

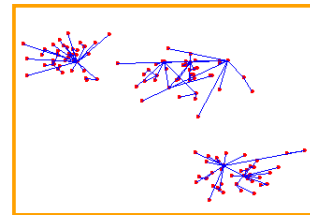
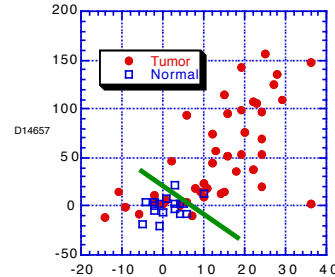
Machine Learning

Robert Stengel

Robotics and Intelligent Systems MAE 345,
Princeton University, 2017

A.K.A.
“Artificial Intelligence”

- **Unsupervised learning**
 - Cluster analysis
 - **Patterns, Clumps, and Joining**
- **Supervised learning**
 - Graph/tree search
 - Hypothesis testing
 - Linear discriminant
 - Nearest neighbor method
- **Estimating classification errors**



Copyright 2017 by Robert Stengel. All rights reserved. For educational use only.
<http://www.princeton.edu/~stengel/MAE345.html>

1

Some Machine Learning Objectives

- Logical Inference
- Classification
- Pattern Recognition
- Image Processing
- System Modeling
- Decision Analysis
- Data Representation
- Linguistic Translation
- “Explainable AI”

2

Old-Fashioned A.I.

- **Expert Systems**
 - **Communication/ Information Theory**
 - **Decision Rules**
 - **Graph and Tree Searches**
 - **Asymmetric Structure**
 - **Explanation Facility**

Trendy A.I.

- **Deep-Learning Neural Networks**
 - **Unsupervised Shallow Networks**
 - **Supervised Shallow Networks**
 - **Back-Propagation**
 - **Associative/ Recurrent Networks**



Explainable Artificial Intelligence?

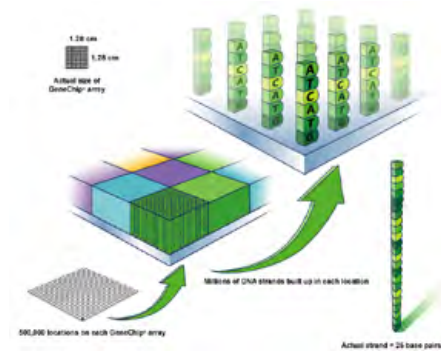
3

Classification Objectives

- **Class comparison**
 - Identify **feature sets** for predefined classes
- **Class prediction**
 - Develop mathematical function/algorithm that predicts class membership in a novel **feature set**
- **Class discovery**
 - Identify **new classes**, sub-classes, or features related to classification objectives

4

DNA Microarray Chip



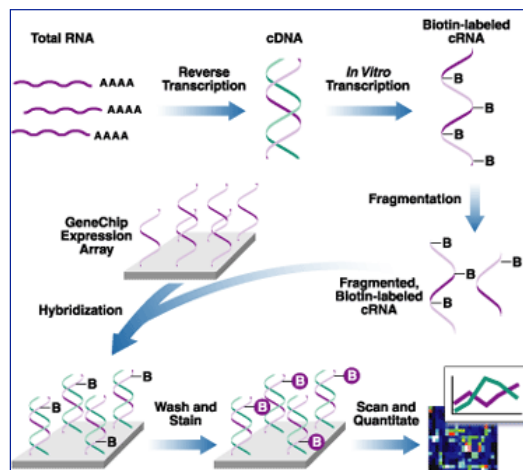
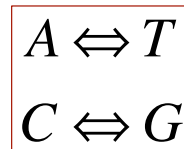
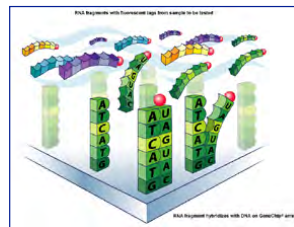
- Glass plate with short strands of synthesized DNA arrayed in spots (probes) across the surface. Typically:
 - A million spots containing different nucleotide sequences
 - Each spot contains 10^6 - 10^7 strands of same sequence
 - 25 nucleotides (base pairs) in each strand
 - Strands are short segments of 20,000 genes
- 10-20 probes (base pairs) per gene

See Supplemental Material for Lecture 15

5

Microarray Processing

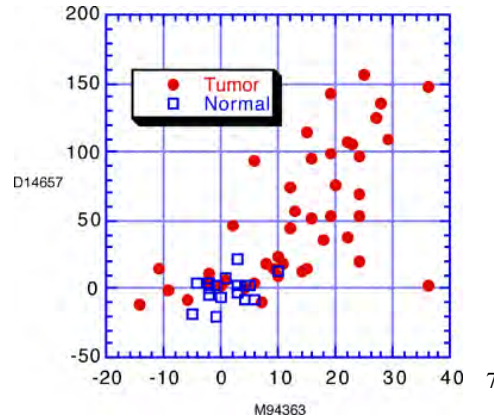
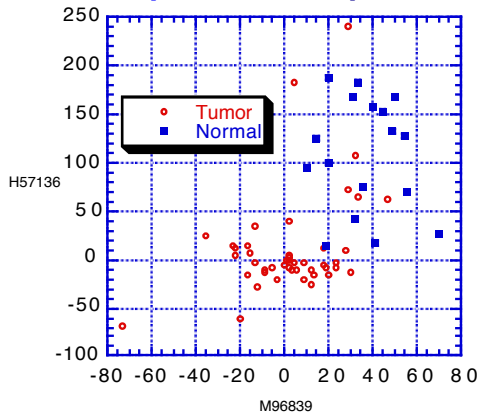
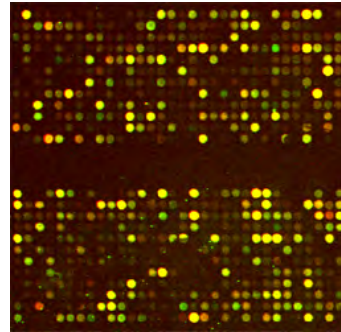
- RNA from biological sample (target) is reverse transcribed to cDNA, transcribed to cRNA, labeled, and hybridized to complementary nucleotides on chip
- Array is washed, stained, and scanned to quantify expression level of genes in sample
- Perfect and mismatched features for each gene in separate probes



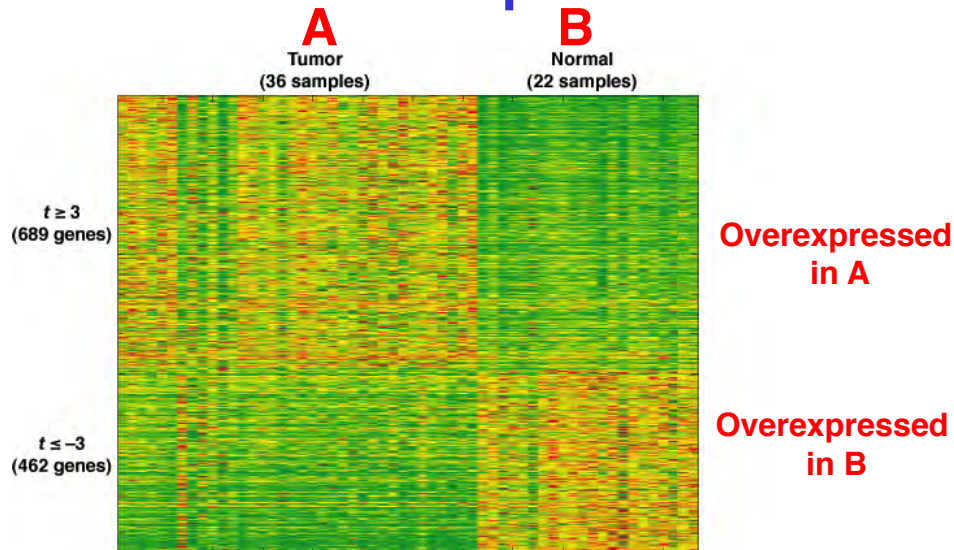
6

Detection of Gene Expression Level in cDNA (from RNA)

- Each tissue sample evaluated by a separate microarray
- Intensity of dot represents over- or under-expression of an RNA gene transcript in the sample



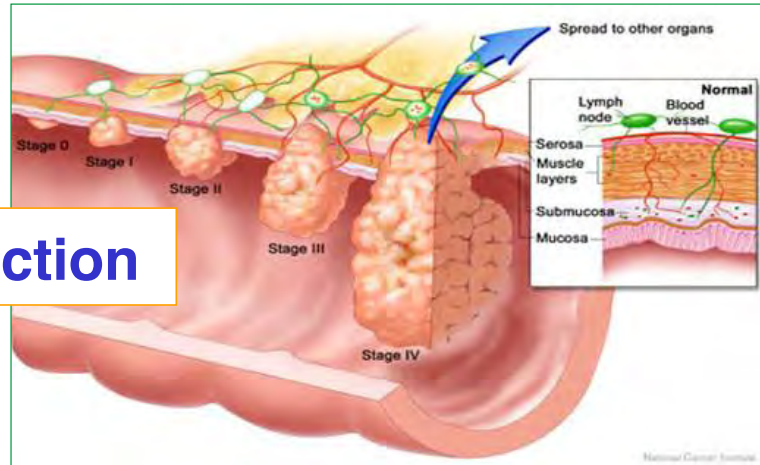
Class Comparison



- Feature sets for predefined classes
 - Group A samples from tumor tissue
 - Group B samples from normal tissue
 - Genes overexpressed in Group A
 - Genes overexpressed in Group B



Class Prediction

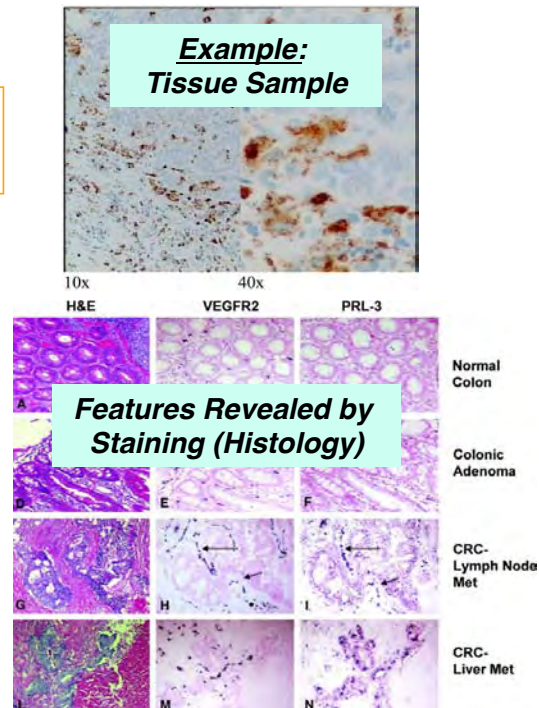


- Algorithm that predicts class membership for a novel feature set
 - Genes of a new sample are analyzed
 - New sample in Group A or Group B?

9

Class Discovery

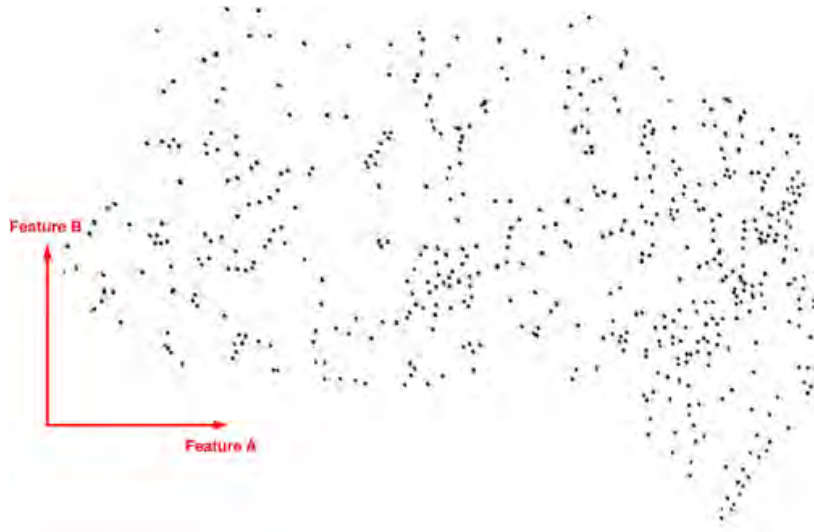
- New features revealed in classification
 - New class in universal set?
 - Novel sample type (e.g., antibody) correlates with group?
 - Novel characteristic (e.g., gender, age, or metastasis) correlates with group?



10

Example for Data Classification

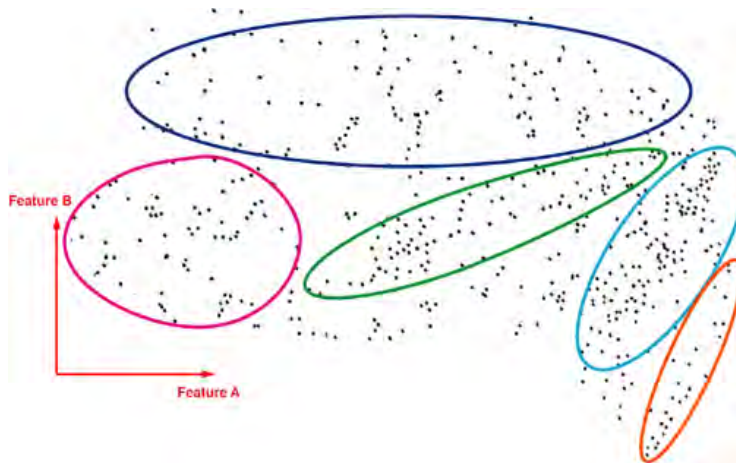
Data set characterized by two features



11

Clustering of Data

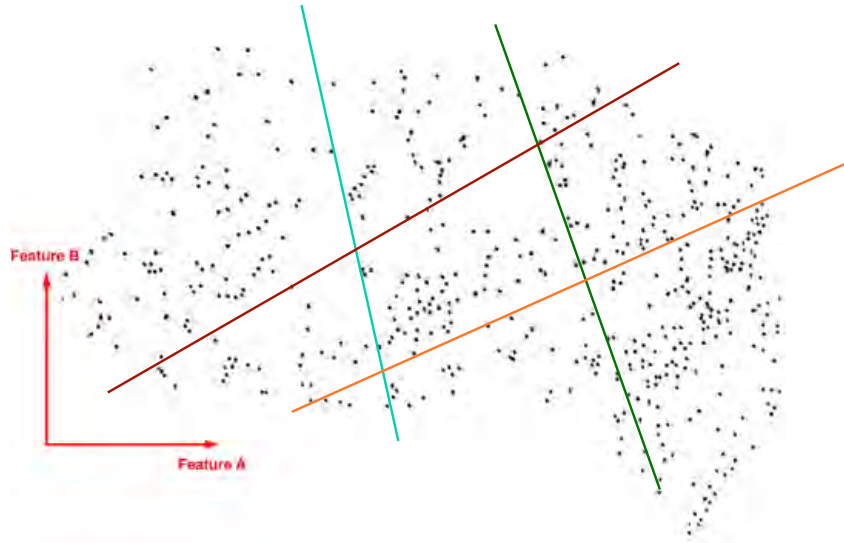
- What characterizes a cluster?
- How many clusters are there?



12

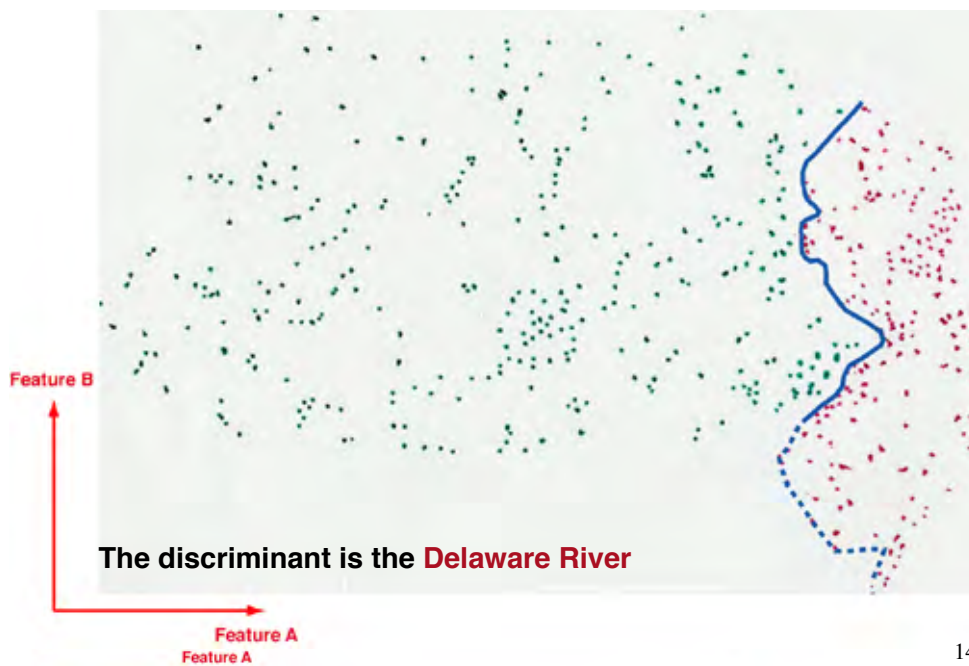
Discriminants of Data

Where are the boundaries between sets?



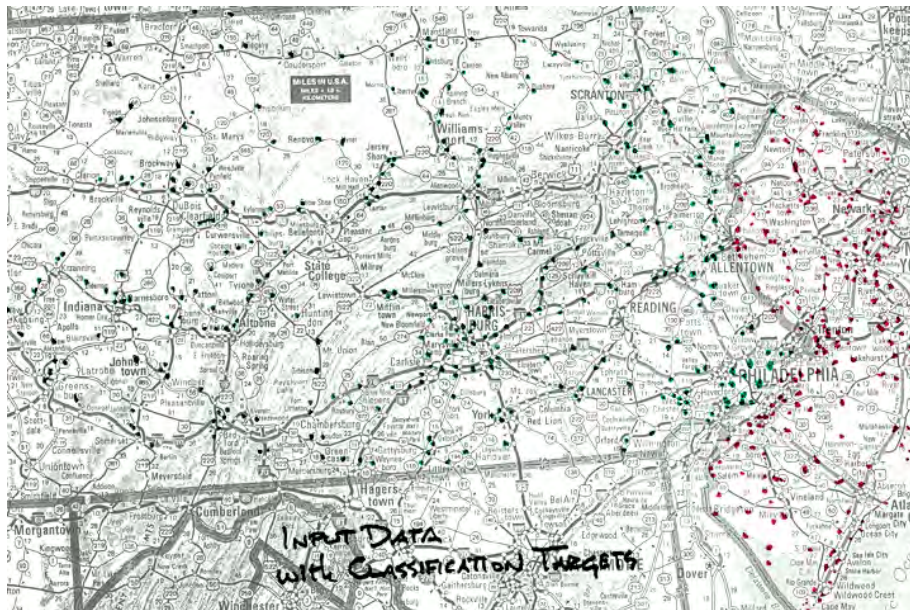
13

The Data Set Revealed



14

Towns and Crossroads of Pennsylvania and New Jersey



15



Choosing Features for Classification

- How many?
- How “strong”?
- Correlation between strong and weak features
- Degree of overlap
- Use of exogenous information for selection
- Statistical significance
- Closeness to boundaries
- To distinguish **New Jersey** from **Pennsylvania**, we could consider
 - Longitude
 - Latitude
 - Altitude
 - Temperature
 - Population
 - # of fast-food stores
 - Cultural factors
 - Zip Code

16

Recall: Membership in a Set

- A = a particular set in U
 - defined in a **list** or **rule**, or a **membership function**
- Universal set = all guests at a party
- Particular sets = distinguishing features of guests

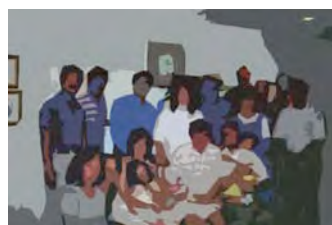


17



Distorted Membership Functions*: Photo

Ambiguity and uncertainty in data sets to be classified

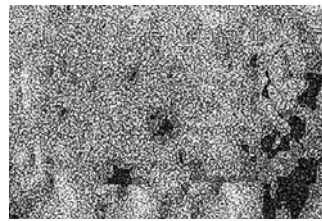
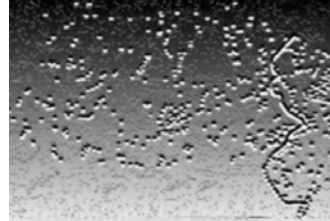
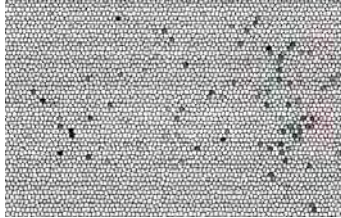


* Photoshop

18



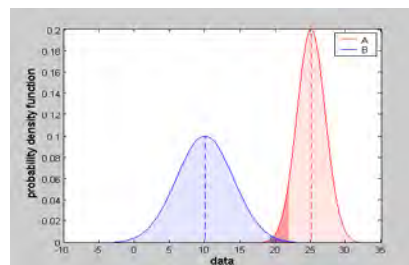
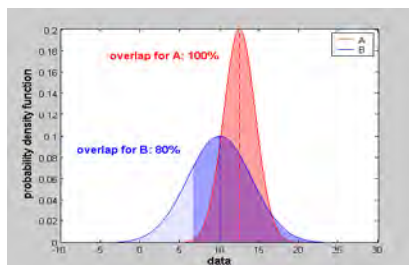
Distorted Membership Functions*: Map



* Photoshop

19

Characteristics of Classification Features

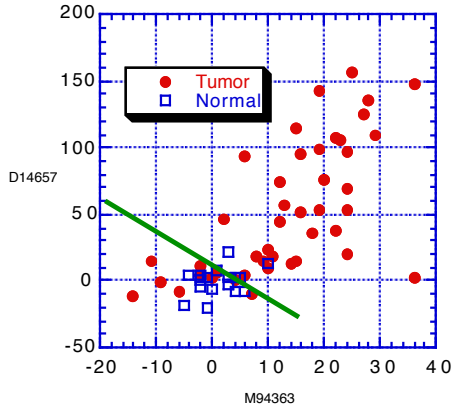


- **Strong feature**
 - Individual feature provides good classification
 - **Minimal overlap** of feature values in each class
 - Significant difference in **class mean values**
 - **Low variance in class**
- **Additional features**
 - **Orthogonal feature** (low correlation) adds **new information** to the set
 - **Co-expressed feature** (high correlation) is **redundant**; **averaging** reduces error

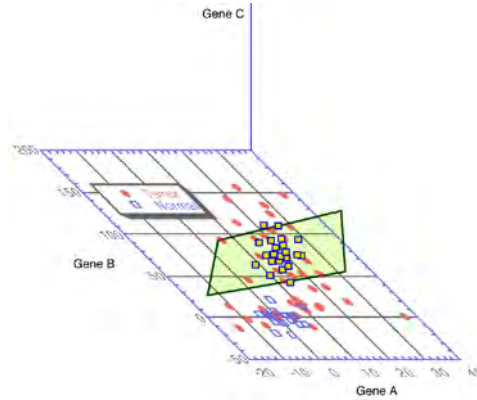
20

Feature Sets

Best **line** or **curve** may classify with significant error

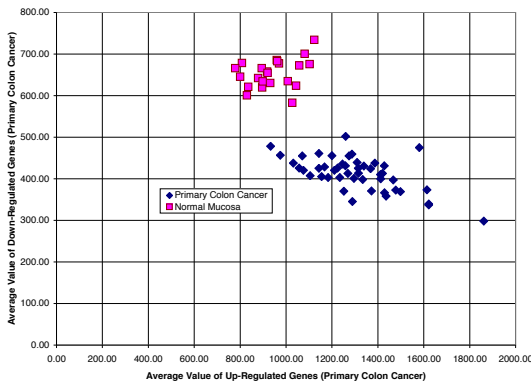


Best **plane** or **surface** classifies with equal or less error

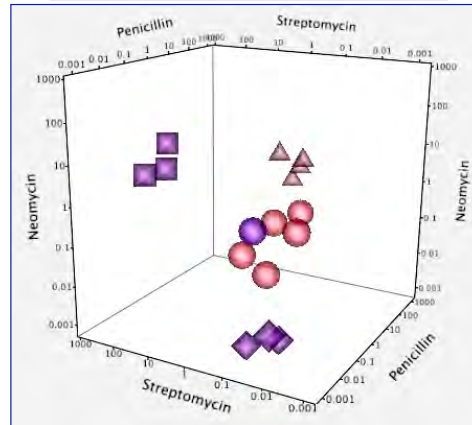


Separable Sets

Gene Analysis (2-D)

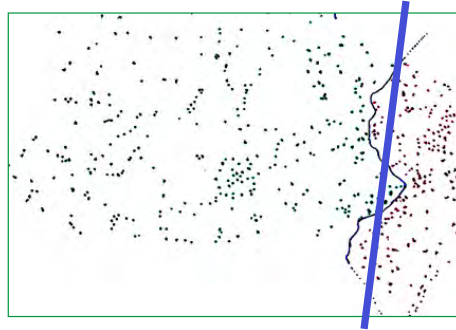


Bacterial Response to Antibiotics (3-D)



Expected Error in Classification

- **Minimum possible error** with statistically optimal discriminant (e.g., Delaware River) **plus**
- **Error due to constraint** imposed by sub-optimal discriminant (e.g., straight vs. curved line) **plus**
- **Error due to sampling** (i.e., number and distribution of points)



23

Errors in Classification

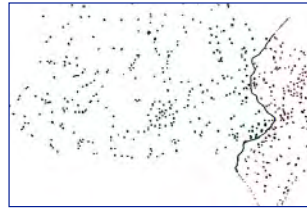
- **Over-/under-fitting**
 - **Excessive/inadequate sensitivity to details** in training data set
 - **Lack of generalization** to novel data



- **Validation**
 - **Train with less** than all available data
 - **Reserve some data** for evaluation of trained classifier
 - **Vary sets** used for training and validation

24

Validation of Classifier



- **Train, Validate, and Test**
- **Reserve some data for evaluation of trained classifier**
- **Train with A, test with B**
 - **A**: Training set (or sample)
 - **B**: Novel set (or sample)
 - Vary sets used for training and validation
- **Leave-one-out validation (combined validation and test)**
 - Remove a single sample
 - Train on remaining samples
 - Does the trained classifier identify the single sample?
 - Repeat with all sets, removing all samples, one-by-one

25

3 x 3 Confusion Matrix

Number of cases predicted to be in each class vs. actual numbers

	True Class		
Predicted Class	Cats	Dogs	Rabbits
Cats	5	2	0
Dogs	3	3	2
Rabbits	0	1	11

- **Interpretation: Actually, there are**
 - **8 cats**: 5 predicted to be cats, 3 to be dogs, and none to be rabbits
 - **6 dogs**: 2 predicted to be cats, 3 to be dogs, and 1 to be rabbit
 - **13 rabbits**: None predicted to be cats, 2 to be dogs, and 11 to be rabbits

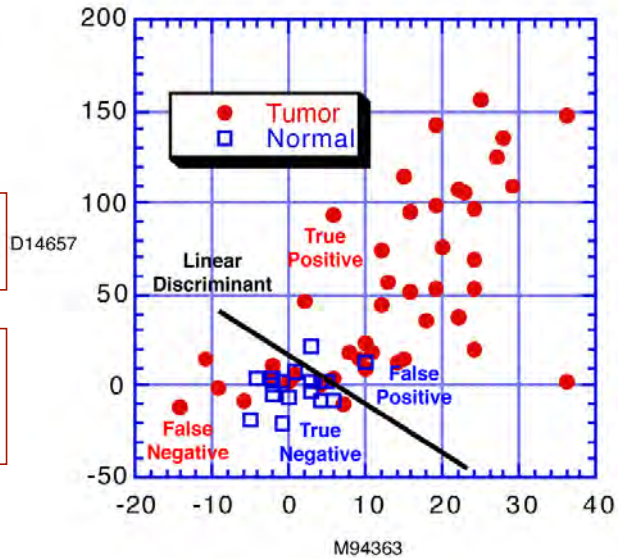
26

Classification of Overlapping Sets

- Tumor = Positive
- Normal = Negative

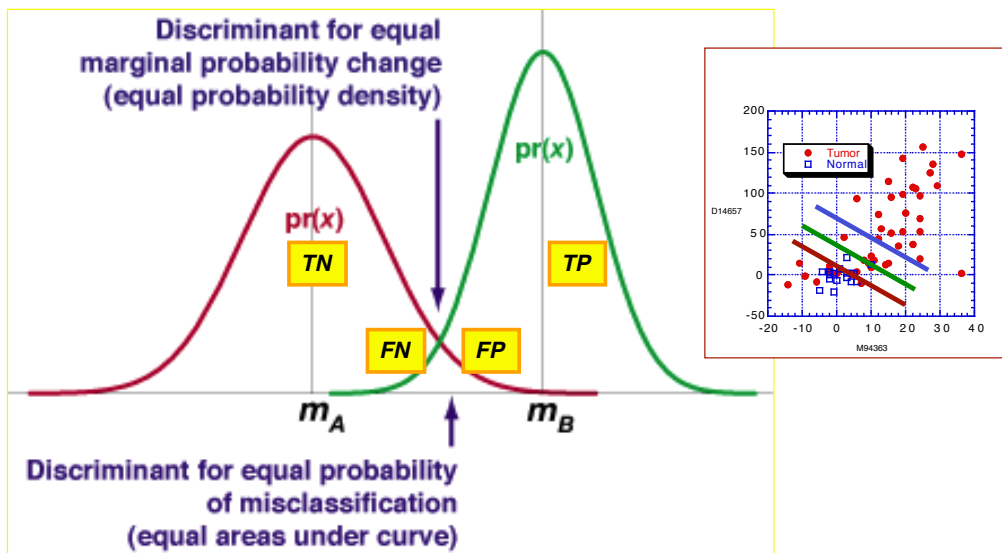
- **Altering discriminant changes classification errors**

- In the example, classification error can never be zero with simple discriminant



27

False Classification May Be Inevitable



28

Categories of Classification Performance

(2 x 2 Confusion Matrix)

Predicted Class	Actual Class		Number in Predicted Class
	Positive	Negative	
Positive	True Positive (TP)	False Positive (FP)	# Predicted(+) = (# TP + # FP)
Negative	False Negative (FN)	True Negative (TN)	# Predicted(-) = (# FN + # TN)
Number in Actual Class	# Actual(+) = (# TP + # FN)	# Actual(-) = (# FP + # TN)	

29

Predicted Class	Actual Class		Number in Predicted Class
	Positive	Negative	
Positive	True Positive (TP)	False Positive (FP)	# Predicted(+) = (# TP + # FP)
Negative	False Negative (FN)	True Negative (TN)	# Predicted(-) = (# FN + # TN)
Number in Actual Class	# Actual(+) = (# TP + # FN)	# Actual(-) = (# FP + # TN)	

Measures of Classification Performance

$$\text{Sensitivity } (\% / 100) = \frac{\# \text{ True Positive}}{\# \text{ Actual Positive}}$$

$$\text{Specificity } (\% / 100) = \frac{\# \text{ True Negative}}{\# \text{ Actual Negative}}$$

30

Measures of Classification Performance

$$\text{Accuracy } (\%/100) = \frac{\# \text{ True Positive} + \# \text{ True Negative}}{\# \text{ Actual Positive} + \# \text{ Actual Negative}}$$

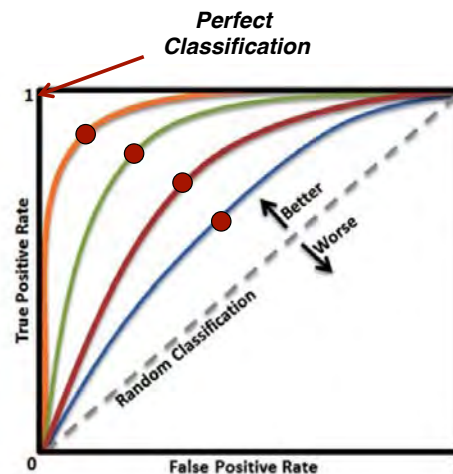
$$\text{Positive Predictive Value } (\%/100) = \frac{\# \text{ True Positive}}{\# \text{ Predicted Positive}}$$

$$\text{Negative Predictive Value } (\%/100) = \frac{\# \text{ True Negative}}{\# \text{ Predicted Negative}}$$

31

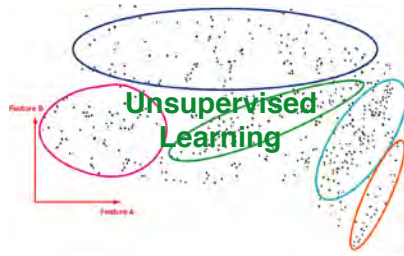
Receiver Operating Characteristic* (ROC) Curve

- Comparison of 4 discriminant functions
- **True positive rate (Sensitivity)** vs. **false positive rate (1 – Specificity)** for a varying parameter (or criterion) of the discriminant function
- The more area under the curve, the better the discriminant function [Ideal AUC = 1]
- For a given discriminant function, choose the criterion that is
 - **closest to (0,1)** [perfect classification]
 - **OR farthest from random**



* Devised during WWII to evaluate radar target detection

32

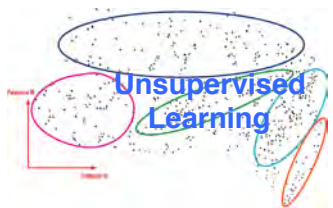


Unsupervised Learning

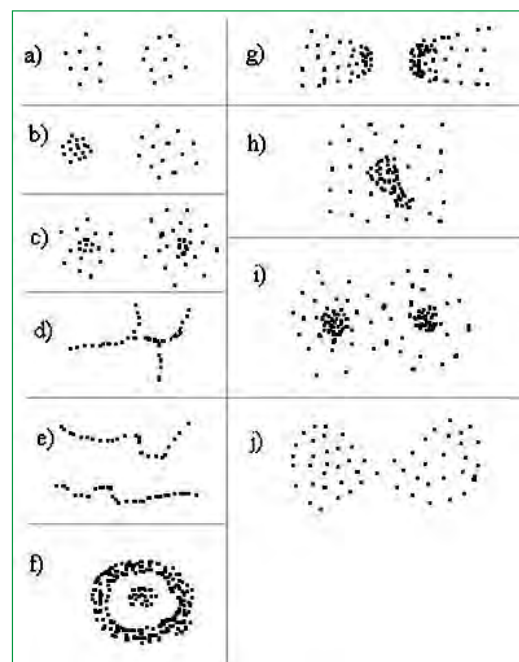
- Learning depends on “closeness” of related features
- Previously unknown correlations or features may be detected
- Meaning of classification revealed after learning via **exogenous knowledge**
- **Same answer given for all questions**

33

Cluster Analysis



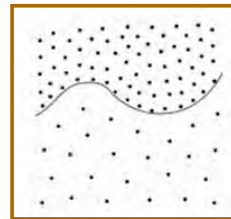
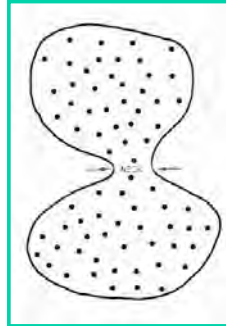
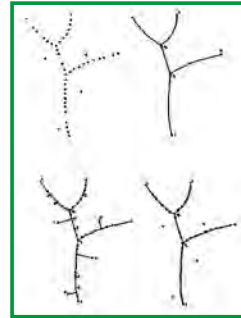
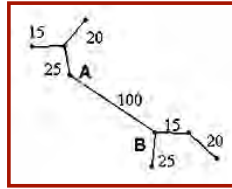
- **Recognize patterns within data sets**
- **Group data points that are close to each other**
 - Hierarchical trees
 - Two-way clustering
 - *k*-means clustering
- **Data compaction**



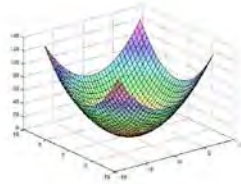
34

Pattern Recognition

- **Minimum spanning tree**
 - Find smallest total edge length
 - Eliminate inconsistent edges (e.g., A-B much longer than others)
 - Delete noisy points (e.g., bubble chamber track at right)
 - Recognize and span gaps
 - Delete necks by diameter comparison
 - Group by similar density
- ... plus other methods of digital image analysis (shapes, edges, ...)



35



Distance Measures Between Data Points

- **Distance between real vectors, \mathbf{x}_1 and \mathbf{x}_2 :**
 - Euclidean distance
 - Weighted Euclidean distance
 - Squared Euclidean distance
 - Manhattan distance
 - Chebychev distance
- **“Distance” between different categories, \mathbf{x}_1 and \mathbf{x}_2 :**
 - Categorical disagreement distance

$$\sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T (\mathbf{x}_1 - \mathbf{x}_2)}$$

$$\sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{Q} (\mathbf{x}_1 - \mathbf{x}_2)}$$

$$(\mathbf{x}_1 - \mathbf{x}_2)^T (\mathbf{x}_1 - \mathbf{x}_2)$$

$$\sum_i |\mathbf{x}_{1_i} - \mathbf{x}_{2_i}|$$

$$\max_i |\mathbf{x}_{1_i} - \mathbf{x}_{2_i}|$$

$$(Number\ of\ \mathbf{x}_{1_i} \neq \mathbf{x}_{2_i}) / i$$

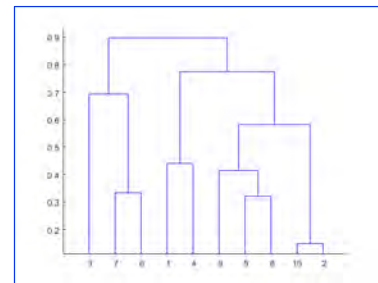
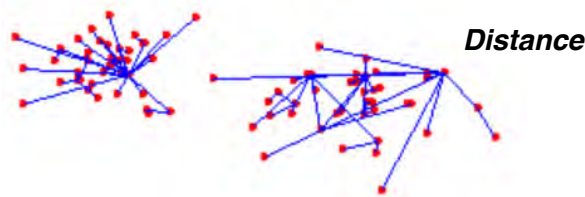


36

Hierarchical Trees (Dendrograms)

- **Classification based on distance between points**
- **Top-down evolution**
 - Begin with 2 best clusters
 - Plot against linkage distance, e.g., distance between centroids
 - Divide each cluster into 2 best clusters until arriving at individuals

$$\text{Cluster Centroid: } \bar{\mathbf{x}} = \frac{\sum_{i=1}^N \mathbf{x}_i}{N}$$

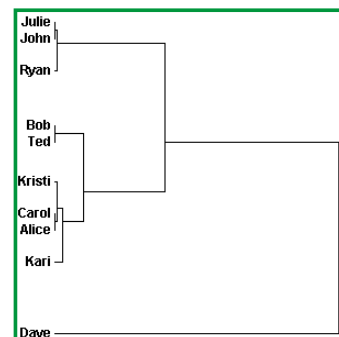
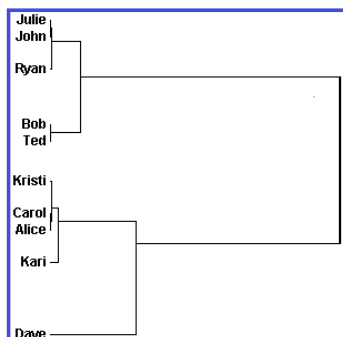
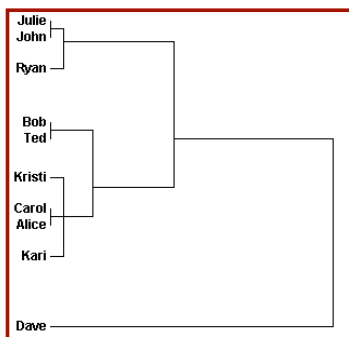


Individuals

37

Hierarchical Trees (Dendrograms)

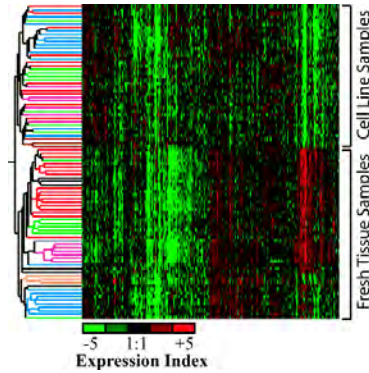
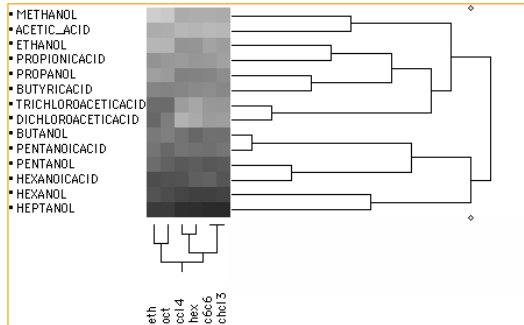
- **Bottom-up evolution**
 - Start with each point in set
 - Link each point to a neighbor
 - **Single linkage:** distance between nearest neighbors in clusters
 - **Complete linkage:** distance between farthest neighbors in clusters
 - **Pair-group average/centroid**
 - Link pairs to closest pairs



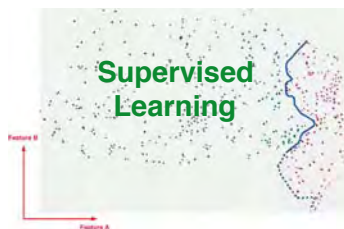
38

Dual Hierarchical Trees

- **Two-way joining**
 - Trees derived from two independent variables
 - Cluster by **feature** and by **sample**
 - Cluster by different components of measurement



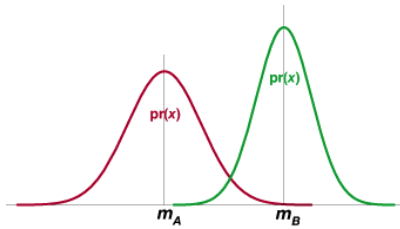
39



Supervised Learning

- Learning depends on prior definition and knowledge of class
- Complex correlation between features is revealed
- Classification is inherent in learning
- **Different answers given for different questions**

40



Simple Hypothesis Test: *t* Test

Is A greater than B?

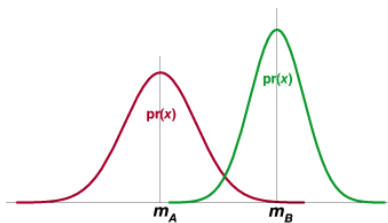
- **Welch's *t* test** compares mean values of two data sets
 - Unequal numbers and variances
 - ***t*** is reduced by uncertainty in the data sets (σ)
 - ***t*** is increased by number of points in the data sets (n)
 - Distributions are not necessarily Gaussian, but classification is based on means and variances

$$t = \frac{(m_A - m_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

- m = mean value of data set
- σ = standard deviation of data set
- n = number of points in data set

- **$|t| > 3$, $m_A \neq m_B$ with $\geq 99.7\%$ confidence** (error probability ≤ 0.003 for Gaussian distributions) [$n > 25$]

41



Analysis of Variance

Variance

$$\sigma_x^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{(N - 1)}$$

F Statistic

$$F_{AB} = \frac{\sigma_{x_1}^2}{\sigma_{x_2}^2} = \frac{\sigma_A^2}{\sigma_B^2}$$

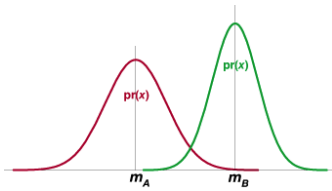
- **F test of two populations**
 - Mean value of secondary importance
 - Populations are **equivalent** if

$$F_{\min} < F_{AB} < F_{\max} \quad \text{or} \quad F_{AB} \approx 1$$

- Populations are **strongly equivalent** if

$$F_{AB} \approx 1 \quad \text{and} \quad t_{AB} \approx 0$$

42



Example of Gene-by-Gene Tumor/Normal Classification by t Test

(Data from Alon *et al*, 1999)

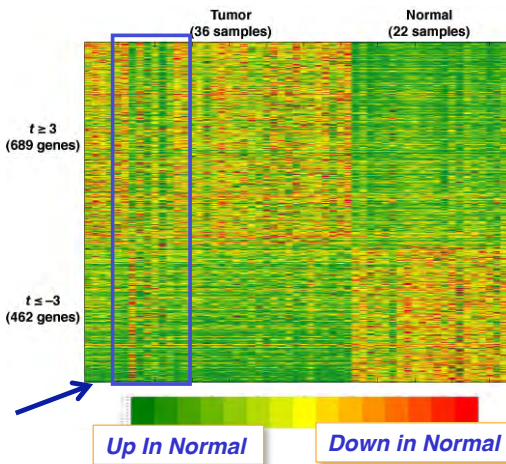
- 58 RNA samples representing tumor and normal tissue
- 1,151 genes are over/under-expressed in tumor/normal comparison, $p \leq 0.003$
- Genetically dissimilar samples are apparent
- *Dimension reduction* by neglecting genes with $|t| < 3$

$$t = (m_T - m_N) / \sqrt{\frac{\sigma_T^2}{36} + \frac{\sigma_N^2}{22}}$$

“Cancer-positive gene sets”

“Cancer-negative gene sets”

Possibly misclassified by pathologist

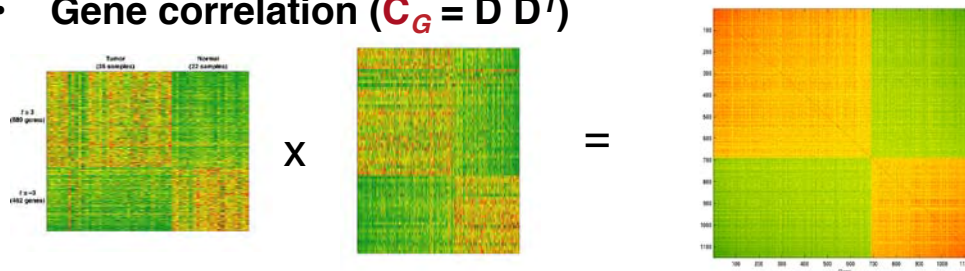


43

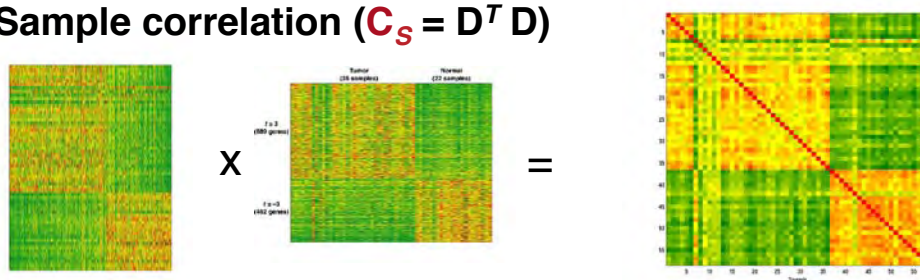
Sample and Gene Correlation Matrices Over Entire Data Set



- Gene correlation ($C_G = D D^T$)



- Sample correlation ($C_S = D^T D$)

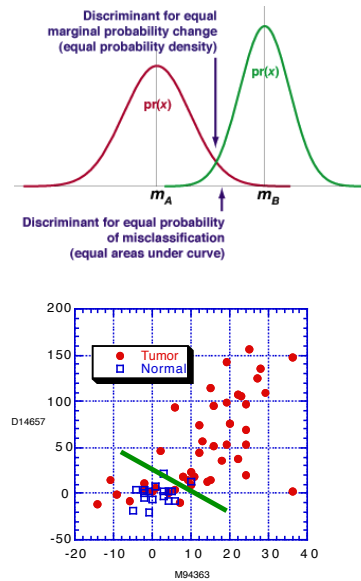


44

Discriminant Analysis

- **Hypothesis test**
 - Are 2 given populations different?
- **Linear discriminant**
 - What is(are) the best line(s)/plane(s)/hyperplane(s) for separating 2 (or k) populations?

$$y = mx + b$$

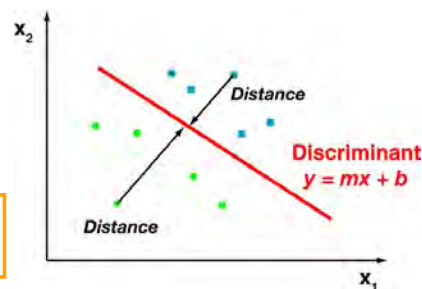


45

Statistical Linear Discriminant

- What is(are) the best line(s)/plane(s)/hyperplane(s) for separating 2 (or k) populations?
 - Fisher's linear discriminant
 - Gradient descent
 - Perceptron
- Nonseparable sets
 - Minimum square error

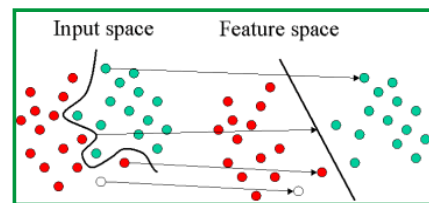
http://en.wikipedia.org/wiki/Linear_discriminant_analysis



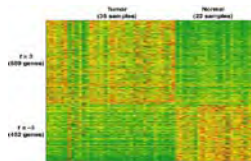
46

Nearest-Neighbor Discriminant Analysis

- Ignore all points except those closest to evolving discriminant
- Support vector machine
 - Linear classifier
 - Reshape the space by transformation (via *kernel*)



47



Ensemble Mean Values Pre-Processing to Reduce Feature Space

- Treat each probe set (row) as a **redundant, corrupted measurement of the same tumor/normal indicator**

$$z_{ij} = k_i y + \varepsilon_{ij}, \quad i = 1, m, \quad j = 1, n$$

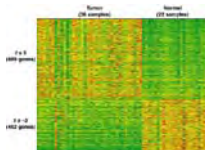
- Compute **column averages for each sample sub-group** (i.e., sum each column and divide by n)

$$\hat{z}_j = \frac{1}{n} \sum_{i=1}^n z_{ij}$$

- **Dimension reduction: Feature space** is reduced from (# samples x # genes) to (# samples)
- Statistics of random variable sums are **~Gaussian by central limit theorem**

A simple step toward deep learning

48



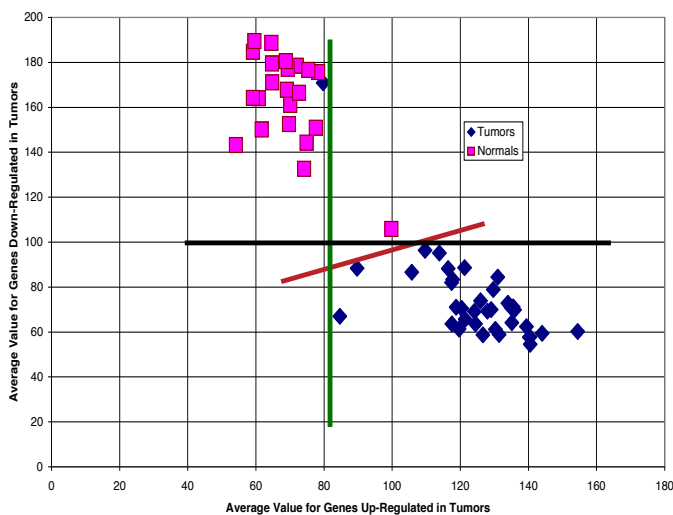
Two-Feature Discriminants for Class Prediction and Evaluation

(Alon, Notterman, 1999, data)

- Scatter plot presents ensemble averages of up genes vs. down genes for each sample

$$\left[\hat{z}_{up_j}, \hat{z}_{down_j} \right]$$

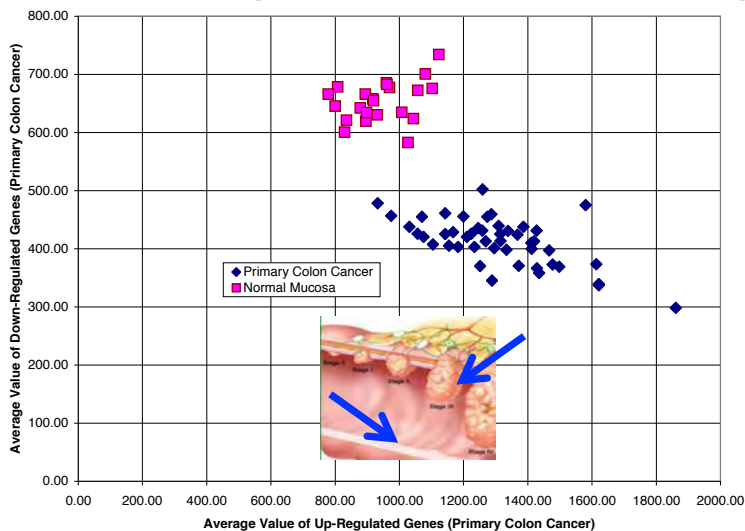
- Classification based on ensemble averages
- Mislabeled samples are identifiable



49

Clustering of Sample Averages for Primary Colon Cancer vs. Normal Mucosa

(NIH PPG data, 2004)

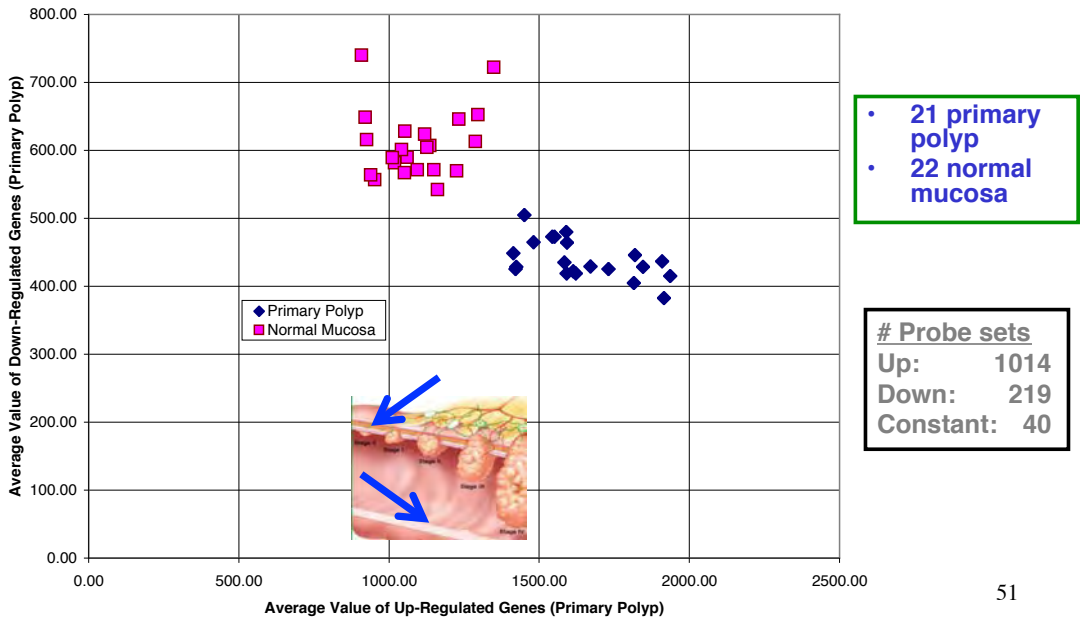


- 144 samples, 3,437 probe sets analyzed
- 47 primary colon cancer
- 22 normal mucosa
- Affymetrix HGU-133A GeneChip
- All transcripts "Present" in all samples

# Probe sets	
Up:	1067
Down:	290
Constant:	19

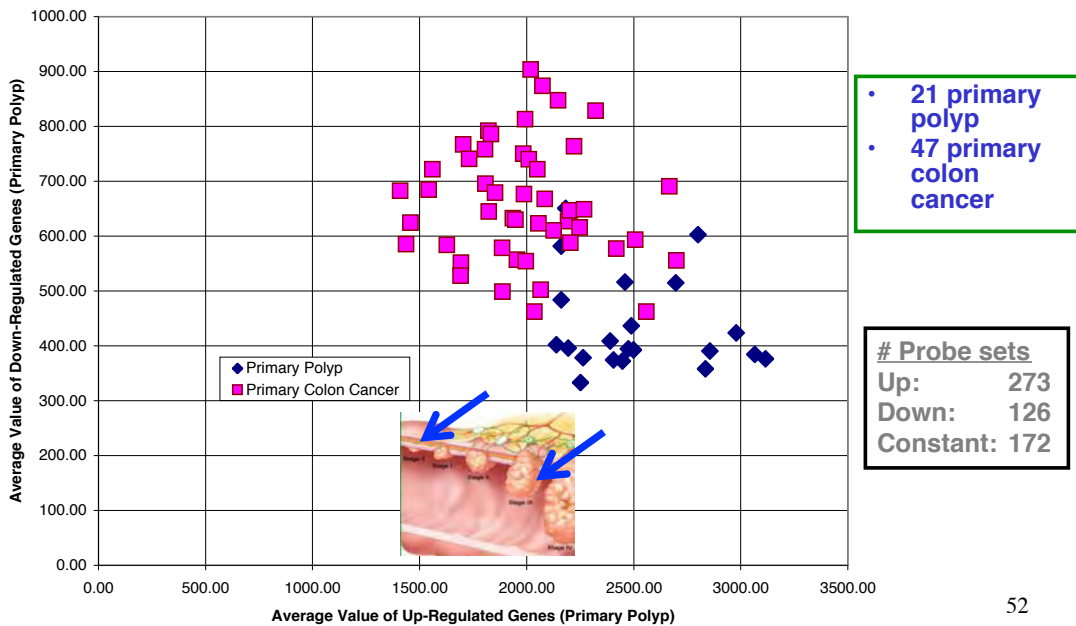
50

Clustering of Sample Averages for Primary Polyp vs. Normal Mucosa



51

Clustering of Sample Averages for Primary Polyp vs. Primary Colon Cancer



52

*Next Time:
Introduction to
Neural Networks*

53

***Supplemental
Material***

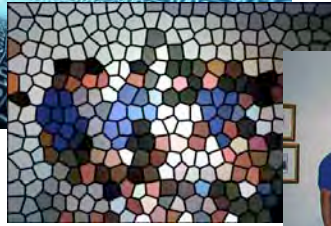
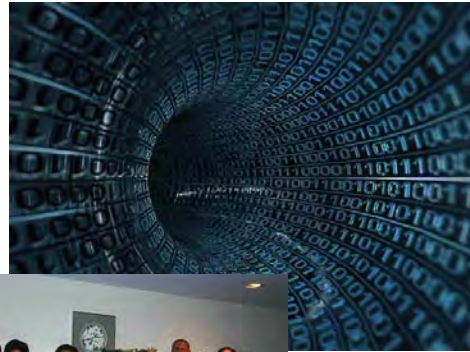
54

“Big Data” and Data Mining

Multi-dimensional classification

Ray of hope ...?

... or infinite harm?



http://en.wikipedia.org/wiki/Big_data