BMEG3105 Lecture 5 Scribing

Scriber: Lai Yuk Fai (1155190593)

Recap from last lecture:

1. Scoring Matrix
   - Why Mismatch? Mutation
   - Why Gap? Insertion/Deletion, Gene Duplication
   - We have a lot of different scoring matrices built from different databases which can be chosen based on our needs.
   - Gap Penalty = -10
2. Transition from Sequence Data to Data Matrix
3. RNA-seq
   - Map the short read to the genome
   - Count the no. of reads → a gene expression matrix
   - Where and how do we get the genome sequence? (4.)
   - How do we map the short reads to the genome? (5.)
4. Genome Assembly
   - Illumina sequencing length = 200 bp
   - Human genome length = 3B bp
   - Two 200 bp short reads can have overlap regions
   - The genome is assembled based on the short reads and overlapping regions (like Jigsaw Puzzle)
5. Mapping
   - Slide each read along the genome and calculate the difference
   - We can use Dynamic Programming
   - Issues: speed, errors, mutations

Today's Agenda:

1. Data Cleaning
   - Data Type: Data Matrix
   - Data Matrix:
     i. Data consists of a collection of records, each of which consists of a fixed set of attributes
     ii. Data set can be represented by an n by m matrix, n rows of objects and m columns of attributes
     iii. If you shuffle the entire column or the entire row at one time, you will not change the data
   - Data quality problems:
     i. Noise: refers to modification of original values (e.g. distortion of a person's voice when talking on a poor phone and "snow" on television screen)
     ii. Outliers: data objects with characteristics that are considerably different than most of the other data objects in the data set
     iii. Missing values: Information is not collected (e.g. people decline to give their age and weight) or attributes may not be applicable to all cases (e.g. annual income is not applicable to children). We handle the missing values by:
       - ◆ Eliminating data objects
       - ◆ Estimating missing values

- Ignoring the missing values during analysis
- Replacing with all possible values (weighted by their probabilities)

iv. Duplicate data: major issue when merging data from heterogeneous sources (e.g. the same person with multiple email addresses)

v. Unnormalized data: attributes not on the similar level of measurement
- Normalization → attributes on the similar level of measurement
  - Min-max normalization:

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

  - Z-score normalization:

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

vi. Categorical data: computers are better at handling numbers. For categorical data, we can use one-hot encoding

- Data cleaning techniques:

i. Denoise data (if applicable)
ii. Remove outliers
iii. Handling missing data
iv. Remove duplicates
v. Categorical data encoding
vi. Data normalization

2. Data Exploration
- Summary statistics: numbers that summarize properties of the data

i. Measures of location:
- mean (most common measure of the location of a set of points and very sensitive to outliers):

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

- median/trimmed mean:

$$median(x) = \begin{cases} x_{(r+1)} \ if\ m\ is\ odd, i.e., m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}\ if\ m\ is\ even, i.e., m = 2r \end{cases}$$

ii. Measure of spread:
- Range: the difference between the max. and the min.
- Variance/ standard deviation (most common measure of the spread of a set of points):

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

- Sensitive to outliers, other measures:
  - Median Absolute Deviation (MAD):

$$median(|x_1 - mean(x)|, \dots, |x_m - mean(x)|)$$

  - Interquartile Range:

$$x_{75\%} - x_{25\%}$$

iii. Percentiles: Given an ordinal or continuous attribute x and a number p between 0 and 100, the p-th percentile is a value of x such that p% of the observed values of x are less than x_p

iv. Frequency and mode:
- The frequency of an attribute value is the percentage of time the value occurs in data set (e.g. data: a representative population of people; attribute: "gender"; frequency of "gender = female" occurs about 50% of the time
- The mode of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

- Exploratory visualization:

i. Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported

ii. Visualization of data is one of the most powerful and appealing techniques for data exploration
- Humans have a well-developed ability to analyze large amounts of information that is presented visually
- Easier to detect general patterns and trends
- Easier to detect outliers and unusual patterns

iii. Visualization techniques:
- Histograms: usually show the distribution of values of a single variable
- 2D Histograms: show the joint distribution of the values of two attributes
- Box Plots: another way of displaying and comparing the distribution of data

3. Clustering (will be discussed in detail in the next lecture)
- Why Clustering?

i. Understanding:
- As a stand-alone tool to get insight into data distribution
- As a pre-processing step for other algorithms

ii. Summarization:
- Reduce the size of large data sets
- Preserve privacy

- What is clustering analysis?

i. Finding groups of objects such that the objects in a group will be similar (or related) to one another and different form (or unrelated to) the objects in other groups

- What are needed to do clustering?

i. Data to be clustered

ii. Similarity measurement

iii. Clustering algorithm

4. Similarity and Dissimilarity
- Similarity: numerical measure of how alike two data objects are ([0,1], higher = more similar)
- Dissimilarity: numerical measure of how different two data objects are (0 = identical, lower = more similar)
- Cosine similarity: (where • indicate vector dot product and $|d|$ is the length of the vector d)

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{(|d_1| * |d_2|)}$$

- Correlation: (measures the linear relationship between objects)

$$\rho_{X,Y} = corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- Euclidean distance: (where m is the number of dimensions (attributes) and p_k and q_k are, respectively, the *k*-th attributes (components) or data objects *p* and *q*)

$$Ed(p, q) = \sqrt{\sum_{k=1}^{m} (p_k - q_k)^2}$$

- Minkowski distance: (where *r* is a parameter, m is the number of dimensions (attributes) and p_k and $q_k$are, respectively, the *k*-th attributes (components) or data objects *p* and *q*)

$$dist(p, q) = (\sum_{k=1}^{m} |p_k - q_k|^r)^{\frac{1}{r}}$$

  i. r = 1 → City block (Manhattan, taxicab, L_1 norm) distance
     - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
  ii. r = 2 → Euclidean distance
  iii. r = ∞ → "Supremum" (L_{max} norm, L_∞ norm) distance
     - This is the maximum difference between any component of the vectors

<u>Potential Project - 1: A pipeline to get the gene expression matrix from reads</u>

1. Find the genome
2. Find the reads
3. Map reads to reference genome
4. Count reads for each gene
5. Use Google to find the software and the data
6. Explain each step in the report to let us know you understand what you are doing

<u>Potential Project - 2-1: Data preprocessing for the gene expression matrix</u>

1. Data collecting and merging
2. Exploration
3. Visualization
4. Data cleaning