Eric Zounes

November 21, 2012

CS434: Assignment 4

1. Association Rules. Given the following database of trasactions:

Transaction ID	Items
1	A,B
2	A, B, C, D
3	A, D,
4	C, D

a. Find all frequent itemsets with minimum support count = 2. Pease provide clear intermediate steps, indicating $C_1, L_1, C_2, L_2, \ldots$ until the algorithm terminates. Be sure to include (and clearly indicate) the results of the self-join, and pruning step in generating candidate sets.

Item	σ
AB	2
AC	1
AD	2
BC	1
BD	1
CD	2

b. Find all association rules with miminum support count = 2 and min_conf = 75.

Item	Conf.
$A \rightarrow B$	$\frac{2}{3}$
$B \rightarrow A$	1
$A \rightarrow D$	$\frac{2}{3}$
$D \rightarrow A$	$\frac{2}{3}$
$C \rightarrow D$	1
$D \rightarrow C$	$\frac{2}{3}$

2. Please consider the data provided on the class website, which is a 2 dimensional data that comes from a Gaussian distributino with a full covariance matrix. Apply PCA to this data. In particular, you need to first estimate the covariance matrix of this data, and the two PCA projection vectors (i.e., the eigen vectors of the covariance matrix).

   a. Report the covariance matrix that you estimated from the data.

```
MATLAB code:
M = csvread('pcs.csv')
V = cov(M)
  = 0.9301    0.8121
0.8121    2.0456
```

   b. Report the projection vectors for PC1 and PC2.

```
 [x1, x2] = eig(V)
  x1 =
     = -0.8849    0.4657
0.4657    0.8849
```

```
x2 =

    0.5027         0
         0    2.4730
```

c. Please plot the data in two difference figures, one in the original 2-d coordinate system, and one in the PCA space(2-d).

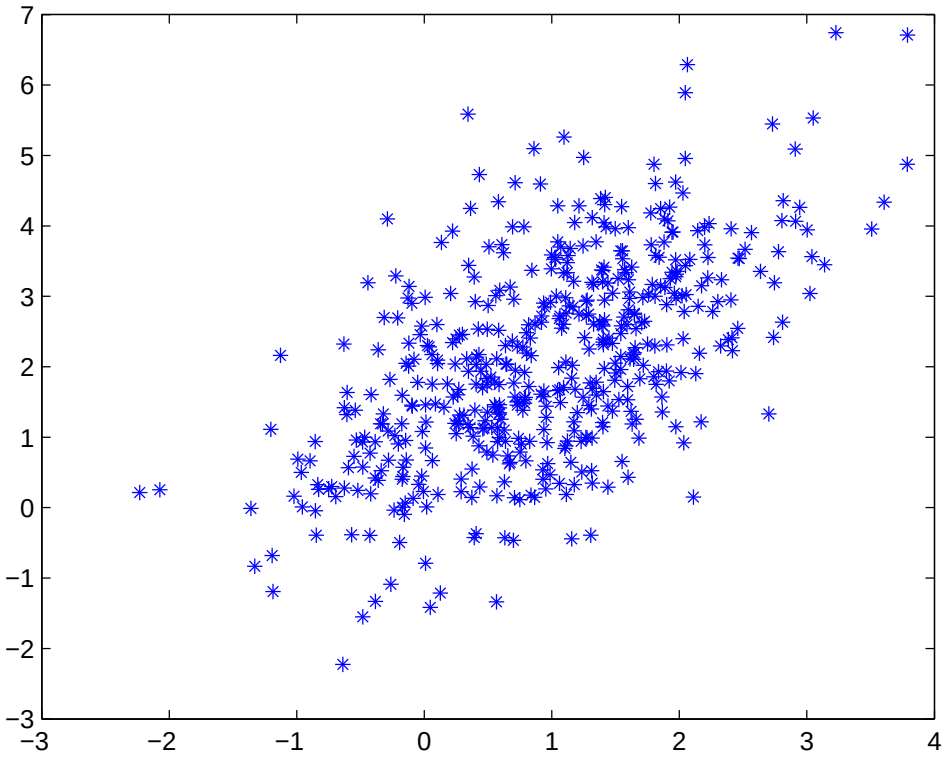

Figure 1: A plot of the data on the original 2d space.

When we find the PCA vectors, we can change the axis of the data plot.

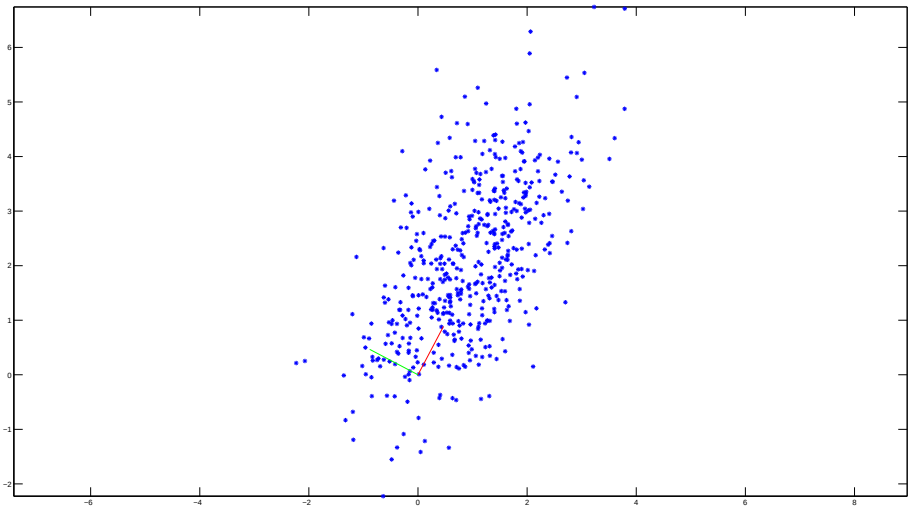

Figure 2: A plot of the data in PCA space.

Given a dataset, PCA seeks to find a small number of dimensions that preserve as much variance in data as possible. Now consider a supervised learning task, can you come up with a situation where applying PCA to the data can significantly worsen the classification performance? In other words, is it possible to have data that is original separable but made non-separable by PCA projection (to lower dimension)? Please provide a visual example

Boosting. Consider a data set $D = x_1, x_2, \ldots, x_{10}$. We apply Adaboost with Decision Stump. Let $w_1, \ldots, w_{10}$ be the weights of the ten training examples respectively.

a. In the first iteration , x_1, x_2,andx_3 are misclassified. Please rank the updated weights $w_1, \ldots, w_{10}$ be the weights of the ten training examples respectively.

$\epsilon_1 = \frac{3}{10}$

$\alpha_1 = \frac{1}{2} ln(\frac{1-.3}{.3})$

$\alpha_1 = .4236$

Updated rank:

$$D_{t+1}(i) = D_t(i) \times \begin{cases} e^{\alpha_t}, h_t(x_t) \neq y_t \\ e^{\alpha_t}, h_t(x_t) = y_t \end{cases}$$

$w_4, w_5, w_6, w_7, w_8, w_9, w_{10}, w_1, w_2, w_3$

b. In the second iteration, x_3 and x_4 are misclassified. Please provide the rank of the updated weights $w_1, \ldots, w_{10}$ in increasing order. Explain your ordering.

$\epsilon_t = x_4 + x_3$

$\epsilon_2 = \frac{1}{10} + .1527$

$\epsilon_2 = .252$ $\alpha_2 = \frac{1}{2} ln(\frac{1-.252}{.252})$

$\alpha_2 = .544$

Updated rank:

$$D_{t+1}(i) = D_t(i) \times \begin{cases} e^{\alpha_t}, h_t(x_t) \neq y_t \\ e^{\alpha_t}, h_t(x_t) = y_t \end{cases}$$

$w_4 = .1723, w_3 = .2631$

$w_5, w_6, w_7, w_8, w_9, w_{10}, w_1, w_2, w_4, w_3$