BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine

Lecture 6: Data Exploration and Data Cleaning

25 September 2024

Lecturer: Yu LI

Scriber: LEE, Michelle (1155179462)

1. Recap

- Different scoring matrix → different alignment score
- Translating sequencing data to data matrix involves mapping short read to the reference genome and counting the number of reads for the gene expression matrix.

2. Genome assembly

- Illumina sequencing length is 200 bp, and it needs to be assembled into the whole genome.
- Two 200 bp reads have overlap region.
- Impossible to sequence the whole genome all at once.
- Needs to be shattered into reads.
- Using the overlapping to assemble the DNA sequences into a whole genome, or a longer sequence.
- Example: 3 bp short reads.
  - Problem: multiple possibilities with the final assembly, e.g., when the last sequences are the poly-A tail AAAAAAA, where short reads AAA cannot determine the exact number of A in the sequenced gene.
  - Solution: produce longer reads.
- Other problems: mutation, conflict (AAT vs AAA), sequencing error, repeats (TGGGTGGGTGGGT), needs for faster algorithm.
- How to map: slide each read along the reference genome and calculate the difference, or we can use dynamic programming for calculating the alignment score for each read.
- For calculating the gene expression: depends on algorithm (make your assumption clear in HW and exams), some software count those reads flanking both the target reference gene, some count only those within the target gene.

3. Data cleaning

- Recap: data matrix
  - Collection of records; each consists of a fixed set of attributes.
  - Can be represented by row(n) x columns(m) matrix, each row for each object and each column for each attribute.
  - Swapping the entire column or the entire row at one time will not change the data.

a. Noise and outliers

- Noise is a modification of the original value, affecting the original values.
- Outliers are data objects with distinct characteristics from other data objects.
- Sometime just random or errors; not useful
- Some gives important information

b. Missing values

- Some information is not collected or not applicable.
- Solution:
  - Eliminate the whole data object that has missing values. Risk of deleting a lot of data.
  - Estimate missing values; make assumptions, e.g., similar height → similar weight.
  - Ignore.
  - Replace with all possible values, weighted by their probabilities.

c. Duplicate data

- Merging two datasets may cause duplicates.

d. Unnormalized data

- Data are incomparable.
- We need comparable data to calculate norms or Euclidean distance.
- The scales of the unnormalized data will cause bias to one of the parameters during calculation. E.g. in gene expression level; to eliminate the technical variation. Like if you sequence one sample too "deep", which generally produces more copies of all reads for that certain cell/system.
- Min-max normalization: ranges are 0 to 1.

$$v' = \frac{v - v^{min}}{v^{max} - v^{min}}$$

- Z-score normalization with gaussian/normal distribution assumption.

$$v' = \frac{v - Mean(v)}{Std(v)}$$

- One-hot encoding, which, for example, splits gender to two attributes and do 0 and 1 for each attribute.

*) Data cleaning order affects the final results.

4. Summary statistics

- Numbers that summarize the properties of the data, includes frequency, location, and spread, or mean and SD.
- Measures of location

$$mean(x) = \frac{1}{m}\sum_{i=1}^{m} x_i$$

  - Mean is the most common measure of the location of a set of points.
  - Sensitive to outliers
  - Alternatives: median/trimmed mean

$$median(x) = \begin{cases} x_{(r+1)} & if\ m\ is\ odd \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & if\ m\ is\ even \end{cases}$$

- Measures of spread
  - Range: difference between max and min.
  - Variance or SD as the most common measure of spread.

$$variance(x) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - mean(x))^2$$

  - Var and SD are sensitive to outliers.
  - Alternatives: median absolute deviation (MAD), interquartile range

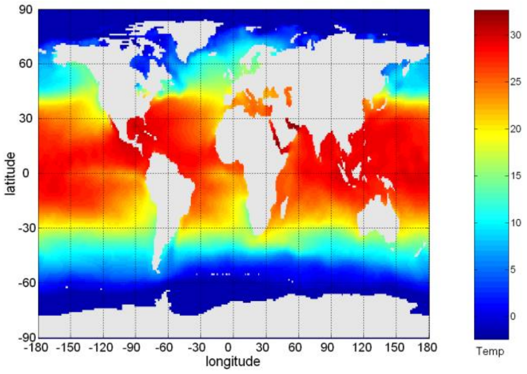
$median(|x_1 - mean(x)|, \dots, |x_m - mean(x)|)$ $x_{75\%} - x_{25\%}$

- Percentiles
  - Applicable to ordinal or continuous attribute *x*.
  - *p* is between 0 and 100.
  - *p*-th is the value of x where *p*% of the observed values of x are less than x_p
  - Example: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
    30% percentile: 4, the sets of values are [1, 2, 3]
- Frequency and mode
  - Frequency: percentage of time a value occurs in the data set
  - Mode: most frequent attribute value occurring in the data set
  - Both typically used with categorical data

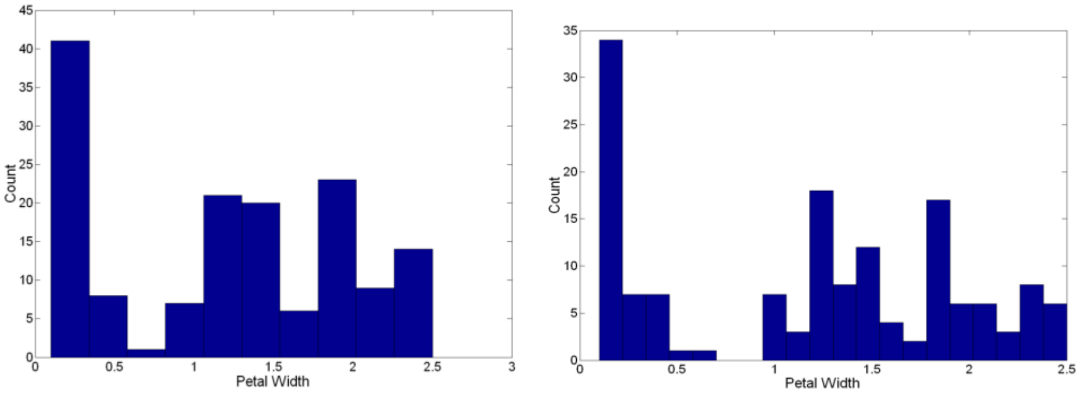
5. Exploratory visualization

- Visualization: conversion of data into visual or tabular format.
- To analyze or report the characteristics of the data and relationships between items/attributes.

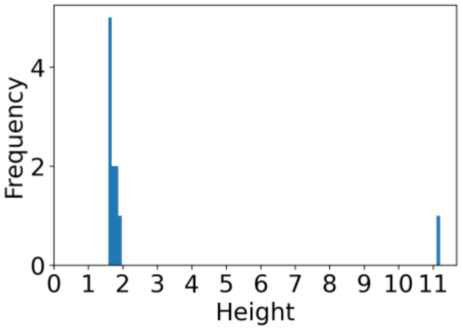
- Can detect general patterns and trends, outliers and unusual patters.
- Example: visualization for sea surface temperature



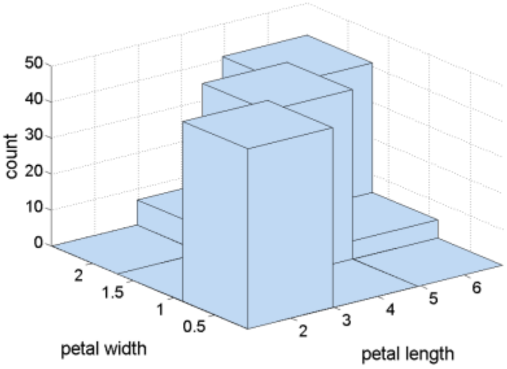
- Histograms
  - Visualize distribution of values of a single variable
  - Divide values into bins and show bar plot for number of objects in each bin
  - Height of bar: number of objects
  - Shape of histogram depends on the number of bins



  - Easy to spot outliers of the data

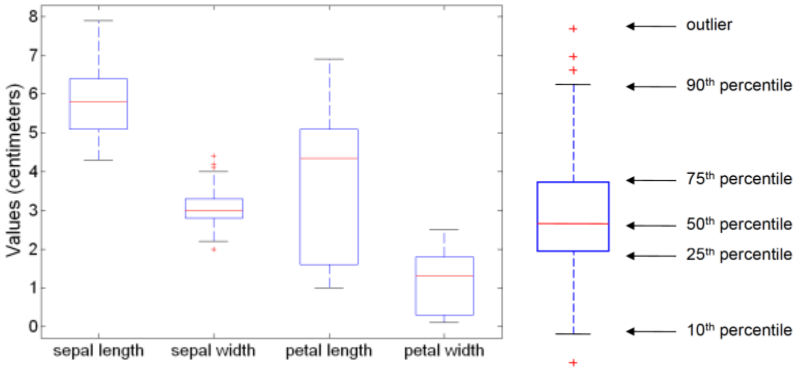


- 2D histograms
  - Joint distribution values from two attributes
  - Example: relationship between petal width and length



- Box plots

- For displaying and comparing distribution of data



Disclaimer: all figures are adapted from Prof. Li BMEG3105 Lecture Notes