Lecture 9: Introduction to Pattern Analysis

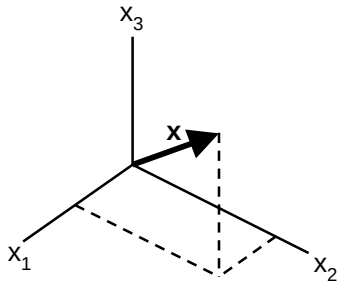
- Features, patterns and classifiers
- Components of a PR system
- An example
- Probability definitions
- Bayes Theorem
- Gaussian densities

Features, patterns and classifiers

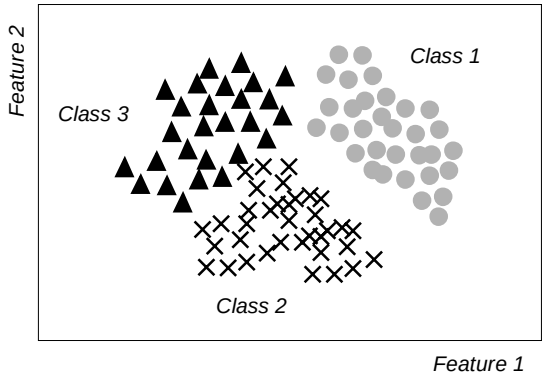
- **Feature**
  - Feature is any distinctive aspect, quality or characteristic
    - Features may be symbolic (i.e., color) or numeric (i.e., height)
  - The combination of *d* features is represented as a *d*-dimensional column vector called a **feature vector**
    - The d-dimensional space defined by the feature vector is called **feature space**
    - Objects are represented as points in feature space. This representation is called a **scatter plot**

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_d \end{bmatrix}$$

Feature vector



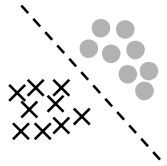
Feature space (3D)



Scatter plot (2D)

Features, patterns and classifiers

- **Pattern**
  - Pattern is a composite of traits or features characteristic of an individual
  - In classification, a pattern is a pair of variables $\{x, \omega\}$ where
    - $\boldsymbol{x}$ is a collection of observations or features (feature vector)
    - ω is the concept behind the observation (label)
- **What makes a "good" feature vector?**
  - The quality of a feature vector is related to its ability to discriminate examples from different classes
    - Examples from the same class should have similar feature values
    - Examples from different classes have different feature values

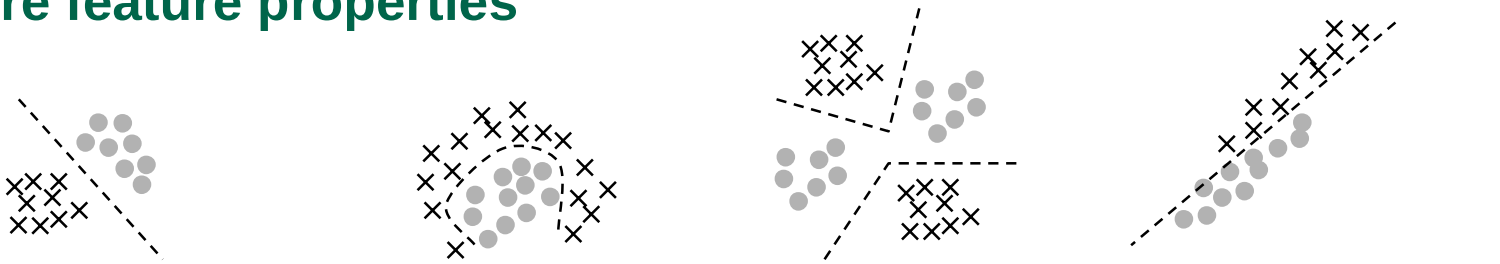


"Good" features

"Bad" features

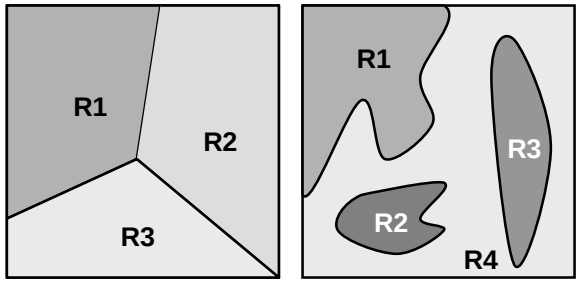
Features, patterns and classifiers

- **More feature properties**



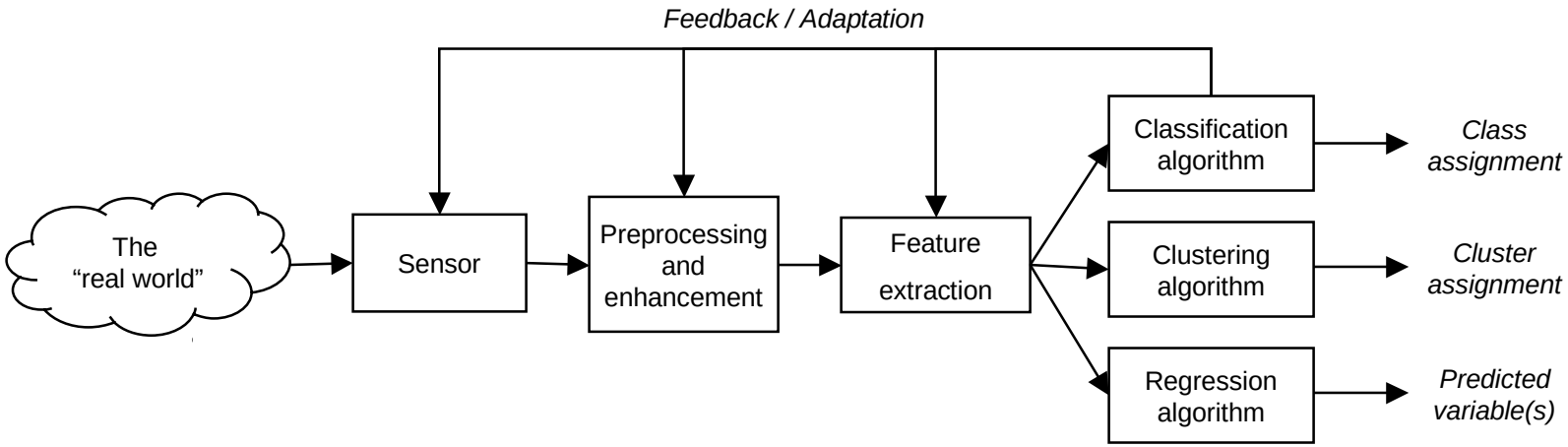
Linear separability *Non-linear separability* *Multi-modal* *Highly correlated features*

- **Classifiers**
  - The goal of a classifier is to partition feature space into class-labeled **decision regions**
  - Borders between decision regions are called **decision boundaries**



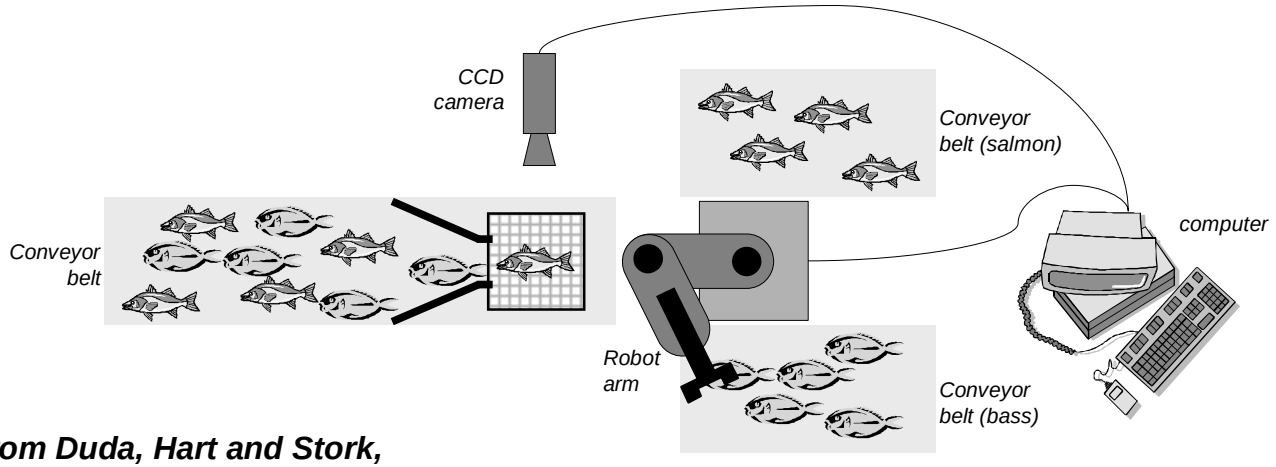
Components of a pattern rec. system

- **A typical pattern recognition system contains**
  - A sensor
  - A preprocessing mechanism
  - A feature extraction mechanism (manual or automated)
  - A classification or description algorithm
  - A set of examples (training set) already classified or described



An example

- **Consider the following scenario***
  - A fish processing plan wants to automate the process of sorting incoming fish according to species (salmon or sea bass)
  - The automation system consists of
    - a conveyor belt for incoming products
    - two conveyor belts for sorted products
    - a pick-and-place robotic arm
    - a vision system with an overhead CCD camera
    - a computer to analyze images and control the robot arm

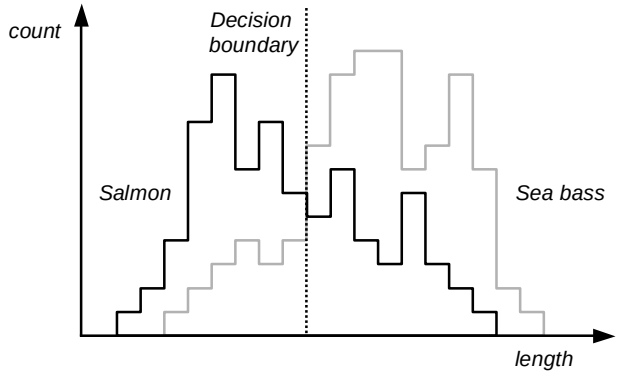


Adapted from Duda, Hart and Stork, Pattern Classification, 2nd Ed.

An example

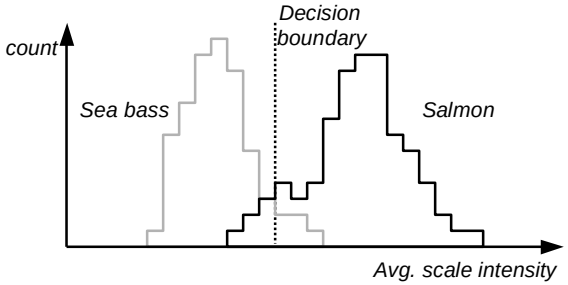
- **Sensor**
  - The camera captures an image as a new fish enters the sorting area
- **Preprocessing**
  - Adjustments for average intensity levels
  - Segmentation to separate fish from background
- **Feature Extraction**
  - Suppose we know that, on the average, sea bass is larger than salmon

- **Classification**
  - Collect a set of examples from both species
    - Plot a distribution of lengths for both classes
  - Determine a decision boundary (threshold) that minimizes the classification error
    - We estimate the system's probability of error and obtain a discouraging result of 40%
  - **What is next?**

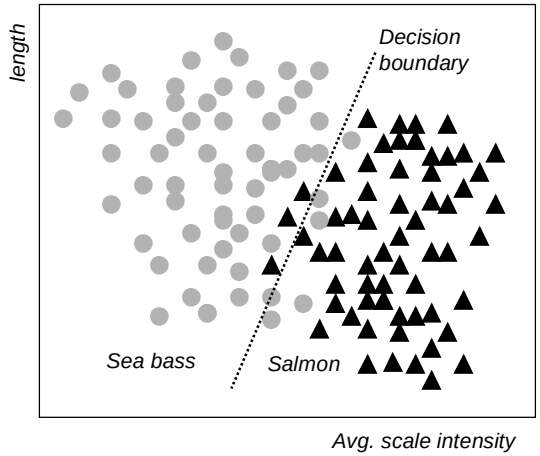


An example

- **Improving the performance of our PR system**
  - Committed to achieve a recognition rate of 95%, we try a number of features
    - Width, Area, Position of the eyes w.r.t. mouth...
    - only to find out that these features contain no discriminatory information
  - Finally we find a “good” feature: average intensity of the scales

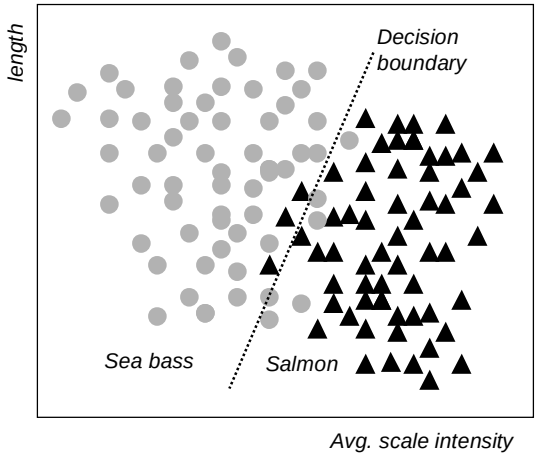


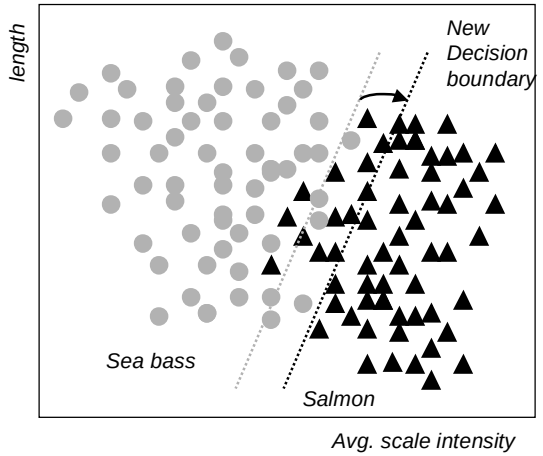
- We combine “*length*” and “*average intensity of the scales*” to improve class separability
- We compute a linear discriminant function to separate the two classes, and obtain a classification rate of 95.7%



An example

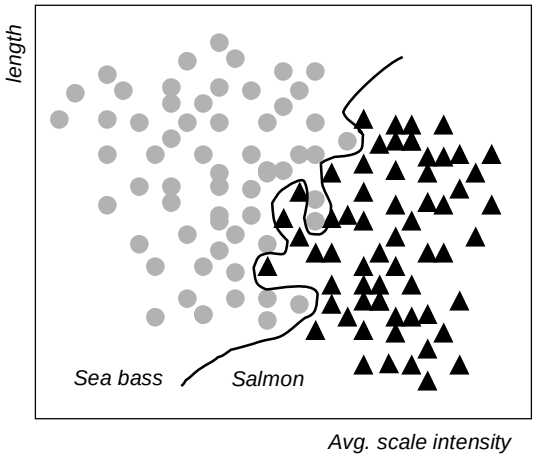
- **Cost Versus Classification rate**
  - Is classification rate the best objective function for this problem?
    - The **cost** of misclassifying salmon as sea bass is that the end customer will occasionally find a tasty piece of salmon when he purchases sea bass
    - The **cost** of misclassifying sea bass as salmon is a customer upset when he finds a piece of sea bass purchased at the price of salmon
  - We could intuitively shift the decision boundary to minimize an alternative cost function





An example

- **The issue of generalization**
  - The recognition rate of our linear classifier (95.7%) met the design specs, but we still think we can improve the performance of the system
    - *We then design an artificial neural network with five hidden layers, a combination of logistic and hyperbolic tangent activation functions, train it with the Levenberg-Marquardt algorithm and obtain an impressive classification rate of 99.9975% with the following decision boundary*

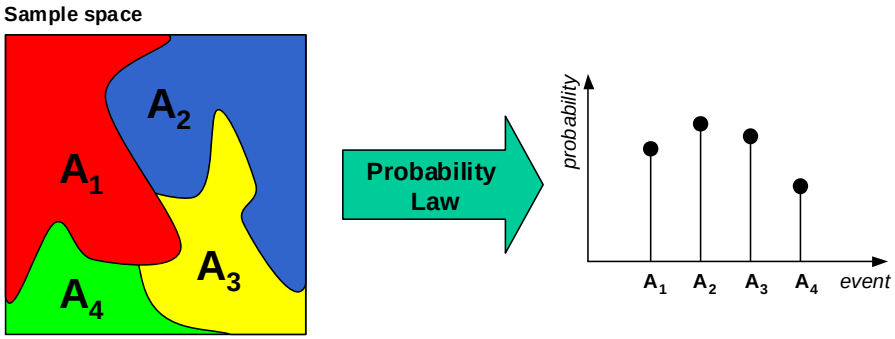


  - Satisfied with our classifier, we integrate the system and deploy it to the fish processing plant
    - A few days later the plant manager calls to complain that the system is misclassifying an average of 25% of the fish
    - **What went wrong?**

Review of probability theory

Probability

- Probabilities are numbers assigned to events that indicate "***how likely"*** it is that the event will occur when a random experiment is performed



Conditional Probability

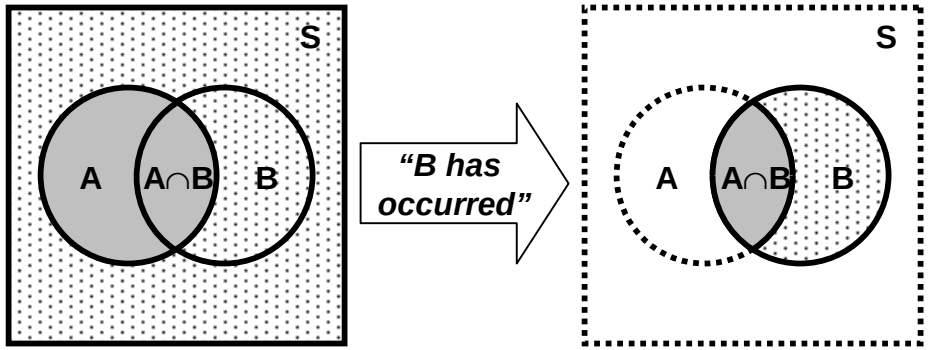
- If A and B are two events, the probability of event A when we already know that event B has occurred P[A|B] is defined by the relation

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} \quad \text{for} \quad P[B] > 0$$

  - P[A|B] is read as the "conditional probability of A conditioned on B", or simply the "probability of A given B"

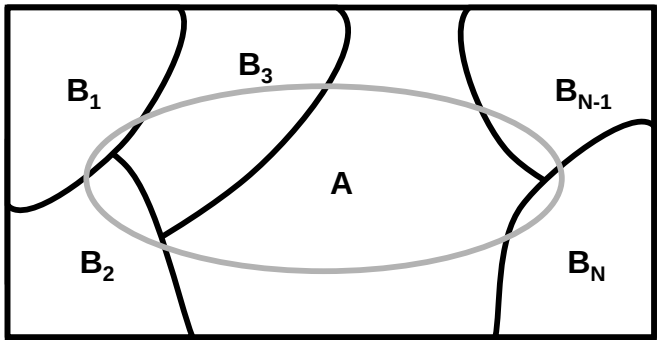
Review of probability theory

- **Conditional probability: graphical interpretation**



- **Theorem of Total Probability**
  - Let $B_1, B_2, \ldots, B_N$ be mutually exclusive events, then

$$P[A] = P[A \mid B_1]P[B_1] + \ldots P[A \mid B_N]P[B_N] = \sum_{k=1}^{N} P[A \mid B_k]P[B_k]$$



Review of probability theory

Bayes Theorem

- Given B_1, B_2, …, B_N, a partition of the sample space S. Suppose that event A occurs; what is the probability of event B_j?
- Using the definition of conditional probability and the Theorem of total probability we obtain

$$P[B_j \mid A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A \mid B_j] \cdot P[B_j]}{\sum_{k=1}^{N} P[A \mid B_k] \cdot P[B_k]}$$

- Bayes Theorem is definitely the fundamental relationship in Statistical Pattern Recognition

Rev. Thomas Bayes (1702-1761)

Review of probability theory

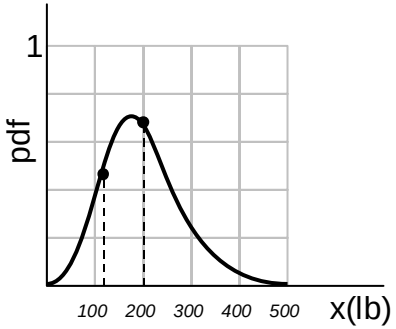
- **For pattern recognition, Bayes Theorem can be expressed as**

$$P(\omega_j \mid x) = \frac{P(x \mid \omega_j) \cdot P(\omega_j)}{\sum_{k=1}^{N} P(x \mid \omega_k) \cdot P(\omega_k)} = \frac{P(x \mid \omega_j) \cdot P(\omega_j)}{P(x)}$$

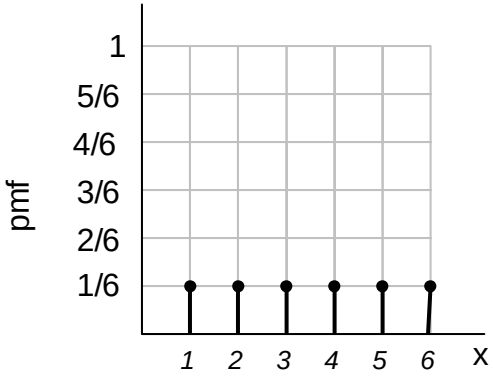
  - where ω_j is the i^{th} class and ***x*** is the feature vector
- **Each term in the Bayes Theorem has a special name, which you should be familiar with**
  - **$P(\omega_i)$** *Prior probability* (of class ω_i)
  - **$P(\omega_i|x)$** *Posterior Probability* (of class ω_i given the observation ***x***)
  - **$P(x|\omega_i)$** *Likelihood* (conditional prob. of ***x*** given class ω_i)
  - **P(x)** A normalization constant that does not affect the decision
- **Two commonly used decision rules are**
  - Maximum A Posteriori (MAP): choose the class ω_i with highest $P(\omega_i|x)$
  - Maximum Likelihood (ML): choose the class ω_i with highest $P(x|\omega_i)$
    - ML and MAP are equivalent for non-informative priors ($P(\omega_i)$=constant)

Review of probability theory

- **Characterizing features/vectors**
  - Complete: Probability mass/density function



pdf for a person's weight



pmf for rolling a (fair) dice

  - Partial: Statistics
    - Expectation
      - The expectation represents the center of mass of a density
    - Variance
      - The variance represents the spread about the mean
    - Covariance (only for random vectors)
      - The tendency of each pair of features to vary together, i.e., to co-vary

Review of probability theory

- **The covariance matrix (cont.)**

$$COV[X] = \Sigma = E[(X-\mu)(X-\mu)^T]$$

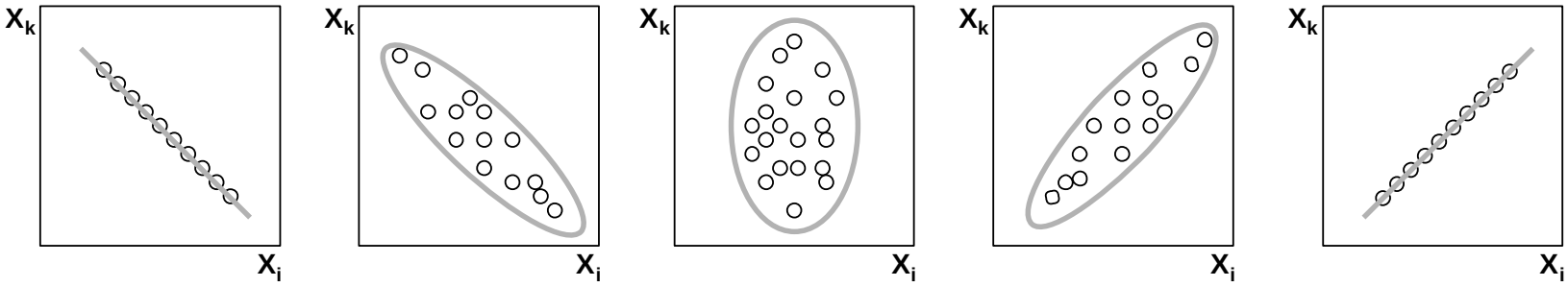
$$= \begin{bmatrix} E[(x_1-\mu_1)(x_1-\mu_1)] & \cdots & E[(x_1-\mu_1)(x_N-\mu_N)] \\ \cdots & \ddots & \\ E[(x_N-\mu_N)(x_1-\mu_1)] & \cdots & E[(x_N-\mu_N)(x_N-\mu_N)] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \cdots & c_{1N} \\ \cdots & \cdots & \\ c_{1N} & \cdots & \sigma_N^2 \end{bmatrix}$$

  - The covariance terms can be expressed as

$$c_{ii} = \sigma_i^2 \quad \text{and} \quad c_{ik} = \rho_{ik}\sigma_i\sigma_k$$

    - where ρ_{ik} is called the correlation coefficient

- **Graphical interpretation**



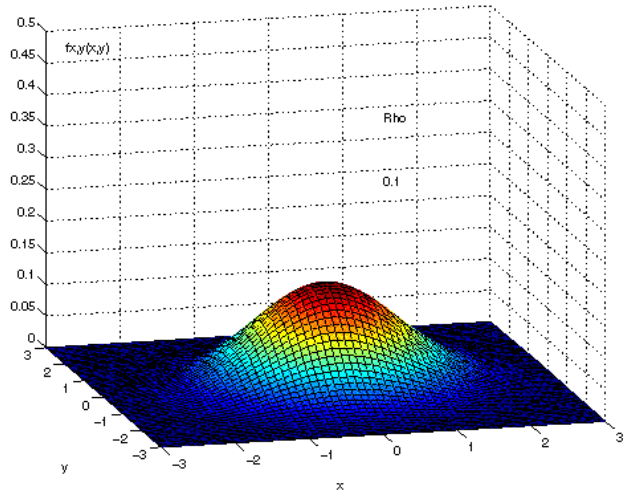
Review of probability theory

- **Meet the multivariate Normal or Gaussian density N(μ,Σ):**

$$f_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right]$$

  - For a single dimension, this expression reduces to the familiar expression

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$



- **Gaussian distributions are very popular**
  - The parameters (μ,Σ) are **sufficient** to uniquely characterize the normal distribution
  - If the $\mathbf{x}_i$'s are mutually **uncorrelated** ($\mathbf{c}_{ik}$=0), then they are also **independent**
    - The covariance matrix becomes diagonal, with the individual variances in the main diagonal
  - Marginal and conditional densities
  - Linear transformations