Week 1

1 Data Cleaning

- Noise (Distortion of values, Addition of spurious examples, Inconsistent and duplicate data) Can be reduced by
  - Using signal and image processing and outlier detection techniques before data mining
  - Using ML algorithms that are more robust to noise
- Missing values
  - Ignore all examples with missing values
  - Estimate the missing values by using the remaining values
    - Nominal attributes: Replace the missing values with the most common value or the most common values among the examples with the same class
    - Numerical attributes: Replace with the average value of the nearest neighbors

2 Data Pre-Processing

- Data aggregation: Combining two or more attributes into one. (Less memory and computational time, more stable data, potential loss of details)
- Dimensionality reduction
- Feature extraction: Creation of features from raw data, may require mapping data to a new space, the new space may reveal important characteristics
- Feature subset selection: Removing irrelevant and redundant features and **selecting a small set of features** that are necessary and sufficient for good classification (Faster building of the classifiers, more compact and easier to interpret classification rules)
  - Brute force: try all combinations of features as input to a ML algorithm
  - Embedded: Some algorithms can automatically select features (e.g. decision trees)
  - Filer: Select features before the ML algorithm is run (information gain, mutual information, odds ratio, relief)
  - Wrapper: Use the ML algorithm as a black box to evaluate different subsets and select the best
  - Feature weighting: Features with higher weights play more important role in the construction of the ML model
- Converting attributes from one type to another:
  - Discretization: Converting numerical attributes into nominal
    - Unsupervised: equal width, equal frequency, clustering

* Supervised:
  Entropy-based: Splits are placed so that they maximizes the purity of the intervals. Evaluate all possible splits and choose the best one with the lowest total entropy, repeat recursively until stopping criteria are satisfied.
  Entropy: Measure of the purity of the dataset S

$$entropy(S) := -\sum_i P_i log_2 P_i,$$

  where P_i is the proportion of examples from class i.
  Total entropy of the split: weighted average of the interval entropies:

$$totalEntropy = \sum_i^n w_i entropy(S_i),$$

  where w_i is the proportion of values in interval i, n is the number of intervals.

- Binarization: Converting categorical and numeric attributes into binary

- Normalization and Standardization: Transform attributes to a new range, e.g. $[0, 1]$, used to avoid dominance of attributes with large values over attributes with small values. (Required for distance-based ML algorithms)
  - Normalization (min-max scaling)

$$x' = \frac{x - min(x)}{max(x) - min(x)},$$

  - Standardization

$$x' = \frac{x - \mu(x)}{\sigma(x)},$$

    where $\mu(x)$ is the mean value of the attribute, $\sigma(x)$ is the standard deviation of the attribute.

3 Similarity Measures

- Distance
  - Euclidean distance (l2 norm): $D(A, B) = \sqrt{(a_1 - b_1)^2 + ... + (a_n - b_n)^2}$
  - Manhattan distance (l1 norm): $D(A, B) = |a_1 - b_1| + ... + |a_n - b_n|$
  - Minkowski distance: $D(A, B) = (|a_1 - b_1|^q + ... + |a_n - b_n|^q)^{1/q}$
  - Weighted distance (requires domain knowledge to assign weights for each attribute): $D(A, B) = \sqrt{w_1|a_1 - b_1|^2 + ... + w_n|a_n - b_n|^2}$
  - Hamming distance (Manhattan distance for binary vectors): $D(A, B) = |a_1 - b_1| + ... + |a_n - b_n|$
  - Simple matching coefficient (SMC): $(f11 + f00)/(f00 + f01 + f10 + f11)$.
    Similarity coefficients: $f00$: number of matching $0 - 0$ bits, $f01$: number of matching $0 - 1$ bits, $f10$: number of matching $1 - 0$ bits, $f11$: number of matching $1 - 1$ bits.
    SMC is not suitable for sparse data.
  - Jaccard coefficient: $J = f11/(f01 + f10 + f11)$
  - Cosine similarity (useful for sparse data): $cos(A, B) = \frac{A \cdot B}{||A||||B||}$

- Correlation: Measures **linear** relationship between numeric attributes

– Pearson correlation coefficient:

$$corr(x, y) = \frac{covar(x, y)}{\sigma(x)\sigma(y)},$$

where $\mu(x) = \frac{\sum_{k=1}^{n} x_k}{n}$, $\sigma(x) = \sqrt{\frac{\sum_{k=1}^{n}(x_k - \mu(x))^2}{n-1}}$, $covar(x, y) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \mu(x))(y_k - \mu(y))$. Range: $[-1, 1]$ (-1 is perfect negative correlation, $+1$ is perfect positive correlation, 0 is no correlation)