Chapter #6
Feature Selection Methods

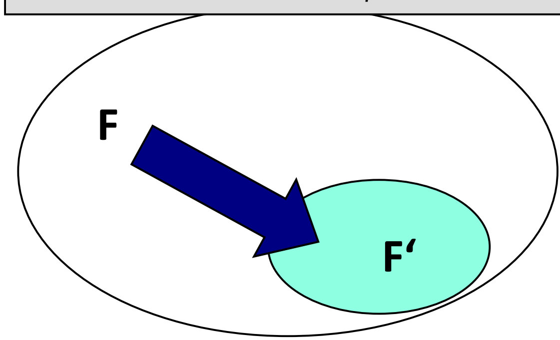
Qiang Yang

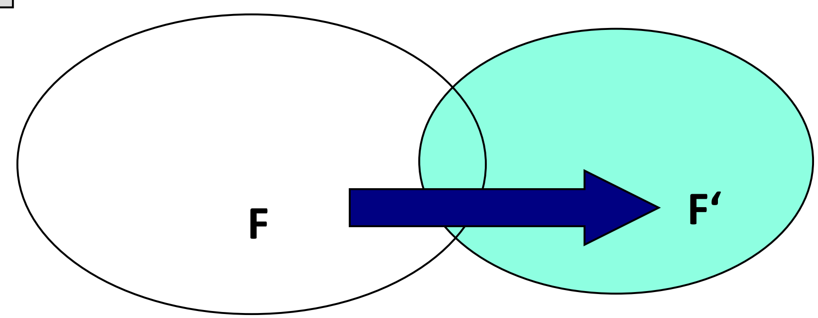
MSC IT 5210

Feature Selection

- Also known as
  - dimensionality reduction
  - subspace learning
  - Two types: subset vs. new features

$$\{f_1,...,f_i,...,f_n\} \xrightarrow[f.selection]{} \{f_{i_1},...,f_{i_j},...,f_{i_m}\}$$





$$\{f_1,...,f_i,...,f_n\} \xrightarrow[f.extraction]{} \{g_1(f_1,...,f_n),...,g_j(f_1,...,f_n),...,g_m(f_1,...,f_n)\}$$

Motivation

The objective of feature reduction is three-fold:

- Improving the accuracy of classification
- Providing a faster and more cost-effective predictors (CPU time)
- Providing a better understanding of the underlying process that generated the data

Filtering methods

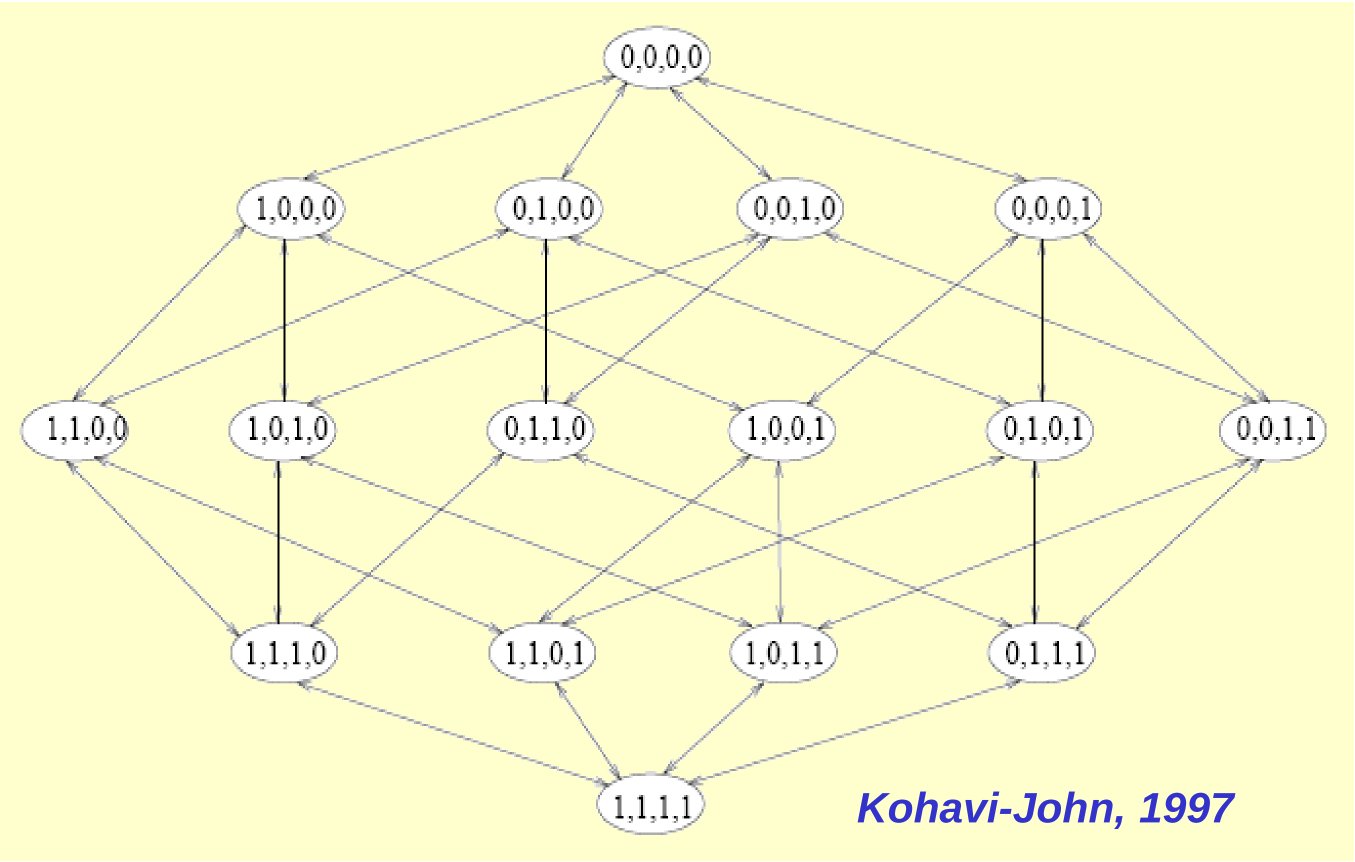
- Assume that you have both the feature Xi and the class attribute Y
- Associate a weight Wi with Xi
- Choose the features with largest weights
  - Information Gain (Xi, Y)
  - Mutual Information (Xi, Y)
  - Chi-Square value of (Xi, Y)

Wrapper Methods

- Classifier is considered a black-box: Say KNN
- Loop
  - Choose a subset of features
  - Classify test data using classifier
  - Obtain error rates
- Until error rate is low enough (< threshold)

- One needs to define:
  - how to search the space of all possible variable subsets ?
  - how to assess the prediction performance of a learner ?

The space of choices is large



n features, 2^n possible feature subsets!

Comparsion of filter and wrapper methods for feature selection:

- Wrapper method (+: optimized for learning algorithm)
  - tied to a classification algorithm
  - very time consuming
- Filtering method (+: fast)
  - Tied to a statistical method
  - not directly related to learning objective

Feature Selection using Chi-Square

- Question: Are attributes A1 and A2 independent?
  - If they are very dependent, we can remove either A1 or A2
  - If A1 is independent on a class attribute A2, we can remove A1 from our training data

Outlook	Tempreature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Chi-Squared Test (cont.)

- Question: Are attributes A1 and A2 independent?
- These features are nominal valued (discrete)
- Null Hypothesis: we expect *independence*

Outlook	Temperature
Sunny	High
Cloudy	Low
Sunny	High

The Weather example: Observed Count

temperature → Outlook	High	Low	Outlook Subtotal
Sunny	2	0	2
Cloudy	0	1	1
Temperature Subtotal:	2	1	Total count in table =3

Outlook	**Temperature**
Sunny	High
Cloudy	Low
Sunny	High

The Weather example: Expected Count

If attributes were *independent*, then the subtotals would be Like this (this table is also known as

temperature → Outlook	High	Low	Subtotal
Sunny	3*2/3*2/3 =4/3=1.3	3*2/3*1/3 =2/3=0.6	2 (prob=2/3)
Cloudy	3*2/3*1/3 =0.6	3*1/3*1/3 =0.3	1, (prob=1/3)
Subtotal:	2 (prob=2/3)	1 (prob=1/3)	Total count in table =3

Outlook	**Temperature**
Sunny	High
Cloudy	Low
Sunny	High

Question: How different between observed and expected?

The chi-squared formula is:

$$\text{Chi-squared } (X^2) = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \ldots\ldots + \frac{(o_n - e_n)^2}{e_n}$$

- X^2=(2-1.3)^2/1.3+(0-0.6)^2/0.6+(0-0.6)^2/0.6+(1-0.3)^2/0.3
- If Chi-squared value is very large, then A1 and A2 are not independent → that is, they are dependent!
- Thus,
  - X^2 value is large → Attributes A1 and A2 are dependent
  - X^2 value is small → Attributes A1 and A2 are independent

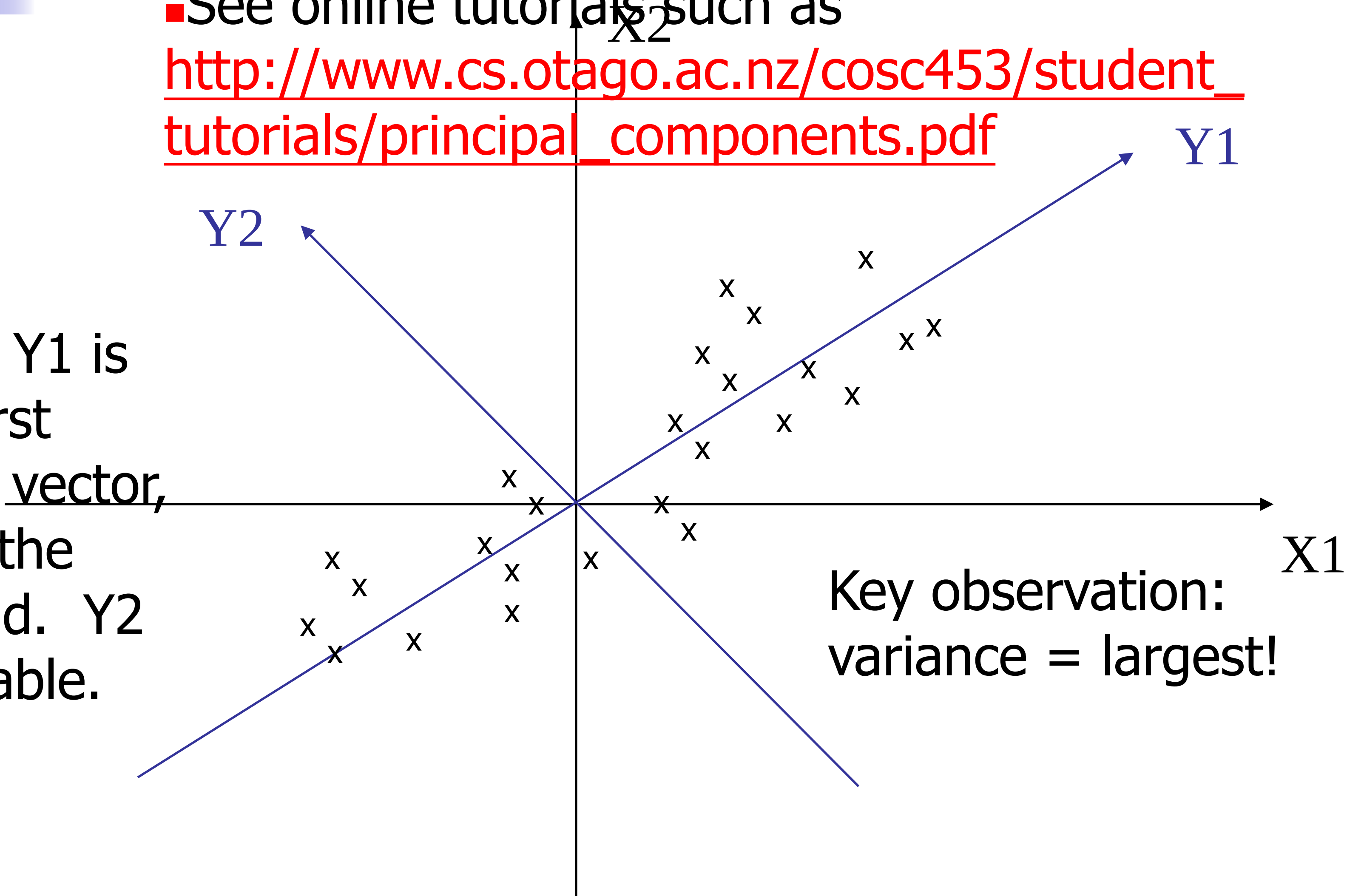
Chi-Squared Table: what does it mean?

- If calculated value is much greater than in the table, then you have reason to reject the independence assumption
  - When your calculated chi-square value is **greater than** the chi^2 value shown in the 0.05 column (3.84) of this table → you are 95% certain that attributes are actually dependent!
  - i.e. there is only a 5% probability that your calculated X^2 value would occur by chance

Degrees of Freedom	Probability, *p*				
	0.99	**0.95**	**0.05**	**0.01**	**0.001**
1	0.000	0.004	3.84	6.64	10.83
2	0.020	0.103	5.99	9.21	13.82

Principal Component Analysis (PCA)

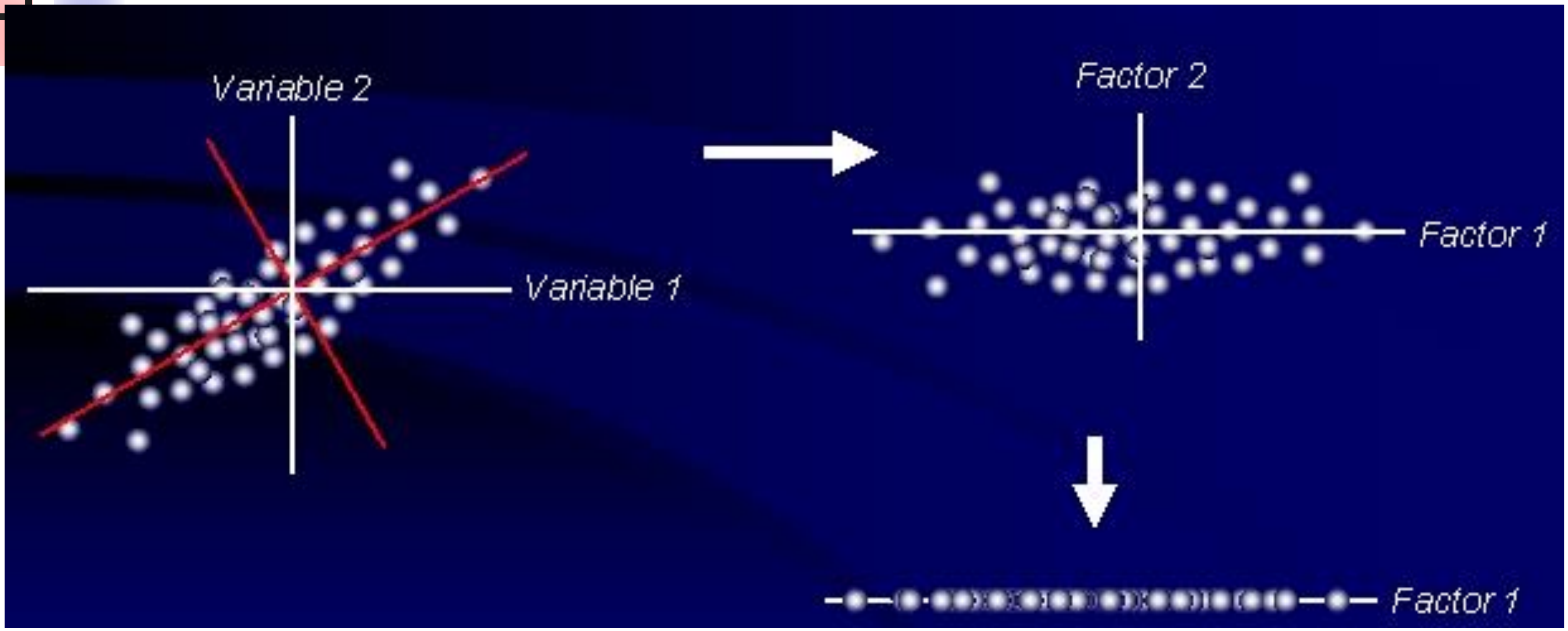
- See online tutorials such as http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf



Note: Y1 is the first eigen vector, Y2 is the second. Y2 ignorable.

Key observation: variance = largest!

Principle Component Analysis (PCA)



Principle Component Analysis: project onto subspace with the most variance (unsupervised; doesn't take y into account)

Principal Component Analysis: one attribute first

- Question: how much spread is in the data along the axis? (distance to the mean)
- Variance=Standard deviation^2

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

Temperature
42
40
24
30
15
18
15
30
15
30
35
30
40
30

Now consider two dimensions

Covariance: measures the correlation between X and Y

- cov(X,Y)=0: independent
- Cov(X,Y)>0: move same dir
- Cov(X,Y)<0: move oppo dir

X=Temperature	Y=Humidity
40	90
40	90
40	90
30	90
15	70
15	70
15	70
30	90
15	70
30	70
30	70
30	90
40	70
30	90

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

More than two attributes: covariance matrix

- Contains covariance values between all possible dimensions (=attributes):

$$C^{nxn} = (c_{ij} \mid c_{ij} = \operatorname{cov}(Dim_i, Dim_j))$$

- Example for three attributes (x,y,z):

$$C = \begin{pmatrix} \operatorname{cov}(x,x) & \operatorname{cov}(x,y) & \operatorname{cov}(x,z) \\ \operatorname{cov}(y,x) & \operatorname{cov}(y,y) & \operatorname{cov}(y,z) \\ \operatorname{cov}(z,x) & \operatorname{cov}(z,y) & \operatorname{cov}(z,z) \end{pmatrix}$$

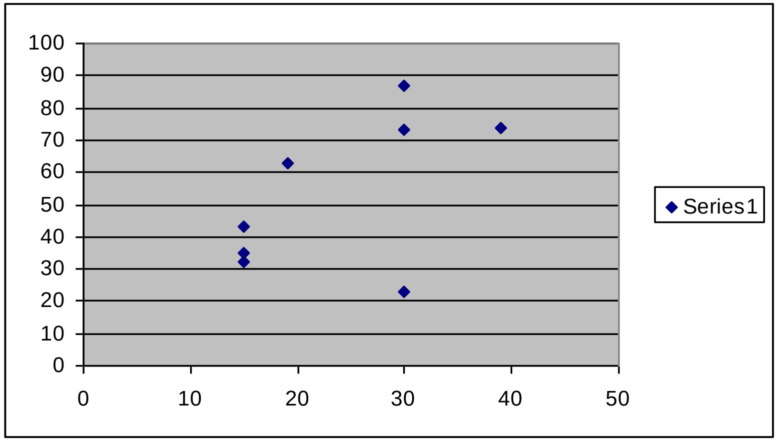
Background: eigenvalues AND eigenvectors

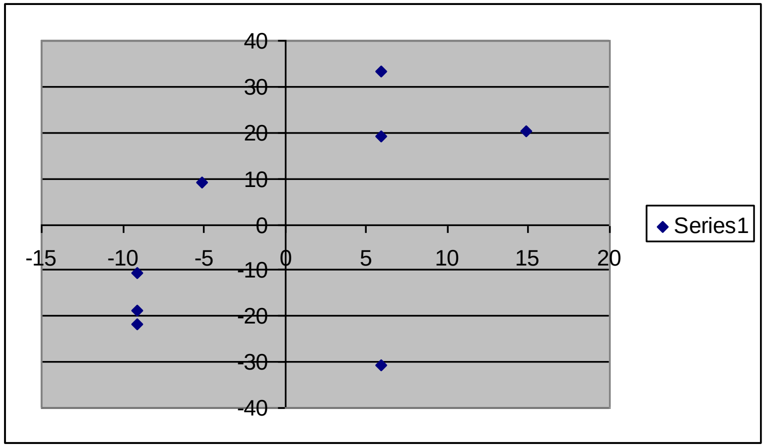
- Eigenvectors **e** : $C\mathbf{e} = \lambda\, \mathbf{e}$
- How to calculate **e** and λ:
  - Calculate $det(C-\lambda I)$, yields a polynomial (degree n)
  - Determine roots to $det(C-\lambda I)=0$, roots are eigenvalues λ
- Check out any math book such as
  - *Elementary Linear Algebra* by Howard Anton, Publisher John,Wiley & Sons
  - Or any math packages such as MATLAB

An Example

Mean1=24.1
Mean2=53.8

X1	X2	X1'	X2'
19	63	-5.1	9.25
39	74	14.9	20.25
30	87	5.9	33.25
30	23	5.9	-30.75
15	35	-9.1	-18.75
15	43	-9.1	-10.75
15	32	-9.1	-21.75
30	73	5.9	19.25





Covariance Matrix

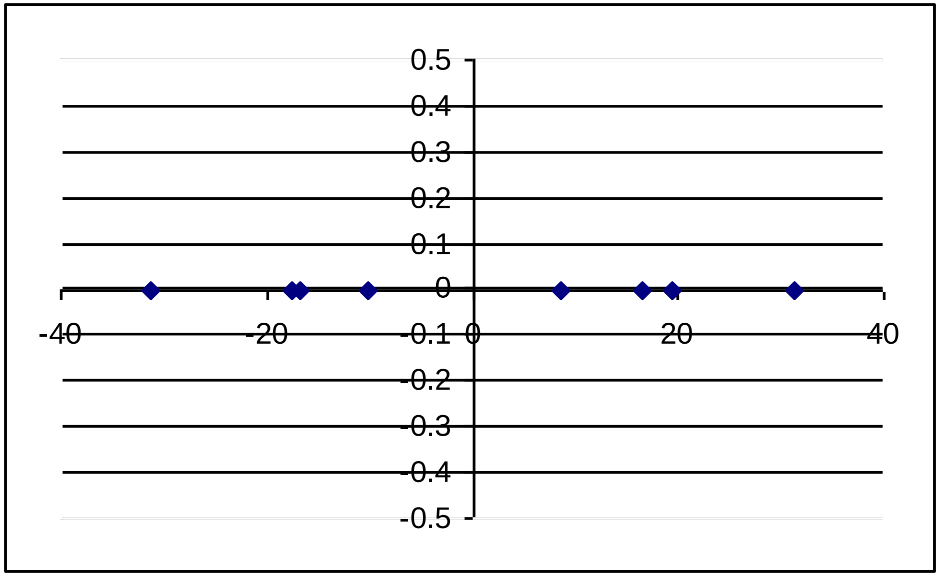
- C=

75	106
106	482

- Using MATLAB, we find out:
  - Eigenvectors:
  - e1=(-0.98, 0.21), $\lambda 1$=51.8
  - e2=(0.21, 0.98), $\lambda 2$=560.2
  - Thus the second eigenvector is more important!

If we only keep one dimension: e2

- We keep the dimension of e2=(0.21, 0.98)
- We can obtain the final data as



yi
-10.14
-16.72
-31.35
31.374
16.464
8.624
19.404
-17.63

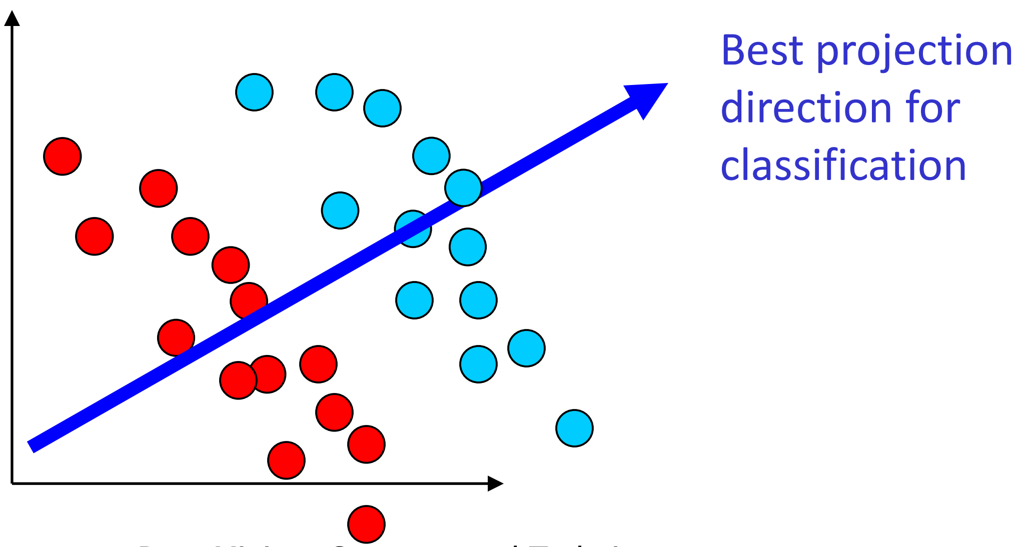
$$y_i = \begin{pmatrix} x_{i1} & x_{i2} \end{pmatrix} \begin{pmatrix} 0.21 \\ 0.98 \end{pmatrix} = 0.21 * x_{i1} + 0.98 * x_{i2}$$

Summary of PCA

- PCA is used for reducing the number of numerical attributes
- The key is in data transformation
  - Adjust data by mean
  - Find eigenvectors for covariance matrix
  - Transform data
- Note: only linear combination of data (weighted sum of original data)

Linear Method: Linear Discriminant Analysis (LDA)

- LDA finds the projection that best separates the two classes
- Multiple discriminant analysis (MDA) extends LDA to multiple classes



PCA vs. LDA

- PCA is unsupervised while LDA is supervised.
- PCA can extract r (rank of data) principles features while LDA can find (c-1) features.
- Both based on SVD technique.