$ is square and invertible, then $\boldsymbol{X}^\dagger = \boldsymbol{X}^{-1}$.

d) Argue that if $\boldsymbol{X}$ is rank $r < p$, then for $\lambda > 0$,

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \lambda} \boldsymbol{v}_i \boldsymbol{u}_i^T.$$

e) Now argue that if $\boldsymbol{X}$ is rank $r < p$,

$$\boldsymbol{X}^\dagger = \sum_{i=1}^{r} \frac{1}{\sigma_i} \boldsymbol{v}_i \boldsymbol{u}_i^T = \boldsymbol{V}\Sigma_r^{-1}\boldsymbol{U}^T$$

where Σ_r^{-1} is a matrix with $1/\sigma_i$ on the diagonal for $i = 1, \ldots, r$, and zero elsewhere.

4. The data file is available with a matrix $\boldsymbol{X}$ of 100 three-dimensional data points. A script is available with code to assist you with visualizing and fitting this data. Use the results of the SVD to find $\boldsymbol{a}$, a basis for the best (minimum sum of squared distances) one-dimensional subspace for the data.

a) Run the code to display the data in Figure the first figure. Use the rotate tool to inspect the scatter plot from different angles. Does the data appear to lie very close to a one-dimensional subspace? Does the data appear to be zero mean?

b) Figure 2 depicts the centered data and the one-dimensional subspace that contains the dominant feature you identified using the SVD. Use the rotate tool to inspect the data and one-dimensional subspace from different angles. Is a one-dimensional subspace a reasonable fit to the data? Comment on the error.

c) Now comment out (insert %) the line of code that subtracts the mean of the data. Does the dominant feature identified by SVD continue to be a good fit to the data? Comment on the importance of removing the mean before performing PCA.

1. Let the n-by-p rank-r $(n > p > r)$ matrix $\boldsymbol{X}$ have SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ where $\boldsymbol{U}$ is n-by-r, $\boldsymbol{\Sigma}$ is r-by-r, and $\boldsymbol{V}$ is p-by-r.

a) Find the SVD of $\boldsymbol{Z} = \boldsymbol{X}^T$ in terms of $\boldsymbol{U}, \boldsymbol{\Sigma}$, and $\boldsymbol{V}$.

b) Find the orthonormal basis for the best rank-1 subspace to approximate the rows of $\boldsymbol{Z}$ in terms of $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{\Sigma}$.

$$\underline{\underline{X}}_{n \times p} = \underline{U}_{n \times r}\ \underline{\Sigma}_{r \times r}\ \underline{V}^T_{r \times p}$$

rank r

Best approx for rows
- Left singular vectors
- (columns of left matrix)

Best approx for columns
- Right singular vectors
- (rows of right matrix)

a.) $\underline{Z} = \underline{X}^T = \underline{V}\ \underline{\Sigma}\ \underline{U}^T$

$p \times n \quad p \times n \quad p \times r \quad r \times r \quad r \times n$

b.) This is given by the first left singular vector of the decomposition.

In the case of $\underline{Z}$, that is the first column of $\underline{V}$.

2. **Uniqueness of solutions and Tikhonov regularization (ridge regression).** The least-squares problem is $\min_w ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2$. Assume $\boldsymbol{X}$ is n-by-p with $p < n$.

a) Under what conditions is the solution to the least-squares problem not unique?

b) The Tikhonov-regularized least-squares problem is

$$\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_2^2$$

Show that this can be written as an ordinary least-squares problem $\min_{\boldsymbol{w}} ||\hat{\boldsymbol{y}} - \hat{\boldsymbol{X}}\boldsymbol{w}||_2^2$ and find $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{X}}$.

c) Use the results from the previous part to determine the conditions for which the Tikhonov-regularized least-squares problem has a unique solution.

a.) $\underset{n\times p}{\underline{X}}$ $\begin{bmatrix} \underline{X} \end{bmatrix}$, $\min_{\underline{w}} \|\underline{y} - \underline{X}\,\underline{w}\|_2^2$

→ The least-squares problem does not have a unique solution when $\underline{X}$ is not full column rank.

- This can occur when $\underline{X}\,\underline{w} = \underline{y}$ has no solution or when $\underline{X}\,\underline{w} = \underline{y}$ has infinitely many solutions.

b.) $\min_{\underline{w}} \|\underline{X}\,\underline{w} - \underline{y}\|_2^2 + \lambda\|\underline{w}\|_2^2$

→ use identity: $\|\underline{a}\|_2^2 + \|\underline{b}\|_2^2 = \left\|\begin{bmatrix} \underline{a} \\ \underline{b} \end{bmatrix}\right\|_2^2$

→ $\left\|\begin{bmatrix} \underline{X}\,\underline{w} - \underline{y} \\ \lambda^{1/2}\,\underline{w} \end{bmatrix}\right\|_2^2$ → $\left\|\underbrace{\begin{bmatrix} \underline{X} \\ \lambda^{1/2}\,\underline{I} \end{bmatrix}}_{\tilde{X}}\underline{w} - \underbrace{\begin{bmatrix} \underline{y} \\ \underline{0} \end{bmatrix}}_{\tilde{y}}\right\|_2^2$

→ $\|\tilde{\underline{X}}\,\underline{w} - \tilde{\underline{y}}\|_2^2$

c.) The Tikhonov-reg LS equation always has a unique solution because $\tilde{\underline{X}}$ is guaranteed to be full rank due to the addition of $\lambda\underline{I}$.

a) Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $p \leq n$, have SVD $\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^T = \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$. Show that

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T = \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_i^2 + \lambda} \boldsymbol{v}_i \boldsymbol{u}_i^T.$$

Hint: Note that $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\Sigma^2\boldsymbol{V}^T$ *and* $\lambda\boldsymbol{I} = \boldsymbol{V}\lambda\boldsymbol{I}\boldsymbol{V}^T$.

b) Using the limit definition of the psuedoinverse above, show that when $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, then $\boldsymbol{X}^\dagger = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

c) Argue that when $\boldsymbol{X}$ is square and invertible, then $\boldsymbol{X}^\dagger = \boldsymbol{X}^{-1}$.

d) Argue that if $\boldsymbol{X}$ is rank $r < p$, then for $\lambda > 0$,

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \lambda} \boldsymbol{v}_i \boldsymbol{u}_i^T.$$

e) Now argue that if $\boldsymbol{X}$ is rank $r < p$,

$$\boldsymbol{X}^\dagger = \sum_{i=1}^{r} \frac{1}{\sigma_i} \boldsymbol{v}_i \boldsymbol{u}_i^T = \boldsymbol{V}\Sigma_r^{-1}\boldsymbol{U}^T$$

where Σ_r^{-1} is a matrix with $1/\sigma_i$ on the diagonal for $i = 1, \ldots, r$, and zero elsewhere.

a.) $(X^TX + \lambda I)^{-1} X^T$

$= (V \Sigma U^T \cancel{U} \Sigma V^T + \lambda I)^{-1} V \Sigma U^T$ (I)

$= (V \Sigma^2 V^T + V \lambda I V^T)^{-1} V \Sigma U^T$

$= (V(\Sigma^2 + \lambda I) V^T)^{-1} V \Sigma U^T$

$= V(\Sigma^2 + \lambda I)^{-1} \cancel{V^T V} \Sigma U^T$ (I)

$= V \underbrace{(\Sigma^2 + \lambda I)^{-1} \Sigma U^T}_{①}$

$= \sum_{i=1}^{p} \frac{\delta_i}{\delta_i^2 + \lambda} v_i u_i^T$

① $\begin{bmatrix} 1/\delta_1^2+\lambda & & 0 \\ & \ddots & \\ 0 & & 1/\delta_p^2+\lambda \end{bmatrix} \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_p \end{bmatrix}$

$= \begin{bmatrix} \delta_1/\delta_1^2+\lambda & & 0 \\ & \ddots & \\ 0 & & \delta_p/\delta_p^2+\lambda \end{bmatrix}$

This is just a scaling matrix that scales each v_i and u_i^T by $\delta_i/\delta_i^2+\lambda$

b.) $$\boldsymbol{X}^{\dagger} = \lim_{\lambda \downarrow 0} (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T.$$

if $\underline{X}^T \underline{X}$ is invertible, then as $\lambda \downarrow 0$,

$$X^{\dagger} = (\underline{X}^T \underline{X} + 0\,\underline{I})^{-1} \underline{X}^T = (\underline{X}^T \underline{X})^{-1} \underline{X}^T.$$

c.) When $\underline{X}$ is square and invertible,

$$\underline{X}^{\dagger} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T = \underline{X}^{-1} \underline{X}^{-T} \underline{X}^T = \underline{X}^{-1}.$$

this becomes $\underline{I}$ because $\underline{X}^{-1} \underline{X} = \underline{I}$, and this property holds for the transpose as well.

$$\underline{X}\,\underline{X}^{-1} = \underline{I} \leftrightarrow (\underline{X}\,\underline{X}^{-1})^T = \underline{I}^T$$

$$\Rightarrow \underline{X}^{-T} \underline{X}^T = \underline{I}.$$

d.) In this case, return to the term:

$$(\underline{X}^T\underline{X} + \lambda \underline{I})^{-1}\underline{X}^T = \underline{V}\,\underbrace{(\underline{\mathcal{E}}^2 + \lambda \underline{I})^{-1}\underline{\mathcal{E}}}_{①}\,\underline{U}^T$$

① $$\begin{bmatrix} 1/\delta_i^2+\lambda & & \underline{0} \\ & \ddots & \\ & & 1/\delta_r^2+\lambda \\ \underline{0} & & \underline{0} \ddots \end{bmatrix} \begin{bmatrix} \delta_i & & \underline{0} \\ & \ddots & \\ & & \delta_r \\ \underline{0} & & \ddots \underline{0} \end{bmatrix} = \begin{bmatrix} \delta_i/\delta_i^2+\lambda & & \underline{0} \\ & \ddots & \\ & & \delta_r/\delta_r^2+\lambda \\ \underline{0} & & \ddots \underline{0} \end{bmatrix}$$

Because now there are only $r < p$ nonzero singular values.

$$(\underline{X}^T\underline{X} + \lambda \underline{I})^{-1}\underline{X}^T = \sum_{i=1}^{r} \frac{\delta_i}{\delta_i^2+\lambda}\,\underline{v}_i\,\underline{u}_i^T$$

e.) $\underline{X}^+$ is the limit as $\underline{\lambda} \downarrow 0$ for the expression above.

$$\underline{X}^+ = \sum_{i=1}^{r} \frac{\delta_i}{\delta_i^2+0}\,\underline{v}_i\,\underline{u}_i^T = \sum_{i=1}^{r} \frac{1}{\delta_i}\,\underline{v}_i\,\underline{u}_i^T = \underline{V}\,\underline{\mathcal{E}}^{-1}\,\underline{U}^T$$

Back to that step, but with $\lambda = 0$ ...

$$(\underline{X}^T\underline{X})^{-1}\underline{X}^T = \underline{V}\,(\underline{\mathcal{E}}^2)^{-1}\underline{\mathcal{E}}\,\underline{U}^T = \underline{V}\,\underline{\mathcal{E}}^{-2}\underline{\mathcal{E}}\,\underline{U}^T = \underline{V}\,\underline{\mathcal{E}}^{-1}\,\underline{U}^T$$

4a.) Yes, the data appears to lie very close to a line (1-d subspace)

No, the data appears to be skewed off the mean in the $-ve$ x_1 direction.

4b.) The line produced captures a lot of the variance in the data.

There is still a decent amount of variance in the x_3 direction.

4c.) No, the line now points in a completely different direction.

Removing the mean before performing PCA is crucial to ensure the correctly-oriented principal components are identified.

$\underline{X}\,\underline{w} = \underline{y}$ $\quad \text{rank}[\underline{X}] \leq \text{rank}[\underline{X} \vdots \underline{y}] \rightarrow$ No solution

$\text{rank}[\underline{X} \vdots \underline{y}] = \text{rank}[\underline{X}] \rightarrow \geq 1$ exact solution

↳ (rank deficient)

- $\text{rank}(\underline{X}) < P \rightarrow \infty$ many solutions
- $\text{rank}(\underline{X}) = P \rightarrow$ unique solution. (full rank)

$\min_{\underline{w}} \| \underline{X}\underline{w} - \underline{y} \|_2^2$

System has an exact, unique solution:

$w_{min} = \underline{X}^{-1}\underline{y}$, unique

system has no solution:

if $\underline{X}$ is full column rank $\rightarrow X^T X$ is invertible $\rightarrow \underline{w}_{min}$ unique

$\underline{X}$ is not full column rank:

if $\underline{X}$ is not full column rank $\rightarrow X^T X$ is not invertible $\rightarrow$