Aslan Oztreves

CS 383

Matthew Burlick

Homework #1

Theory Problems

1. (15 points) Consider the following data:

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

(a) Compute the information gain for each feature. You could standardize the data overall, although it won't make a difference. (13pts).

(b) Which feature is more discriminating based on results in Part (a) (2pt)?

a)

$$H\big(P(v_1), \dots, P(v_n)\big) = \sum_{i=1}^{n} \big(-P(v_i) log_n P(v_i)\big)$$

$$E\big(H(A)\big) = \sum_{i=1}^{k} \frac{p_i + n_i}{p + n} H\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

$$IG(A) = H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - E(H(A))$$

Feature 1:

$$E(1) = \frac{1+1}{5+5} H\left(\frac{1}{1+1}, \frac{1}{1+1}\right) + 4 \times \frac{1+0}{5+5} H\left(\frac{1}{1+0}, \frac{0}{1+0}\right) + 4 \times \frac{0+1}{5+5} H\left(\frac{0}{0+1}, \frac{1}{0+1}\right) = 0.2$$

$$H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$$

$$H(0,1) = -1\, log 1 - 0 log 0 = 0$$

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

Feature 2:

$$E(2) = \frac{2+1}{5+5} H\left(\frac{2}{2+1}, \frac{1}{1+2}\right) + 8 \times \frac{1+0}{5+5} H\left(\frac{1}{1+0}, \frac{0}{1+0}\right) = 0.2755$$

$$H\left(\frac{2}{3}, \frac{1}{3}\right) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3} log \frac{1}{3} = 0.9183$$

$$IG(2) = 0.9183 - 0.2755 = 0.6428$$

b)

IG(1) > IG(2) thus we should choose feature 1.

2. (15 points) In principle component analysis (PCA) we are trying to maximize the variance of the data after projection while minimizing how far the magnitude of w, $|w|$ is from being unit length. This results in attempting to find the value of w that maximizes the equation

$$w^T \Sigma w - \alpha(w^T w - 1)$$

where Σ is the covariance matrix of the observable data matrix X.

One problem with PCA is that it doesn't take class labels into account. Therefore projecting using PCA can result in worse class separation, making the classification problem more difficult, especially for linear classifiers.

To avoid this, if we have class information, one idea is to separate the data by class and aim to find the projection that maximize the distance between the means of the class data after projection, while minimizing their variance after projection. This is called **linear discriminant analysis** (LDA).

Let C_i be the set of observations that have class label i, and μ_i, σ_i be the mean and standard deviations, respectively, of those sets. Assuming that we only have two classes, we then want to find the value of w that maximizes the equation:

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

Which is equivalent to

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

Show that to maximize we must solve an eigen-decomposition problem, i.e $Aw = bw$. In particular what are A and b for this equation.

$$= \frac{\partial}{\partial \omega} \omega^T (\mu_1 - \mu_2)^T (\mu_1 - \mu_2)\omega - \lambda(\omega^T(\sigma_1^T \sigma_1 + \sigma_2^T \sigma_2)\omega)$$

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

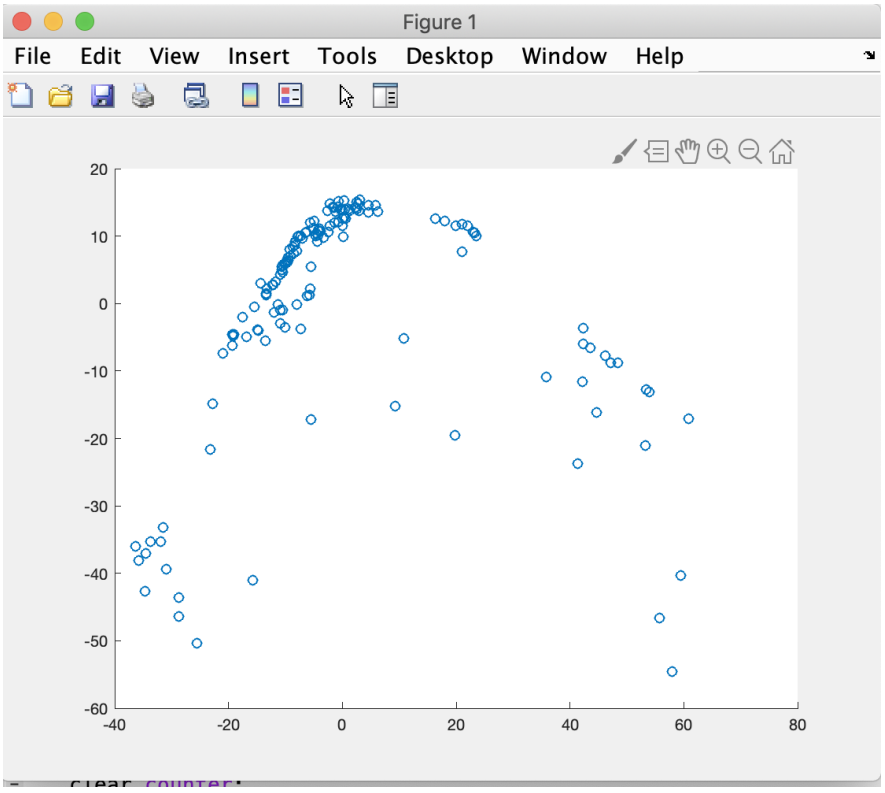
$$(\mu_1 - \mu_2)^T (\mu_1 - \mu_2)\omega = \lambda(\sigma_1^T \sigma_1 + \sigma_2^T \sigma_2)\omega$$

$$(AA^T)^T = (A^T)^T A^T = AA^T$$

$$A\omega = \lambda(\sigma_1^T \sigma_1 + \sigma_2^T \sigma_2)\omega$$

$$A\omega = b\omega$$

Visualization of the PCA results



1) Number of principle components needed to represent %95 of information, k is 37.
2) Visualization of primary principle component
3) Visualization of the reconstruction of the first person using
    a) Original image
    b) Single principle component
    c) K principle component