3- data preprocessing

data quality = accuracy, completeness, consistency, timeliness, believability, interpretability

data cleaning

- missing values = ignore or fill with "unknown", mean, or most probable value
- noisy = errors / outlier (ex: age=-2)
  - ↳ solve by regression, clustering, inspection, or binning (partition then smooth)

↳ data discrepancy (inconsistency) detection = use metadata, check rules

↳ data scrubbing (cleaning/format) = use domain knowledge to detect errors and make corrections

↳ data auditing (detetleme) = analyze data to discover rules to detect violators (correlation, clustering)

↳ data migration = moving data from one location to another

data integration combining data from multiple sources into a coherent store, redundant data is common

- object identification (entity identification) = the same obj/att may have different names in different databases
- derivable data = one att may be a derived att in another table

↳ correlation analysis for nominal data

larger χ^2 val, more likely related variables

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

↳ correlation analysis for numeric data

Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

r>0 → positive correlation r<0 → negative correlation r=0 → independent

different

↳ correlation as linear relationship

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

$$correlation(A,B) = A' \bullet B'$$

* correlation does not imply causality: # cars and # hospitals are correlated, because both are causally linked to the third variable: population

↳ covariance for numeric data

$cov_{A,B} > 0$ → A, B tend to be larger than expected vals

$cov_{A,B} < 0$ → if A is larger than exp, B is smaller than exp

$cov_{A,B} = 0$ → independent (with some additional assumptions)

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

↳ $E(A \cdot B) - \bar{A}\bar{B}$

Correlation coefficient: $r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

data reduction reducing the size of dataset while still preserving the most important information

↳ dimensionality reduction = larger # dimension, more sparse data → curse of dimentionality

- wavelet transform = mapping data to a new space (fourier transform in img compression)
  - removal of outliers, efficient (ocw) complexity), only applicable to low dimensional data
  - DWT (digital wavelength transform) = store only the strongest of the wavelet coefficient

similar to DFT (fourier transform), but less space, lossy → skipped

ex: S = [2, 2, 0, 2, 3, 5, 4, 4] → length must be 2^n, add 0 to end if necessary

↳ [2, 1, 4, 4] coefficients [0, -1, -1, 0]

↳ [1.5, 4] [0.5, 0]

↳ [2.75] (Smoothing) [-1.25] (difference)

[2.75, -1.25, 0.5, 0, 0, -1, -1, 0]

- principle component analysis = transforms large number of correlated variables into a smaller set of correlated variables called principle components (variance) 3d → 2d
- data → standardization → covariance matrices → eigen val and vector calculation → choosing components, and form feature vec → new dataset
- attribute subset selection = 2^d possible attribute combinations of d attributes
  - best step-wise selection = best is picked first, next best condition to the best
  - step wise elimination = repeatedly eliminate worst
- attribute creation / feature generation = att extraction, mapping, att construction

↳ numerosity reduction = reducing data volume by choosing smaller form of data repository

- parametric method = data fits some model, estimate model parameters, store only the parameters, discard data → regression (multiple reg $Y = b_0 + b_1 x_1 + b_2 x_2$)
- non-parametric method = no models, histograms, clustering, sampling (→ without replacement, no duplicates)
- data cube aggregation = representing the original data set by aggregating at multiple layers of a cube, condensing data into a more manageable format

↳ data compression = string, audio/video compression (loseless - lossy)

- dimentionality and numerocity reduction are also data compression

data transformation mapping entire set of values

↳ normalization = organization of data to appear similar accross all fields and records

- min-max normalization = to [new_min, new-max] → $v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$
- z-score normalization = $v' = \frac{v - \mu_A}{\sigma_A}$ μ = mean σ = std dev
- normilizing by decimal scaling = $v' = \frac{v}{10^j}$ Where j is the smallest integer such that $Max(|v'|) < 1$

↳ discretization = dividing the range of a continuous attribute into intervals

↳ reducing data size, can be performed recursively

- binning = top-down split, unsupervised
  - ↳ equal width partitioning = $w = \frac{max - min}{N}$, skewed data is not handled well
  - ↳ equal depth (frequency) = each containing almost same number of samples
- histogram analysis = top-down split, unsupervised, classification
- clustering analysis = top-down split or bottom up merge, unsupervised
- decision-tree analysis = top-down split, supervised
- correlation (χ^2) analysis = bottom up merge, unsupervised
- concept hierarchy generation = organization of data into a tree like structure, where each level of hierarchy represents a concept that is more general than the level below it. → street (674,339) < city (3567) < state (365) < country (15 distinct val)