**BIOST 2078 Introductory high-throughput genomic data analysis II: theories and algorithms**

**Date: 8/21/2013**

**Lecturer: George C. Tseng (****ctseng@pitt.edu****)**

**Course hours: Wed and Fri 3:30-4:45PM at Crabtree A425.**

**TA: Ying Ding (****dingying85@gmail.com****)**

**TA Office hour: TBD**

**TA Office: 325A Parran Hall**

**Summary of course:** This course is a graduate level course to introduce theories and algorithms for statistical analysis of high-throughput genomic data. Most of the methods are not covered in other standard biostatistical courses. Emphases will be given to high-dimensional data analysis and theories behind the commonly used methods. This course is designed for graduate students who already have sufficient statistical background, have basic knowledge of various high-throughput genomic experiments and wish to learn advanced statistical theories for bioinformatics and genomics research.

**Targeted audience**: This course is designed for students who have taken the “Introductory high-throughput genomic data analysis I: data mining and applications” course (or have basic bioinformatics knowledge) and are interested in more in-depth theories and algorithms of high-throughput data analysis methods. The students are expected to have solid statistical inference training and R programming experiences.

Schedule of Sessions and Assignments (subject to change)

|  |  |
| --- | --- |
| 8/28 | Short intro. of the course |
| 8/30 | *Experimental design and data preprocessing* experimental design, quality diagnosis, filtering, normalization, missing value imputation |
| 9/4 | *Dimension reduction* eigen-decomposition, principal component analysis (PCA) |
| 9/6 | *Dimension reduction* eigen-decomposition, principal component analysis (PCA)***Homework assignment 1 distributed*** |
| 9/11 | *Dimension reduction* multi-dimensional scaling (MDS), partial least squares (PLS) |
| 9/13 | *Hypothesis testing and permutation analysis* |
| 9/18 | *Differential analysis and multiple comparison* |
| 9/20 | *Supervised machine learning* Bayes classification rule, logistic regression. linear and quadratic discriminant analysis (LDA, QDA). |
| 9/25 | *Supervised machine learning* linear and quadratic discriminant analysis (LDA, QDA). classification and regression tree (CART). Resampling methods, such as Bagging, Boosting and random forest. |
| 9/27 | *Supervised machine learning support vector machines (SVM)****Homework assignment 2 distributed*** |
| 10/2 | *Supervised machine learning cross-validation, feature selection, performance assessment, over-fitting and under-fitting, common mistakes and how to choose a classification method.* |
| 10/4 | *Unsupervised machine learning* basic concept, hierarchical clustering, K-means, model-based clustering. |
| 10/9 | *Unsupervised machine learning* penalized K-means*;* tight cluste*ring; Bayesian model-based clustering* |
| 10/11 | *Unsupervised machine learning* selection of number of clusters; clustering evaluation |
| 10/16 | *Regularization and sparse methods* ridge regression, lasso, elastic net |
| 10/18 | *Regularization and sparse methods* general regularization methods, sparse PCA |
| 10/23 | *Regularization and sparse methods* sparse K-means, nearest shrunken centroids***Homework assignment 3 distributed*** |
| 10/25 | *Genomic meta-analysis* Basic concept, hypothesis setting, methods to combine effect sizes. |
| 10/30 | *Genomic meta-analysis* Methods to combine p-values. |
| 11/1 | *Genomic meta-analysis* Statistical properties (power, admissibility), microarray meta-analysis, GWAS meta-analysis |
| 11/6 | Pathway (gene set) analysis Fisher’s exact test, KS-test |
| 11/8 | Pathway (gene set) analysisGSEA, hypothesis setting, meta-analysis for pathway analysis.***Homework assignment 4 distributed*** |
| 11/13 | *Graphic models* Intro of graphic models. |
| 11/15 | *Graphic models* Hidden Markov model. |
| 11/20 | *Graphic models* Bayesian network. |
| 11/22 | *Biological network analysis* Various network measures, small-world network, scale-free network***Homework assignment 5 distributed*** |
| 11/27 | *Thanksgiving recess (no class)* |
| 11/29 | *Thanksgiving recess (no class)* |
| 12/4 | *Dynamic programming* example in sequence alignment |
| 12/6 | *Selected topics* Basic information theory (entropy, mutual information, KL-divergence); Commonly used norms, distance measures and correlation measures of different data types |

**Grades**: Since this is a special topic course and the teaching objective is mainly to enhance your research capability and independent thinking, there will be no exam in this course. Instead, we have multiple homework assignments and one final project (a bigger computing homework project).

Homework 1: 20

Homework 2: 20

Homework 3: 20

Homework 4: 20

Homework 5: 20