DSC 190 - Homework 04

Due: Wednesday, April 27

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. (Extra Credit)

We'd like to hear how things are going. Please consider filling out the feedback form at

`https://forms.gle/i5UyG9SA3oBWmJnN8`.

Your feedback is especially important for this class, since it is only the second offering. Your honest opinion will help make this class better. The feedback form is totally anonymous (as long as you don't enter your email!)

If you fill out this form, respond to this question and let us know and you'll get 2 points of extra credit towards this homework.

Problem 2.

As a data scientist, there will be many times when you will be working with massive, high dimensional data sets consisting of hundreds of thousands or even millions of points. This is not one of those times.

In this problem, we'll work with the following data set of three points:

$$x^{(1)} = (1, 3)^T$$
$$x^{(2)} = (-3, -9)^T$$
$$x^{(3)} = (2, 6)^T$$

a) Compute the sample covariance[1] matrix by hand. Show the calculations for each entry of the matrix.

Solution: We have three numbers to compute: the variance of feature 1, the variance of feature two, and the covariance.

The variance of the first feature:

$$\begin{aligned}\sigma_{11} &= \frac{1}{3}(1^2 + (-3)^2 + 2^2) \\ &= \frac{1}{3}(1 + 9 + 4) \\ &= \frac{14}{3}\end{aligned}$$

[1] Use the version of the sample covariance defined in lecture, not the one that divides by $n - 1$.

The variance of the second feature:

$$\begin{aligned}\sigma_{22} &= \frac{1}{3}(3^2 + (-9)^2 + 6^2) \\ &= \frac{1}{3}(9 + 81 + 36) \\ &= \frac{126}{3} \\ &= 42\end{aligned}$$

And the covariance:

$$\begin{aligned}\sigma_{12} &= \frac{1}{3}(1 \times 3 + (-3) \times (-9) + 2 \times 6) \\ &= \frac{1}{3}(3 + 27 + 12) \\ &= \frac{42}{3} \\ &= 14\end{aligned}$$

Therefore, the covariance matrix is

$$\begin{pmatrix} \frac{14}{3} & 14 \\ 14 & 42 \end{pmatrix}$$

b) What is the top eigenvector of the covariance matrix? You do not need to calculate the eigenvector explicitly, but you should justify your answer.

Hint: plot the data.

Solution: If we plot the data, we see that the three points are collinear – they all fall on the line of slope 3 through the origin. In other words, all three points are in the direction $(1, 3)^T$. This the direction of maximum variance, which we know is what the top eigenvector of the covariance matrix encodes.

Therefore, the top eigenvector of the covariance matrix is $(1, 3)^T$. Or, if you prefer normalized eigenvectors: $\frac{1}{\sqrt{10}}(1, 3)^T$.

c) What is the eigenvalue associated with the top eigenvector?

Solution: We know that the top eigenvector $\vec{u} = (1, 3)^T$. Since it is an eigenvector, $C\vec{u} = \lambda\vec{u}$. So we can multiply C and $\vec{u}$ to find λ.

We have:

$$\begin{aligned}C\vec{u} &= \begin{pmatrix} \frac{14}{3} & 14 \\ 14 & 42 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} \\ &= \begin{pmatrix} \frac{14}{3} + 42 \\ 14 + 126 \end{pmatrix} \\ &= \begin{pmatrix} \frac{140}{3} \\ 140 \end{pmatrix} \\ &= \frac{140}{3} \begin{pmatrix} 1 \\ 3 \end{pmatrix}\end{aligned}$$

So the top eigenvalue is 140/3.

d) Reduce the dimensionality of each point above by carrying out PCA by hand. Be sure to use the normalized eigenvector. Show your calculations.

Hint: one of your new features should be equal to $-3\sqrt{10}$.

Solution: To compute the new representation $z^{(i)}$ of point $x^{(i)}$, we carry out $x^{(i)} \cdot \vec{u}$, where $\vec{u}$ is the (normalized) top eigenvector of C. Therefore:

$$\begin{aligned} z^{(1)} &= x^{(1)} \cdot u \\ &= (1,3)^T \cdot \frac{1}{\sqrt{10}}(1,3)^T \\ &= \frac{1}{\sqrt{10}}(1+9) \\ &= \sqrt{10} \end{aligned}$$

$$\begin{aligned} z^{(2)} &= x^{(2)} \cdot \vec{u} \\ &= (-3,-9)^T \cdot \frac{1}{\sqrt{10}}(1,3)^T \\ &= \frac{1}{\sqrt{10}}(-3-27) \\ &= \frac{-30}{\sqrt{10}} \\ &= -3\sqrt{10} \end{aligned}$$

$$\begin{aligned} z^{(3)} &= x^{(3)} \cdot \vec{u} \\ &= (2,6)^T \cdot \frac{1}{\sqrt{10}}(1,3)^T \\ &= \frac{1}{\sqrt{10}}(2+18) \\ &= \frac{20}{\sqrt{10}} \\ &= 2\sqrt{10} \end{aligned}$$

e) The result of PCA is a data set consisting of three numbers. Compute the variance of these three numbers.

Hint: the result should be familiar.

Solution: The variance of the new features is

$$\begin{aligned}\frac{1}{3}\left[(z^{(1)})^2+(z^{(2)})^2+(z^{(3)})^2\right] &= \frac{1}{3}\left[\left(\sqrt{10}\right)^2+\left(-3\sqrt{10}\right)^2+\left(2\sqrt{10}\right)^2\right]\\ &= \frac{1}{3}\left[10+90+40\right]\\ &= \frac{140}{3}\end{aligned}$$

Which is, coincidentally, the top eigenvalue of C.