

# UNSUPERVISED LEARNING 2011

## LECTURE : INTRODUCTION

Rita Osadchy

Some of the slides are due to O. Veksler and Z. Ghahramani

# About the course

- ◎ Course Homepage:

[http://www.cs.haifa.ac.il/~rita/uml\\_course/course.html](http://www.cs.haifa.ac.il/~rita/uml_course/course.html)

- ◎ Office hours: request meeting by email

- ◎ Contact:

- You contact me by email: [rita@cs.haifa.ac.il](mailto:rita@cs.haifa.ac.il)
- I contact you by email: All announcement and guidelines will be distributed by email.

You must send me an email by  
November 14 from your **active address**  
with the subject “UML course contact”.

- Those who do not send their contact address on time will not be added to the contact list!!!

# About the course

## ⦿ **Mandatory Prerequisites**

The course assumes some basic knowledge of probability theory and linear algebra; for example, you should be familiar with

- Joint and marginal probability distributions
- Normal (Gaussian) distribution
- Expectation and variance
- Statistical correlation and statistical independence
- Eigen value decomposition

Links to tutorial in the course homepage.

**Suggested Prerequisites:** Introduction to Machine Learning

# About the course

- ◎ **Textbook:**

Pattern Recognition and Machine Learning,  
by Christopher Bishop. Springer, August  
2006.

- ◎ Course material: lecture notes and  
reading material in:

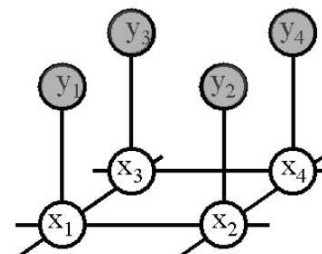
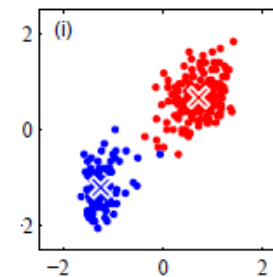
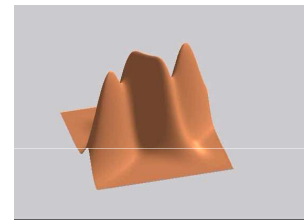
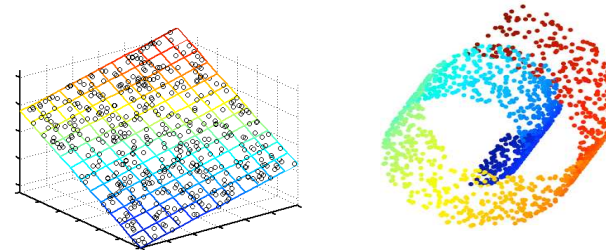
[http://www.cs.haifa.ac.il/~rita/UML\\_course/course.htm](http://www.cs.haifa.ac.il/~rita/UML_course/course.htm)

# About the course

- Grading: Exam

# Syllabus at glance

- Dimensionality Reduction
  - Linear
  - Manifold Mapping
  - KPCA
- Density Estimation
  - Mixture of Gaussians
  - EM
  - Factor Analysis
- Clustering
  - K-means
  - Spectral Clustering
- Introduction to Graphical Models
  - Belief Propagation
  - MRF



More topics (see the extended syllabus) if time permits.

# Goal

- **Machine learning** is a study of algorithms that improve their performance at some task with experience.
- **Machine learning** is an interdisciplinary field focusing on both the mathematical foundations and practical applications of systems that learn, reason and act.
- The goal of this course: to introduce basic concepts, models and algorithms in machine learning with particular emphasis on **unsupervised learning**.

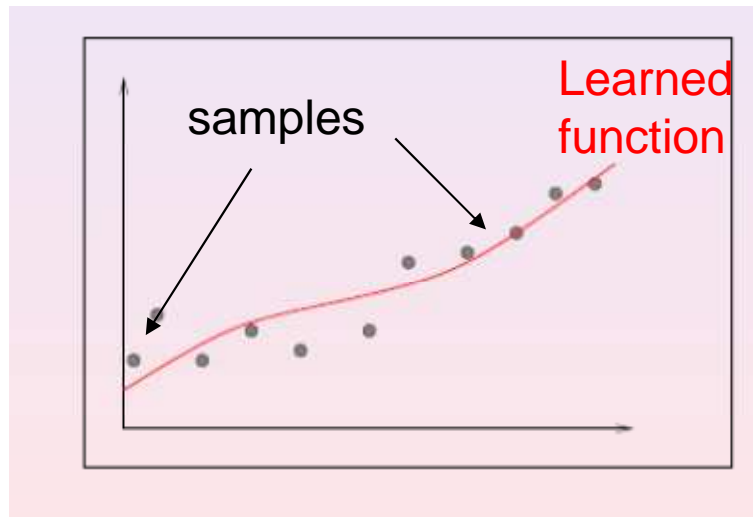
# Types of Learning Problems

Assume that we are given a set of training inputs (e.g., a series of sensory inputs)  $x_1, x_2, \dots, x_n$

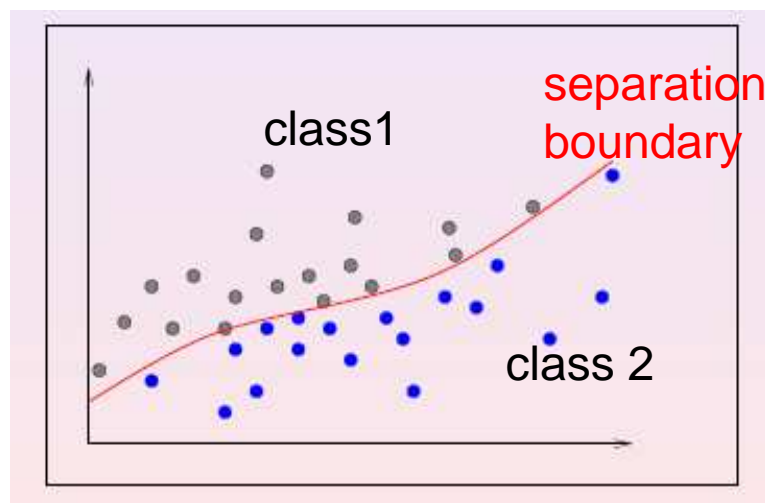
- ⦿ **Supervised learning:** We are also given the desired outputs  $y_1, y_2, \dots, y_n$  and the goal is to learn to **produce the correct output** given a new input.
- ⦿ **Unsupervised learning:** The goal is to **build a model of  $x$**  that can be used for reasoning, decision making, predicting things, communicating etc.
- ⦿ **Reinforcement learning:** where we only get feedback in the form of how well we are doing (e.g., the outcome of the game). The goal is to learn to act in a way that **maximises rewards** in the long term.



# Two kinds of Supervised Learning



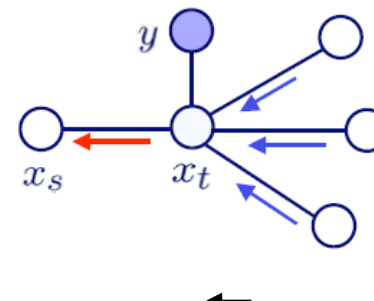
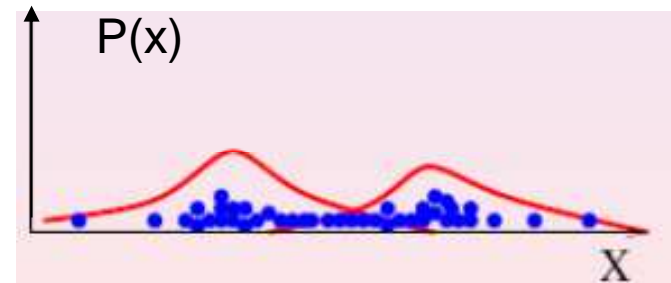
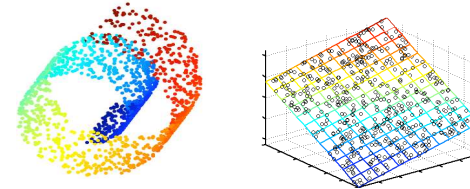
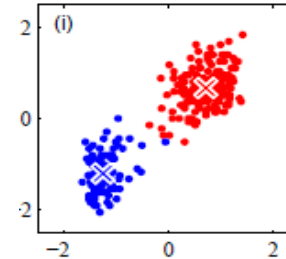
- **Regression:** Learn a continuous input-output mapping from a limited number of examples in order to predict the output accurately for new inputs.



- **Classification:** outputs are discrete variables (category labels). Learn a decision boundary that separates one class from the other to classify new inputs correctly

# Goals of Unsupervised Learning

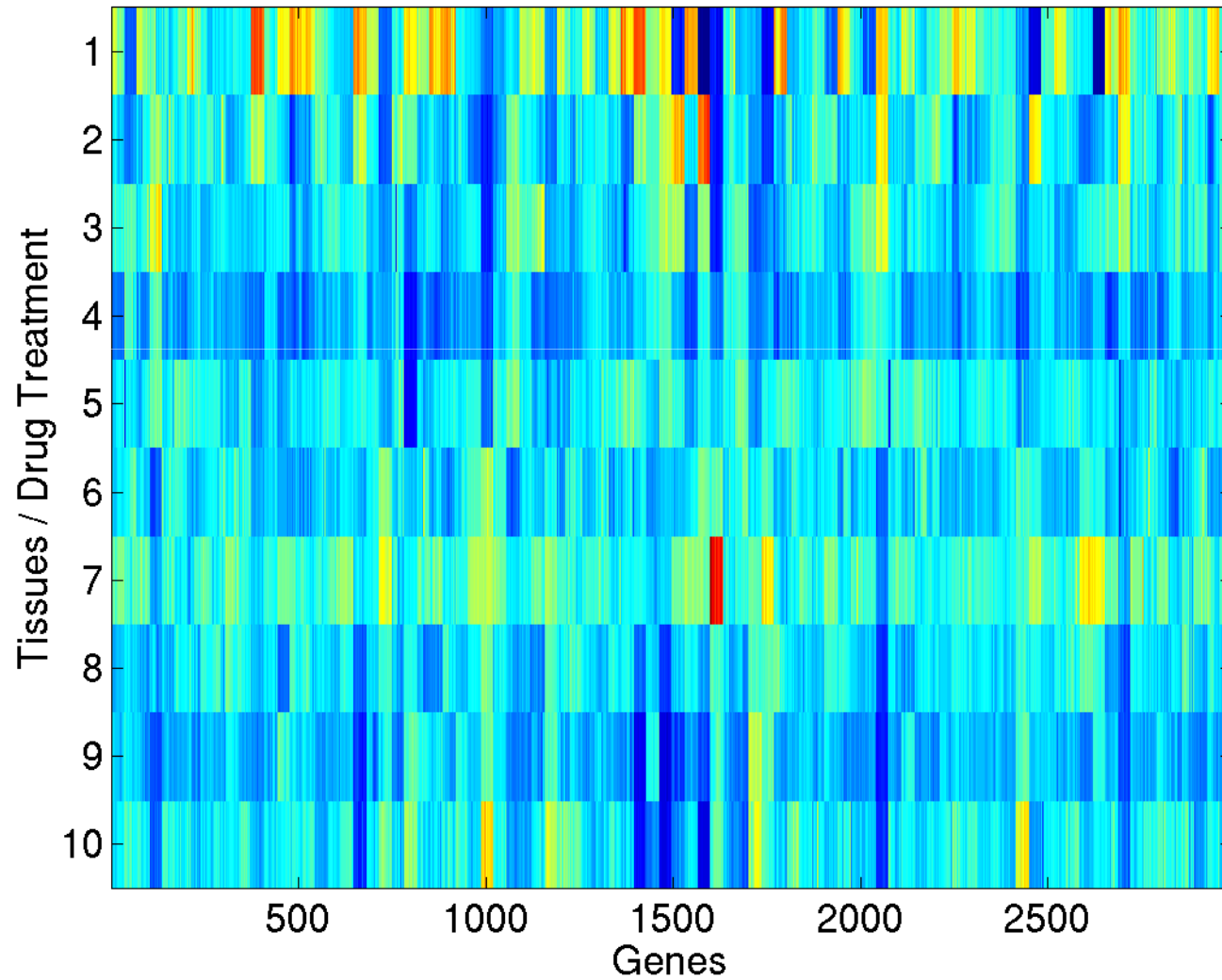
- **Clustering:** discover “clumps” of points
- **Embedding:** discover low-dimensional manifold or surface near which the data lives.
- **Density Estimation.** Find a function  $f$  such  $f(X)$  approximates the probability density of  $X$ ,  $p(X)$ , as well as possible.
- Finding good explanations (**hidden causes**) of the data;



# Uses of Unsupervised Learning

- ⦿ data compression
- ⦿ outlier detection
- ⦿ help classification
- ⦿ make other learning tasks easier
- ⦿ use as a theory of human learning and perception

# Applications



# Applications

## 🕒 Profiling Web Users

The screenshot shows a Google search results page for the query "machine vision and applications". The search bar at the top contains the text "mach[ne vision and applications]". Below the search bar, there are several search suggestions: "machine vision and applications", "machine learning course stanford", "machu picchu", and "machine gun kelly". The search results are organized into a sidebar on the left and a main content area on the right. The sidebar includes categories like "Everything", "Images", "Maps", "Videos", "News", "Shopping", and "More". The main content area displays several search results, including "Machine Vision and Applications - Springer", "Machine Vision and Applications (Editorial Board)", "Machine Vision and Applications - Editorial Manager", "Machine vision - Wikipedia, the free encyclopedia", "Computer vision - Wikipedia, the free encyclopedia", and "Machine Vision and Applications". On the right side of the page, there are several advertisements for machine vision systems, such as "Machine Vision Cameras", "Cognex Vision Systems", and "Optical Inspection System".

**Google** Search

mach[ne vision and applications]

machine vision and applications Remove

machine learning course stanford Remove

machu picchu

machine gun kelly

Advanced search

**Everything**

Images

Maps

Videos

News

Shopping

More

**All results**

Related searches

Timeline

More search tools

**Machine Vision and Applications - Springer**  
www.springer.com › Home › Computer Science › Image Processing +7  
Sponsored by the International Association for Pattern Recognition, this journal publishes high-quality, technical contributions in **machine vision** research and ...

**Machine Vision and Applications (Editorial Board)**  
www.springer.com › Home › Computer Science › Image Processing +7  
Editorial Board - Sponsored by the International Association for Pattern ...

Show more results from springer.com

**Machine Vision and Applications - Editorial Manager**  
www.editorialmanager.com/mvap/ +7

**Machine vision - Wikipedia, the free encyclopedia**  
en.wikipedia.org/wiki/Machine\_vision +7  
Jump to **Applications: Applications**. The primary uses for **machine vision** are automatic inspection and robot guidance. Common MV **applications** include ...

**Computer vision - Wikipedia, the free encyclopedia**  
en.wikipedia.org/wiki/Computer\_vision +7  
Jump to **Applications for computer vision**: One of the most prominent **application** fields is medical **computer vision** or ...

Show more results from wikipedia.org

**Machine Vision and Applications**  
www.informatik.uni-trier.de/~ley/db/journals/mva/index.html +7  
**Machine Vision and Applications** **Machine Vision and Applications** @

Ads

**Machine Vision Cameras**  
www.net-gmbh.com +7  
USB, IEEE1394, GigE, Smart and Line scan camera for **vision applications**

מערכות ראייה ממוחשבת  
www.visionsystem.com +7  
מדיטל - Keyence Israel 073-2000208  
מצלמות עם בקרה ללא צורך במחשב

**Cognex Vision Systems**  
www.cognex.com +7  
**Vision** systems for inspecting, guiding, gauging & identifying.

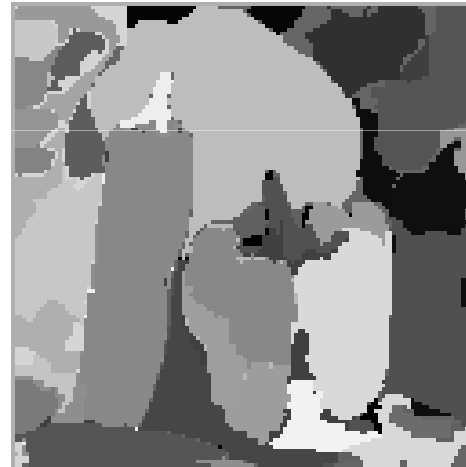
פתחנות ראייה ממוחשבת  
www.vibe-technologies.co.il +7  
אפליקציות עיבוד תמונה, בקרה אוטומטית  
לייעוץ ממוחשב לחץ כאן

**Optical Inspection System**  
www.brossh.com +7  
Non contact 2D and 3D Inspection  
Metrology and Defects

**Machine Vision**

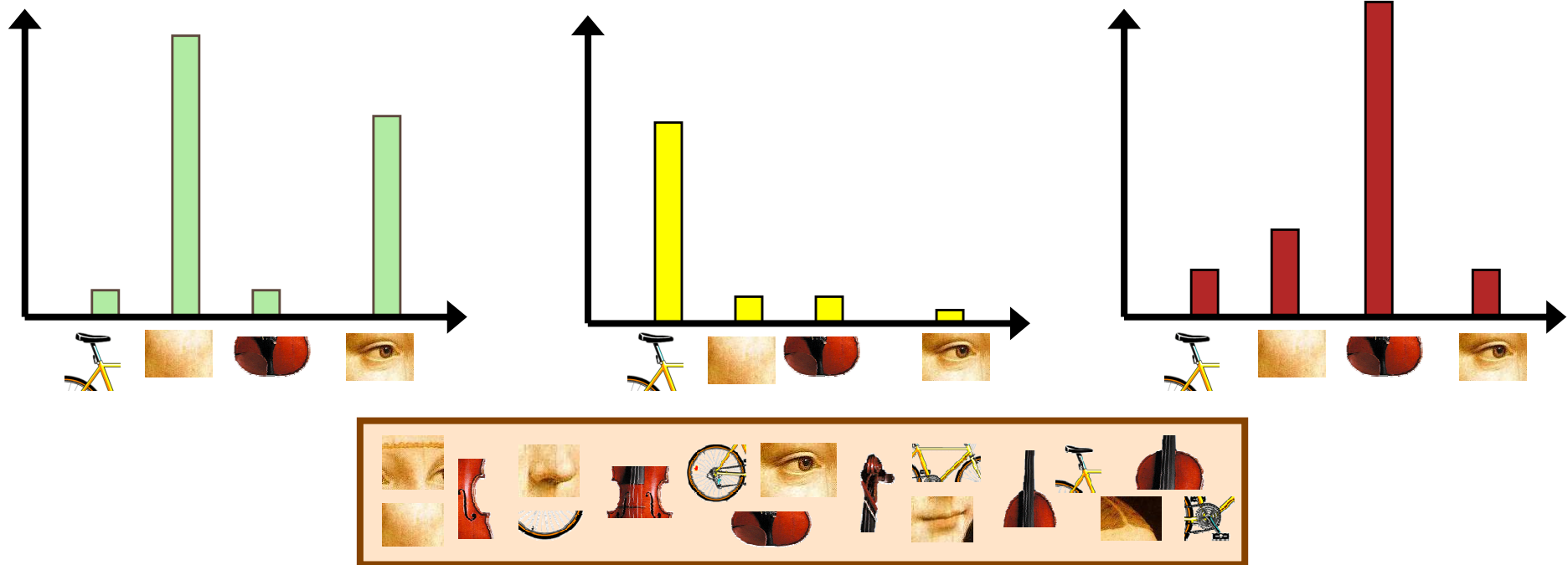
# Applications

- Image Processing



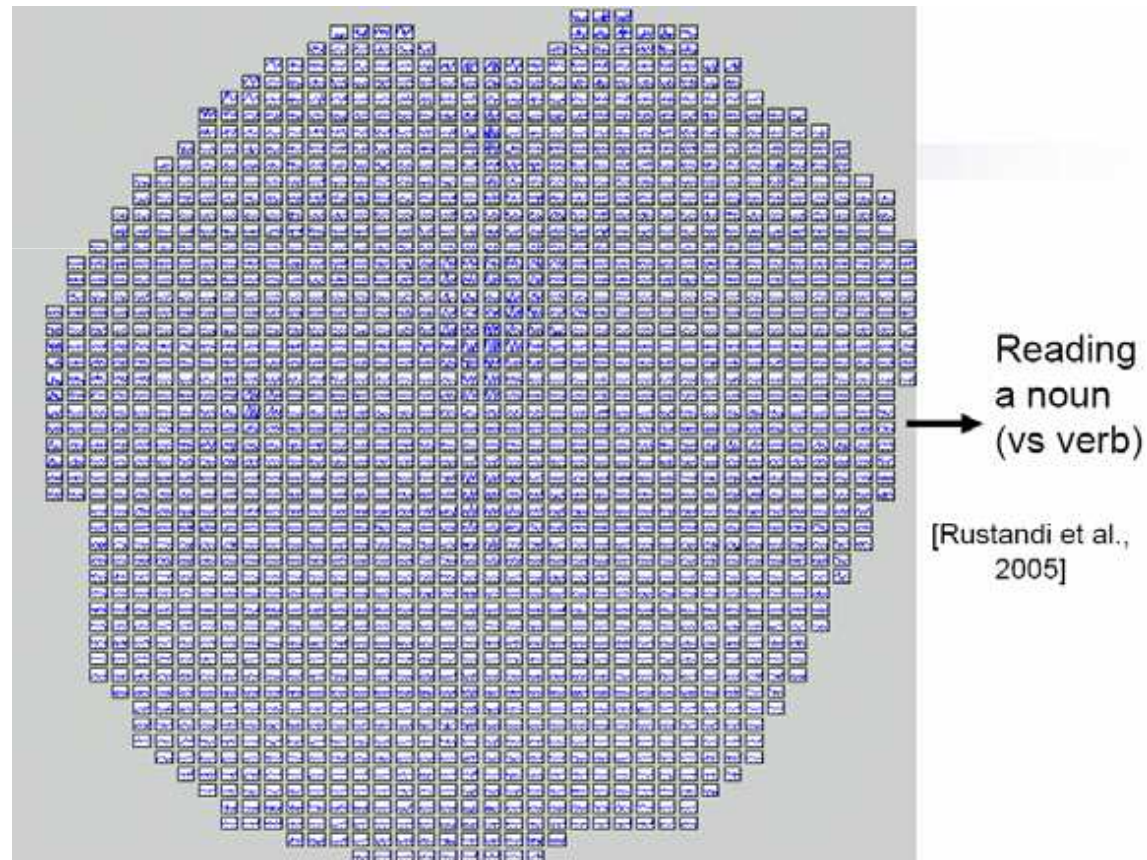
# Applications

- Visual Object Categorization



# Applications

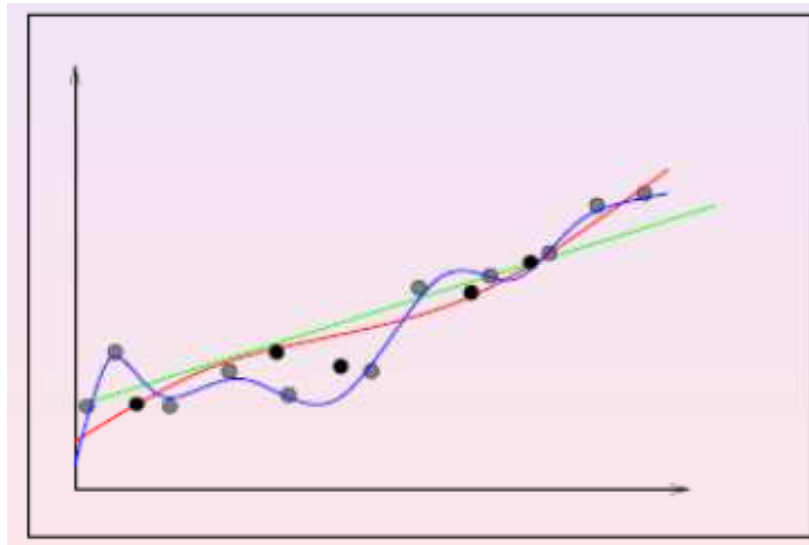
- Understanding Brain Activities





# Why Learning is Difficult?

- Given a finite amount of training data, you have to derive a relation for an infinite domain
- In fact, there is an infinite number of such relations

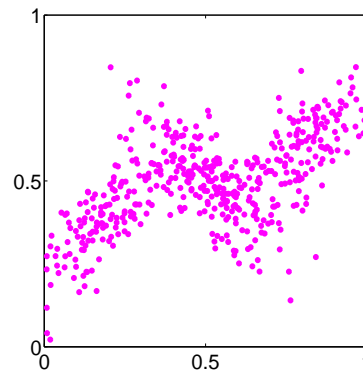
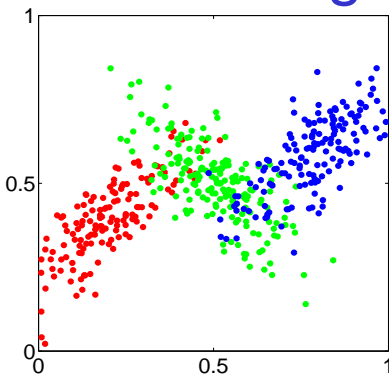


- ... the hidden test points...

# What about Unsupervised Learning

- ☺ Obtaining unlabelled data is easier (at least for some applications)
  - More training data – better models
- ☹ We know very little about the data.
  - Less than in supervised learning:

Build classifier for 3 classes, given the training data



# A Recap from Introductory Course

- The next several slides summarize few topics from the Introductory course in Machine Learning.
- These definitions and formulations are the basis required to understand the material in this course.
- For more details see:  
[http://www.cs.haifa.ac.il/~rita/ml\\_course/course.html](http://www.cs.haifa.ac.il/~rita/ml_course/course.html)

# Basic Rules of Probability

Probabilities are non-negative  $P(x) \geq 0 \forall x$ .

Probabilities normalise:  $\sum_{x \in \mathcal{X}} P(x) = 1$  for distributions if  $x$  is a discrete variable and  $\int_{-\infty}^{+\infty} p(x) dx = 1$  for probability densities over continuous variables

The **joint probability** of  $x$  and  $y$  is:  $P(x, y)$ .

The **marginal probability** of  $x$  is:  $P(x) = \sum_y P(x, y)$ , assuming  $y$  is discrete.

The **conditional probability** of  $x$  given  $y$  is:  $P(x|y) = P(x, y)/P(y)$

**Bayes Rule:**

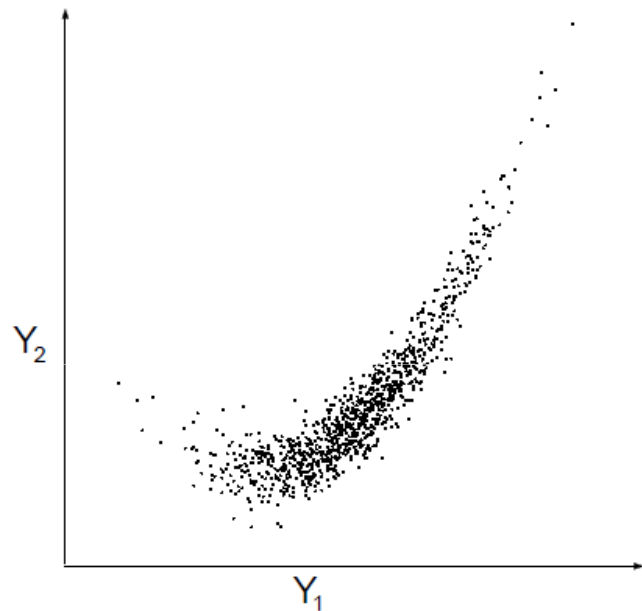
$$P(x, y) = P(x)P(y|x) = P(y)P(x|y) \quad \Rightarrow \quad \boxed{P(y|x) = \frac{P(x|y)P(y)}{P(x)}}$$

**Warning:** I will not be obsessively careful in my use of  $p$  and  $P$  for probability density and probability distribution. Should be obvious from context.

# Probability Density estimation

- Parametric methods – assume we know the shape of the distribution, but not the parameters  $\theta$ .
  - Maximum Likelihood Estimation
  - Bayesian Estimation
- Non parametric methods – the form of the density is entirely determined by the data without any model.

# Simple Statistical Modeling : modeling correlations



Assume:

- we have a data set  $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$
- each data point is a vector of  $D$  features:  
 $\mathbf{y}_i = [y_{i1} \dots y_{iD}]$
- the data points are i.i.d. (independent and identically distributed).

One of the simplest forms of unsupervised learning: model the **mean** of the data and the **correlations** between the  $D$  features in the data

We can use a multivariate Gaussian model:

$$p(\mathbf{y}|\mu, \Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(\mathbf{y} - \mu)^\top \Sigma^{-1}(\mathbf{y} - \mu) \right\}$$

# Maximum Likelihood Parameter Estimation

- ⦿ Parameters  $\theta$  are unknown but fixed (i.e. not random variables).
- ⦿ Given the training data, choose the parameter value  $\theta$  that makes the data most probable (i.e., maximizes the probability of obtaining the sample that has actually been observed)

# Maximum Likelihood Estimation

- Consider the following function, which is called **likelihood of  $\theta$**  with respect to the set of samples  $D$

$$p(\mathbf{D} | \theta) = \prod_{k=1}^{k=n} p(\mathbf{x}_k | \theta) = F(\theta)$$

- **Maximum likelihood estimate** (abbreviated **MLE**) of  $\theta$  is the value of  $\theta$  that maximizes the likelihood function  $p(\mathbf{D} | \theta)$

$$\hat{\theta} = \underset{\theta}{\operatorname{arg\,max}}(p(\mathbf{D} | \theta))$$



# ML Estimation of a Gaussian

Data set  $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ , likelihood:  $p(Y|\mu, \Sigma) = \prod_{n=1}^N p(\mathbf{y}_n|\mu, \Sigma)$

Maximize likelihood  $\Leftrightarrow$  maximize log likelihood

**Goal:** find  $\mu$  and  $\Sigma$  that maximise log likelihood:

$$\begin{aligned}\mathcal{L} &= \log \prod_{n=1}^N p(\mathbf{y}_n|\mu, \Sigma) = \sum_n \log p(\mathbf{y}_n|\mu, \Sigma) \\ &= -\frac{N}{2} \log |2\pi\Sigma| - \frac{1}{2} \sum_n (\mathbf{y}_n - \mu)^\top \Sigma^{-1} (\mathbf{y}_n - \mu)\end{aligned}$$

**Note:** equivalently, minimise  $-\mathcal{L}$ , which is *quadratic* in  $\mu$

**Procedure:** take derivatives and set to zero:

$$\frac{\partial \mathcal{L}}{\partial \mu} = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{N} \sum_n \mathbf{y}_n \quad (\text{sample mean})$$

$$\frac{\partial \mathcal{L}}{\partial \Sigma} = 0 \quad \Rightarrow \quad \hat{\Sigma} = \frac{1}{N} \sum_n (\mathbf{y}_n - \hat{\mu})(\mathbf{y}_n - \hat{\mu})^\top \quad (\text{sample covariance})$$

# Bayesian Parameter Estimation

- $\theta$  is a random variable with prior  $p(\theta)$ 
  - Unlike MLE case,  $p(x|\theta)$  is a conditional density
- The training data  $D$  allow us to convert  $p(\theta)$  to a posterior probability density  $p(\theta|D)$  .
  - After we observe the data  $D$ , using Bayes rule we can compute the posterior  $p(\theta|D)$
- But  $\theta$  is not our final goal, our final goal is the unknown  $p(\mathbf{x})$
- Therefore a better thing to do is to maximize  $p(\mathbf{x}/D)$ , this is as close as we can come to the unknown  $p(\mathbf{x})$  !

# Bayesian Estimation: Formula for $p(\mathbf{x} | \mathbf{D})$

- From the definition of joint distribution:

$$p(\mathbf{x} | \mathbf{D}) = \int p(\mathbf{x}, \theta | \mathbf{D}) d\theta$$

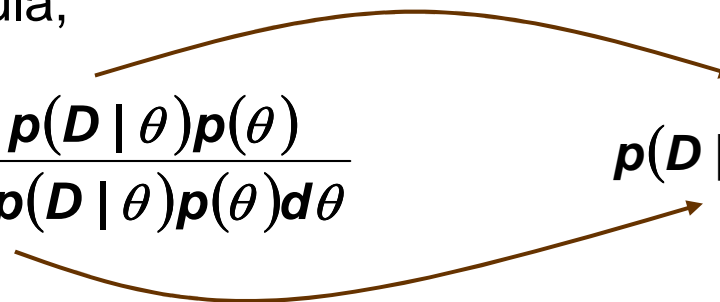
- Using the definition of conditional probability:

$$p(\mathbf{x} | \mathbf{D}) = \int p(\mathbf{x} | \theta, \mathbf{D}) p(\theta | \mathbf{D}) d\theta$$

- But  $p(\mathbf{x} | \theta, \mathbf{D}) = p(\mathbf{x} | \theta)$  since  $p(\mathbf{x} | \theta)$  is completely specified by  $\theta$

$$p(\mathbf{x} | \mathbf{D}) = \int \overset{\text{known}}{p(\mathbf{x} | \theta)} \overset{\text{unknown}}{p(\theta | \mathbf{D})} d\theta$$

- Using Bayes formula,

$$p(\theta | \mathbf{D}) = \frac{p(\mathbf{D} | \theta) p(\theta)}{\int p(\mathbf{D} | \theta) p(\theta) d\theta} \qquad p(\mathbf{D} | \theta) = \prod_{k=1}^n p(\mathbf{x}_k | \theta)$$


# Bayesian Estimation vs. MLE

*support  $\theta$  receives from  
the data*

$$p(\mathbf{x} | D) = \int \underbrace{p(\mathbf{x} | \theta)}_{\substack{\text{proposed model with} \\ \text{certain } \theta}} \underbrace{p(\theta | D)}_{\substack{\text{support } \theta \text{ receives from} \\ \text{the data}}} d\theta$$

- The above equation implies that if we are less certain about the exact value of  $\theta$ , we should consider a weighted average of  $p(\mathbf{x}/\theta)$  over the possible values of  $\theta$ .
- Contrast this with the MLE solution which always gives us a single model:

$$p(\mathbf{x} | \hat{\theta})$$

# Recommended Reading

- More basic topics from the introductory course that we will not talk about in this course, but are related to unsupervised learning
  - Parzen Window
  - Hierarchical Clustering
  - Naïve Bayese Model