Bioengineering at GMU

- Bioengineering Department demographics
  - Established in 2011
  - 13 core faculty members; 5 affiliate GMU faculty
  - BS program: Over 200 undergraduate BioE majors
  - PhD program started in Spring 2015: Currently 10 PhD students
  - MS program approval underway: To be launched in Spring 2017

- Highest average GPA in Volgenau School of Engineering

- Three concentrations:
  - Biomedical Signals and Systems
  - Health Care Informatics
  - Pre-Health

- Exchange program with Carlos III University of Madrid

- BENG 499: Bioengineering World Health
  - 4 BioE majors worked in a hospital in Guatemala in Fall 2015
AFTER MASON

- Twenty-two just graduated this May
  - Over 90% had some research experience
  - Over 60% received an undergraduate research award
- Our department has now graduated 68 students
- The majority graduated from the Biomedical Signals and Systems concentration
- Overall GPA: 3.34
- After graduation
  - ~20% are in graduate school
  - ~10% are in research related jobs
  - Industry
Research Areas

- Biomechanics
- Biomedical imaging
- Computational Neural Modeling
- Electrophysiology
- Fluid dynamics
- Neuroengineering
- Nanoscale technology
- Prosthetics
- Sensorimotor integration
- Tissue engineering

Funding from NIH, DoD, NSF:
Total amount from active awards $9M

FY 2015 research expenditures:
$1.26M or $140k per faculty member
Workshop Objectives

• Introduce the fundamental data mining concepts and techniques for biomedical data, important for many biomedical engineering problems.

• Gain hands-on experience using Matlab on data analysis by applying an algorithm to analyze and interpret biomedical data.
Outline

• Introduction
• Classification for biomedical problems
• Artificial Neural Network
  – Application in cancer prediction
Data Mining Definition

Mining in Data is the *non-trivial* process of identifying
- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.
What is (not) Biomedical Data Mining?

- Look up a patient’s name in the directory
- Certain diseases are more prevalent for certain population (e.g. myopia, heart disease, joint problem)
- Query a Web search engine for information about “lung cancer”
- Group together similar documents returned by search engine according to their context (e.g. globe eye, globe)
Data Mining and Medicine

- 1854 Broad Street cholera outbreak
- Miasma theory
- John Snow: contaminated water

https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak
The future of medical data mining

• 10 biggest killers in US

<table>
<thead>
<tr>
<th>Cause</th>
<th>Deaths</th>
<th>Drugs</th>
<th>$ Billions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Disease</td>
<td>739,000</td>
<td>Cardiovascular</td>
<td>16.5</td>
</tr>
<tr>
<td>Cancer</td>
<td>538,000</td>
<td>Anticancer</td>
<td>4.1</td>
</tr>
<tr>
<td>Stroke</td>
<td>158,000</td>
<td>Anticoagulants</td>
<td>0.3</td>
</tr>
<tr>
<td>Lung Disorders</td>
<td>105,000</td>
<td>Respiratory</td>
<td>4.4</td>
</tr>
<tr>
<td>Accidents</td>
<td>90,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pneumonia &amp; Flu</td>
<td>84,000</td>
<td>Antibiotics/Antibacterials</td>
<td>9.5</td>
</tr>
<tr>
<td>Diabetes</td>
<td>59,000</td>
<td>Insulin</td>
<td>0.8</td>
</tr>
<tr>
<td>AIDS</td>
<td>43,000</td>
<td>Antivirals</td>
<td>0.3</td>
</tr>
<tr>
<td>Suicide</td>
<td>31,000</td>
<td>Antidepressants and Antipsychotics</td>
<td>4.6</td>
</tr>
<tr>
<td>Homicide</td>
<td>25,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Leads to:
  – Question based answers
  – Anomaly based discovery
  – New Knowledge discovery
  – Informed decisions
  – Probability measures
  – Predictive modeling
  – Decision support
  – Improved health
  – Personalized medicine
Biomedical Data

• Wide Range of Data
  – Clinical Data on Patients
  – Diagnostic Data (Scans, Labs, EKG, etc.)
  – Population Data (Public Health Surveillance)
  – Research on Genomic and Biological Data

• Any Data Involved in
  – Care of Patients
  – Medical and Clinical Research
Major Issues in Biomedical Data Mining

• Heterogeneity of medical data
  – Volume and complexity
  – Physician’s interpretation
  – Poor mathematical categorization
  – Canonical Form
  – Solution: Standard vocabularies, interfaces between different sources of data integrations, design of electronic patient records

• Ethical, Legal and Social Issues
  – Data Ownership
  – Lawsuits
  – Privacy and Security of Human Data
  – Expected benefits
  – Administrative Issues
Supervised versus Unsupervised Learning

- **Supervised learning (classification)**
  - Supervision: Training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on training set

- **Unsupervised learning (clustering)**
  - Class labels of training data are unknown
  - Given a set of measurements, observations, etc., need to establish existence of classes or clusters in data

Classification

Learn a computational model (classifier) to assign data points into 2+ classes

Many approaches: Support Vector Machine, Decision Trees, Neural Networks, ...

## Training Dataset

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Hereditary</th>
<th>Vision</th>
<th>Risk of Condition X</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>&lt;=30</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>P2</td>
<td>&lt;=30</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>P3</td>
<td>&gt;40</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>P4</td>
<td>31…40</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>P5</td>
<td>31…40</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>P6</td>
<td>31…40</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>P7</td>
<td>&gt;40</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>P8</td>
<td>&lt;=30</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>P9</td>
<td>&lt;=30</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>P10</td>
<td>31…40</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>P11</td>
<td>&lt;=30</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>P12</td>
<td>&gt;40</td>
<td>no</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>P13</td>
<td>&gt;40</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>P14</td>
<td>31…40</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
</tbody>
</table>
Class Overlap

- Consider two class case
  - data from D1 and D2 may overlap
    - features = \{age, body temperature\}, classes = \{flu, not-flu\}
  - common in practice
  - not linearly separable
Not linearly separable

- If such linear decision surface does not exist, the data is mapped into a much higher dimensional space (“feature space”) where the separating decision surface is found.
Example: Clinical decision support system

- Classification and prediction
  - Prognosis, diagnosis and treatment planning

![Diagram showing patient, biopsy, gene expression profile, classifier model, and cancer classification.]
Artificial Neural Networks

Biologically realistic?
- Nonlinear summations
- Spike train vs. single output response
- Updated sequentially vs. asynchronously
- Threshold at which they fire changes over time
- Positive and negative weights – excitatory and inhibitory

http://www.scienceclarified.com/scitech/Artificial-Intelligence/Mind-Versus-Metal.html
Feed-Forward Networks

- Each unit receives input only from units in the immediately preceding layer
Feed-Forward Networks

- Predictions are fed forward through the network to classify
Back Propagation Networks

- Estimate edge weights in the reversed direction
NN Learning Algorithms

- Hebbian’s rule [Hebb 1949]: cells that fire together, wire together
  - When a synapse fires from an input neuron, and correlates to the firing of the output neuron, the dendrite connection to the input axon grows stronger.

Pavlov’s experiment

\[ \Delta \theta_{ij} = \alpha x_i x_j \]

http://www.simplypsychology.org/pavlov.html
http://www.thomelston.com/rhythms/
Tumor diagnosis using ANN

- Small, round blue-cell tumors (SRBCTs) of childhood
  - Four distinct categories have similar appearances in tourine histology and make it hard to discriminate
- Accurate tumor diagnosis is essential for identifying effective therapy
- Gene-expression profiling using cDNA microarrays and Artificial Neural Networks (ANNs)

cDNA microarray

- Commonly known as DNA chip or biochip, is a collection of microscopic DNA spots attached to a solid surface
- Measure the expression levels of large numbers of genes simultaneously

https://en.wikipedia.org/wiki/DNA_microarray
Comparison of Typical Gene Pairs in Microarray Data
Matlab Example

• [http://mason.gmu.edu/~qxing/](http://mason.gmu.edu/~qxing/)

• cDNA microarray with 6,567 genes

• 63 training examples
  – Tumor biopsy material

• Filtering for a minimal level of expression
  – 2,308 genes

• Principle Component Analysis (PCA) further reduced the dimensionality
  – 10 dominant PCA components were used
Analysis

1. Filter genes (88 x 2308)
2. Reduce dimensionality PCA (88 x 10)
3. Random partition 63 training experiments into 3 groups
   - 1/3
   - 1/3
   - 1/3
4. Select validation group
   - 2/3 training
   - 1/3 validation
5. Train
6. Re-select (x 3)
7. Repartition (x 1250)
8. Rank genes (sensitivity measurement)
9. Minimize number of genes
Evaluation – Confusion Matrix

• A table visualizing performance of a classifier
  – Number of cases predicted to be in each class vs. actual numbers

• Accuracy is not always reliable: 990 positives and 10 negatives
Cancer prediction using ANN

• Matlab example (example_geneCancer.m)
  – Training accuracy, testing accuracy

• Vary the number of reduced genes between 1 and 30.
  – Line 57: nGenes = 30;
  – Do you observe an optimal number?

• Vary the neural network complexity
  – Line 66: net = newff(geneexpTrain_low, targets_train, []);
  – Does the perform get improved with more neurons?
Acknowledgement

- Dr. Laurence Bray

- “Introduction to Data Mining”, Pang-Ning Tan, Michael Steinbach, Vipin Kumar
  http://www-users.cs.umn.edu/~kumar/dmbook/index.php#item4

- Presentation summarized by Marcilio Souto