Bioinformatics and Biomarker Discovery
Part 1: Foundations

Limsoon Wong

3 September 2014

NUS
National University of Singapore

Themes of Bioinformatics

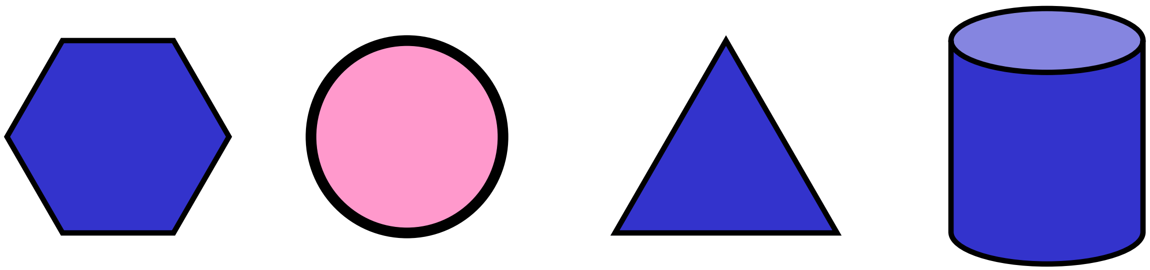
Bioinformatics =
Data Mgmt +
Knowledge Discovery +
Sequence Analysis +
Physical Modeling + ….

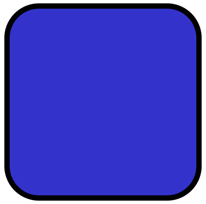
Knowledge Discovery =
Statistics + Algorithms + Databases

Applications include diagnosis, prognosis, & treatment optimization, often thru biomarker discovery

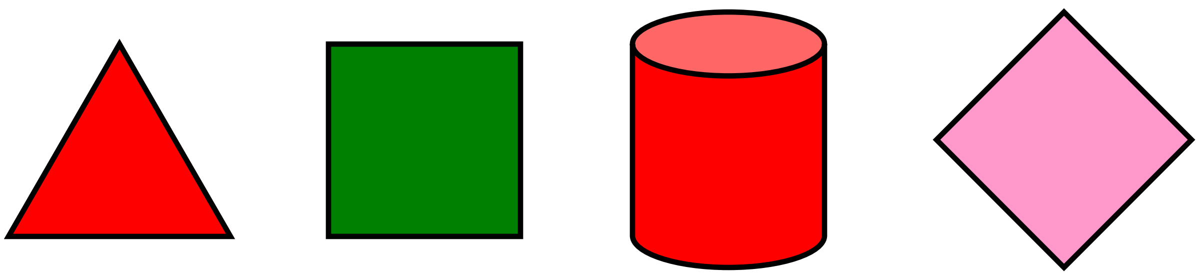
What is Knowledge Discovery?

Jonathan's blocks





Jessica's blocks



Whose block is this?

Jonathan's rules : Blue or Circle
Jessica's rules : All the rest

What is Knowledge Discovery?




Question: Can you explain how?

Key Steps of Knowledge Discovery

- **Training data gathering**

- **Feature generation**
  - k-grams, colour, texture, domain know-how, ...

- **Feature selection**
  - Entropy, $\chi 2$, CFS, t-test, domain know-how...

- **Feature integration**
  - SVM, ANN, PCL, CART, C4.5, kNN, ...

What is Accuracy?

What is Accuracy?

	predicted as positive	predicted as negative
positive	TP	FN
negative	FP	TN

$$\text{Accuracy} = \frac{\text{No. of correct predictions}}{\text{No. of predictions}}$$

$$= \frac{TP + TN}{TP + TN + FP + FN}$$

Examples (Unbalanced Population)

classifier	TP	TN	FP	FN	Accuracy
A	25	75	75	25	50%
B	0	150	0	50	75%
C	50	0	150	0	25%
D	30	100	50	20	65%

- **Clearly, D is better than A**
- **Is B better than A, C, D?**

Exercise: What is B's Prediction strategy?

What is Sensitivity (aka Recall)?

	predicted as positive	predicted as negative
positive	TP	FN
negative	FP	TN

$$\underset{\text{wrt positives}}{\text{Sensitivity}} = \frac{\text{No. of correct positive predictions}}{\text{No. of positives}}$$

$$= \frac{TP}{TP + FN}$$

Sometimes sensitivity wrt negatives is termed **specificity**

What is Precision?

	predicted as positive	predicted as negative
positive	TP	FN
negative	FP	TN

$$\underset{\text{wrt positives}}{\text{Precision}} = \frac{\text{No. of correct positive predictions}}{\text{No. of positives predictions}}$$

$$= \frac{TP}{TP + FP}$$

Unbalanced Population Revisited

classifier	TP	TN	FP	FN	Accuracy	Sensitivity	Precision
A	25	75	75	25	50%	50%	25%
B	0	150	0	50	75%	0%	ND
C	50	0	150	0	25%	100%	25%
D	30	100	50	20	65%	60%	38%

- **What are the sensitivity and precision of B and C?**
- **Is B better than A, C, D?**

Comparing Prediction Performance

- **Accuracy is the obvious measure**
  - But it conveys the right intuition only when the positive and negative populations are roughly equal in size

- **Recall and precision together form a better measure**
  - But what do you do when A has better recall than B and B has better precision than A?

So let us look at some alternate measures ….

Adjusted Accuracy

- **Weigh by the importance of the classes**

Adjusted accuracy = α * Sensitivity + β * Specificity

where $\alpha + \beta = 1$

typically, $\alpha = \beta = 0.5$

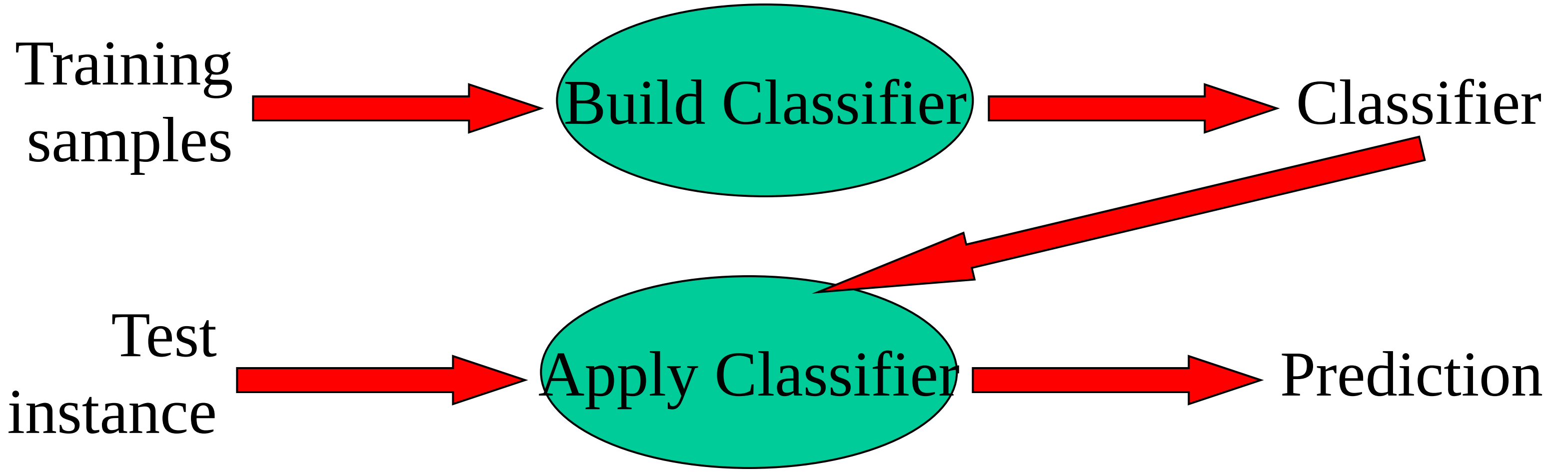
classifier	TP	TN	FP	FN	Accuracy	Adj Accuracy
A	25	75	75	25	50%	50%
B	0	150	0	50	75%	50%
C	50	0	150	0	25%	50%
D	30	100	50	20	65%	63%

But people can't always agree on values for α, β

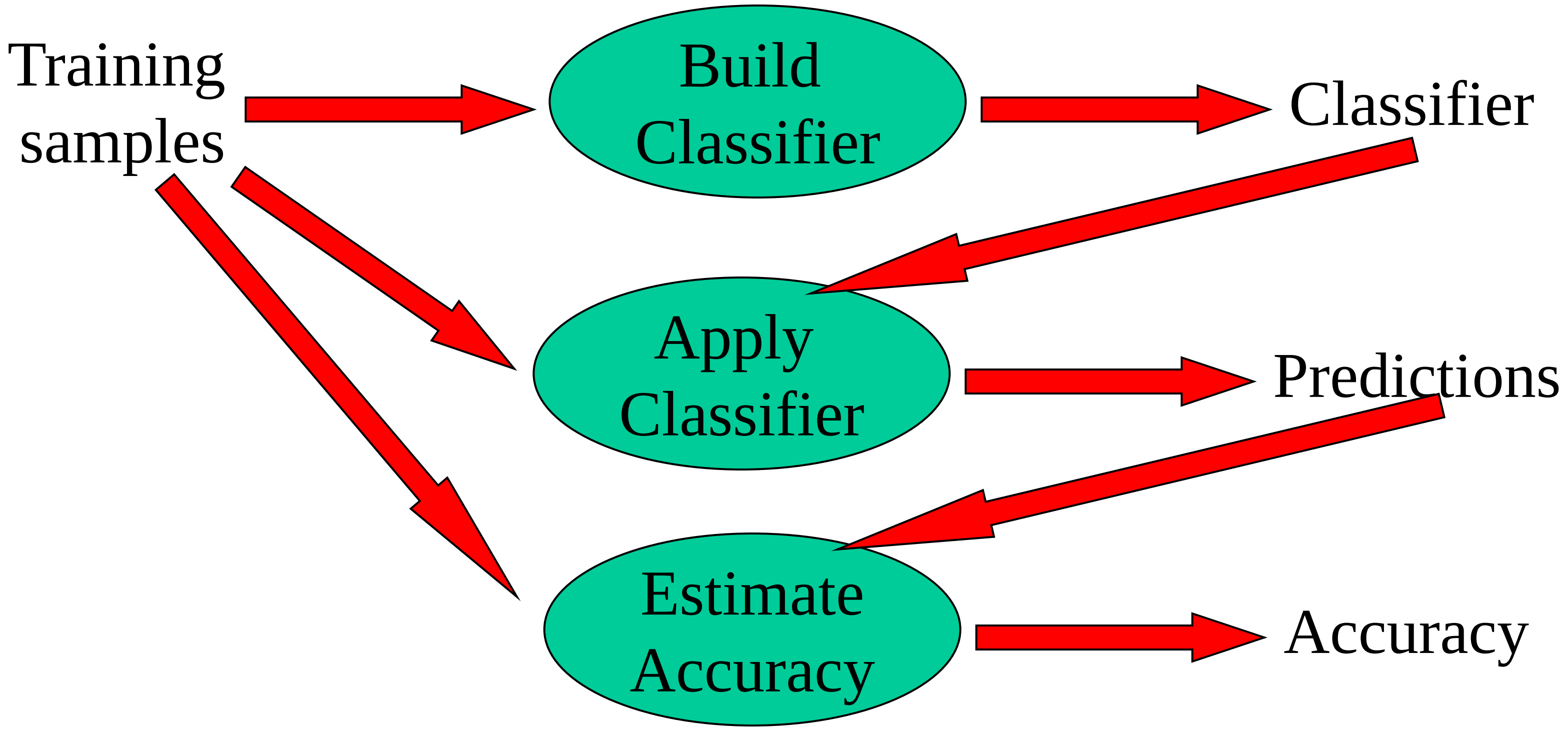
What is Cross Validation?

National University of Singapore

Construction of a Classifier

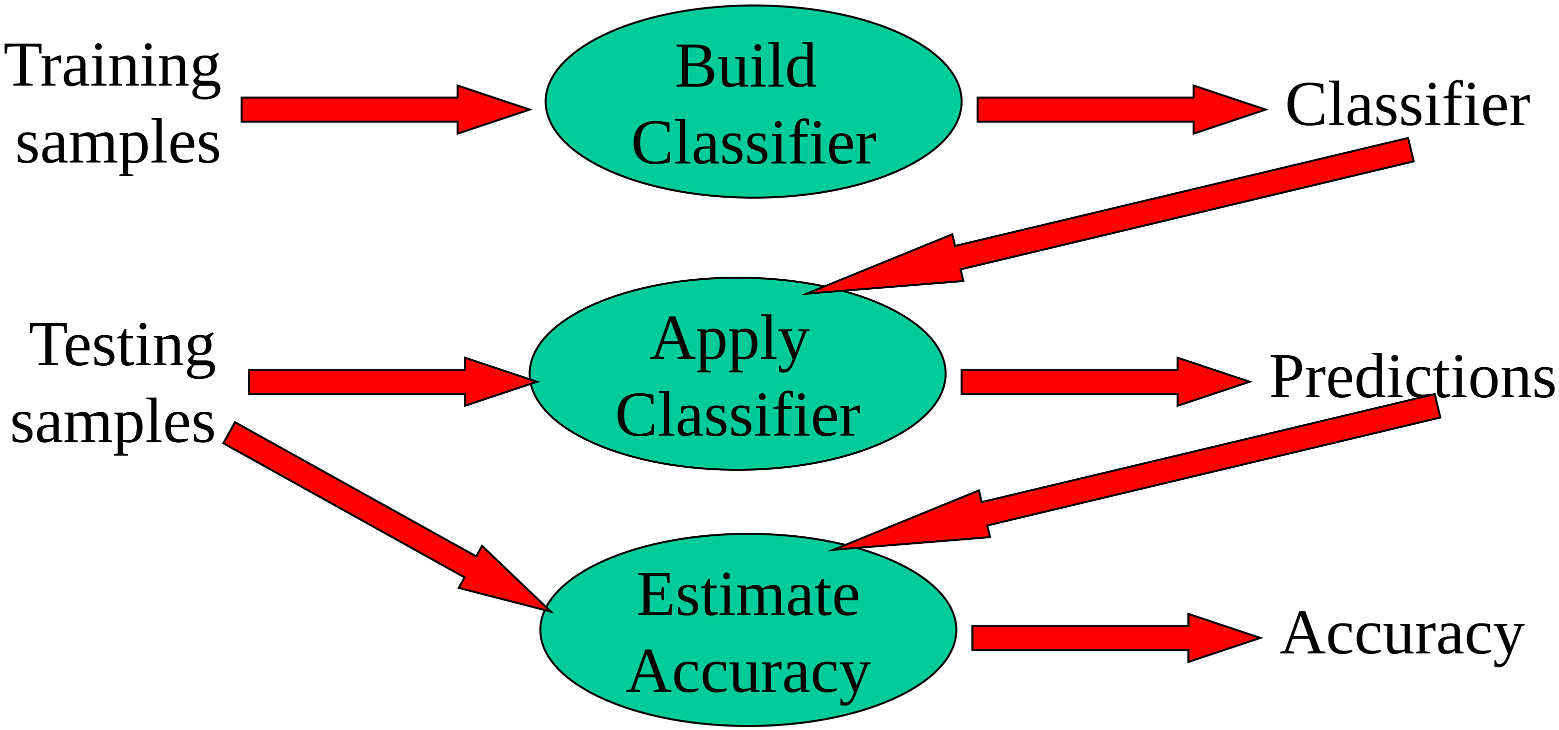


Estimate Accuracy: Wrong Way



Exercise: Why is this way of estimating accuracy wrong?
Think of what will happen in the case of 1-NN classifier.

Estimate Accuracy: Right Way



Testing samples are NOT to be used during “Build Classifier”

Cross Validation

1.Test	2.Train	3.Train	4.Train	5.Train
1.Train	2.Test	3.Train	4.Train	5.Train
1.Train	2.Train	3.Test	4.Train	5.Train
1.Train	2.Train	3.Train	4.Test	5.Train
1.Train	2.Train	3.Train	4.Train	5.Test

- **Divide samples into k roughly equal parts**
- **Each part has similar proportion of samples from different classes**
- **Use each part to test other parts**

What is Feature Selection?

Curse of Dimensionality

- **Given a sample space of p dimensions/features**
- **It is possible that some features are irrelevant**
- **Irrelevant features can confuse a classifier algorithm (or the human analyst!)**
- **Need to find ways to separate those dimensions (aka features) that are relevant (aka signals) from those that are irrelevant (aka noise)**

Signal Selection (Basic Idea)

- **Choose a feature w/ low intra-class distance**
- **Choose a feature w/ high inter-class distance**

Signal Selection (e.g., t-statistics)

The t-stats of a signal is defined as

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

where σ_i^2 is the variance of that signal in class i, μ_i is the mean of that signal in class i, and n_i is the size of class i.

Exercise: Look up other feature selection methods.

Self-fulfilling Oracle

- **Construct artificial dataset with 100 samples, each with 100,000 randomly generated features and randomly assigned class labels**
- **Select 20 features with the best t-statistics (or other methods)**
- **Evaluate accuracy by cross validation using only the 20 selected features**
- **The resultant estimated accuracy can be ~90%**
- **But the true accuracy should be 50%, as the data were derived randomly**

Exercise: What went wrong?

Original photographer unknown/
See also www.cs.gmu.edu/~jessica/DimReducDanger.htm

Concluding Remarks

What have we learned?

- **Methodology of data mining**
  – Feature generation, feature selection, feature integration
- **Evaluation of classifiers**
  – Accuracy, sensitivity, precision
  – Cross validation
- **Curse of dimensionality**
  – Feature selection concept
  – Self-fulfilling oracle

References

- John A. Swets, Measuring the accuracy of diagnostic systems, *Science* 240:1285--1293, June 1988
- Trevor Hastie et al., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2001. Chapters 1, 7
- Lance D. Miller et al., Optimal gene expression analysis by microarrays, *Cancer Cell* 2:353--361, 2002
- David Hand et al., *Principles of Data Mining*, MIT Press, 2001
- Jinyan Li et al., Data Mining Techniques for the Practical Bioinformatician, *The Practical Bioinformatician*, Chapter 3, pages 35—70, WSPC, 2004