BMEG3105 Student Name: Chail Kamalpreet Kaur SID:1155191208 (Fri Lecture)

Lecture Title: Multi-Omics & Cancer Genome Overview (LT14)

Date:25 October 2024

Presented by: Professor Yu Li

Institution: Department of Computer Science and Engineering, The Chinese University of Hong Kong

Topics Covered:

1. Logistic Regression:

- Formula: $\frac{1}{1+e^{-(w_h H + w_w W + w_0)}} \geq 0.5$

- Deep Learning for Disease Screening:

- The complexity of relationships among variables can exceed simple logistic regression.

- Solutions include increasing model capacity through:

- More nodes

- More layers

- Non-linear functions

2. Model Underfitting and Overfitting

- Underfitting occurs when a model is too simple to capture the underlying patterns in the data. This results in poor performance on both training and test data.
- Overfitting happens when a model is too complex and starts to fit noise in the training data. This leads to excellent performance on training data but poor generalization to new, unseen data.

To address overfitting:

- Increase the amount of training data
- Reduce model complexity
- Use regularization techniques
- Implement early stopping during training
- Use techniques like dropout to reduce co-adaptation of neurons

- Evaluating Overfitting:

- Use train-validation-test splits (e.g., 70%-15%-15%).

- Cross-validation techniques, such as 5-fold validation and leave-one-out.

3. Multi-omics Overview

- A longitudinal approach to precision health, integrating large-scale biological data.

- Omics refers to large-scale studies of biological molecules that translate into the structure, function, and dynamics of organisms.
- Examples include genomics, transcriptomics, proteomics, metabolomics, etc.

Multi-omics approaches aim to combine these different data types to gain a more comprehensive understanding of biological systems. This is particularly relevant in complex diseases like cancer, where multiple factors contribute to disease development and progression.

Importance of Multi-omics:

- Provides a holistic view of biological processes and disease mechanisms.

- Enables the integration of various data types, such as demographic, clinical, and environmental factors.

Several pipelines for different omics data types:

- Genome pipeline: DNA sequencing → Read alignment → Variant calling
- Epigenome pipeline: Various assays (e.g., ChIP-seq, ATAC-seq) → Peak calling → Differential analysis
- Transcriptome pipeline: RNA sequencing → Read alignment → Gene expression quantification → Differential expression analysis

4. Statistical Testing

- T-test: Used to compare means between two groups, considering both the difference in means and the variability of the data.
- P-value: The probability of obtaining results at least as extreme as the observed results, assuming the null hypothesis is true.
- Different types of t-tests: Paired vs. unpaired, one-tailed vs. two-tailed.

Gene enrichment analysis, which aims to identify biological pathways associated with a set of genes. This involves:

- Using databases like KEGG (Kyoto Encyclopedia of Genes and Genomes) to define gene sets associated with pathways.

- Applying statistical tests like Fisher's exact test to determine if there's a significant overlap between genes of interest and pathway-associated genes.

5.Cancer Genomics Overview

Cancer genomics:

- Cancer is defined as a disease where some of the body's cells grow uncontrollably and spread to other parts of the body.
- Cancer is a leading cause of death worldwide, making it a critical area of research.
- Cancer is often considered a genomic disease, involving changes in DNA that affect cell growth and division.

Various omics approaches to studying cancer:

- Genome: Identifying genetic variants and conducting genome-wide association studies.
- Epigenome: Studying changes in DNA methylation and histone modifications.
- Transcriptome: Analyzing differential gene expression and identifying gene fusions.