Welcome to:

Machine Learning Approaches for Pattern Recognition

Unit objectives

After completing this unit, you should be able to:

- Understand the concept of neural networks and kernel methods
- Learn example of sparse kernel machines and graphical models
- Gain knowledge on sampling methods for pattern recognition
- Understand pattern recognition in sequential data

Neural networks

- A neural network is an algorithm series that attempts to understand simple connections in a data set using a computer that simulates the role of the brain.

- In this case, neural networks refer to biological or artificial neuronal structures.

- In the world of finance, neural networks are helpful in developing processes, including time series forecasts, algorithmic trading, and classification of stocks, credit risk modeling and the creation of proprietary price indices.

How neural networks learn?

Supervised	Unsupervised	Semi-Supervised	Reinforcement
• Data has **known labels** or output	• Labels or output unknown • Focus on **finding patterns and gaining insight** from the data	• Labels or output known for **a subset of** data • A blend of supervised and unsupervised learning	• Focus on **making decisions** based on previous experience • Policy-making with feedback
• Insurance underwriting • Fraud detection	• Customer clustering • Association rule mining	• Medical predictions (where tests and expert diagnoses are expensive, and only part of the population receives them)	• Game AI • Complex decision problems • Reward systems

Figure: Type of learning

Source: https://images.app.goo.gl/yUyUcXku5R6fg9sK9

Neural networks examples

- Handwriting recognition.
- Financial-market forecasting.
- Social media images analysis.
- Diagnostic imaging for diagnosing cancer.

Neural networks use cases

- Space.
- Automotive.
- Electronics.
- Manufacturing.
- Mechanics.
- Robotics.
- Telecom.
- Banking.

Kernel methods

- Kernels or kernel approaches (also called kernel functions) are collections of different forms of algorithms used to evaluate trends. Use a linear classifier, they are utilized to solve a nonlinear problem.

- Kernel techniques are employed in SVM (Support Vector Machines) and is used in problems of classification and regression.

- The SVM utilizes what is called a "kernel trick" where the data is converted and an appropriate boundary for potential outputs is sought.

Self evaluation: Exercise 11

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 11: Random forest.

Sparse kernel machines use cases

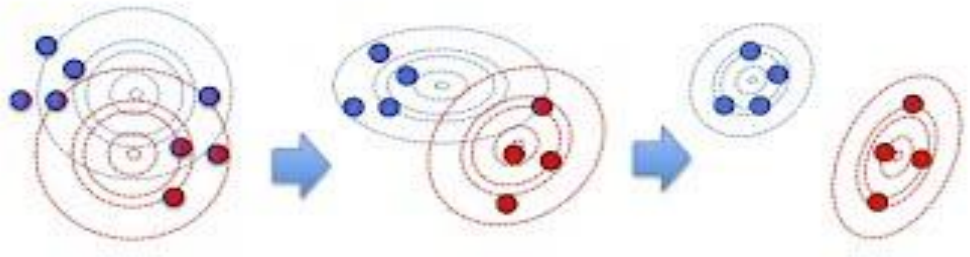
- Facial recognition.
- Categorization of text and hypertext.
- Identification of photographs.
- Bioinformatics.
- Protein folding and remote detection of homology.
- Recognition of handwriting.
- GPC.

Graphical models

- Some popular machine learning problems involve the classification of discrete, different data points.
- For example, if a picture contains a cat or a dog, predict which digit it will be from 0 to 9, if a handwritten character is included.
- There are some things that do not fit into the above structure. Each word is defined by its part of expression (a noun, a pronoun, a verb, an adjective, etc.) for example given a sentence "Machine learning I like.
- "As this basic illustration alone demonstrates, each term "learning" may be considered as a meaning-based noun or verb.
- This role is essential for many more complicated text activities such as the language, voice-to - text translation, etc.

Mixture models and EM

Gaussian Mixture Model



- Data with D attributes, from Gaussian sources $c_1 \dots c_k$
  - how typical is $\mathbf{x}_i$ under source **c**: $P(\vec{x}_i \mid c) = \frac{1}{\sqrt{2\pi|\Sigma_c|}} \exp\left\{-\frac{1}{2}(\vec{x}_i - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x}_i - \vec{\mu}_c)\right\}$
    $\sum_a \sum_b (x_{ia} - \mu_{ca}) \left[\Sigma_c^{-1}\right]_{ab} (x_{ib} - \mu_{cb})$
  - how likely that $\mathbf{x}_i$ came from **c**: $P(c \mid \vec{x}_i) = \frac{P(\vec{x}_i \mid c) P(c)}{\sum_{c=1}^{k} P(\vec{x}_i \mid c) P(c)}$
  - how important is $\mathbf{x}_i$ for source **c**: $w_{ic} = P(c \mid \vec{x}_i) / \left(P(c \mid \vec{x}_1) + \dots + P(c \mid \vec{x}_n)\right)$
  - mean of attribute **a** in items assigned to **c**: $\mu_{ca} = w_{c1} x_{1a} + \dots + w_{cn} x_{na}$
  - covariance of **a** and **b** in items from **c**: $\Sigma_{cab} = \sum_{i=1}^{n} w_{ci} \left(x_{ia} - \mu_{ca}\right)\left(x_{ib} - \mu_{cb}\right)$
  - prior: how many items assigned to **c**: $P(c) = \frac{1}{n}\left(P(c \mid \vec{x}_1) + \dots + P(c \mid \vec{x}_n)\right)$

Figure: Mixture Models and EM

Source: https://images.app.goo.gl/9TQmWETJYvpDJi7w5

Bayesian networks: Directed graphical models

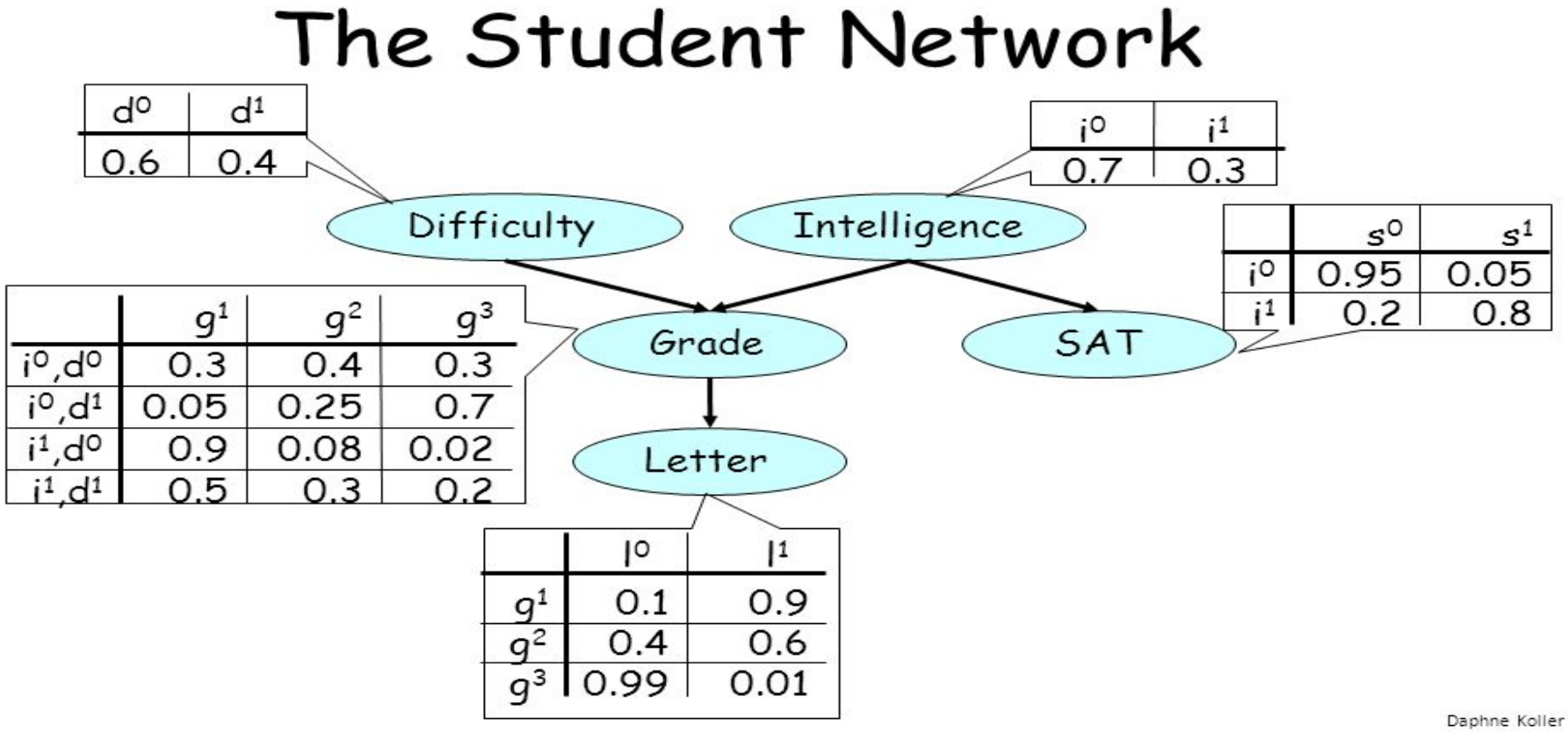


Figure: Bayesian networks: Directed graphical models

Source: https://images.app.goo.gl/qECFiT5QUHJAS7U38

Conditional probability distributions

- CPDs with "difficulties" and "knowledge," since they do not depend on the other variables, are relatively simple.

- In theory, the tables reflect the probability of other variables, as values 0 or 1.

- The numbers in each table must be 1, as you may have noted.

Potential functions

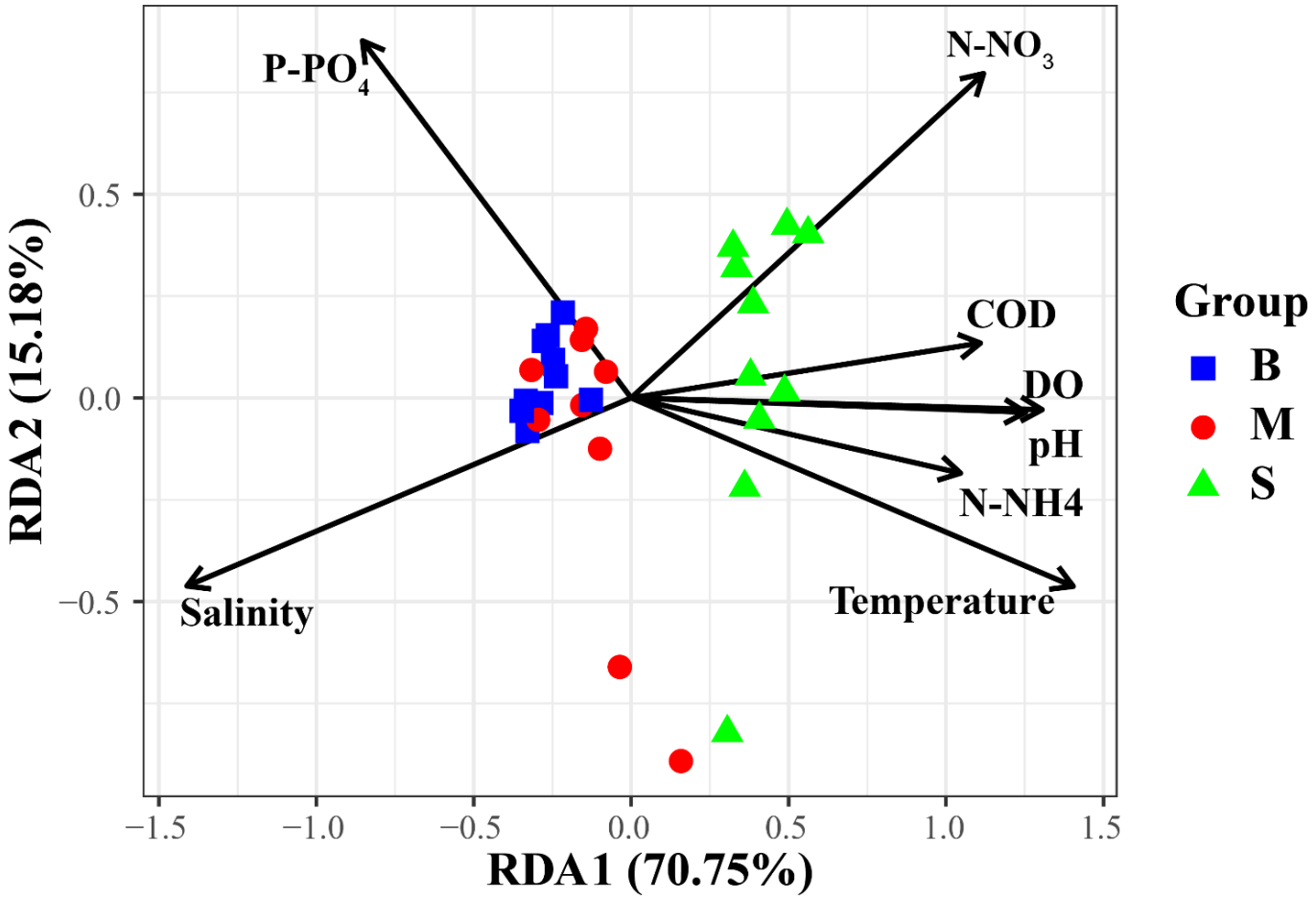


Figure: Potential Functions

Source: https://images.app.goo.gl/crfvoAiNu3bruX8E8

Self evaluation: Exercise 12

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 12: SVM.

Conditional independences

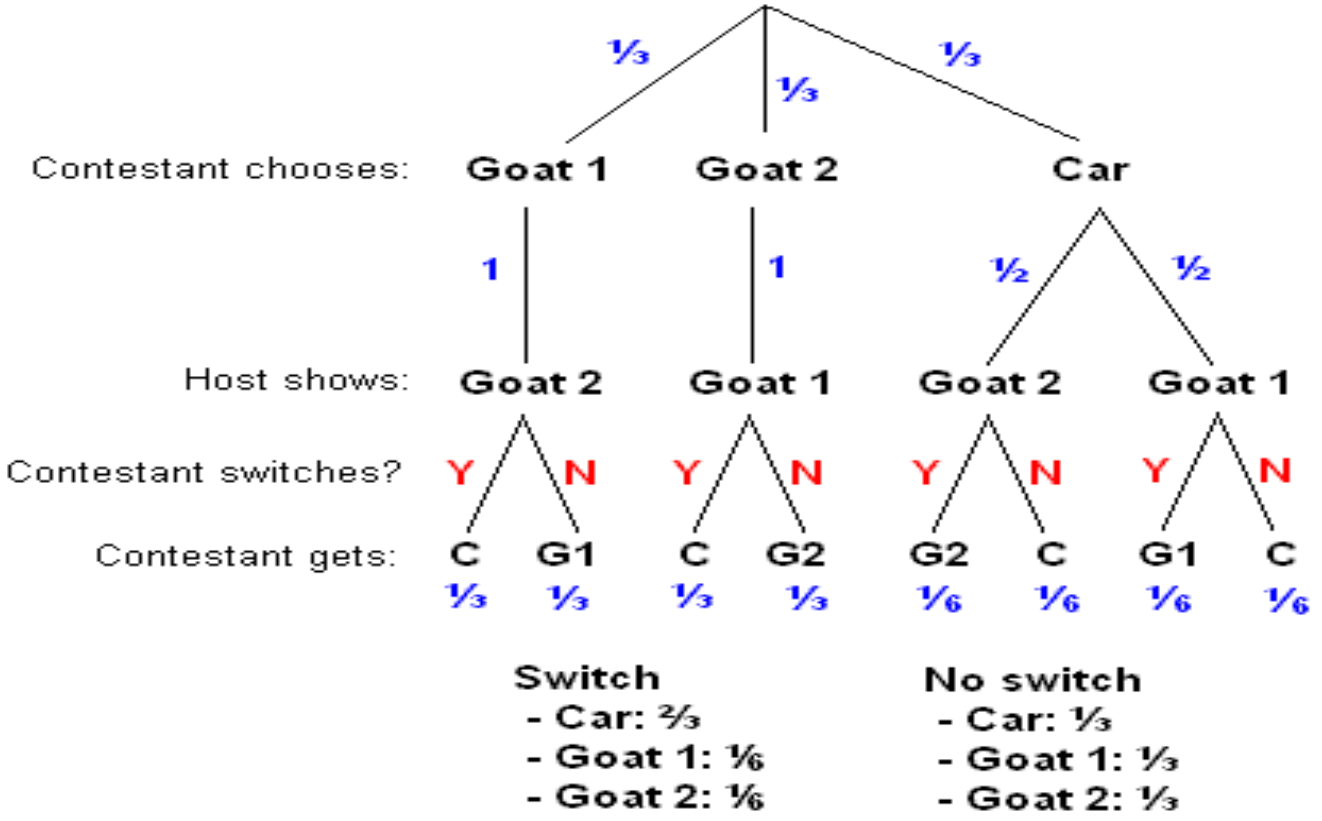


Figure: Conditional Independences

Source: https://images.app.goo.gl/DEUAMZEwBBKR1CN58

Sampling methods for pattern recognition

- Sampling is a method to gather population data based on statistics from a portion of the community (sample) without the need to look at individuals.

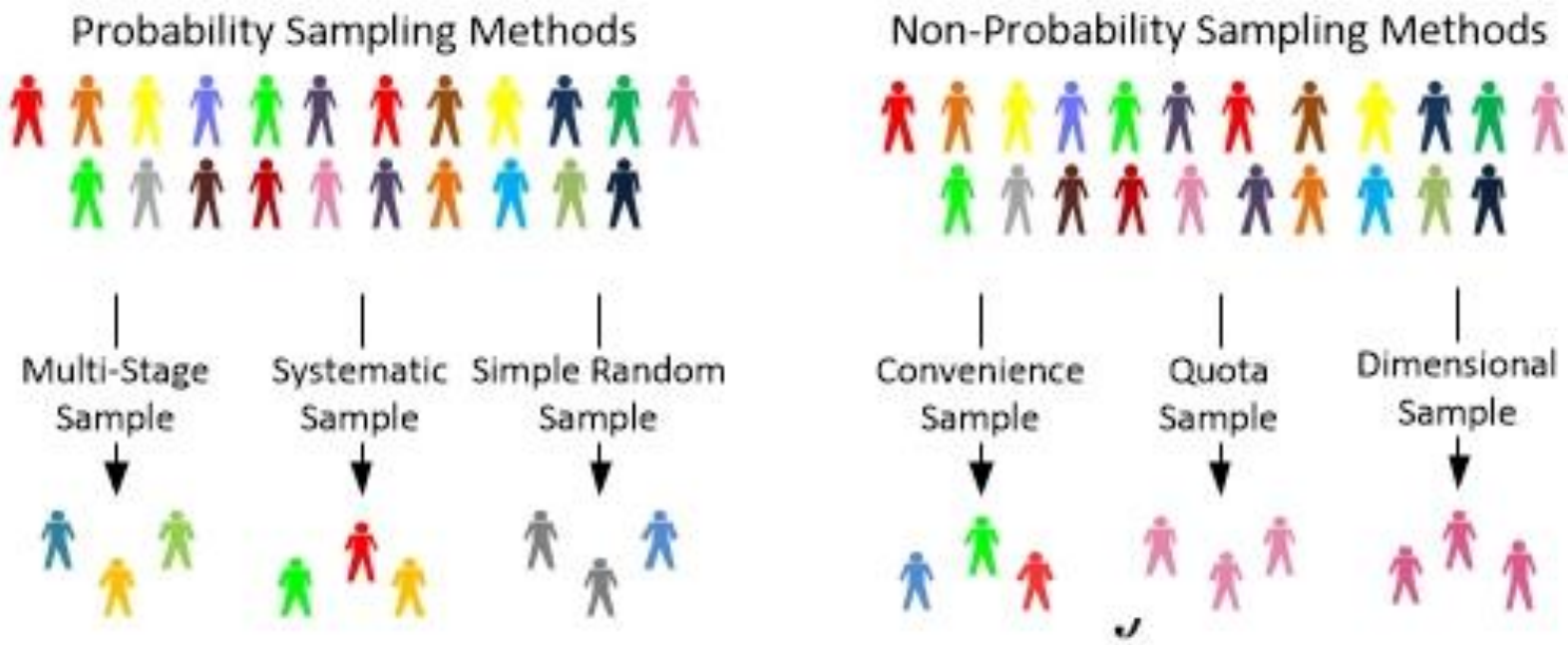


Figure: Sampling Methods

Source: https://images.app.goo.gl/xbLo6pyTzPRrszEs9

Continuous latent variables

	Manifest variables	
Latent variables	**Continuous**	**Categorical**
Continuous	Factor analysis	Item response theory
Categorical	Latent profile analysis	Latent class analysis

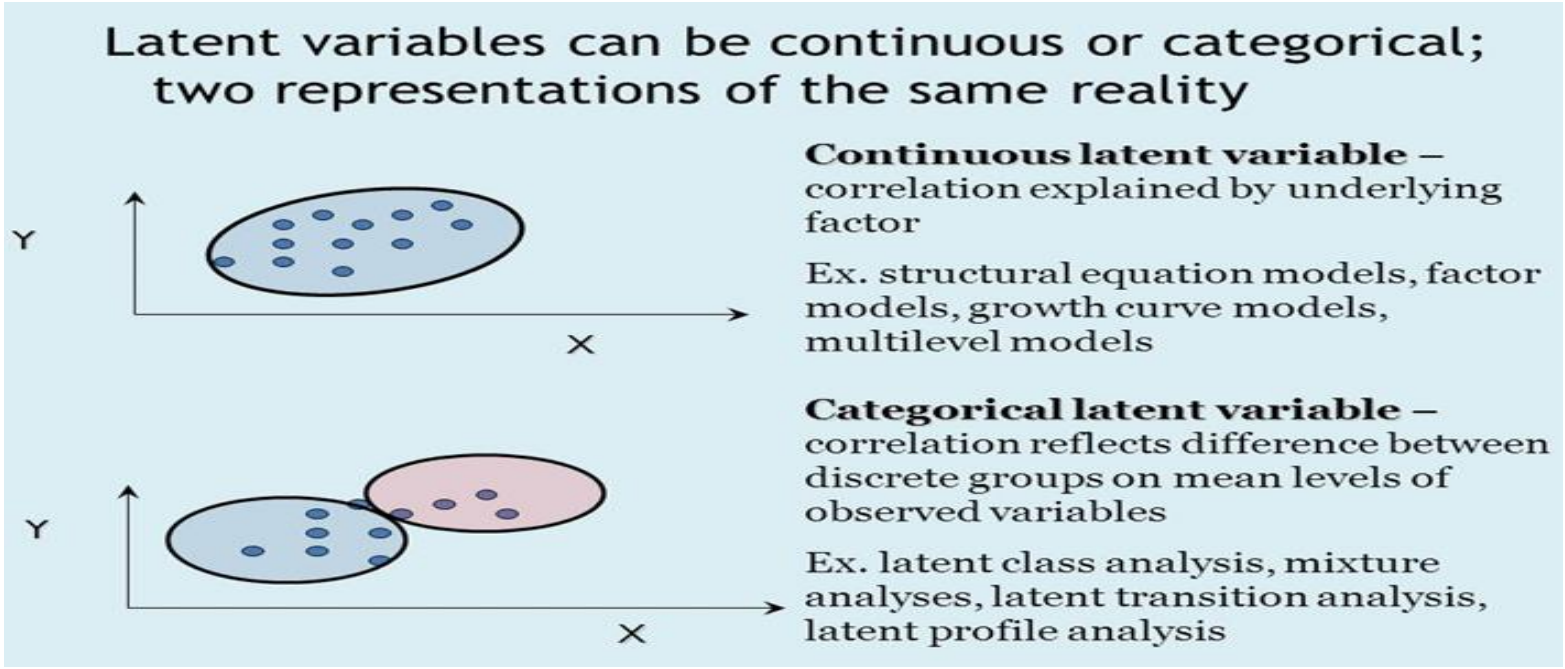


Figure: Continuous Latent Variables

Source: https://images.app.goo.gl/PjPtHiRmSCW6J7ej6

Self evaluation: Exercise 13

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 13: Local Outlier Factor (LOF).

Combining models for pattern recognition

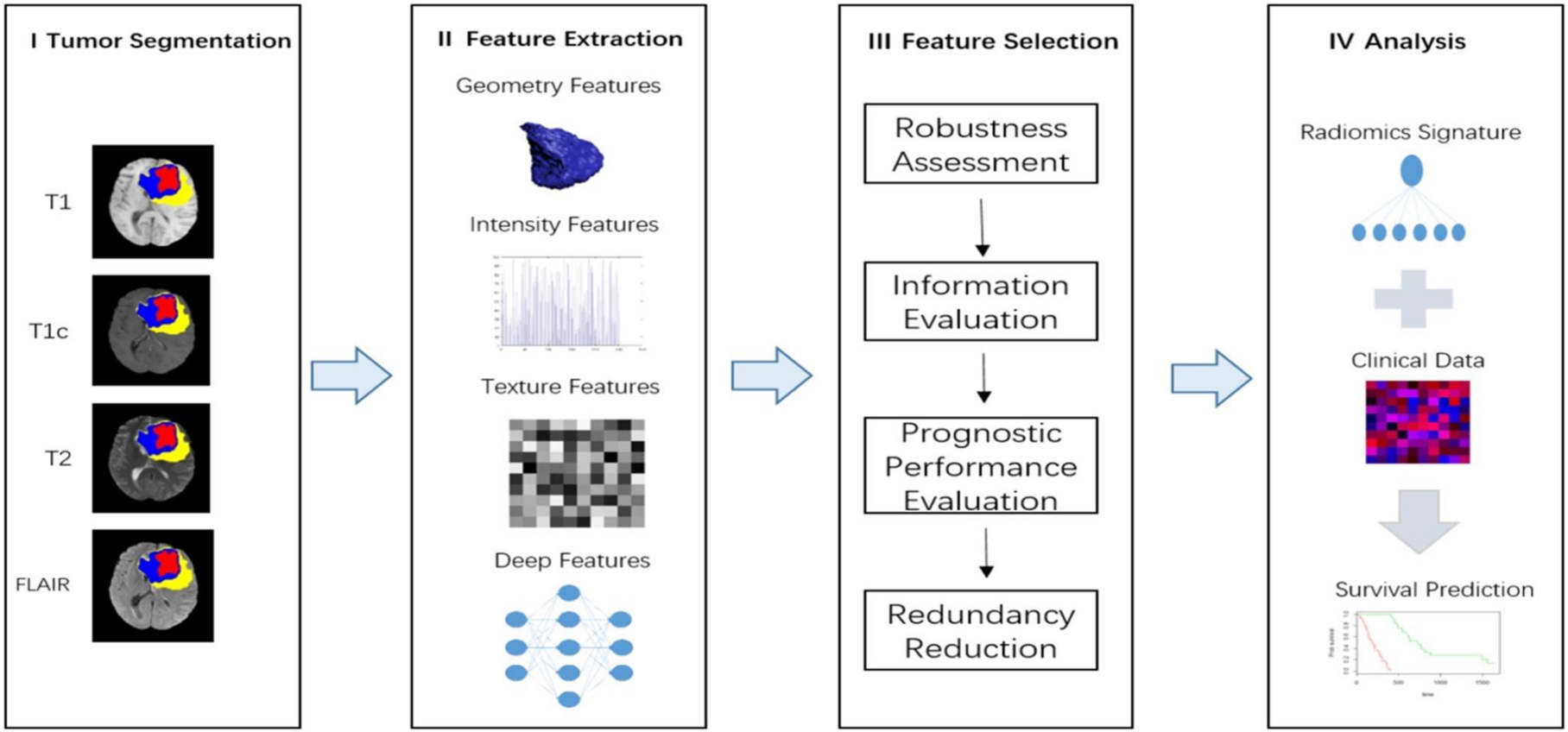


Figure: Combining Models for Pattern Recognition

Source: https://images.app.goo.gl/zct722Vc3vptf1KU9

Markov chain Monte Carlo

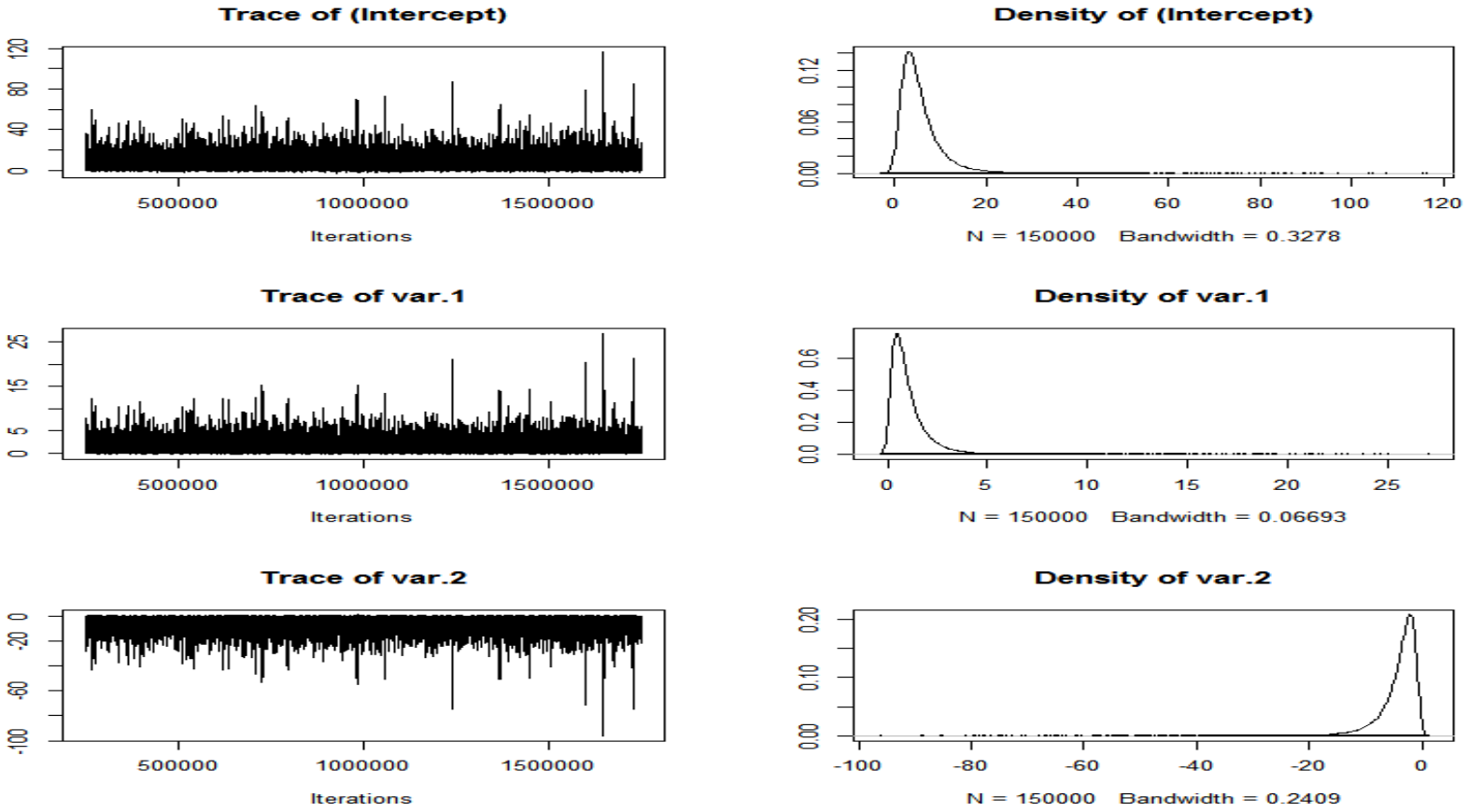


Figure: Markov chain Model

Source: https://images.app.goo.gl/9pTxRVtBURiuar6n8

The K-means algorithm

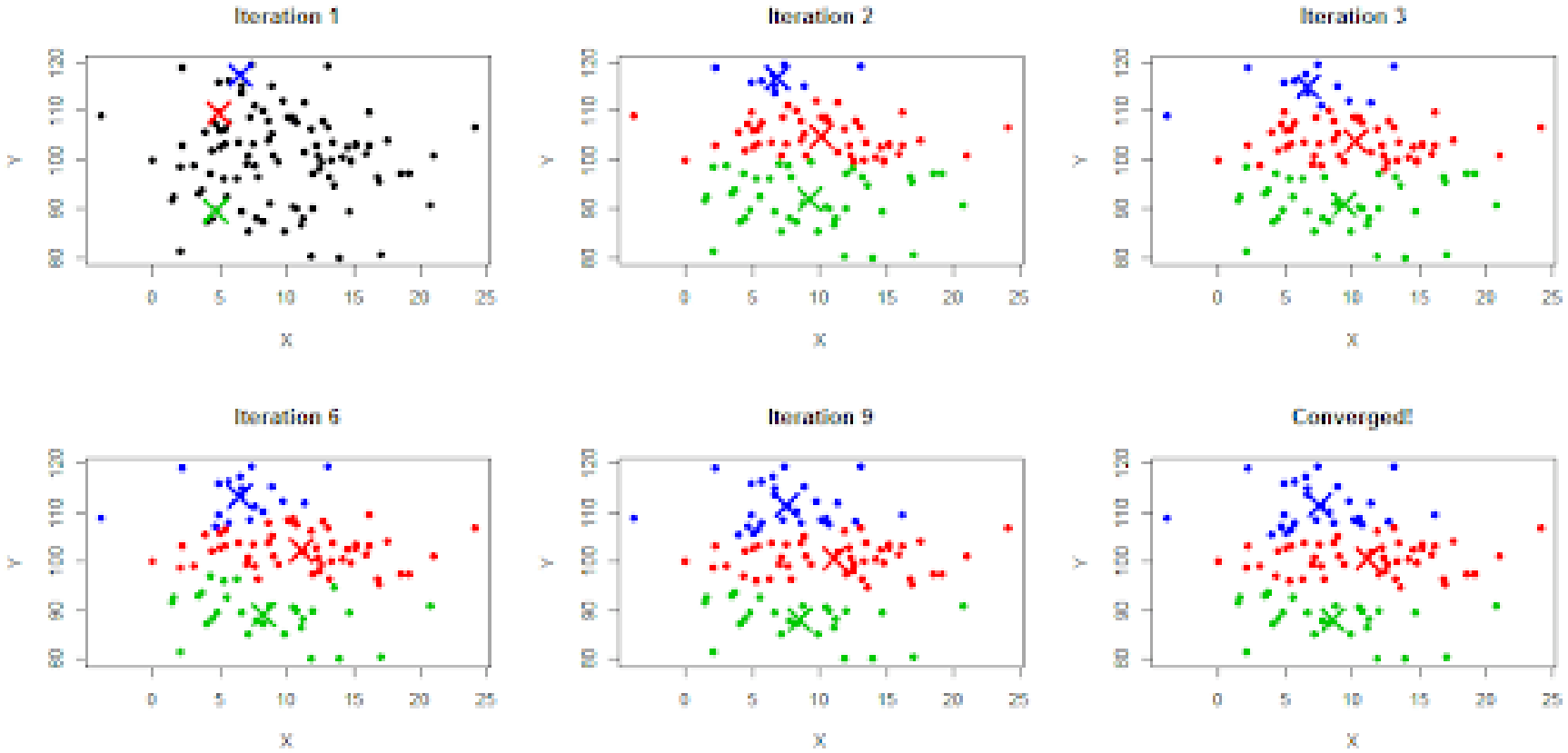


Figure: K-Means Algorithm

Source: https://images.app.goo.gl/27BUWSqpXkXMvG5u5

Applications of K-means

- Classification document.
- Delivery store optimization.
- Crime locations identifying.
- Segmentation of customers.
- Statistical review football team.
- Detection of fraud in insurance.
- Rideshare data analysis.

Checkpoint (1 of 2)

Multiple choice questions:

1. High entropy means that the partitions in classification are
   a) Pure
   b) Not pure
   c) Useful
   d) Useless

2. The attributes have 3, 2, 2, and 2 possible values each. The class has 3 possible values. How many maximum possible different examples are there?
   a) 12
   b) 24
   c) 48
   d) 72

3. Which of the following is NOT supervised learning?
   a) PCA
   b) Decision tree
   c) Linear regression
   d) Naive Bayesian

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. High entropy means that the partitions in classification are
   a) Pure
   b) Not pure
   c) Useful
   d) Useless

2. The attributes have 3, 2, 2, and 2 possible values each. The class has 3 possible values. How many maximum possible different examples are there?
   a) 12
   b) 24
   c) 48
   d) 72

3. Which of the following is NOT supervised learning?
   a) PCA
   b) Decision tree
   c) Linear regression
   d) Naive Bayesian

Checkpoint (2 of 2)

Fill in the blanks:

1. High entropy means that the partitions in classification are _____.
2. _____ is NOT supervised learning.
3. The _____ methods recognize clusters based on density function distribution.
4. Attributes are statistically _____ of one another given the class value.

True or False:

1. Stochastic gradient descent performs less computation per update than batch gradient descent. True/False
2. To classify job applications into two categories and to detect the applicants who lie in their applications using density estimation to detect outliers we can use generative classifiers. True/False
3. A good way to pick the number of clusters k, used for k-Means clustering is to try multiple values of k and choose the value that minimizes the distortion measure. True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. High entropy means that the partitions in classification are **not pure**.
2. **PCA** is NOT supervised learning.
3. The **density-based clustering** methods recognize clusters based on density function distribution.
4. Attributes are statistically **dependent** of one another given the class value.

True or False:

1. Stochastic gradient descent performs less computation per update than batch gradient descent. **True**
2. To classify job applications into two categories and to detect the applicants who lie in their applications using density estimation to detect outliers we can use generative classifiers. **True**
3. A good way to pick the number of clusters k, used for k-Means clustering is to try multiple values of k and choose the value that minimizes the distortion measure. **False**

Question bank

Two mark questions:

1. Define precision and recall.
2. Explain how a ROC curve works.
3. How is KNN different from k-means clustering?
4. What is the difference between supervised and unsupervised machine learning?

Four mark questions:

1. What's the trade-off between bias and variance?
2. Why is "Naive" Bayes naive?
3. Describe the difference between L1 and L2 regularization.
4. What is the difference between type I and type II error?

Eight mark questions:

1. What is a Fourier transform?
2. What is the difference between a generative and discriminative model?

Unit summary

Having completed this unit, you should be able to:

- Understand the concept of neural networks and kernel methods
- Learn example of sparse kernel machines and graphical models
- Gain knowledge on sampling methods for pattern recognition
- Understand pattern recognition in sequential data