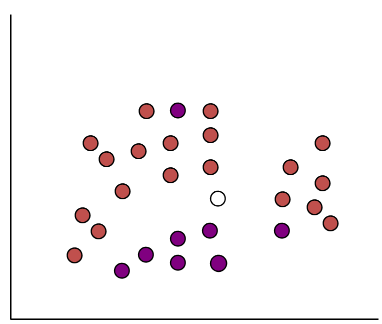
Linear Classification (Part I: Intro and Fisher’s LDA)

nanopoulos@ismll.de

The task of classification

Learn a method for predicting the instance class from pre-labeled (classified) instances

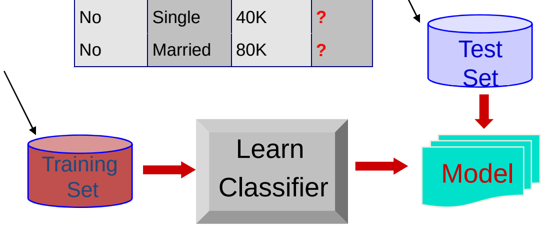


Classification Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Outline

- Applications of classification
- Linear classification
- Fisher's linear discriminant

4

Classification: Application 1

Direct Marketing

- Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
- Approach:
  - Use the data for a similar product introduced before.
  - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
    - Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 2

Fraud Detection

- Goal: Predict fraudulent cases in credit card transactions.
- Approach:
  - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc
  - Label past transactions as fraud or fair transactions. This forms the class attribute.
  - Learn a model for the class of the transactions.
  - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

Customer Attrition/Churn:

- Goal: To predict whether a customer is likely to be lost to a competitor.
- Approach:
  - Use detailed record of transactions with each of the past and present customers, to find attributes.
    - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
  - Label the customers as loyal or disloyal.
  - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 4

Sky Survey Cataloging

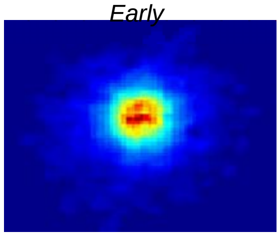
- Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
- Approach:
  - Segment the image.
  - Measure image attributes (features) - 40 of them per object.
  - Model the class based on these features.
  - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classifying Galaxies

Courtesy: http://aps.umn.edu

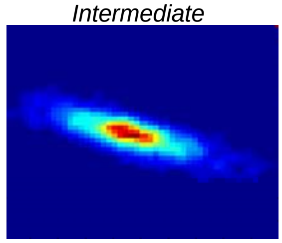
Early



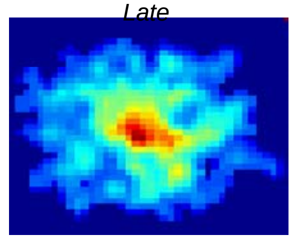
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate



Late



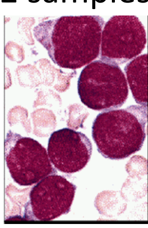
Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

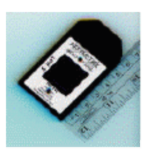
Biology: Molecular Diagnostics

Leukemia: Acute Lymphoblastic (ALL) vs Acute Myeloid (AML)

72 samples, about 7,000 genes

ALL





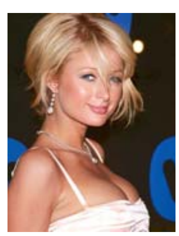
AML

Image classification

Genre classification

Top Artists tagged "brutal death metal"

1	Paris Hilton	718
2	Nile	528
3	Cannibal Corpse	474
4	Suffocation	281
5	Aborted	259
6	Cryptopsy	241
7	Dying Fetus	181
8	Deicide	170
9	Devourment	166
10	Behemoth	142
11	Prostitute Disfigurement	140
12	Disgorge	139
13	Hate Eternal	133
13	Deeds of Flesh	133
15	Decapitated	128
15	Krisiun	128
17	Necrophagist	109
18	Barney	108
19	Brodequin	102
20	Gorgasm	101



Outline

- Applications of classification
- Linear classification
- Fisher's linear discriminant

Linear classification

Two classes: C_1, C_2

x is the input vector, **w** the model's parameters

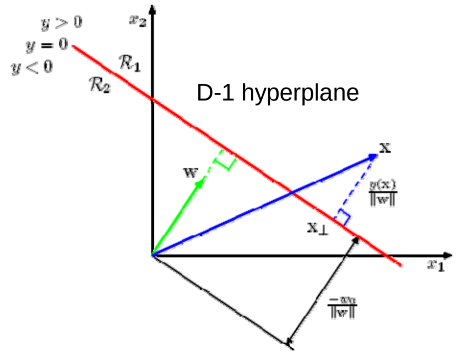
$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

Assign to C_1 if y(**x**) >= 0

Else, assign to C_2

y(**x**) = 0 defines the decision boundary, which is a line

Illustration of decision boundary



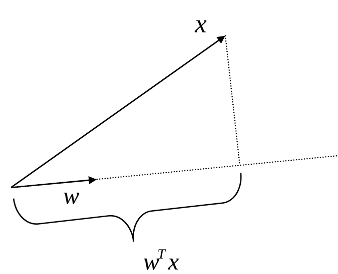
$\mathbf{x}_a,\mathbf{x}_b$ on the boundary: $y(\mathbf{x}_a)-y(\mathbf{x}_b) = \mathbf{w}^T(\mathbf{x}_a-\mathbf{x}_b)=0$

w is orthogonal to the decision boundary and determines its direction

Linear classification as dim reduction

$$y = \mathbf{w}^T\mathbf{x} \begin{cases} \text{classify } y \geqslant -w_0 \text{ as class } \mathcal{C}_1 \\ \text{otherwise } \mathcal{C}_2 \end{cases}$$

y(**x**) is the projection of **x** on **w**



Find **w** so as to maximize the separation of the two classes

Separating the class means

Class C_1 has N_1 points and C_2 N_2 points

Their means are: $\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n, \qquad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n$

Project means: $m_k = \mathbf{w}^{\mathrm{T}} \mathbf{m}_k$

Choose **w** to maximize: $m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)$

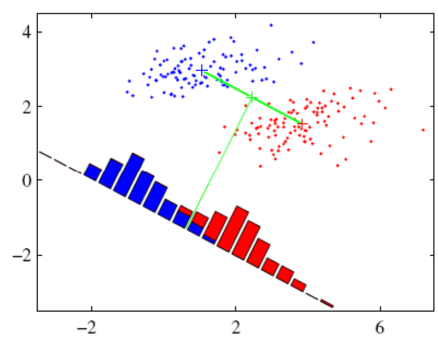
From training set we want to find out a direction ***w*** where the separation between the projections of class means is high

17

Maximizing the separation of means

The line joining the means defines the direction of greatest means spread (why?)

but gives high class overlap



18

Outline

- Applications of classification
- Linear classification
- Fisher's linear discriminant

Fisher's Linear Discriminant

Maximize a function that:

- Gives large separation between projected means and
- Giving small variance within each class (minimize class overlap)

Fisher's criterion

Within class variance: $s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$
(where $y_n = \mathbf{w}^{\mathrm{T}}\mathbf{x}_n$)

Total within-class variance: $s_1^2 + s_2^2$

Find **w** that maximizes: $J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$

J(**w**) as a function of **w**

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}$$

between-class covariance matrix
$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$

total *within-class* covariance matrix

$$\mathbf{S}_{\mathrm{W}} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}$$

Maximizing J(**w**)

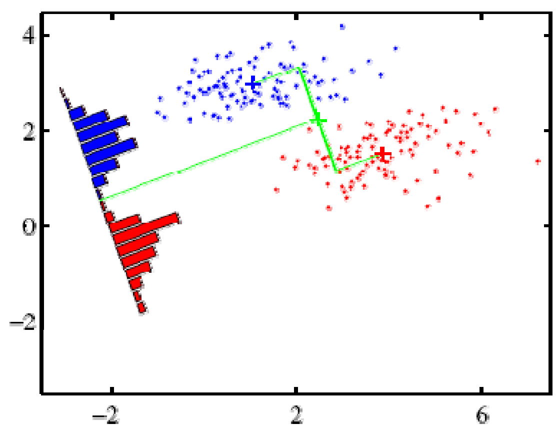
Derivative of dJ/d**w** = 0 gives (how?):

$$(\mathbf{w}^T \mathbf{S}_B \mathbf{w}) \mathbf{S}_W \mathbf{w} = (\mathbf{w}^T \mathbf{S}_W \mathbf{w}) \mathbf{S}_B \mathbf{w}$$

We just need the direction, omit the scalars:

$$\mathbf{w} \propto \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1)$$

What does this look like?



Rotate (by S_w^{-1}) the line joining the means

But, how to classify?

So far we got the direction of the decision boundary

We need to decide the threshold w_0

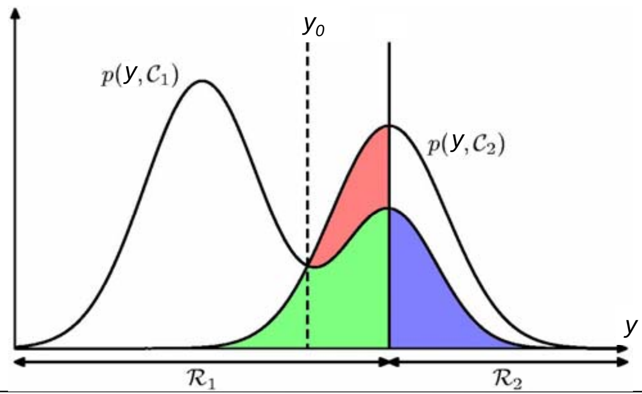
Remember $y = \mathrm{w}^{\mathrm{T}}\mathrm{x}$ — classify $y \geqslant -w_0$ as class $\mathcal{C}_1$; otherwise $\mathcal{C}_2$

How? **Decision theory**

25

Deciding the threshold

Find all the projections *y* and the value y_0 that minimizes the misclassification rate



26

Relation between Fisher's LD and min SSE

Linear regression: minimize SSE for target

Linear classification (Fisher LD): max class separation

Are those two related?

"Magic" targets

For C_1 let target be N/N_1

For C_2 let target be $-N/N_2$

$$\sum_{n=1}^{N} t_n = N_1 \frac{N}{N_1} - N_2 \frac{N}{N_2} = 0$$

SSE: $E = \frac{1}{2} \sum_{n=1}^{N} \left(\mathbf{w}^{\mathrm{T}} \mathbf{x}_n + w_0 - t_n \right)^2$

dE/dw = 0

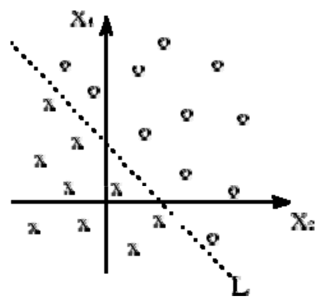
$$\sum_{n=1}^{N} \left(\mathbf{w}^{\mathrm{T}} \mathbf{x}_n + w_0 - t_n \right) \mathbf{x}_n = 0$$

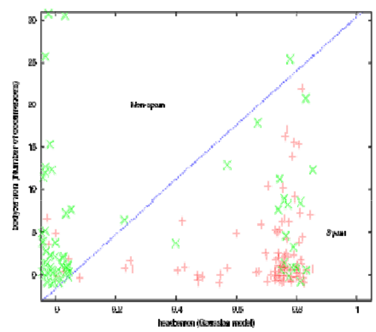
$$\left(\mathbf{S}_{\mathrm{W}} + \frac{N_1 N_2}{N} \mathbf{S}_{\mathrm{B}} \right) \mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2)$$

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1} (\mathbf{m}_2 - \mathbf{m}_1)$$

Conclusion

Linear classification works well when data are linearly separable





29

But don't forget…

The result does not only depend on the classification method

It also depends on the features

Example:

- C_1 "sexy", C_2 "not so sexy"
- x_1 is the hair color, x_2 is the bust size
- If blonde and rich bust, then C_1

30


GOLF WITH FRIE