Math 343 Lab 6: Statistics

Objective

In this lab, we explore the use of linear algebra in statistics. Specifically, we discuss the notion of correlation and relate it to the inner product.

Shifting Data by the Mean

Consider the table below representing students scores in a class.

Student	Homework	Exam 1	Exam 2	Final
S1	89	91	77	75
S2	67	72	76	66
S3	72	77	69	70
S4	56	60	55	61
S5	92	98	89	86
S6	83	88	90	84
S7	45	60	55	48
Average	72	78	73	70

We can shift our data set by subtracting each column by its average value. This makes it so that the average of each column in the matrix below is zero. Hint: If W represents the matrix of scores, the following Matlab command will subtract out the average. Why does this work?

```
>> X = W - ones(7,1)*mean(W)
```

$$X = \begin{bmatrix} 17 & 13 & 4 & 5 \\ -5 & -6 & 3 & -4 \\ 0 & -1 & -4 & 0 \\ -16 & -18 & -18 & -9 \\ 20 & 20 & 16 & 16 \\ 11 & 10 & 17 & 14 \\ -27 & -18 & -18 & -22. \end{bmatrix}$$

Angles Between Vectors

Inner products give information about lengths of vectors and angles between vectors. Recall that the angle θ between two vectors is given by

$$\cos\theta = \frac{\langle x, y\rangle}{\|x\|\|y\|},$$

where the norm (or length) of a vector is $\|x\| = \sqrt{\langle x, x\rangle}$. Note that the usual inner product in $\mathbb{R}^n$ is given by

$$\langle x, y\rangle = x^T y.$$

Alternatively, if we can first divide our vectors by their length (these are called unit vectors) and then take the inner product

$$\cos\theta = \langle \frac{x}{\|x\|}, \frac{y}{\|y\|}\rangle,$$

Hence, we can find the cosine of the angles between our data columns in X by dividing each column by its length. There are a couple of ways of doing this. The following is an interesting way of doing it. Be sure to explore why it works:

```
>> Y = X ./ (ones(7,1)*sqrt(diag(X'*X)).')
```

Hence we have:

$$Y = \begin{bmatrix} 0.3985 & 0.3533 & 0.1139 & 0.1537 \\ -0.1172 & -0.1631 & 0.0854 & -0.1230 \\ 0 & -0.0272 & -0.1139 & 0 \\ -0.3750 & -0.4892 & -0.5124 & -0.2767 \\ 0.4688 & 0.5435 & 0.4555 & 0.4919 \\ 0.2578 & 0.2718 & 0.4839 & 0.4304 \\ -0.6329 & -0.4892 & -0.5124 & -0.6764 \end{bmatrix}$$

Finally, we get the cosines of the angles between columns by computing $Y^T Y$. In Matlab, that's just

```
>> Y' * Y
```

This yields

$$Y^TY = \begin{bmatrix} 1.0000 & 0.9778 & 0.8901 & 0.9491 \\ 0.9778 & 1.0000 & 0.9098 & 0.9249 \\ 0.8901 & 0.9098 & 1.0000 & 0.9277 \\ 0.9491 & 0.9249 & 0.9277 & 1.0000 \end{bmatrix}.$$

Note that the (j, k) entry of Y^TY corresponds to the cosine of the angle between the j^{th} and k^{th} columns. We remark that the diagonals are always equal to one because the angle between a vector and itself is zero and the cosine of zero is one.

Correlation

We remark that the cosine of the angle between two vectors is sometimes called the correlation coefficient. Two columns are said to be

- Perfectly correlated if the cosine of the angle between them is one.
- Positively correlated if the cosine of the angle between them is between zero and one.
- Uncorrelated if the cosine of the angle between them is zero.
- Negatively correlated if the cosine of the angle between them is between negative one and zero.
- Perfectly anticorrelated if the cosine of the angle between them is negative one.

The notion of correlation is important in establishing the relationships between measurements. For example, there is a high correlation between those who smoke and those who get lung cancer. It's important to understand that the high correlation along does not necessarily imply that smoking causes lung cancer, it only links them statistically. This is the famous issue of causation vs. correlation. For example, there is a high correlation between crime rates and sales of ice cream. This doesn't mean that ice cream causes crime or that increases in crime makes people want to eat more ice cream, but both rates do go up in the summer.

Assignment

Problem 1. *Download the file lab6.txt from the following link:*

`http://www.math.byu.edu/~jeffh/teaching/m343/labs/lab6.txt`

You can load this datafile by typing

```
load lab6.txt
```

Note that the data is available in the matrix `lab6`*. This consists of two columns. Find the correlation coefficient of the two columns. Then plot the original data and see if the value that you got is reasonable. Finish off the problem with a discussion of what you've learned.*