INTRODUCTION TO IMAGE CLASSIFICATION

Lecture 7

Concept of Image Classification

Image classification - assigning pixels in the image to categories or classes of interest

Examples: built-up areas, waterbody, green vegetation, bare soil, rocky areas, cloud, shadow, …

Concept of Image Classification

Image classification is a process of mapping numbers to symbols

$f(x): x \rightarrow \Delta; x \in R^n, \Delta = \{c_1, c_2, \ldots, c_L\}$

Number of bands = n;

Number of classes = L

f(.) is a function assigning a pixel vector x to a single class in the set of classes Δ

Concept of Image Classification

- In order to classify a set of data into different classes or categories, the relationship between the data and the classes into which they are classified must be well understood
- To achieve this by computer, the computer must be ***trained***
- Training is key to the success of classification
- Classification techniques were originally developed out of research in ***Pattern Recognition*** field

Concept of Image Classification

Computer classification of remotely sensed images involves the process of the computer program ***learning*** the relationship between the data and the information classes

Important aspects of accurate classification

- Learning techniques
- Feature sets

Types of Learning

- **Supervised Learning**
  - Learning process designed to form a mapping from one set of variables (data) to another set of variables (information classes)
  - A teacher is involved in the learning process
- **Unsupervised learning**
  - Learning happens without a teacher
  - Exploration of the data space to discover the scientifc laws underlying the data distribution

Features

- Features are attributes of the data elements based on which the elements are assigned to various classes.

- E.g., in satellite remote sensing, the features are measurements made by sensors in different wavelengths of the electromagnetic spectrum – visible/ infrared / microwave/texture features …

Features

- In medical diagnosis, the features may be the temperature, blood pressure, lipid profile, blood sugar, and a variety of other data collected through pathological investigations
- The features may be qualitative (high, moderate, low) or quantitative.
- The classification may be presence of heart disease (positive) or absence of heart disease (negative)

Supervised Classification

- The classifier has the advantage of an analyst or domain knowledge using which the classifier can be guided to learn the relationship between the data and the classes.

- The number of classes, prototype pixels for each class can be identified using this prior knowledge

Partially Supervised Classification

When prior knowledge is available

- For some classes, and not for others,
- For some dates and not for others in a multitemporal dataset,

Combination of supervised and unsupervised methods can be employed for ***partially supervised classification*** of images

Unsupervised Classification

- When access to domain knowledge or the experience of an analyst is missing, the data can still be analyzed by numerical exploration, whereby the data are grouped into subsets or ***clusters*** based on statistical similarity

Supervised vs. Unsupervised Classifiers

Supervised classification generally performs better than unsupervised classification IF good quality training data is available

Unsupervised classifiers are used to carry out preliminary analysis of data prior to supervised classification

Role of Image Classifier

The image classifier performs the role of a **discriminant** – discriminates one class against others

Discriminant value highest for one class, lower for other classes (**multiclass)**

Discriminant value positive for one class, negative for another class (**two class)**

Discriminant Function

$g(c_k, \mathbf{x})$ is **discriminant function**, relating feature vector $\mathbf{x}$ and class c_k, k=1,…,L

Denote $g(c_k,x)$ as $g_k(x)$ for simplicity

Multiclass Case

$g_k(\mathbf{x}) > g_l(\mathbf{x})$, l = 1,…,L, l ≠k $\mathbf{x} \in c_k$

Two Class Case

$g(\mathbf{x}) > 0$ $\mathbf{x} \in c_1$; $g(\mathbf{x}) < 0$ $\mathbf{x} \in c_2$

Example of Image Classification

Multiple Class Case

Recognition of characters or digits from bitmaps of scanned text

Two Class Case

Distinguishing between text and graphics in scanned document

Prototype / Training Data

- Using domain knowledge (maps of the study area, experienced interpreter), small sets of sample pixels are selected for each class.

- The size and spatial distribution of the samples are important for proper representation of the total pixel population in terms of the samples

Statistical Characterization of Classes

Each class has a conditional probability density function (pdf) denoted by $p(\mathbf{x} \mid c_k)$

The distribution of feature vectors in each class c_k is indicated by $p(\mathbf{x} \mid c_k)$

We estimate $P(c_k \mid \mathbf{x})$, the conditional probability of class c_k given that the pixel's feature vector is $\mathbf{x}$

Supervised Classification Algorithms

- There are many techniques for assigning pixels to informational classes, e.g.:
  - Minimum Distance from Mean (MDM)
  - Parallelpiped
  - Maximum Likelihood (ML)
  - Support Vector Machines (SVM)
  - Artificial Neural Networks (ANN)
  - …

Supervised Classification Principles

- The classifier learns the characteristics of different thematic classes – forest, marshy vegetation, agricultural land, turbid water, clear water, open soils, manmade objects, desert etc.

- This happens by means of analyzing the statistics of small sets of pixels in each class that are reliably selected by a human analyst through experience or with the help of a map of the area

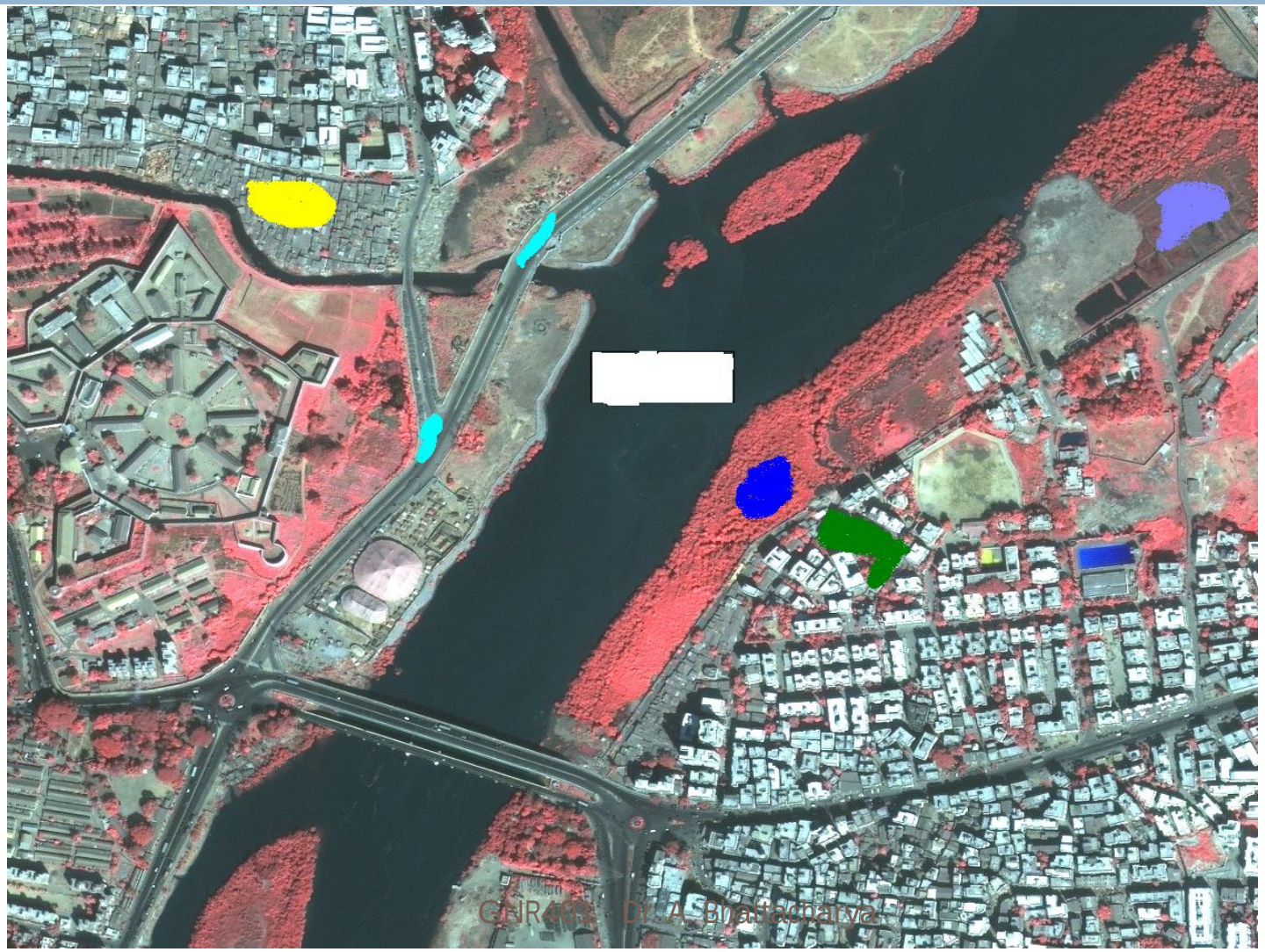
Supervised Classification Principles

- **Typical characteristics of classes**
  - Mean vector
  - Covariance matrix
  - Minimum and maximum gray levels within each band
  - Conditional probability density function $p(C_i|\mathbf{x})$ where C_i is the i^{th} class and $\mathbf{x}$ is the feature vector

- Number of classes L into which the image is to be classified should be specified by the user

Prototype Pixels for Different Classes

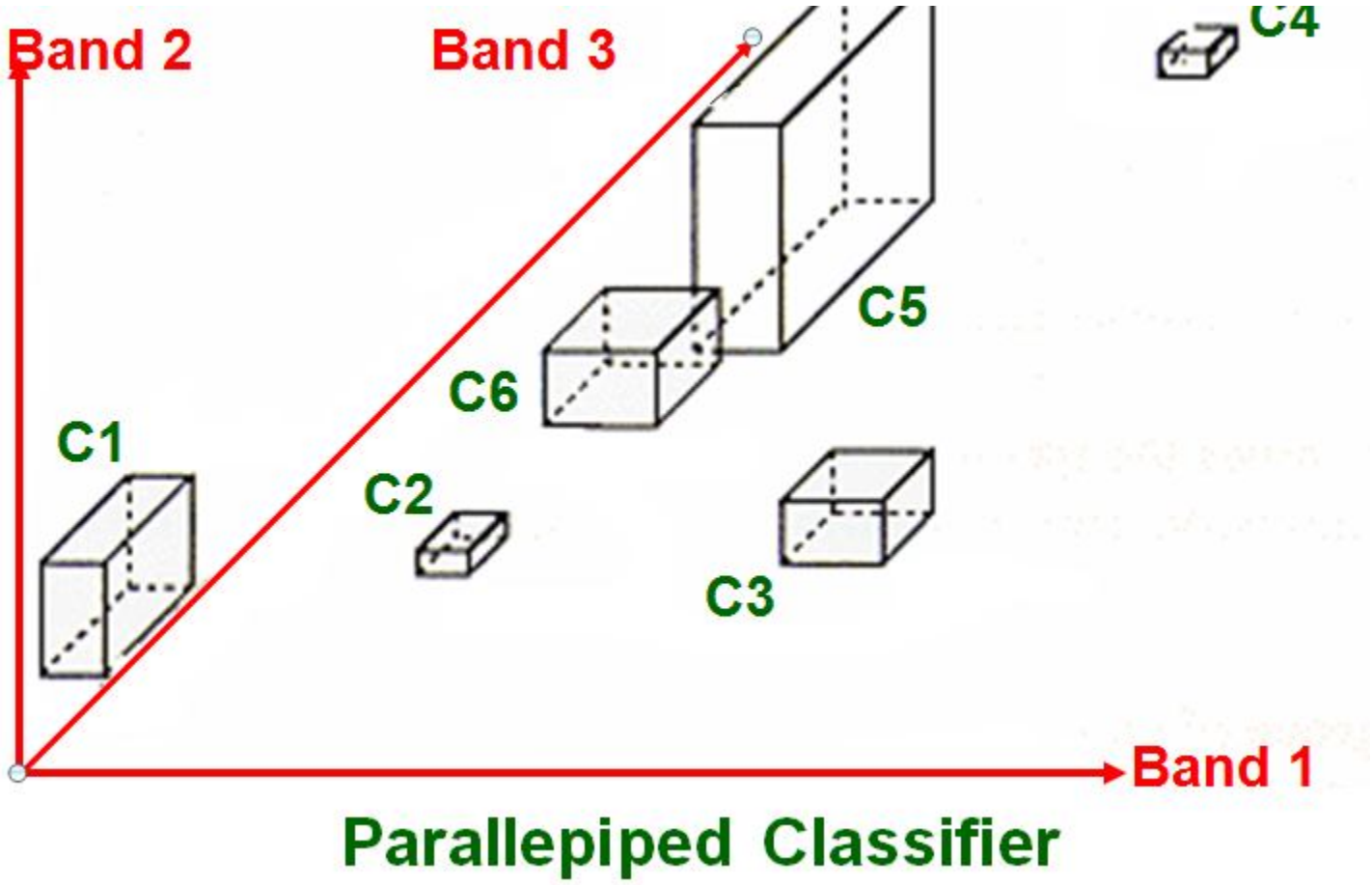
- The prototype pixels are ***samples*** of the population of pixels belonging to each class
- The size and distribution of samples are formally governed by the mathematical theory of sampling
- There are several criteria for choosing the samples belonging to different classes



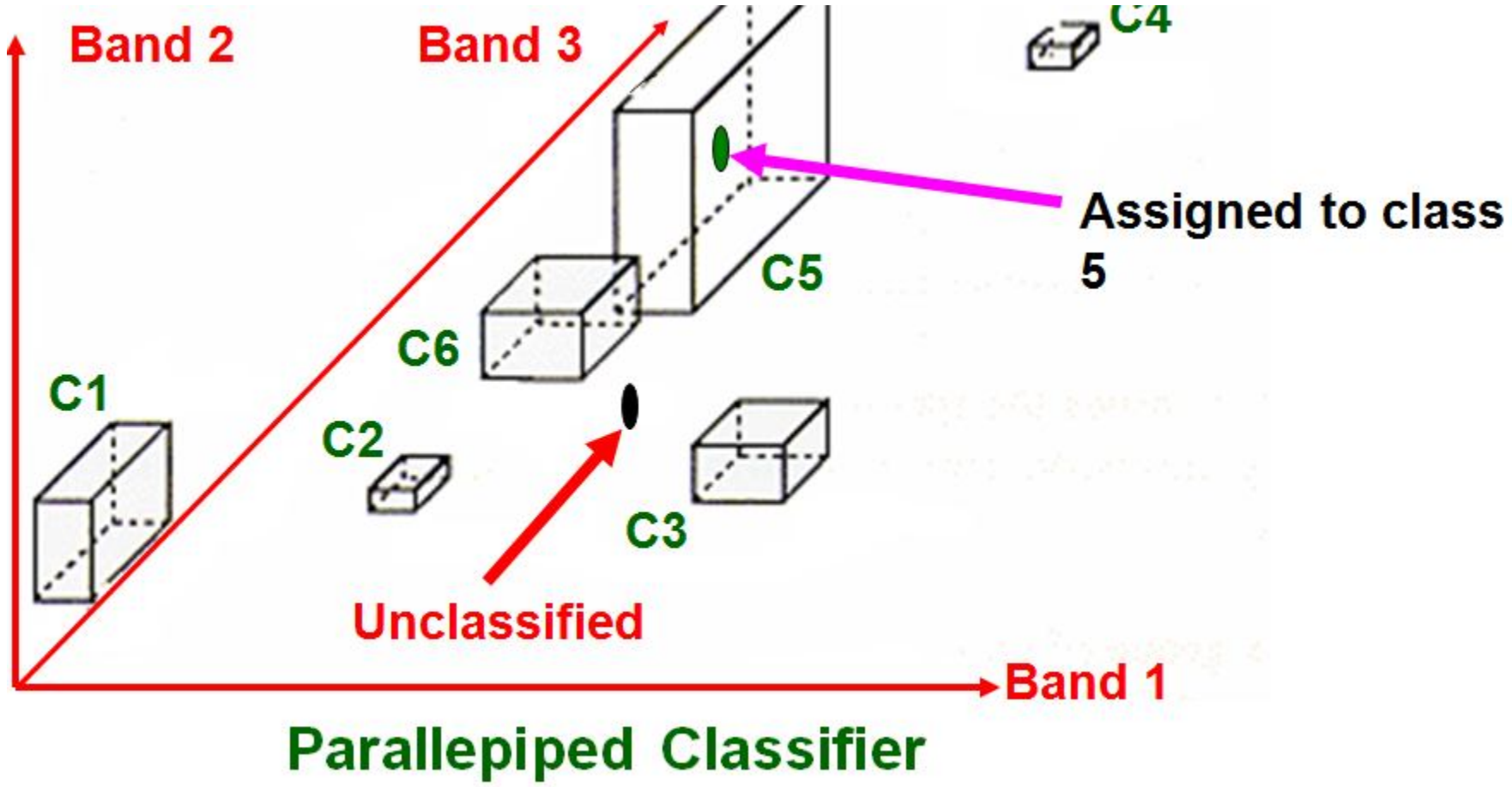
Parallelepiped Classifier - Example of a Supervised Classifier

- Assign ranges of values for each class in each band
  - Really a “feature space” classifier
  - Training data provide bounds for each feature for each class
  - Results in bounding boxes for each class
  - A pixel is assigned to a class only if its feature vector falls within the corresponding box

Parallelepiped Classifier



Parallelepiped Classifier



Advantages/Disadvantages of Parallelpiped Classifier

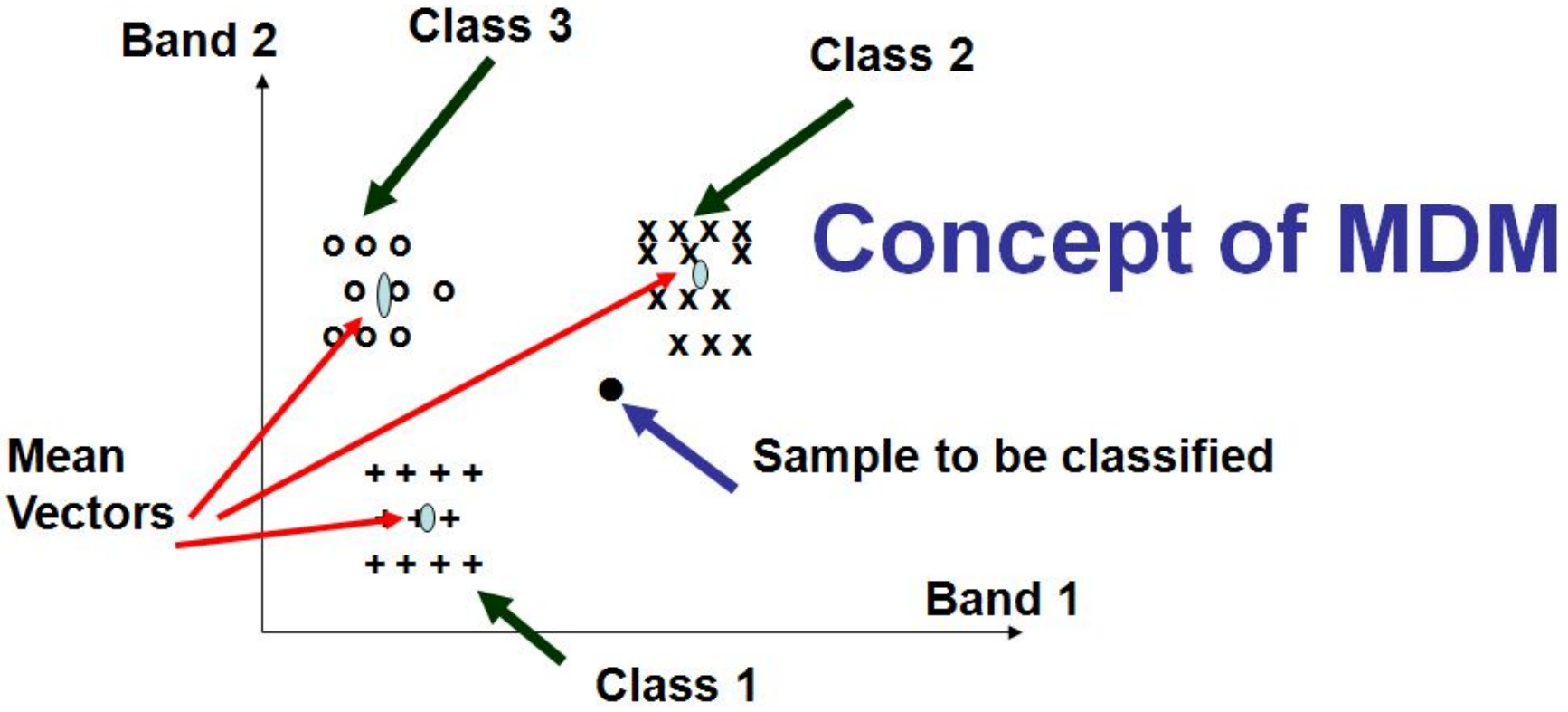
- Does NOT assign every pixel to a class. Only the pixels that fall within ranges.
- Fastest method computationally
- Good for helping decide if you need additional classes (if there are many unclassified pixels)
- Problems when class ranges overlap—must develop rules to deal with overlap areas.

Minimum Distance Classifier

- Simplest kind of supervised classification
- The method:
  - Calculate the mean vector for each class
  - Calculate the statistical (Euclidean) distance from each pixel to class mean vector
  - Assign each pixel to the class it is closest to

Minimum Distance Classifier

- 2-D Feature Space



Minimum Distance Classifier

- Algorithm
- Estimate class mean vector and covariance matrix from training samples
  - $m_i = \Sigma_{j \in Ci} \; X_j$; $C_i = E\{(X - m_i)(X - m_i)^T\} \mid X \in C_i\}$
  - Compute distance between X and m_i
  - $X \in C_i$ if $d(X, m_i) \leq d(X, m_j) \; \forall j$
  - Compute $P(C_k \mid X)$ =
  - Leave X unclassified if $\max_k P(C_k \mid X) < T_{min}$

Minimum Distance Classifier

- Normally classifies every pixel no matter how far it is from a class mean (still picks closest class) unless the T_{min} condition is applied

- Distance between X and m_i can be computed in different ways – Euclidean, Mahalanobis, city block, …

Maximum Likelihood Classifier

- Calculates the likelihood of a pixel being in different classes conditional on the available features, and assigns the pixel to the class with the highest likelihood

Likelihood Calculation

- The likelihood of a feature vector **x** to be in class Ci is taken as the conditional probability $P(C_i|\mathbf{x})$.
- We need to compute $P(C_i|\mathbf{x})$, that is the conditional probability of class C_i given the pixel vector **x**.
- It is not possible to directly estimate the conditional probability of a class given the feature vector. Instead, it is computed indirectly in terms of the conditional probability of feature vector **x** given that it belongs to class C_i.

Likelihood Calculation

$P(C_i|\mathbf{x})$ is computed using Bayes' Theorem in terms of $P(\mathbf{x}|C_i)$

$P(C_i|\mathbf{x}) = P(\mathbf{x}|C_i)\,P(C_i) \,/\, P(\mathbf{x})$

$\mathbf{x}$ is assigned to class C_j such that

$P(C_j|\mathbf{x}) = \text{Max}_i\, P(C_i|\mathbf{x})$, i=1…K, the number of classes.

$P(C_i)$ is the prior probability of occurrence of class i in the image

$P(\mathbf{x})$ is the multivariate probability density function of feature $\mathbf{x}$.

Likelihood Calculation

- P(**x**) can be ignored in the computation of Max{P(C_i|**x**)}
- If P(**x**|C_j) is not assumed to have a known distribution, then its estimation is said to be non-parametric estimation.
- If P(**x**|C_j) is assumed to have a known distribution, then its estimation is said to be parametric.
- The training data **x** with the class already given, can be used to estimate the conditional density function P(**x**|C_i)

Likelihood Calculation

- P($\mathbf{x}|C_i$) is assumed to be multivariate Gaussian distributed in practical parametric classifiers.
- Gaussian distribution is mathematically simple to handle.
- Each class conditional density function P($\mathbf{x}|C_i$) is represented by its mean vector $\boldsymbol{\mu}_i$ and covariance matrix Σ_i

$$p(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{L/2}|\Sigma_i|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}$$

Assumption of Gaussian Distribution

- Each class is assumed to be multivariate normally distributed
- That implies each class has a mean μ_i that has the highest likelihood of occurrence
- The likelihood function decreases exponentially as the feature vector **x** deviates from the mean vector μ_i
- The rate of decrease is governed by the class variance; Smaller the variance, steep will be the decrease, and larger the variance, slower will be the decrease.

Likelihood Calculation

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{L}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

- We assume that the covariance matrices for each class are different.
- The term $(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)$ is known as the Mahalanobis distance between x and μ_i (after Prof. P.C. Mahalanobis, famous Indian statistician and founder of Indian Statistical Institute)

Interpretation of Mahalanobis distance

- The Mahalanobis distance between two multivariate quantities x and y is

$$d_M(x,y) = (x-y)^t \Sigma^{-1} (x-y)$$

- If the covariance matrix is $k.I$, (I is the unit matrix) then the Mahalanobis distance reduces to a scaled version of the Euclidean distance.
- Mahalanobis distance reduces the Euclidean distance according to the extent of variation within the data, given by the covariance matrix Σ

Advantages/Disadvantages of Maximum Likelihood Classifier

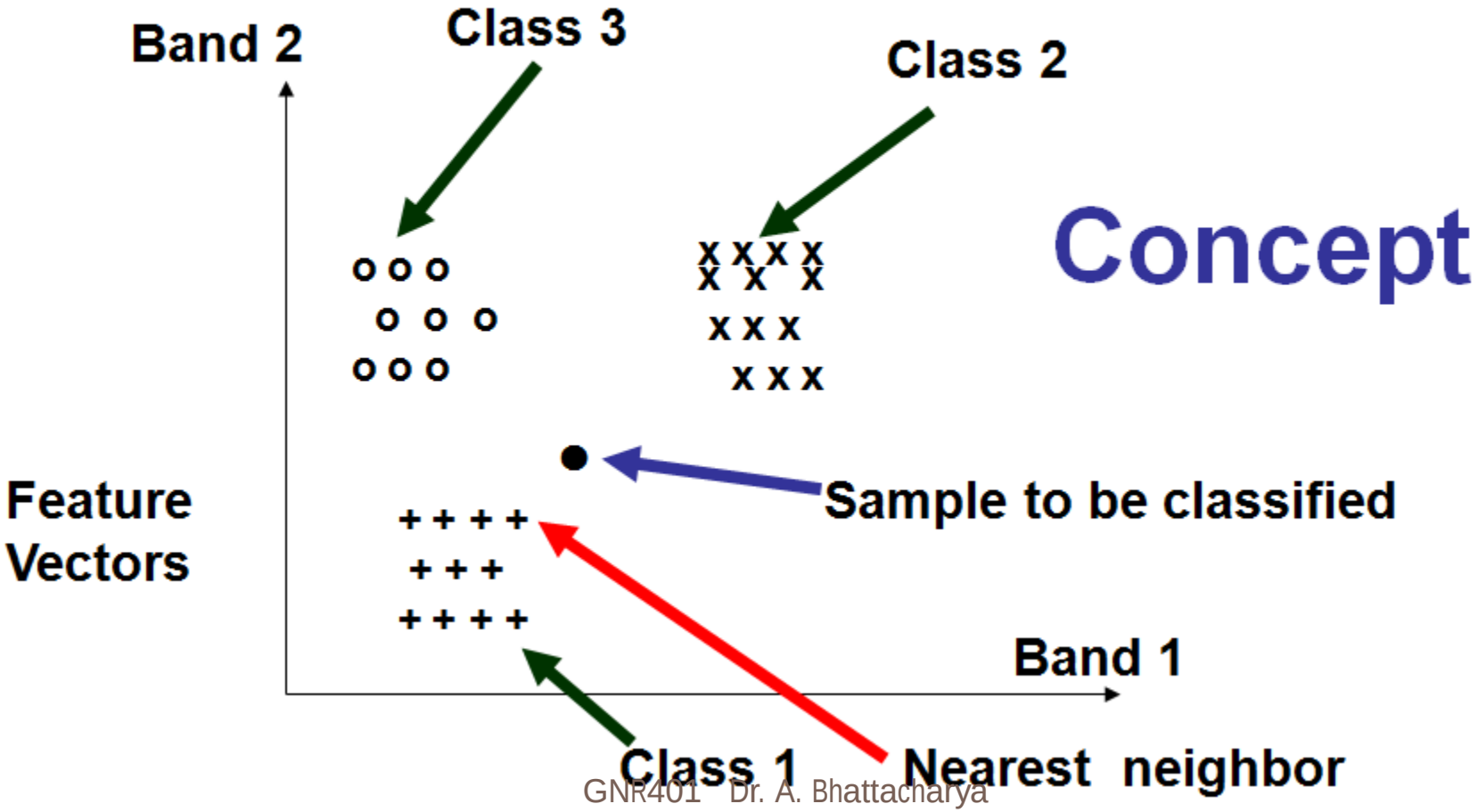
- Normally classifies every pixel no matter how far it is from a class mean
- Slowest method – more computationally intensive
- Normally distributed data assumption is not always true, in which case the results are not likely to be very accurate
- Thresholding condition can be introduced into the classification rule to separately handle ambiguous feature vectors

Nearest-Neighbor Classifier

Non-parametric in nature

- The algorithm is:
  - Find the distance of given feature vector **x** from ALL the training samples
  - **x is assigned to the class of the nearest training sample (in the feature space)**
  - **This method does not depend on the class statistics like mean and covariance.**

- 2-D Feature Space



K-NN Classifier

- K-nearest neighbour classifier
- Simple in concept, time consuming to implement
- For a pixel to be classified, find the K closest training samples (in terms of feature vector similarity or smallest feature vector distance)
- Among the K samples, find the most frequently occurring class C_m
- Assign the pixel to class C_m

K-NN Classifier

- Let k_i be number of samples for class C_i (out of K closest samples), i=1,2,…,L (number of classes)
- Note that $\sum_i k_i = K$

- The discriminant for K-NN classifier is
- $g_i(x) = k_i$
- The classifier rule is
- *Assign x to class C_m if $g_m(x) > g_i(x)$, for all i, $i \neq m$*

K-NN Classifier

- It is possible to find more than one class whose training samples are closest to the feature vector of pixel **x**. Therefore the discriminant function is refined further as

$$g_i(x) = \frac{\sum_{j=1}^{k_i} 1/d(x, x_i^j)}{\sum_{l=1}^{L} \sum_{j=1}^{k_i} 1/d(x, x_l^j)}$$

The distances of the nearest neighbours to the feature vector of the pixel to be classified are taken into account

K-NN Classifier

- If the classes are in different proportions in the image, then the prior probabilities can be taken into account:

$$g_i(x) = \frac{k_i\, p(\omega_i)}{\sum_{l=1}^{L} k_l\, p(\omega_l)}$$

- For each pixel to be classified, the feature space distances to all training pixels are to be computed before the decision is made, due to which this procedure is extremely computation intensive, and is not used when the dimensionality (number of bands) of the feature space is large, e.g., with hyperspectral data.

Spectral Angle Mapper

- Given a large dimensional data set, computing the covariance matrix, its inverse, and the distance for each pixel
- $(X - \mu)^T \Sigma^{-1} (X - \mu)$ is highly time consuming and if the covariance matrix is close to singular then its inverse can be unstable, leading to erroneous results
- In such cases, alternate methods can be applied, such as Spectral Angle Mapper

S.A.M. Principle

- If each class is represented by a vector $\mathbf{v}_i$, then the angle between the class vector and the pixel feature vector $\mathbf{x}$ is given by
- $\cos\theta = [\mathbf{v}_i.\mathbf{x}] / [|\mathbf{v}_i||\mathbf{x}|]$
- For small values of θ, the value of $\cos\theta$ is large
- The likelihood of $\mathbf{x}$ to belong to different classes can be ranked according to the value of $\cos\theta$.

S.A.M. Advantage

- The value of the vector would not be greatly affected by minor changes in $\mathbf{v}_i$ or $\mathbf{x}$.

- The computation is simpler compared to the Mahalanobis distance computation involved in ML method

Feature Data

- Given a set of training samples, a set of test samples, N band training data, the given image is classified after training the classifier using training data
- The classification result is verified using test data
- If some bands are discarded, how is the result affected?

Feature Selection/Reduction

- Feature selection – selecting a subset of features based on some criteria
- Feature reduction – discarding some features after performing some transformation on the input data; e.g., PCT
- Reduced number of bands also reduces need for large training data

Mixed Pixels

- When spatial resolution is coarse, one pixel may contain parts of many landuse classes
  - e.g., tree, bare soil, grass
- Classification is estimating proportions of different classes within a pixel
- The problem is called *mixture modeling*

Mixed Pixels

- Important when spectra of different classes are compared as in hyperspectral remote sensing
- Reference spectra are drawn from single classes
- Most pixel spectra are mixtures of more than one pure class
- Mixture modeling estimates the relative proportion of each class assuming a particular model for mixing

Feature Selection/Reduction

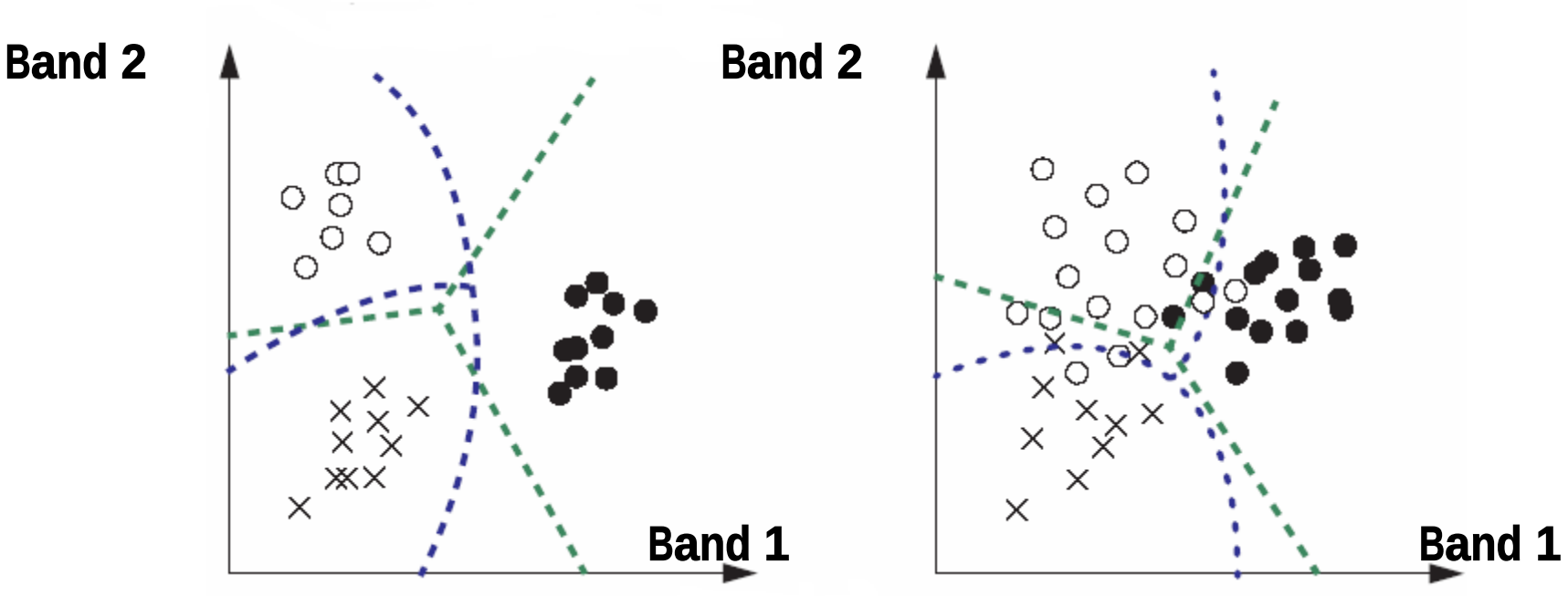
- Feature selection – selecting a subset of features based on some criteria
- Feature reduction – discarding some features after performing some transformation on the input data; e.g., PCT
- Reduced number of bands also reduces need for large training data

Motivation to Consider Feature Selection

- Image classification
  - each element should have useful features
  - discriminate between elements of different classes
- Discrimination power of features

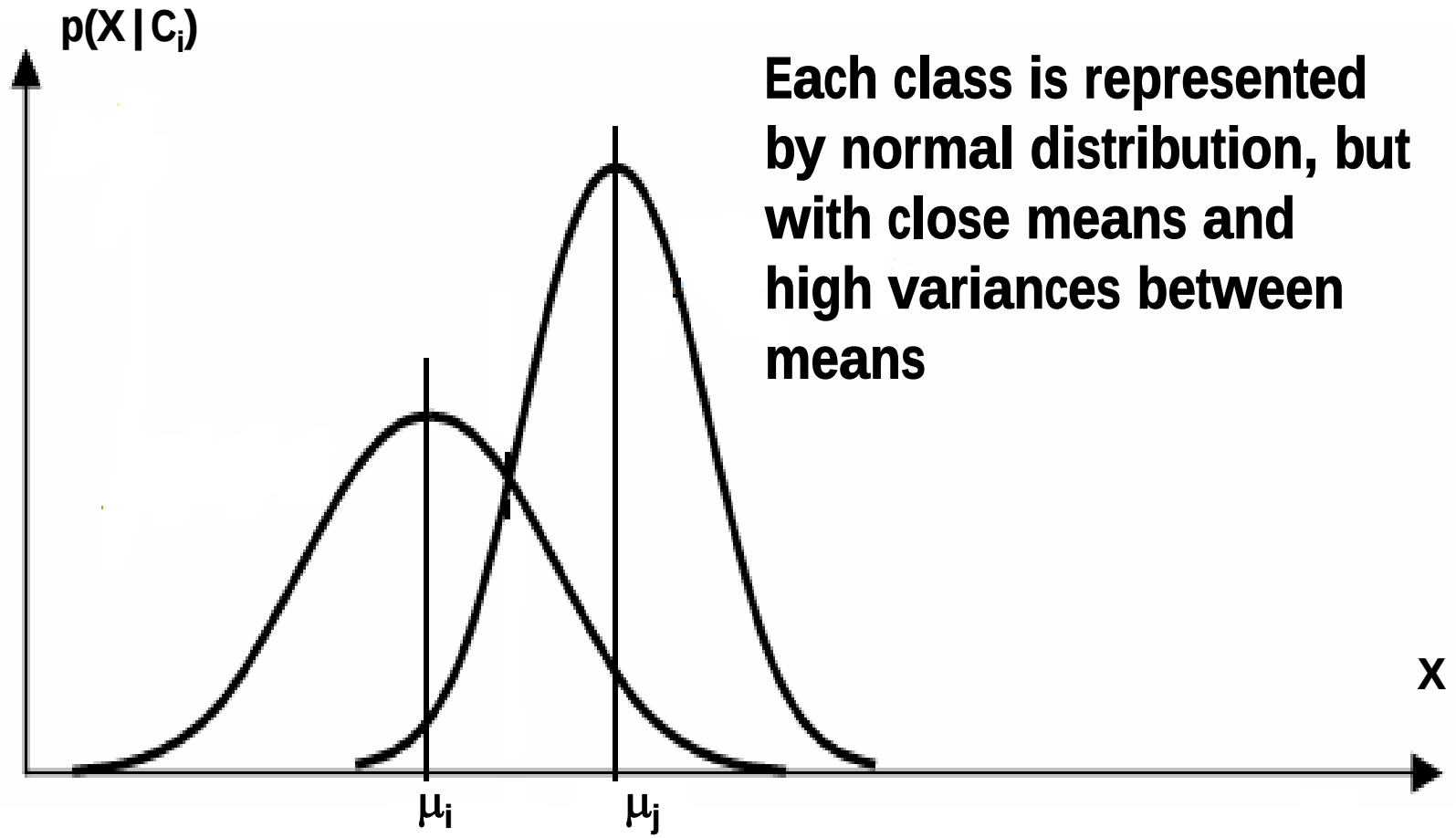
Class separability in feature space

- Two generic cases



Well separated classes **Poorly separated classes**

Low separability (with one feature)



Important Features

- Shape Features
- Spectral Features
- Texture Features
- Transform Features

Unsupervised Classification

- When access to domain knowledge or the experience of an analyst is missing, the data can still be analyzed by numerical exploration, whereby the data are grouped into subsets or ***clusters*** based on statistical similarity

Unsupervised Classification

- *Unsupervised classification* is also known as learning without teacher
- In the absence of reliable training data it is possible to understand the structure of the data using statistical methods such as *clustering algorithms*
- *Popular clustering algorithms are k-means and ISODATA.*

Clustering Algorithms

- All feature vectors are points in an L-dimensional space where L is the number of bands *(The letter K is reserved for the number of clusters!)*

- It is required to find which sets of feature vectors tend to form clusters

- ***Members of a cluster are more similar to each other than to members of another cluster – In other words, they possess low intra-cluster variability and high inter-cluster variability***

K-Means

- Iterative algorithm
- Number of clusters K is known by user
- Most popular clustering algorithm
- Initialize randomly K cluster mean vectors
- Assign each pixel to any of the K clusters based on minimum feature distance
- After all pixels are assigned to the K clusters, each cluster mean is recomputed.
- Iterate till cluster mean vectors stabilize

ISODATA ALGORITHM

- Iterative Self-Organizing Data Analysis Technique (the last A added to make the acronym sound better)
- Developed in biology in the 1960's by Ball and Hall
- See Tou and Gonzalez's classic "Pattern Recognition Principles" for an excellent exposition to clustering algorithms

User specified parameters for ISODATA

- Generalization of K-Means algorithm
- Consists of many user-specified parameters
  - Minimum size of cluster
  - Maximum size of cluster
  - Maximum intra-cluster variance
  - Minimum separation between pairs of clusters
  - Maximum number of clusters
  - Minimum number of clusters
  - Maximum number of iterations