CSE514 Fall 2022 Data Mining
Characteristics of Data

Data Mining is all about the data, so how do we describe it and what factors affect its analysis?

- Size of dataset: number of data points/objects/entities/observations/samples
- Dimensions of data: number of features/variables that describe each data point
  - Curse of dimensionality: as the number of dimensions increases, statistical power tends to decrease
- Quality of data: presence of missing/erroneous values in the dataset
- Data type:
  - Quantitative/Numerical - data that can be expressed in terms of numbers. These values can be ordered or ranked, and the distance between values can be measured
    - Discrete: numerical data that can be counted, i.e. integer values
    - Continuous: numerical data that can't be counted, i.e. real values
  - Qualitative/Categorical- data that can be recorded under names/labels There is no measurable distance between these values
    - Nominal: categorical data that cannot be ordered/ranked, ex. gender
    - Ordinal: categorical data that can be ordered/ranked, ex. military rank
  - Complex – data that has a composite of quantitative and qualitative values, and/or inherent structure to its features, such as images and audio recordings

Data Preprocessing: Cleaning, selecting, transforming, and otherwise processing raw data can change *all* these factors. Preparing data for analysis is a large part of data mining and has a large impact on final results. To summarize the basic techniques:

1. **Data Cleaning**
   a. Missing Data: Samples/features with many missing values can be dropped from the set, or the missing values can be filled in (i.e. imputed)
   b. Noisy Data: Data could be noisy in the sense that they are measured with low precision, or in the sense that there's a degree of randomness in its generation. Binning/clustering can be used to compress values into a smaller set of options, or a line/curve can be fit to the data to smooth it out.
2. **Data Transformation**
   a. Normalization: Scaling data values into a specific range/distribution can help make data more comparable across features
   b. Feature Extraction: New features can be constructed from old ones
   c. Discretization: Replacing numerical values by interval levels or distinct concepts
3. **Data Reduction**
   a. Feature Selection: Some features can be dropped
   b. Dimensionality Reduction: Includes several encoding techniques to compress the size of data

Distance and/or similarity of data samples – To measure the distance between two vectors of numerical values, popular methods include:

- L^p-norm, or p-norm, for $p = 1, 2, \ldots$
  Consider a K dimensional vector space V, where $\vec{x} = (x_1, x_2, \ldots x_K) \in V$.
  For simplicity, we write $\vec{x}$ as x.

  In mathematics, a norm (typically written as $\|x\|$) is a function $f : V \to \mathbb{R}$ that satisfies:
  1. $f(x) > 0$ for $x \in V$ and $x \neq 0$
  2. $f(x + y) \leq f(x) + f(y)$ for $x, y \in V$, i.e. triangle inequality
  3. $f(\lambda x) = |\lambda| f(x)$ for all $\lambda \in \mathbb{R}$ and $x \in V$, i.e. positive homogeneity

  To specify a *p*-norm :
  $$\|x\|_p = (\sum_{k=1}^{K} |x_k|^p)^{1/p}$$
  Common p-norms worth mentioning:
  1. The 1-norm, i.e. Manhattan:
     $$\|x\|_1 = \sum_{k=1}^{K} |x_k|$$
     Visualize: Work commute calculated as walking along right-angle arranged sidewalks and taking the elevator up/down as needed.
  2. The 2-norm, i.e. Euclidean:
     $$\|x\|_2 = (\sum_{k=1}^{K} |x_k|^2)^{1/2}$$
     Visualize: Work commute as if riding a zip-line from your home to your office.
  3. The ∞-norm, i.e. the sup-norm or the uniform norm:
     $$\|x\|_\infty = \max \{|x_k|;\ k = 1, 2, \ldots, K\}$$
     Visualize: Work commute summarized as just the longest component

- Dot product
  This is again in K dimensional vector space, where $x = (x_1, x_2, \ldots x_K) \in V$

  The dot product of $x, y \in V$ is
  $$(x \cdot y) = x^T y = \sum_{k=1}^{K} x_k y_k$$
  From a geometry point of view, dot product computes the product of two vectors' magnitude and the cosine of the angle between them.
  Cosine: ranges from -1 to 1, positive if the vectors are pointing in the same direction, negative if the vectors point opposite directions
  Magnitude: the larger the values in the vector, the larger the product

Distance and/or similarity of data features – To measure the similarity between two values of two features, a popular method is Pearson's correlation coefficient (PCC)

Measures how two variables X and Y are correlated with each other.
Consider the observation sets $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$
The mean of X :

$$\bar{x} = \left(\sum_{k=1}^{n} x_k \right) / n$$

The standard deviation of X :

$$s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

Covariance between X and Y:

$$cov(X, Y) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

PCC of X and Y:

$$r = \frac{cov(X, Y)}{s_x s_y} = \frac{1}{n-1} \sum_{k=1}^{n} \left(\frac{x_k - \bar{x}}{s_x}\right)\left(\frac{y_k - \bar{y}}{s_y}\right)$$

In other words, Pearson's correlation is the covariance of two variables, normalized by their spread from their means.

Data sources: There are many sources of publicly available data of the web for testing and developing data mining algorithms. A brief sampling:

- https://archive.ics.uci.edu/ml/datasets.php
  Oldest and best-known UCI Machine Learning Repository with many curated datasets good for initial method development and testing
- http://networkdata.ics.uci.edu/resources
  UCI Network Data Repository for network analysis
- https://blog.bigml.com/2013/02/28/data-data-data-thousands-of-public-data-sources/
  Links and descriptions of various public datasets
- https://www.kdnuggets.com/2015/04/awesome-public-datasets-github.html
  Links and descriptions of various public datasets on GitHub