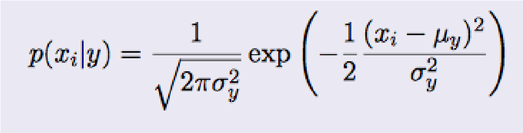
- Intro to Data Mining
  - Definition
    - The exploration and analysis of large quantities of data in order to discover valid, novel and potentially useful and ultimately understandable patterns in data
  - History
    - 1763 – Thomas Bayes published a paper regarding a theorem for relating current probability to prior probability
    - 1805 – Adrien-Marie Legendre and Carl Friedrich Gauss applied regression to determine orbits of bodies about the Sun. The goal of regression analysis was to estimate the relationship among variables and the specific method they used in this case was the method of least squares
    - 1936 – Alan Turing introduced the idea of a Universal Machine capable of performing computations like modern day computers
    - 1943 – Warren McCulloch and Walter Pitts created a conceptual model of neural network. They stated that a neuron can do 3 things: receive inputs, process inputs, and generate output
    - 1965 – Lawrence Fogel formed a new company called Decision Science, Inc. for applications of evolutionary programming.
    - 1970s – Sophisticated database management systems emerged making it possible to store and query terabytes and petabytes of data.
    - 1975- John Henry Holland wrote Adaptation in Natural and Artificial Systems, a ground-breaking book on genetic algorithms
    - 1980s – HNC trademarked the phrase “database mining”. It was meant to protect a product called DataBase Mining Workstation.
    - 1989 – Gregory Piatetsky-Shapiro coined the term Knowledge Discovery in Databases (KDD)
    - 1990s – The term data mining appeared in the database community
    - 1992 – Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik suggested an improvement on the original support vector machine which allows for the creation of nonlinear classifiers
    - 1993 – Gregory Piatetsky-Shapiro starts the newsletter Knowledge Discovery Nuggets
    - 2001 – William Cleveland introduced data science as an independent discipline
    - 2003 – Michael Lewis published Moneyball and changed the way many major league front offices do business
    - 2015 – DJ Patil became the first Chief Data Scientist at the White House
  - Models in Data Mining
    - The relationships and summaries derived through the process are called models or patterns
    - Examples are:
      - Linear equations
      - Rules
      - Graphs
      - Trees
      - Clusters

- Data mining is referred to as secondary data analysis because the data was collected for another purpose
- Dataset (Data Matrix)
  - A set of measurements taken from some environment or process
    - If we have a collection of *n* objects and *d* measurements on those objects, we can think of our data as an *n*x*d* data matrix
    - The *n* rows are called individuals, entities, cases, objects, or records
    - The *d* columns are called variables, features, attributes, or fields
  - The analysis of a single attribute is referred to as **univariate analysis**, whereas the simultaneous analysis of two attributes is called **bivariate analysis**, and the simultaneous analysis of more than two attributes is called **multivariate analysis**
- Main Types of Data
  - Quantitative
    - Measured on a numerical scale and in principle can take any value examples
  - Categorical variables
    - Can only be certain discrete variables
    - Ordinal
      - Possessing a natural order
    - Nominal
      - Simply naming categories

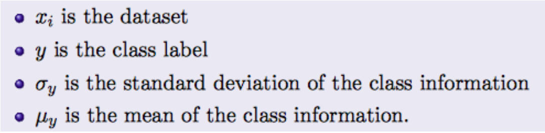
- Basic Mathematical Concepts
  - Trace of a matrix
    - The sum of the diagonal entries
- Measurement and Data
  - Data
    - A collection of objects and their attributes
      - An attribute is a property of characteristic of an object
      - A collection of attributes describes an object
    - The relationships between objects are represented by numerical relationships between variables. These numerical representations are stored in the data set
  - Types of Measurement
    - Nominal – Qualitative variables that do not have a natural order
    - Ordinal – Qualitative variables that have a natural order
    - Interval – Measurements where the difference between two values is meaningful
    - Ratio – Measurements where both difference and ration are meaningful
  - Types of Attributes
    - Discrete Attribute
      - A variable or attribute is discrete if it can take a finite or countably infinite set of values. A discrete variable is often represented as an integer-valued variable
    - Continuous Attribute
      - A variable or attribute is continuous if it can take any value in a given range with the range possibly being infinite
  - Distance Measures
    - Many data mining techniques rely on knowing the similarity or dissimilarity of two objects
      - A metric space is a dissimilarity measure

- Visualizing and Exploring Data
  - Exploratory Data Analysis combines
    - Graphical methods
    - Data transformations
    - Statistics and mathematics
  - Skewness
    - Measures whether or not a distribution has a simple long tail
      - Right-skewed – long tail extends in the direction of increasing values
      - Left-skewed – long tail extends in the direction of decreasing values
      - Symmetric distributions have 0 skew
  - Types of Data Visuals
    - Histogram plot
    - Box Plot
      - The box extends from the lower to upper quartile values of the data with a line at the median
    - Stack plot
    - Pie Chart
- Overview of Data Mining Algorithms
  - A data mining algorithm is a well-defined procedure that takes data as input and produces output in the form of models or patterns
    - **Well-defined** – procedure can be precisely encoded as a finite set of rules
    - **Algorithm** – procedure terminates after a finite number of steps and produces an output
    - **Computational Method** – has all the properties of an algorithm except guaranteeing finite termination
    - **Model Structure** – a global summary of the data set
    - **Pattern Structure** – statements about restricted regions of the space
  - Components of a Data Mining Algorithm
    - Task
      - Visualization, classification, clustering, regression
    - Structure (functional form) of model or pattern
      - Linear regression, hierarchical clustering
    - Score function
      - To judge quality of fitted model or pattern
      - Generalization performance on unseen data
    - Search or Optimization method
      - Steepest descent
    - Data Management technique
      - storing, indexing, and retrieving data
  - Some Data Mining Algorithms
    - CART – Classification and Regression Trees
      - Produces classification and regression models with a tree-based structure
      - Is a flexible tool for classification problems. It is popular for its adaptability and ability to perform well with little to no tuning
      - A recursive algorithm, at each iteration it finds the best splitting of data which could increase the probability of predicting the target values

            - Classification Aspect of CART
                - Task – prediction
                - Model Structure – tree
                - Score Function – Cross-validated Loss Function
                - Search Method – greedy local search
                - Data Management Method – Unspecified
        - Artificial Neural Networks
            - Mathematical model of human nervous systems
            - Essential Characteristics
                - Training
                - Input data
                - Input nodes
                - Layers
                - Weights
                - Targets
                - Loss Function
                - Optimizer function
                - Predictions
        - Multilayer Perceptron (MLP)
            - Feedforward MLP are the most widely used models in the general class of artificial network models
            - Provides a nonlinear mapping from a real valued input vector to a real valued output
- Conditional Probability, Bayes Theorem, Naïve Bayes Classifier
    - Conditional Probability
        - A measure of the probability of an event given that another event has occurred
        - The probability of A given B is defined as the quotient of the probability of A and B and the probability of B
        - Independent events
            - A and B are independent if the probability of A and B is the probability of A times the probability of B
        - Addition Law
            - The probability of A or B is the probability of A plus the probability of B minus the probability of A and B
    - Bayes' Theorem
        - The probability of A given B is the probability of B given A multiplied by the probability of A all divided by the probability of B
    - Naïve Bayes Classifier
        - Classification technique based on Bayes' Theorem with an assumption of independence among predictors
        - Assumes the presence of a particular feature in a class in unrelated to the presence of any other feature
        - Naïve Bayes Model
            - $$pr(x_k|c_k) = p(x_1, ..., x_k|c_k) = \prod_{i=1}^{p} pr(x_i|c_k), 1 \leq k \leq m.$$

- Gaussian Naïve Bayes
  - 

    $$p(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu_y)^2}{\sigma_y^2}\right)$$

    - x_i is the dataset
    - y is the class label
    - σ_y is the standard deviation of the class information
    - μ_y is the mean of the class information.
- Pros and Cons
  - Pros
    - Easy and fast to predict class of test data set
    - When assumption of independence holds, a Naïve Bayes classifier performs better compared to other models
    - Performs well in case of categorical input variables compared to numerical variables
  - Cons
    - If a categorical variable has a category in the test data set which is not observed in the training set the model will be unable to make a prediction
    - Known as a bad estimator, so probability outputs may not be taken too seriously
    - Limited by assumption of independent predictors

❖ Classification Algorithms and Clustering

- Classification Algorithm
  - A well-defined procedure that takes data as input and produces output in the form of models or patterns
  - Examples
    - Logistic Regression
    - Naïve Bayes classifier
    - Support Vector Machines
    - Decision Trees
    - Boosted Trees
    - Random Forest
    - Neural Networks
    - Nearest Neighbor
- Regression
  - The goal of **regression** is to predict the value of one or more continuous target variables given the value of a *n*-dimensional vector of input variables.
  - Linear Regression (Predictive Learning Model)
    - Statistical method for analyzing a data set with the following assumptions
      - Target variable is binary
      - Predictive features are interval (continuous) or categorical
      - Features are independent of one another
      - Sample size is adequate (50 records per predictor
  - Application of Regression
    - Trend lines
      - Model variation in some quantitative data with passage of time

      - Economics
        - Used to predict consumption spending, fixed investment spending, inventory investment, purchases of a country’s exports, spending on imports, etc.
      - Finance
        - Used to analyze capital price asset models and quantify the systematic risks of an investment
      - Biology
        - Used to model causal relationships between parameters in biological systems
    - Random Forest
      - Ensemble learning method for classification and regression that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees
    - k-Nearest Neighbor (kNN)
      - a supervised classification algorithm that takes a set of observations and uses them to learn how to label other observations
  - How to Build Classification Algorithms
    - Extract and assemble features to be used for prediction
    - Determine the size and shape and pre-process the dataset
    - Develop targets for the training
    - Train a model
    - Assess performance on test data
- Feature Selection
  - The main objective of feature selection is to improve the accuracy of the classification model
  - Types of Feature Selection
    - **Filter Methods** – apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset
    - **Wrapper methods** – consider the selection set of features as a search problem where different combinations are prepared, evaluated, and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on a model
    - **Embedded methods**