CS5785 Prelim

Wednesday March 26, 2014

On this exam you are allowed to use a calculator and one 8.5" by 11" sheet of notes. The total number of points possible is 30. Write your answers in a blue book or on separate blank sheets of paper. In order to get full credit you must show all your work. Good luck!

1. Given: a set of N observations $x_i \in \mathbb{R}^p$, $i = 1, \ldots, N$. Let $\mathbf{X}$ denote the matrix formed by stacking up all the observations into the rows of an $N \times p$ matrix. Assume the data is centered.

    (a) (2 pts.) Write down the expression for the covariance matrix $\mathbf{S}$ as a sum of outer products.

    (b) (2 pts.) Express $\mathbf{S}$ as a matrix product.

    (c) (2 pts.) Show that the eigenvalues of $\mathbf{S}$ are nonnegative.

    (d) (1 pt.) If some of the eigenvalues are zero, what does it mean about our observations?

    (e) (1 pt.) What does the leading eigenvector of $\mathbf{S}$ represent?

    (f) (2 pts.) What is the locus of values x such that $x^\top \mathbf{S}^{-1} x =$ constant for $p = 2$ and $\mathbf{S}$ is invertible?

    (g) (2 pt.) Explain how and why one could use $\mathbf{S}$ to preprocess observations for k-means.

2. (4 pts.) Compare and contrast Principal Components Analysis and Fisher Linear Discriminants.

3. Given: a matrix $\mathbf{X}$ and its SVD $\mathbf{X} = \mathbf{UDV}^\top$.

    (a) (3 pts.) Suppose $\mathbf{X}$ represents a *movie× people* ratings matrix. What do $\mathbf{U}$, $\mathbf{D}$ and $\mathbf{V}$ represent?

    (b) (3 pts.) Now let $\mathbf{X}$ be the clown face image. What does the SVD represent in this case?

    (c) (2 pts.) In either case, suppose we randomly shuffle the entries in $\mathbf{X}$. Describe (or sketch) what impact this operation has on the entries of $\mathbf{D}$.

    (d) (1 pt.) Repeat question 3c assuming we only shuffle the rows (but not the columns) of $\mathbf{X}$.

4. Refer to the coronary heart disease example in Fig. 1.

    (a) (1 pt.) What are the tick marks at the top and bottom of the left plot?

    (b) (1 pt.) How were the "CHD" and "no CHD" curves obtained?

    (c) (2 pts.) How was the curve on the right computed? What does the dashed line represent?

    (d) (1 pt.) What is causing the wiggle around 190 mm Hg?

Figure 1