DSC 140B - Homework 04

Due: Wednesday, May 3

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope at 11:59 PM.

Problem 1.

Suppose you have a data set of points X in $\mathbb{R}^{100}$ and wish to use PCA to reduce the dimensionality to 50. Consider these two approaches:

- Approach 1: Run PCA once to go directly from $\mathbb{R}^{100}$ to $\mathbb{R}^{50}$, constructing a new data set Z_1.
- Approach 2: First run PCA with $k = 75$ to create an intermediate data set Z' of points in $\mathbb{R}^{75}$, then run PCA with $k = 50$ on Z' to create a new data set Z_2.

Is there any difference between the two approaches? The correct answer is: no, there is not. That is, $Z_1 = Z_2$. You will show this below.

In this problem, assume that X is an $n \times d$ matrix of n data points in $\mathbb{R}^d$; furthermore, assume the data are centered. Let C be the covariance matrix of the original data. Let C' be the covariance matrix of Z' (the intermediate data in approach #2). Let U_{75} be a 100×75 matrix consisting of the top 75 eigenvectors of C, and let U_{50} be a 100×50 matrix consisting of the top 50 eigenvectors of C. Then the new PCA features in approach 1 are $Z_1 = XU_{50}$, and the intermediate PCA features in approach 2 are $Z' = XU_{75}$.

Throughout this problem you may assume for simplicity that all eigenvalues are unique.

a) Recall that C' is the covariance matrix of Z', the intermediate data in approach #2. Show that C' is a diagonal matrix.

Hint: $C' = \frac{1}{n}(Z')^T Z'$. Also remember that for general matrices AB, $(AB)^T = B^T A^T$.

Solution:

$$C' = \frac{1}{n}(Z')^T Z'$$

Making the substitution $Z' = XU_{75}$:

$$= \frac{1}{n}(XU_{75})^T (XU_{75})$$
$$= \frac{1}{n} U_{75}^T X^T X U_{75}$$

Since $X^T X = C$:

$$= \frac{1}{n} U_{75}^T C U_{75}$$

Now, the columns of U_{75} are all eigenvectors of C. This means that $U_{75}^T C U_{75} = D$, where D is a diagonal 75×75 matrix consisting of the eigenvalues of C. Therefore, we have shown that C' is diagonal.

b) The data set Z_2 is computed by multiplying the intermediate data set Z' by a 75×50 matrix U' consisting of the top 50 eigenvectors of the covariance matrix C'.

Argue that U' is the matrix where entry $u'_{ii} = 1$ and all other entries are zero. That is, it is a kind of rectangular identity matrix.

Solution: We have shown that the covariance matrix is diagonal. Therefore its eigenvectors are all standard basis vectors (we saw this in lecture: diagonal covariance matrices have axis-aligned eigenvectors).

Each eigenvector is a vector in $\mathbb{R}^{75}$, so gathering the top 50 such eigenvectors results in an 75×50 matrix U' described above.

c) Using what we have learned above, show that $Z_2 = XU_{50}$, and is therefore equal to Z_1.

Hint: $Z_2 = Z'U'$. Start by substituting for both U' and Z'.

Solution:

$$\begin{aligned} Z_2 &= Z'U' \\ &= XU_{75}U' \end{aligned}$$

But U' is something like an indentity matrix, as described above. When we multiply $U_{75}U'$, the result is a 100×50 matrix consisting of the first 50 columns of U_{75}. But this is just U_{50}. Therefore:

$$\begin{aligned} &= XU_{50} \\ &= Z_1 \end{aligned}$$