Chris Kasper
CS 383-002
Homework #1

Theory Questions

1a.
Group data together

$$= \begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \\ -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Compute information gain for feature 1

Split into subsets
$f_1 = \{-2, -5, -3, 0, -8, 1, 5, -1, 6\}$

$s_1 = \begin{bmatrix} -2 & 1 \\ -2 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $s_2 = [-5 \quad 4]\,[1]$ $s_3 = [-3 \quad 1]\,[1]$ $s_4 = [0 \quad 3]\,[1]$

$s_5 = [-8 \quad 11]\,[1]$ $s_6 = [1 \quad 0]\,[0]$ $s_7 = [-5 \quad 4]\,[0]$ $s_8 = [-1 \quad -3]\,[0]$

$s_9 = [6 \quad 1]\,[0]$

Entropy of Initial System
$$H\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{5}{10} * \log_2\left(\frac{5}{10}\right) + -\frac{5}{10} * \log_2\left(\frac{5}{10}\right) = 1$$

Entropy of subsets
$$H(s_1) = H\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2} * \log_2\left(\frac{1}{2}\right) + -\frac{1}{2} * \log_2\left(\frac{1}{2}\right) = 1$$

$$H(s_2) = H(s_3) \ldots = H(s_9) = H(0,1) = -0 * \log_2(1) + -1 * \log_2(1) = 0$$

$s_2 - s_9$ *have entropy of* 0 *(all deterministic), each with weight of* $\frac{1}{10}$

$s_1\ has\ entropy\ of\ 1\ (equally\ random), with\ a\ weight\ of \frac{2}{10}$

$$E\big(H(f_1)\big) = \frac{2}{10} * 1 + \frac{1}{10} * 0 + \frac{1}{10} * 0 + \frac{1}{10} * 0 + \frac{1}{10} * 0 + \frac{1}{10} * 0 + \frac{1}{10} * 0 + \frac{1}{10} * 0 \ + \frac{1}{10} * 0$$

$$E\big(H(f_1)\big) = 0.2$$

$$IG(f_1) = 1 - 0.2 = 0.8$$

Compute information gain for feature 2

Split into subsets
$f_2 = \{1, -4,3,11,5,0, -1, -3\}$

$$s_1 = \begin{bmatrix} -2 & 1 \\ -3 & 1 \\ 6 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad s_2 = [-5 \quad 4]\,[1] \quad s_3 = [0 \quad 3]\,[1] \quad s_4 = [-8 \quad 11]\,[1]$$

$$s_5 = [-2 \quad 5]\,[0] \quad s_6 = [1 \quad 0]\,[0] \quad s_7 = [5 \quad -1]\,[0] \quad s_8 = [-1 \quad -3]\,[0]$$

Entropy of Initial System

$$H\left(\frac{1}{2}, \frac{1}{2}\right) = \ -\frac{5}{10} * \log_2\left(\frac{5}{10}\right) + \ -\frac{5}{10} * \log_2\left(\frac{5}{10}\right) = 1$$

Entropy of subsets

$$H(s_1) = H\left(\frac{2}{3}, \frac{1}{3}\right) = \ -\frac{2}{3} * \log_2\left(\frac{2}{3}\right) + \ -\frac{1}{3} * \log_2\left(\frac{1}{3}\right) = 0.9182$$

$$H(s_2) = H(s_3) \ldots = H(s_8) = H(0,1) = \ -0 * \log_2(1) + \ -1 * \log_2(1) = 0$$

$s_2 - s_8\ have\ entropy\ of\ 0\ (all\ deterministic), each\ with\ weight\ of \frac{1}{10}$
$s_1\ has\ entropy\ of\ 0.9182, with\ a\ weight\ of \frac{3}{10}$

$$E\big(H(f_2)\big) = \frac{3}{10} * 0.9182 + \frac{1}{10} * 0 + \frac{1}{10} * 0 + \frac{1}{10} * 0 + \frac{1}{10} * 0 + \frac{1}{10} * 0 + \frac{1}{10} * 0 + \frac{1}{10} * 0$$

$$E\big(H(f_2)\big) = 0.2754$$

$$IG(f_2) = 1 - 0.2754 = 0.7246$$

1b. Feature 1 is more discriminating then Feature 2 because Feature 1 has more information gain.

2. Using equation:

$$w^T(\mu_1 - \mu_2)^T(\mu_1 - \mu_2)w - \lambda(\, w^T(\sigma_1^T\sigma_1 + \sigma_2^T\sigma_2)w)$$

Take derivative with respect to w

Use rule: $\frac{d}{dw}\left(X^T A^T AX\right) = 2A^T AX$

Split into 2 sections

Left section:

$$\frac{d}{dw}\left(w^T(\mu_1 - \mu_2)^T(\mu_1 - \mu_2)w\right) => 2(\mu_1 - \mu_2)^T(\mu_1 - \mu_2)w$$

Right section:

$$\frac{d}{dw}\left(\, \lambda(\, w^T(\sigma_1^T\sigma_1 + \sigma_2^T\sigma_2)w)\right)$$

$$= \lambda\frac{d}{dw}(\, w^T(\sigma_1^T\sigma_1 + \sigma_2^T\sigma_2)w)$$

$$= \lambda\frac{d}{dw}(w^T\sigma_1^T\sigma_1 w + w^T\sigma_2^T\sigma_2 w)$$

$$= \lambda(2\sigma_1^T\sigma_1 w + 2\sigma_2^T\sigma_2 w)$$

Combine both:

$$2(\mu_1 - \mu_2)^T(\mu_1 - \mu_2)w - \lambda(2\sigma_1^T\sigma_1 w + 2\sigma_2^T\sigma_2 w)$$

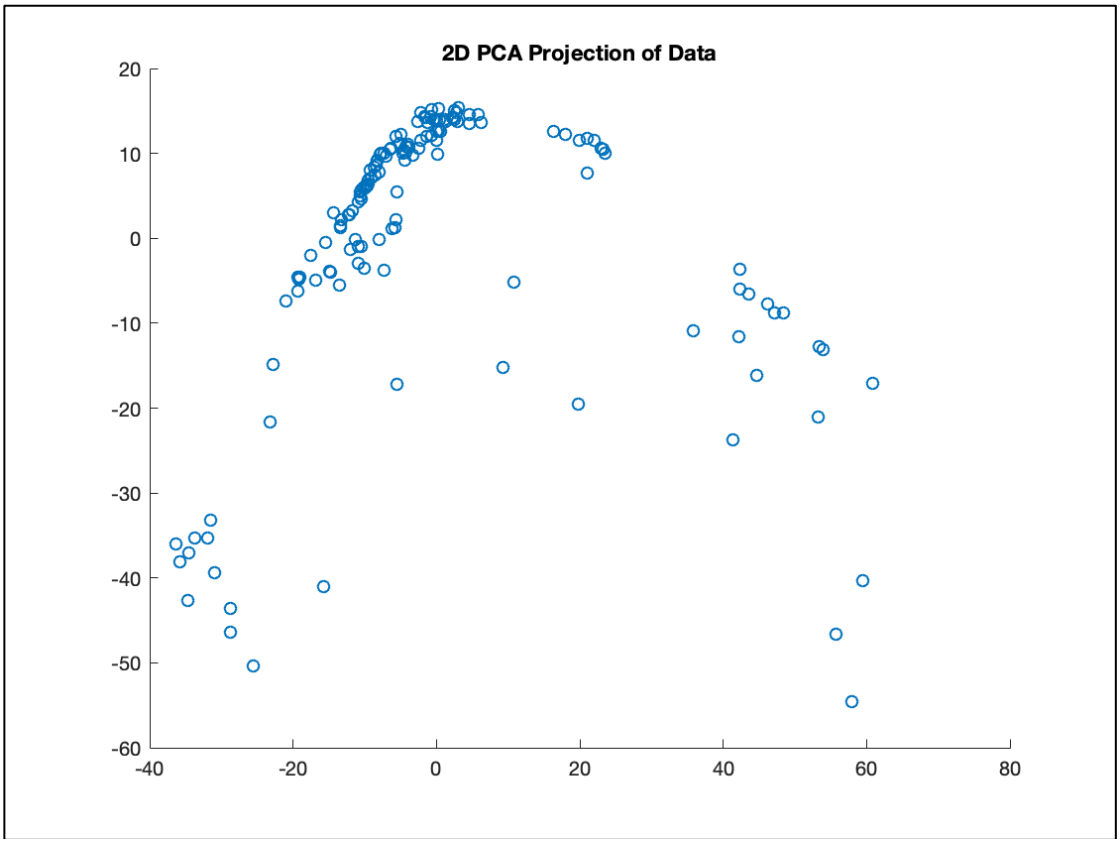
$$2(\mu_1 - \mu_2)^T(\mu_1 - \mu_2)w = \lambda(2\sigma_1^T\sigma_1 w + 2\sigma_2^T\sigma_2 w)$$

$$2(\mu_1 - \mu_2)^T(\mu_1 - \mu_2)w = (\lambda * \sigma_1^T\sigma_1 + \lambda * \sigma_2^T\sigma_2)2w$$

$$(\mu_1 - \mu_2)^T(\mu_1 - \mu_2)w = (\lambda * \sigma_1^T\sigma_1 + \lambda * \sigma_2^T\sigma_2)w$$

Which is in the form of... $Aw = bw$

Part 2: Dimensionality Reduction via PCA



Part 3: Eigenfaces

Number of principle components needed to represent 95% of information, k = 37

Primary Principle Component

Original Image

Single PC Reconstruction

k PC Reconstruction