Concepts in Pattern Recognition

Ed Schofield

Overview

- Applications of pattern recognition (5 min)
- Case 1: Classifying skin in images (10)
- Case 2: Forecasting financial markets (5)
- Formalization of concepts (5)
- Past and future (10)

Applications of pattern recognition

- Computer vision
- Speech processing
- Financial forecasting
- Data mining

Example 1: Computer vision

- Visual speech recognition (lip-reading) [Stork 96]
- Face recognition [Turk 91]
- Autonomous helicopter flight [CMU RI]
- Surveillance
- Virtual reality

Example 2: Speech processing

- Transcribing meetings and broadcasts
- Retrieving spoken information
- Identifying a speaker by their voice

Example 3: Financial forecasting

- Managing the risk of a portfolio of currencies [Risk, Hull]
- Forecasting economic trends
- Forecasting stock and option prices [Tech.Anal.]

Example 4: Data mining

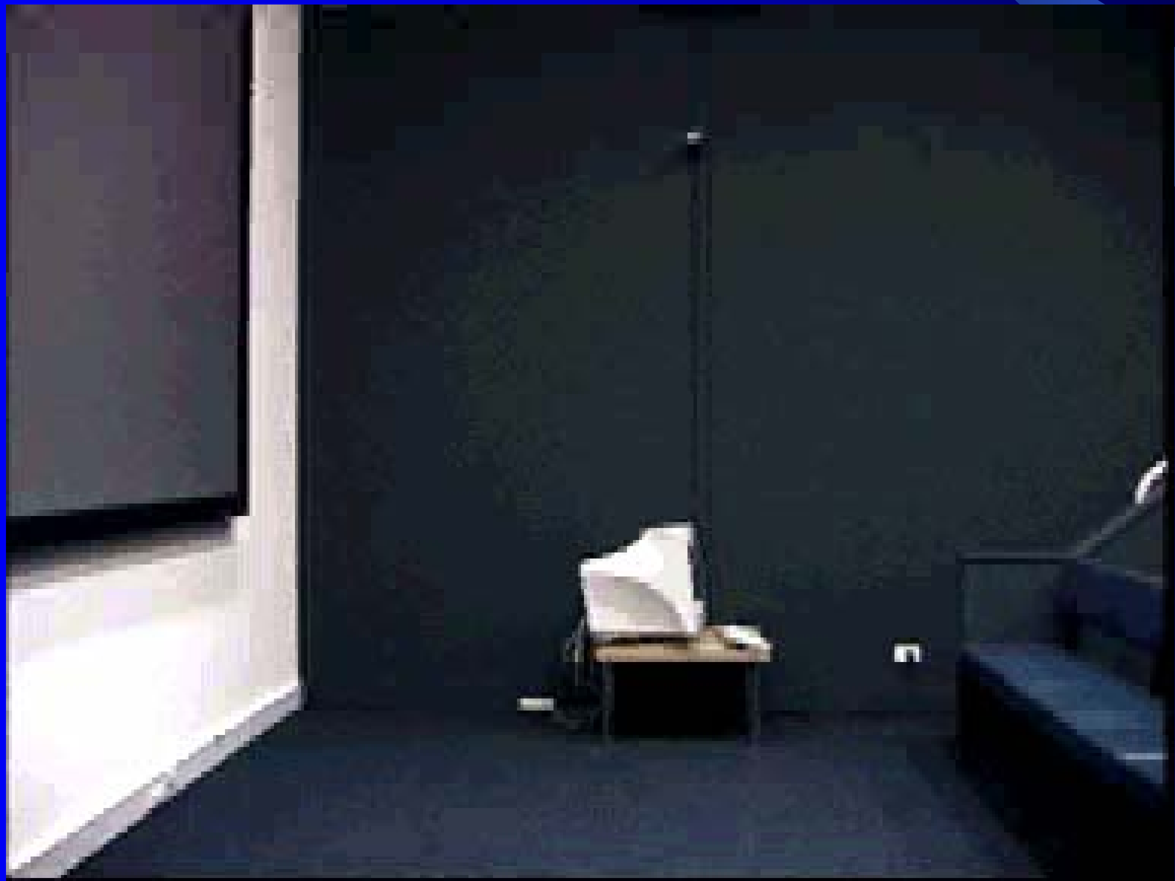
- Web searching
- Semantic indexing

Progress

- ✓ Applications of pattern recognition (5 min)
- Case 1: Classifying skin in images (12)
- Case 2: Forecasting financial markets (5)
- Formalization of concepts (5)
- Past and future (10)

Classifying skin in images

Locate this person's hands and head in 3D



Applications of skin detection

- Face and gesture recognition
- Lip-reading, video-conferencing
- Image classification

The color of skin

What is the probability that a given pixel represents skin?

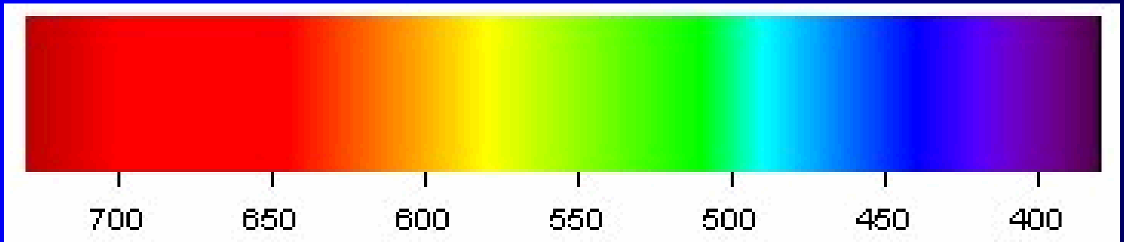
The classifier should be invariant to illumination and ethnicity.

Some models of skin-color

- Gaussians or mixtures of Gaussians [Jebara 97].
- Histograms [Jones 97].

Diversion—what is color?

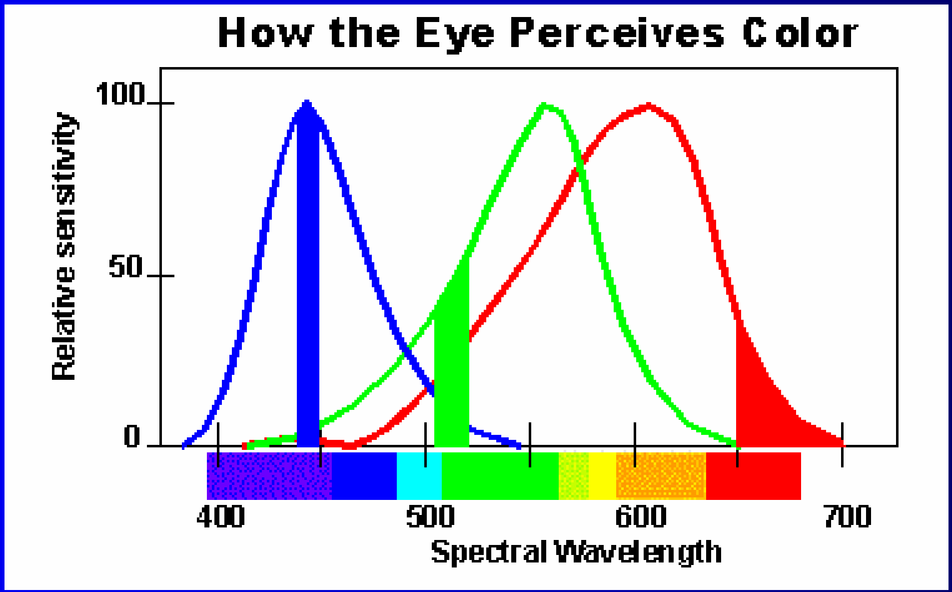
Some colors are represented in the electromagnetic spectrum.



Others, like brown, are not.

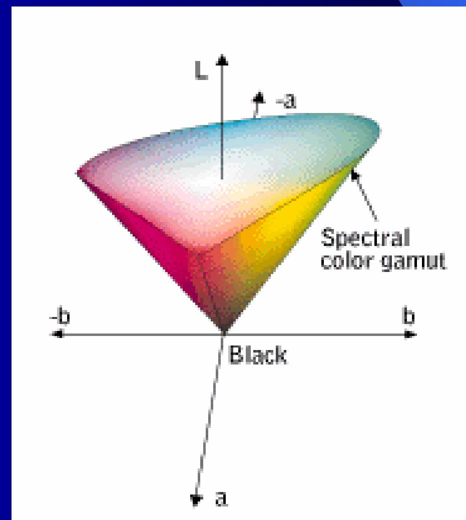
Color spaces

The eye has cones that respond to three ranges of wavelengths [Young-Helmholtz]



Color spaces (2)

- The dimensionality of color space is 3 [Poynton 97].
- Common color spaces are: RGB, HSV, CIE



More about skin

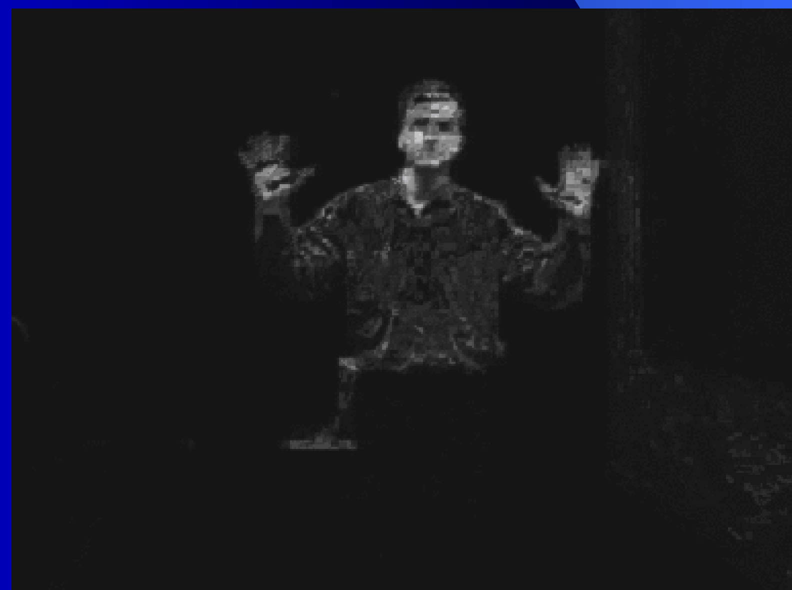
How separable are the classes of ‘skin’ and ‘non-skin’ in color space?

“Very.” [Jones 97]

Interesting fact: hue is largely invariant between ethnicities.

A method for classifying skin

1. Manually classify a large training set
2. Store histogram models of skin
3. Classify pixels using Bayes' rule



Other steps

- Clustering regions of probable skin
- Tracking motion through time

An application to speech

A binary classifier could distinguish accents–

- Austrian from German accents
- American from British accents

A recognizer could then choose the best speech model automatically.

Progress

- ✓ Applications of pattern recognition (5 min)
- ✓ Case 1: Classifying skin in images (12)
- Case 2: Forecasting financial markets (5)
- Formalization of concepts (5)
- Past and future (10)

Forecasting financial markets

- Securities prices are a chaotic dynamical system.
- We cannot expect to forecast prices accurately.
- But we can aim to forecast a *distribution* of future prices.

Tasks for financial forecasting

- What is an upper bound on the risk of a portfolio of securities?
- What is an optimal trading strategy?
- The problem: Given time series $\{x_t^{(1)}\}_{t \in T}$ and related series $\{y_t^{(1)}\}_{t \in T}, \ldots, \{y_t^{(N)}\}_{t \in T}$

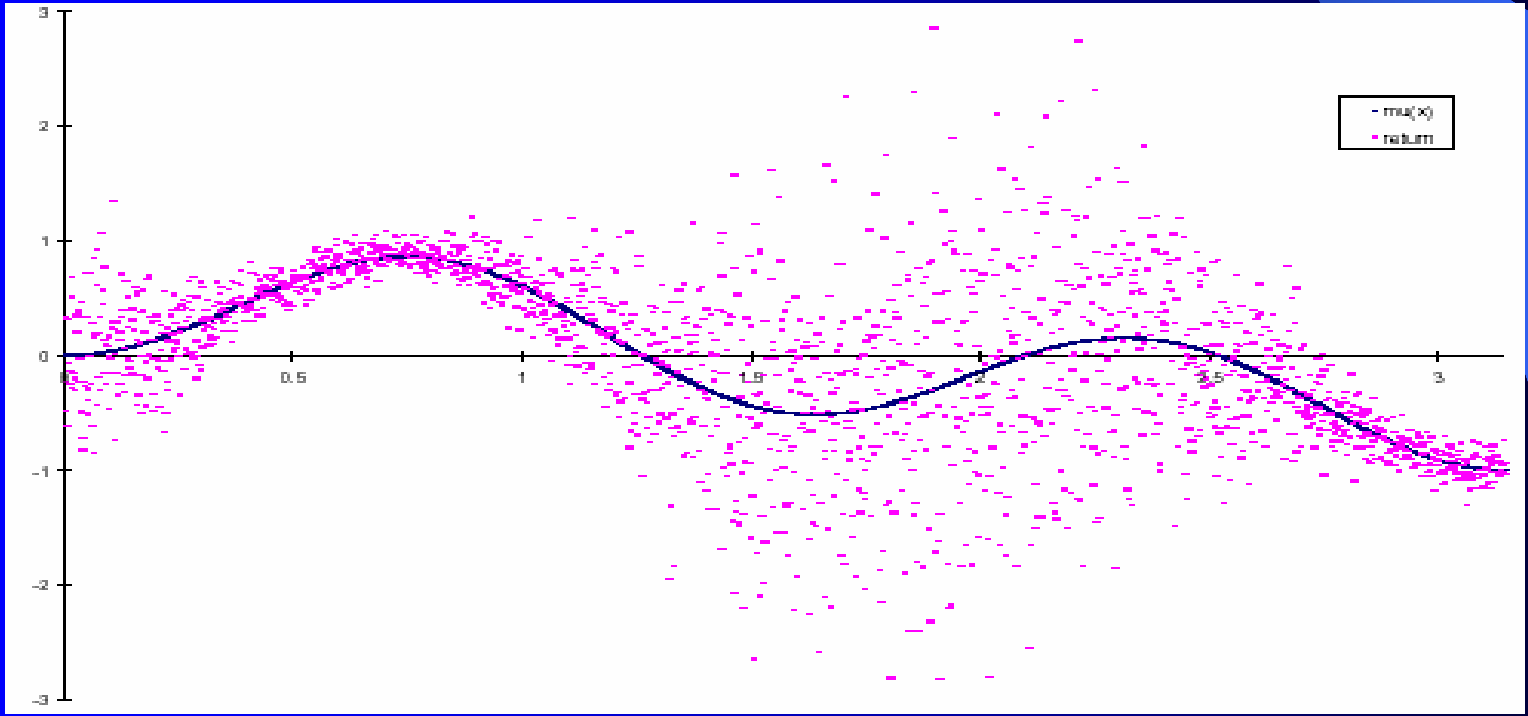
The distribution of prices

- The prices of stocks, bonds, currencies, commodities, and their derivatives, are approximately log-Normal. [Hull]
- Quote: If $X \sim \mathrm{lN}(\mu,\sigma^2)$ then $\log X \sim \mathrm{N}(\mu,\sigma^2)$.

A useful fact!

Density estimation

- Now we want to estimate the mean and variance, where $\mu(\underline{x};t)$ and $\sigma^2(\underline{x};t)$ are functions of all available data.



Estimating density parameters

How are μ and σ^2 constrained?

- Univariate case: $\mu \in \Re$, $\sigma^2 \in [0,\infty)$.
- Multivariate case: $\underline{\mu} \in \Re^n$, Σ must be positive definite.

Learning algorithms must be designed for these constraints, as in [Williams 95].

Financial forecasting: review

- We can model securities prices as log-Normal distributions. [Hull]
- We can estimate the density parameters ($\underline{\mu}(\cdot; t)$, $\Sigma(\cdot; t)$) using a learning algorithm with constrained outputs [e.g. Williams 95].
- Our prediction for $X_{t+1} := \log(S_{t+1})$ is then

$$N(\underline{\mu}(\underline{x}; t+1), \Sigma(\underline{x}; t+1))$$

Progress

- ✓ Applications of pattern recognition (5 min)
- ✓ Case 1: Classifying skin in images (12)
- ✓ Case 2: Forecasting financial markets (5)
- Formalization of concepts (5)
- Past and future (10)

A common thread

- What do the previous two examples have in common?

The first was a *classification* task.

The second was a *regression* task.

A general formulation

Suppose we are given observed data

$(x_1, t), \ldots, (x_m, t_m) \in X \times \{\pm 1\}$ classification

or

$(x_1, t), \ldots, (x_m, t_m) \in X \times \Re$ regression

Classification is a discrete case of the regression problem.

We wish to infer the function y that underlies the observed data t.

Parametric approaches to regression

We express the unknown function $y(\cdot)$ in terms of a function $y(\cdot; \mathbf{w})$ with parameters $\mathbf{w}$.

We then infer the parameters $\mathbf{w}$.

Examples of parametrizations for non-linear regression

Feedforward neural network:

$$y(\mathbf{x};\mathbf{w}) = \sum_{h=0}^{H} w_h \tanh\left(\sum_{i=1}^{I} w_{hi} x_i + w_{h0}\right) + w_0$$

Fourier series:

$$y(x;\mathbf{w}) = w_0/2 + \sum_{k=1}^{\infty} w_k \cos {}^{\pi kx}\!/_{m} + \sum_{k=1}^{\infty} w_k \sin {}^{\pi kx}\!/_{m}$$

Diversion: a new concept

Fourier series extend naturally to more general *function spaces*.

Generalized Fourier series

Choose a set $\{\varphi_k(x)\}$ that is orthogonal and linearly independent.

Then $\{\varphi_k(x)\}$ is the basis of a function space.

If $y \in \mathrm{span}<\varphi_k>$ we can now parametrize y as:

$$y(\mathbf{x};\mathbf{w}) = \sum_k w_k \varphi_k(\mathbf{x})$$

Orthogonal functions

Definition:

The set of functions $\{\varphi_k(\mathbf{x})\}$ for k=1,2,… is **orthogonal** on [a,b] if the inner product

$$<\varphi_m(x),\ \varphi_n(x)> \ = 0$$

whenever $m \neq n$.

$$\left(\text{Here } \langle \varphi_m, \varphi_n \rangle \text{ denotes } \int_a^b \varphi_m(\mathbf{x})\varphi_n(\mathbf{x})d\mathbf{x}. \right)$$

Function spaces

What have we achieved?

- Our non-linear problem is now linear.
- We can use all the tools of linear algebra.

Review: a mathematical formulation

- Classification is just a sub-problem of regression.

1. Parametrize the unknown function as:

$$y(\mathbf{x};\mathbf{w}) = \sum_k w_k \varphi_k(\mathbf{x})$$

2. Infer the parameters **w** using a learning algorithm.

Progress

- ✓ Applications of pattern recognition (5 min)
- ✓ Case 1: Classifying skin in images (12)
- ✓ Case 2: Forecasting financial markets (5)
- ✓ Formalization of concepts (5)
- Past and future (10)

Machine learning: past and future

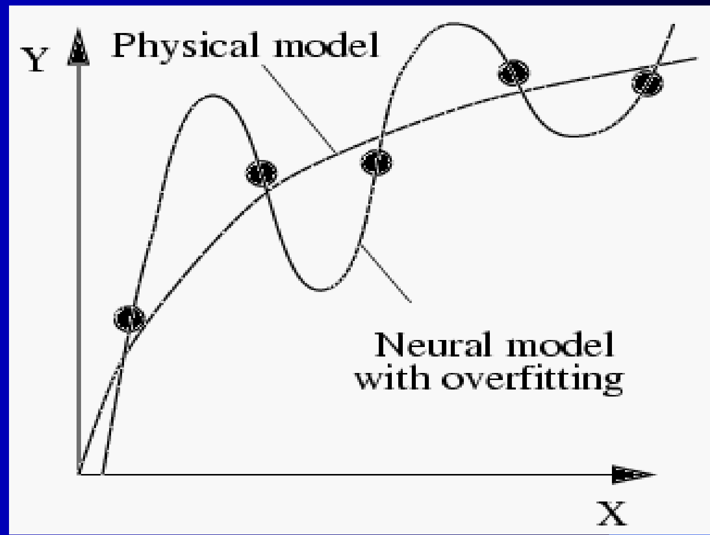
- Neural networks have generated much interest.
- Neural networks have solved some useful problems. [NN FAQ]
- Other learning methods can be even better.

What do neural networks do?

Approximate arbitrary functions from training data.

What is wrong with neural networks?

1. The ‘overfitting’ problem



2. Domain knowledge is hard to utilize.

3. We have no bounds on generalization performance.

One idea for machine learning

Let the number of hidden units $\rightarrow \infty$.

The implicit Bayesian prior is then a class of Gaussian Process [Neal 96].

This suggests discarding parametrized networks and working directly with GPs. See [MacKay 97].

Gaussian processes

- … are probability distributions on a space of functions.
- … are smoothing devices.
- … are well-understood [Thiele 1880 !!].
- … cannot yet be applied to more than ≈10 000 data points [MacKay 97].

A second idea for learning

Find the right balance between …

- … the fit to the training set …

and

- … the ‘learning capacity’ of the machine.

‘Learning capacity’

A botanical example!

A formal definition:

- ‘VC dimension’ [Vapnik 95]

Introducing ‘support vector machines’

- An outgrowth of statistical learning theory.
- Developed by a Russian mathematician, Vapnik. [Vapnik 95]
- Could be applied in many pattern-recognition contexts.

Classifying with SV machines

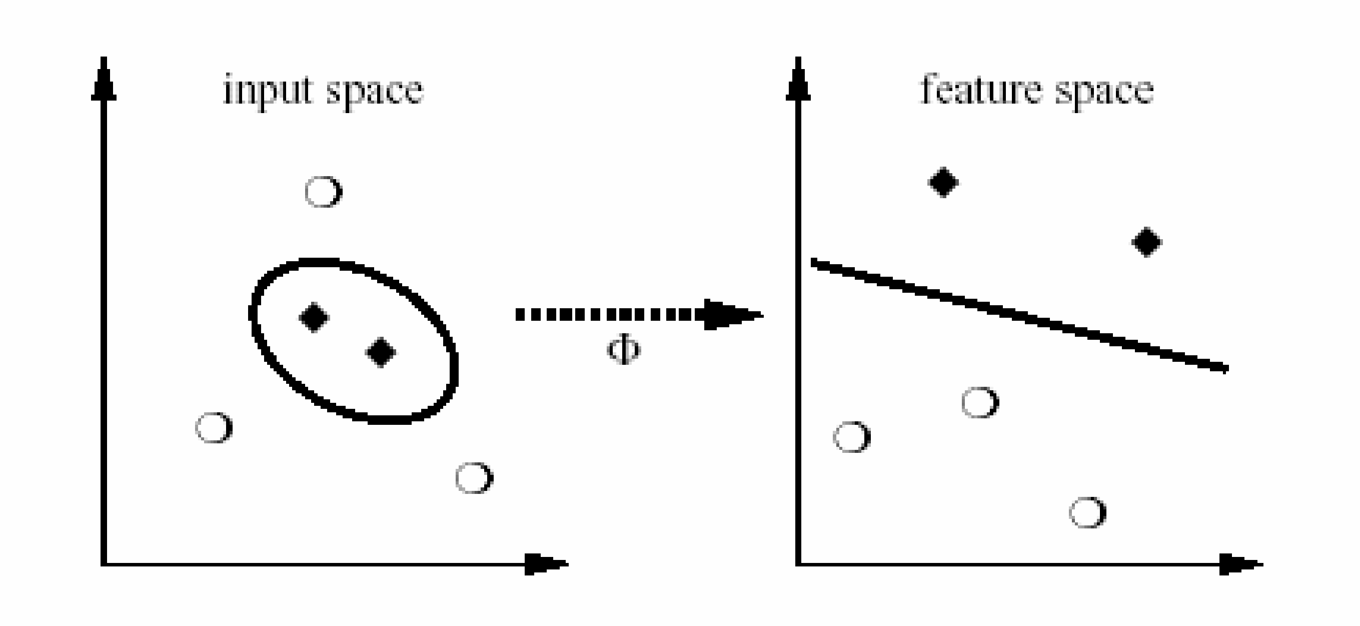
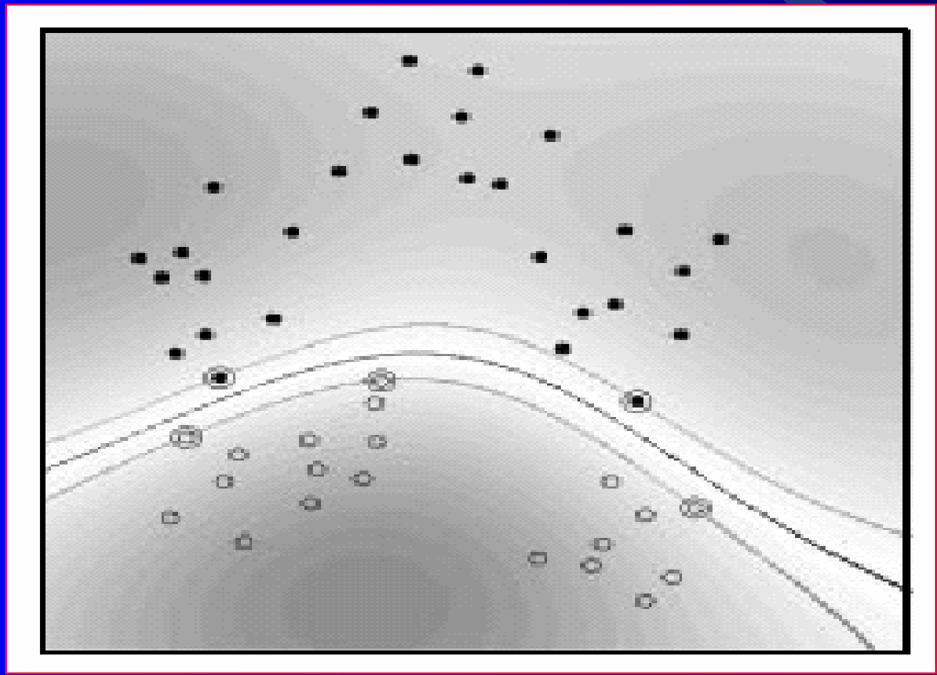


Diagram from [Schoelkopf 00]

An example of a SV classifier

[Schoelkopf 00]



Only some of the training data, called ‘support vectors’, are relevant.

A third idea for learning

How do we recognize a new pattern $\mathbf{x}$?

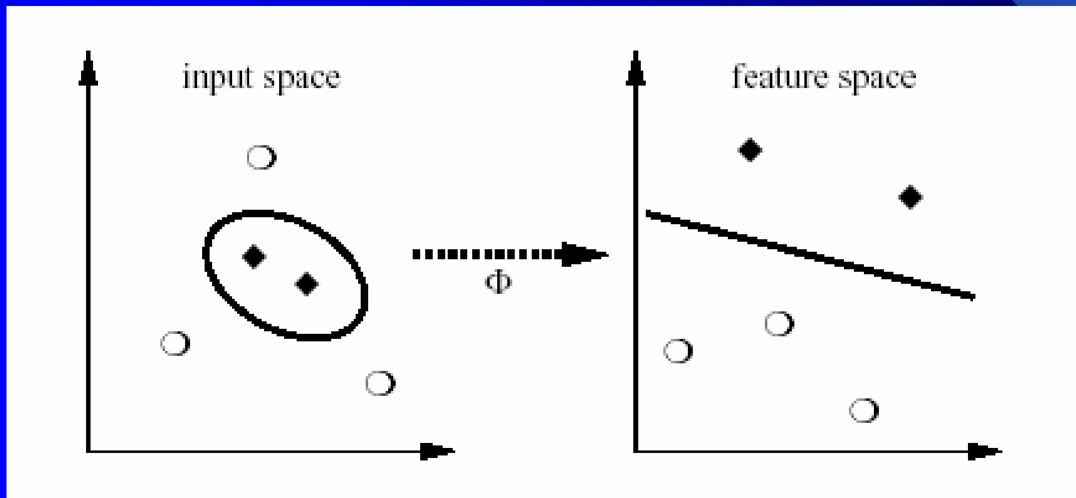
- We choose t so that $(\mathbf{x}, t)$ is similar to the training data.

- So: define a measure of similarity:

$$k : X \times X \rightarrow \Re$$

Features spaces

- As before, define a mapping Φ:



$$\Phi : \underbrace{X}_{\text{input space}} \rightarrow \underbrace{F}_{\text{feature Hilbert space}}$$

Reducing the dimensionality

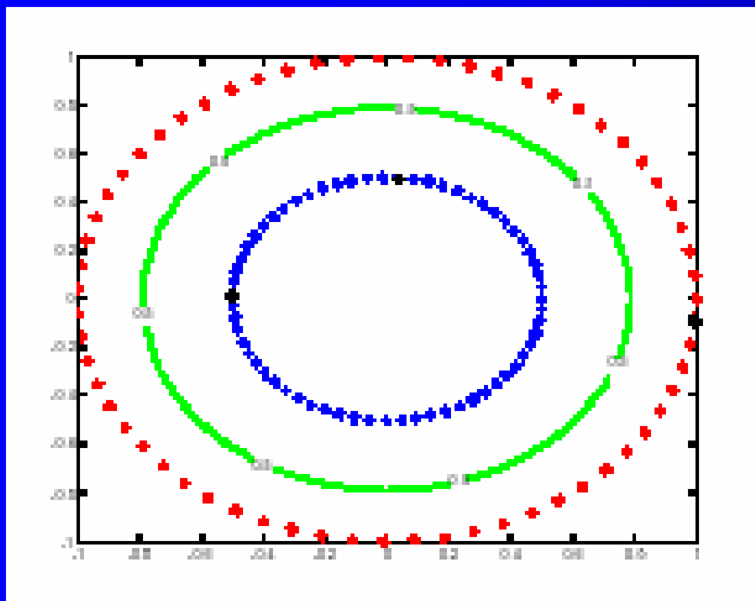
The transformed data $\mathbf{x} = \Phi(x)$ lie in a subspace of F.

What is the dimension of the subspace?

$$\dim S = \text{rank}(\text{kernel matrix of } \Phi)$$

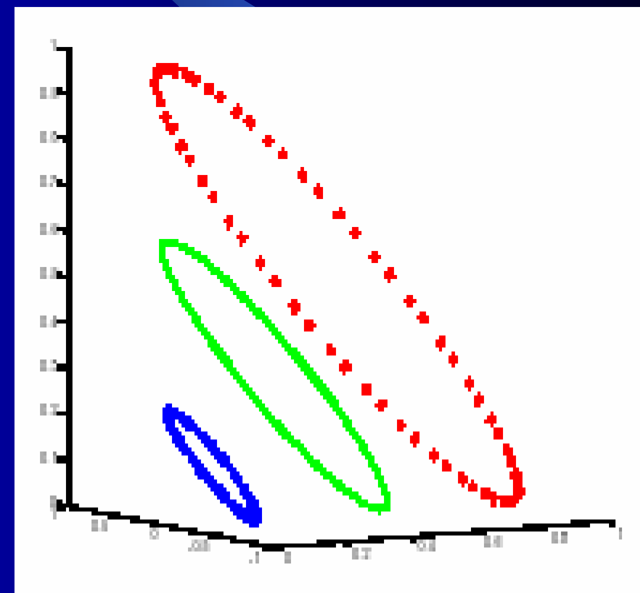
generally *much* smaller than the training set

An illustration



2-D Input space X

Φ



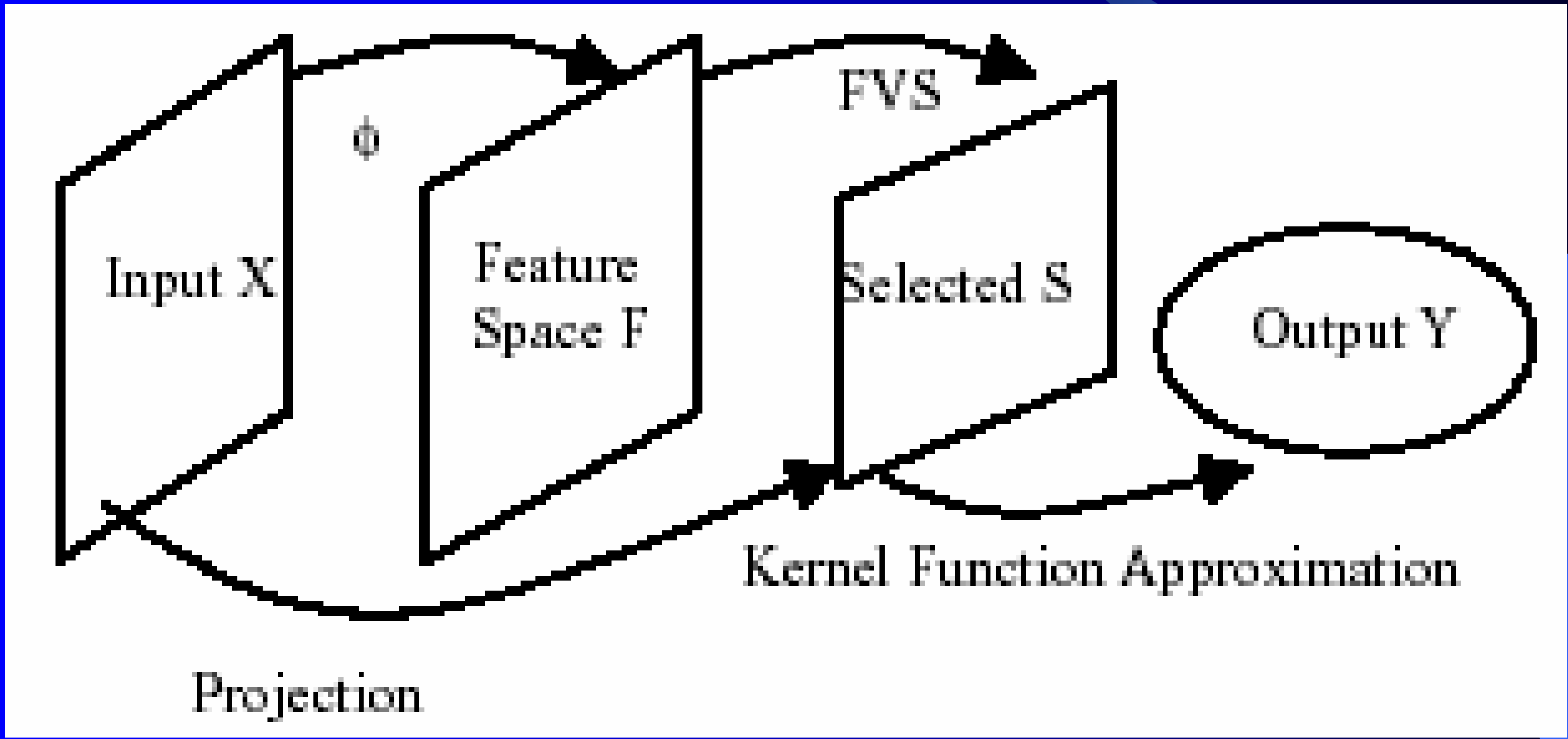
3-D feature space F

Feature vector selection

A new idea:

- Use kernels to preserve the geometrical structure [Schoelkopf 98].
- Project the training examples onto a lower-dimensional subspace of features [Baudat 00].
- Use classical regression techniques.

Feature vector selection (2)



Does this work?

- Yes! [Baudat 00]

Kernel methods can also be integrated with other tools –

(Linear discriminant analysis [Fukanaga 90]

Principal component analysis [ibid] …)

– in extracting features.

Review: past and future

- Neural networks may be superseded by new mathematical structures.

Recent concepts include:

- Gaussian processes [MacKay 97]
- Support vector machines [Burges 98]
- Kernel methods [Schoelkopf 00]

Summary

Pattern recognition …

- has important applications
- can be mathematically elegant
- will develop dramatically this decade.

Where to Get More Information

- Tutorials and papers at kernel-machines.org
- Slides and references at edschofield.com