Unsupervised Analysis: Introduction

About the Instructors

Genevera Allen:

- Rice University - Departments of Statistics, CS, and ECE & Baylor College of Medicine - Neurological Research Institute.
- Research:
  - Graphical Models, Multivariate Analysis, Statistical Machine Learning, Big Data, Neuroscience, Genomics, Data Integration.

  `http://www.stat.rice.edu/~gallen/`
- Fun facts. . .

About the Instructors

Yufeng Liu:

- University of North Carolina, Chapel Hill - Departments of Statistics and Operations Research, Genetics, & Biostatistics.
- Research:
  - Statistical Machine Learning and Data Mining; High-dimensional Data Analysis; Nonparametric Statistics and Functional Estimation; Bioinformatics; Design and Analysis of Experiments.

  `http://www.unc.edu/~yfliu/`
- Fun facts. . .

Statistical Machine Learning

- "Learn" from current data to make predictions about the future.
  Examples?

- Intersection of: Computer Science, Statistics, Applied Math.

Big Data

Big Data - BIG in Volume, Variety and/or Velocity (or Complexity!).

Common Big Data themes in Statistical Learning:

- Big n. Large number of observations.
  - Examples: Internet data, financial transactions, climate data, etc.
- Big p. Large number of features relative to observations. (High-dimensional data).
  - Examples: Medical data - genomics, neuroimaging, medical imaging, etc.

Big Biomedical Data

Examples:

- High-throughput Genomics ("Omics").
  - RNA-sequencing, microarrays, methylation arrays, CGH-arrays, exome sequencing, mass spectrometry, NMR spectroscopy, etc.
- Neuroimaging / neural recordings.
  - MRI, Functional MRI (fMRI), EEG, MEG, DTI, ECoG, PET, etc.
- Electronic Health Records.
- Medical Imaging.

Data Matrix

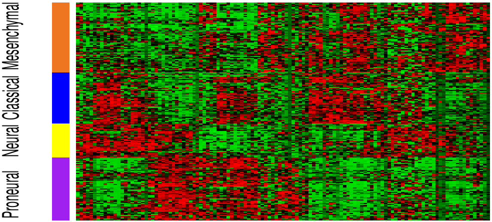
Data Matrix:

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & & \ddots & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

- Rows: n observations / samples / subjects.
- Columns: p features / variables.

Example: Omics Data

Gene Expression Data (Microarray)



- Rows (observations): Subjects ($n \approx 100 - 500$).
- Columns (features): Genes ($p \approx 500 - 20,000$).
- Measurement: Gene expression levels (loosely, how much a gene is turned off or on in a sample).

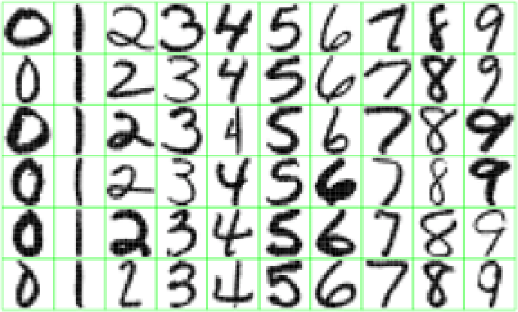
Example: Text Mining

	data	R	big	cluster	shiny	fast	plot
doc 1	57	1	43	2	0	22	4
doc 2	17	29	2	3	35	6	44
doc 3	47	33	0	0	24	3	19
doc 4	23	0	0	31	0	7	2
doc 5	40	5	28	9	0	21	6
doc 6	8	10	7	46	12	17	9

(Bag-of-Words Format)

- Rows (observations): Documents ($n \approx 500 - 100,000$).
- Columns (features): Words ($n \approx 100 - 50,000$).
- Measurement: Count of how many times words appeared in documents.

Example: Image Data



(Handwritten Digits Data)

- Rows (observations): Digits ($n \approx 10,000$).
- Columns (features): Pixels ($p = 256$).
  - Each digit image is converted to a 16×16 grayscale image. The 256 total pixels are vectorized to form the features.
- Measurement: Normalized grayscale intensity of each pixel.

Unsupervised vs. Supervised Learning

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & & \ddots & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

- Rows: n observations / samples / subjects.
- Columns: p features / variables.

Supervised Learning:

$$\boldsymbol{y} = (y_1, y_2, \dots y_n)^T$$

- $\boldsymbol{y}$ - n labels / outcomes associated with each observation.

Unsupervised Learning: No outcomes / labels!

Supervised Learning

Main Goal

Prediction!

- Given: $(Y_n^{train}, \boldsymbol{X}_{n \times p}^{train})$ (Training Data).
- Training: Use training data to find $\hat{f}()$ that maps $\boldsymbol{X}$ to Y: $Y = \hat{f}(\boldsymbol{X}) + \epsilon$.
- Prediction: Given new $\boldsymbol{X}_{m \times p}^{test}$, predict $Y_{m \times 1}^{test}$: $\hat{Y}^{test} = \hat{f}(\boldsymbol{X}^{test})$.

Examples?

Secondary Goals:

- Feature Selection - What features are associated with the outcome?
- Others?

Unsupervised Learning

No labels! What is the goal?

Main Goal

Find some structure that characterizes the data.

(Or, find structure in training data that we expect to be present in future data.)

- Find patterns. (PCA, ICA, NMF, MDS)
- Dimension reduction. (PCA)
- Group observations / Group features / Group both. (Clustering)
- Find associations / relationships between features or observations. (Individualized Treatment Rules; Graphical or Network Models)
- Filter features. (Association testing)

Unsupervised Learning

Challenges:

- Difficult to validate unsupervised learning results.
- No validation or test labels to measure prediction accuracy.
- What is meaningful structure in data?

Uses:

- Data pre-processing / compression / denoising.
- Exploratory data analysis.
  - Need to use multiple unsupervised learning techniques as each gives slightly different "insights" into data.
- Data visualization.

Unsupervised Learning

How is it used in Big Biomedical Data?

Case Study: BRCA gene expression data.

- Data Visualization.
  - Cluster heatmap, graphical models, MDS, PCA.
- Exploratory Analysis.
  - Clustering / dimension reduction to find cancer subtypes.
- Gene Selection.
  - Large-scale hypothesis testing to find genes associated with subtypes.
- Gene Interactions.
  - Graphical models.

This Course

1. Lecture 1 - Dimension Reduction - PCA.
2. Lecture 2 - Dimension Reduction - PCA, NMF, ICA, MDS, Others.
3. Lab 1 - Dimension Reduction.
4. Lecture 3 - Clustering - Intro and K-means.
5. Lecture 4 - Clustering - Hierarchical, and other techniques.
6. Lab 2 - Clustering.
7. Lecture 5 - Individualized Treatment Rules.
8. Lecture 6 / Lab 3 - Large-Scale Hypothesis Testing.
9. Lecture 7 - Graphical Models.
10. Lecture 8 / Lab 4 - Best Practices & BRCA case study.