Stat 5650

Statistical Learning and Data Mining I

Homework #1

Due: *Wednesday, January* 29.

The purpose of this homework is to get you comfortable using R to carry out basic matrix and vector calculations.

The turtle data contains measurements (length, width, height) on male and female turtles of the same species. If you are familiar with SAS but not R, you can check many of the calculations in SAS but please do the calculations in R.

1. Graphically summarize the distributions of the three variables using boxplots, histograms, and normal quantile plots. Do the summaries for the combined data and for each gender of turtle.

2. Compute the covariance matrices and the correlation matrices for the male and female turtles, and visually compare them. (Later we will determine how to formally compare covariance matrices and mean vectors for different groups.)

3. For one of the genders, compute the matrix $T = \frac{1}{n-1}\left(\boldsymbol{Y}^T\boldsymbol{Y} - n\overline{\boldsymbol{y}}\overline{\boldsymbol{y}}^T\right)$ and compare it with the covariance matrix you previously obtained.

4. For one of the covariance matrices, compute the eigenvalues and eigenvectors of the inverse of the covariance matrix. What is the relationship between the eigenvalues and eigenvectors of a covariance matrix and its inverse?

5. Letting $\boldsymbol{u}_1$ denote the eigenvector corresponding to the largest eigenvector, λ_1. Verify that $\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1\boldsymbol{u}_1$ and $\boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1$.

6. The length, width, and height are all measured in the same units. Is there any reason we might prefer to use the correlation matrices over the covariance matrices if we were to carry out principal components analysis on these data?

7. Compute the eigenvalues and eigenvectors for the covariance and the correlation matrix for one of the genders of turtle, and compare the eigenvectors. In each case, how many principal components would you recommend retaining?