Institute for Infocomm Research
A*STAR

CS2220: Computation Foundation in Bioinformatics

Limsoon Wong

Institute for Infocomm Research

Lecture slides for 14 January 2003

Copyright ©2003 limsoon wong

Course Plan

Institute for Infocomm Research
A*STAR

week	date	topics	lecturer
1	7-Jan-03	introduction and background	ken sung
2	14-Jan-03	data mining experiment basics	limsoon wong
3	21-Jan-03	gene feature recognition	limsoon wong
4	28-Jan-03	data mining techniques	jinyan li
5	4-Feb-03	machine learning 1	phil long
6	11-Feb-03	machine learning 2	phil long
7	18-Feb-03	microarray and proteomic data analysis	limsoon wong
8	25-Feb-03	recess	
9	4-Mar-03	term test	
10	11-Mar-03	sequence similarity	ken sung
11	18-Mar-03	suffix tree and applications	ken sung
12	25-Mar-03	physical mapping and sequence assembly	ken sung
13	1-Apr-03	structure prediction and comparison	ken sung
14	8-Apr-03	phylogenetic tree	ken sung

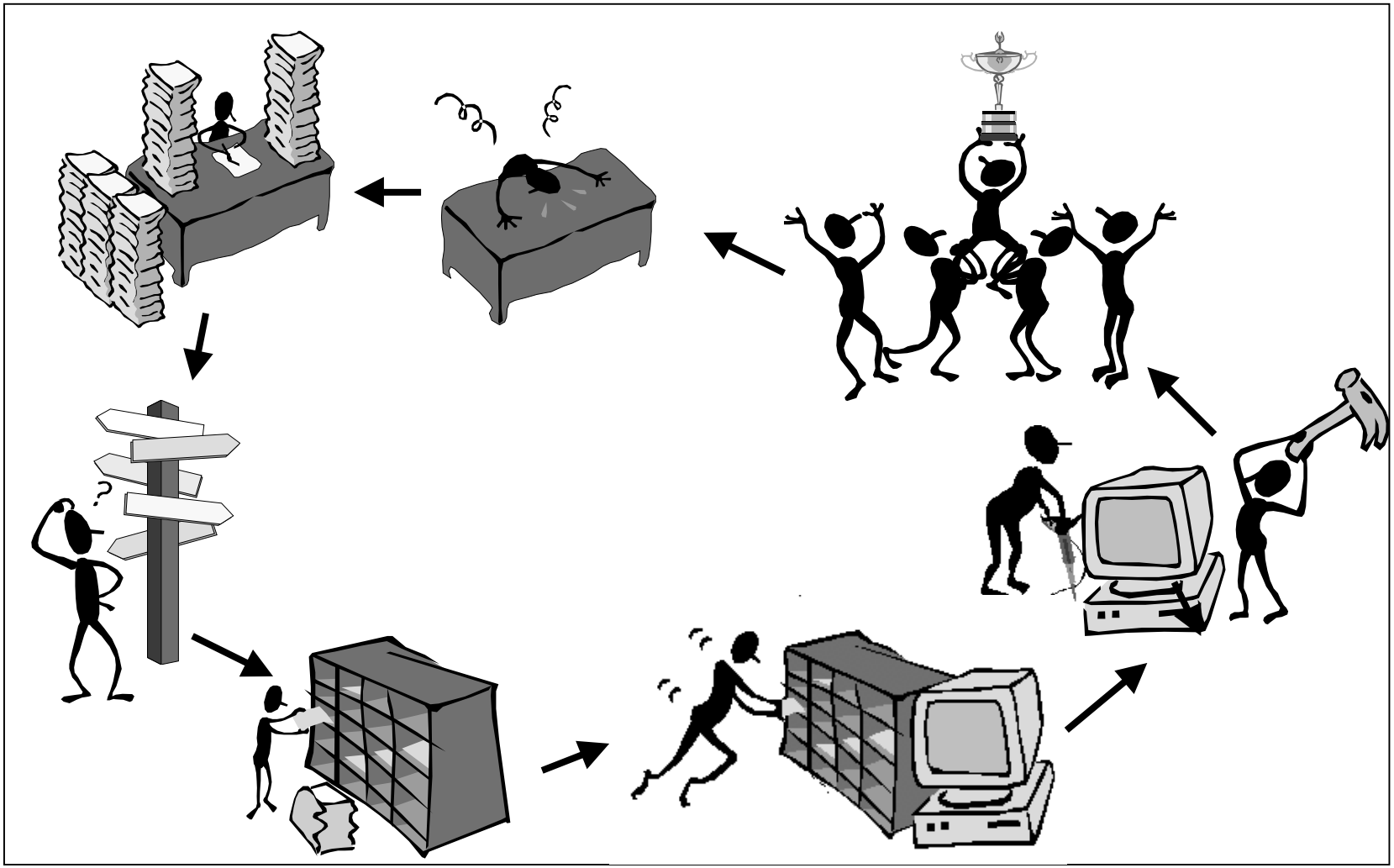
Copyright ©2003 limsoon wong

Institute for Infocomm Research
A*STAR

Essence of Bioinformatics

Copyright ©2003 limsoon wong

What is Bioinformatics?



Copyright ©2003 limsoon wong

Themes of Bioinformatics

Bioinformatics =

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

Knowledge Discovery =

Statistics + Algorithms + Databases

Copyright ©2003 limsoon wong

Benefits of Bioinformatics

To the patient:

Better drug, better treatment

To the pharma:

Save time, save cost, make more $

To the scientist:

Better science

Copyright ©2003 limsoon wong

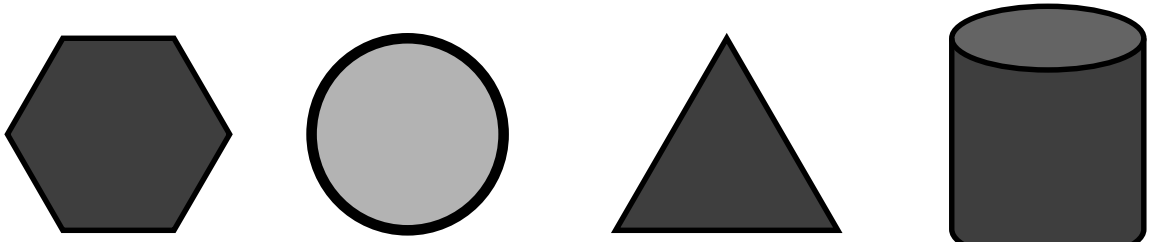
Institute for Infocomm Research
A*STAR

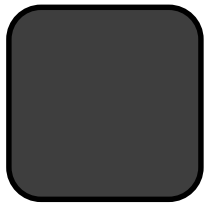
Essence of Knowledge Discovery

Copyright ©2003 limsoon wong

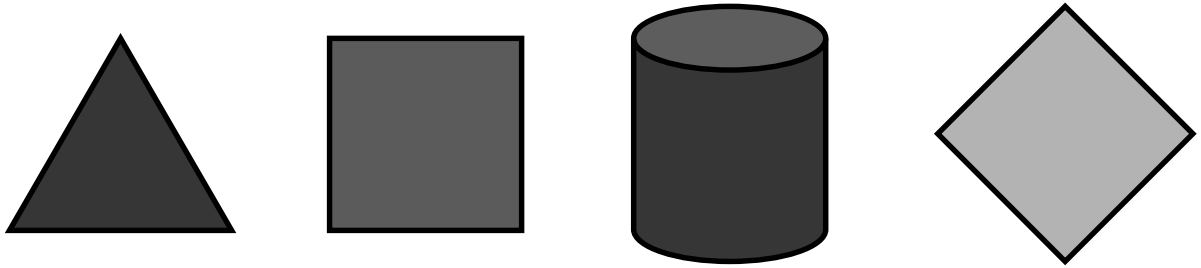
What is Datamining?

Jonathan's blocks





Jessica's blocks

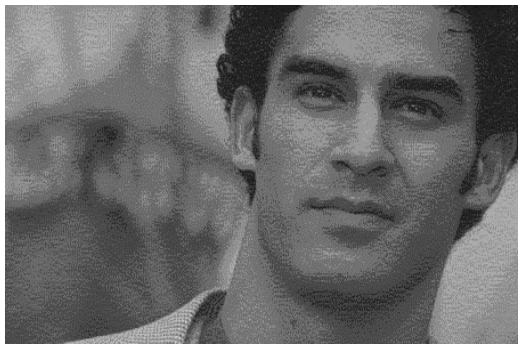


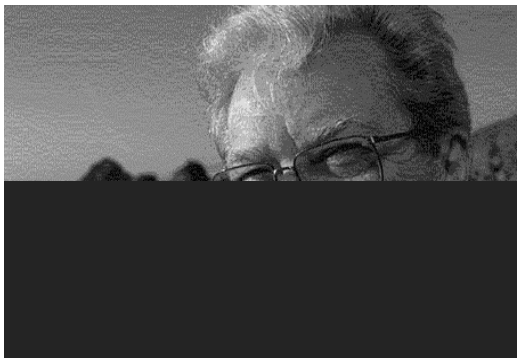
Whose block is this?

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

Copyright ©2003 limsoon wong

What is Datamining?



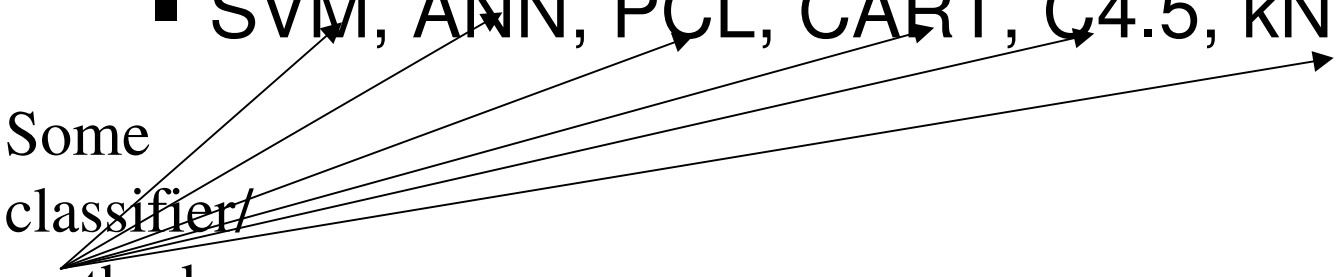


Question: Can you explain how?

Copyright ©2003 limsoon wong

The Steps of Data Mining

- 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, ...



Copyright ©2003 limsoon wong

Institute for Infocomm Research
A*STAR

What is Accuracy?

Copyright ©2003 limsoon wong

What is Accuracy?

Institute for Infocomm Research
A*STAR

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

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

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

Copyright ©2003 limsoon wong

Examples (Balanced Population)

Institute for Infocomm Research
A*STAR

classifier	TP	TN	FP	FN	Accuracy
A	25	25	25	25	50%
B	50	25	25	0	75%
C	25	50	0	25	75%
D	37	37	13	13	74%

- Clearly, B, C, D are all better than A
- Is B better than C, D?
- Is C better than B, D?
- Is D better than B, C?

Accuracy may not tell the whole story

Copyright ©2003 limsoon wong

Examples (Unbalanced Population)

Institute for Infocomm Research

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?

high accuracy is meaningless if population is unbalanced

Copyright ©2003 limsoon wong

What is Sensitivity (aka Recall)?

Institute for Infocomm Research
A*STAR

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

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

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

Sometimes sensitivity wrt negatives is termed **specificity**

Copyright ©2003 limsoon wong

What is Precision?

Institute for Infocomm Research
A*STAR

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

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

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

Copyright ©2003 limsoon wong

Abstract Model of a Classifier

Institute for Infocomm Research
A*STAR

- Given a test sample *S*
- Compute scores *p(S), n(S)*
- Predict *S* as negative if $p(S) < t * n(s)$
- Predict *S* as positive if $p(S) \geq t * n(s)$

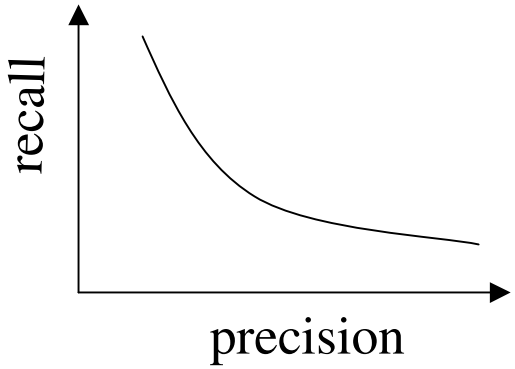
t is the decision threshold of the classifier

changing *t* affects the recall and precision, and hence accuracy, of the classifier

Copyright ©2003 limsoon wong

Precision-Recall Trade-off

- A predicts better than B if A has better recall and precision than B
- There is a trade-off between recall and precision



- In some applications, once you reach a satisfactory precision, you optimize for recall
- In some applications, once you reach a satisfactory recall, you optimize for precision

Copyright ©2003 limsoon wong

Comparing Prediction Performance

Institute for Infocomm Research

- 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 ….

Copyright ©2003 limsoon wong

F-Measure (Used By Info Extraction Folks)

Institute for Infocomm Research
A*STAR

- Take the harmonic mean of recall and precision

$$F = \frac{2 * \text{recall} * \text{precision}}{\text{recall} + \text{precision}} \qquad \text{(wrt positives)}$$

classifier	TP	TN	FP	FN	Accuracy	F-measure
A	25	75	75	25	50%	33%
B	0	150	0	50	75%	undefined
C	50	0	150	0	25%	40%
D	30	100	50	20	65%	46%

Does not accord with intuition

Copyright ©2003 limsoon wong

Adjusted Accuracy

Institute for Infocomm Research
A*STAR

- Weigh by the importance of the classes

$$\text{Adjusted accuracy} = \alpha * \text{Sensitivity} + \beta * \text{Specificity}$$

$$\text{where } \alpha + \beta = 1$$

$$\text{typically, } \alpha = \beta = 0.5$$

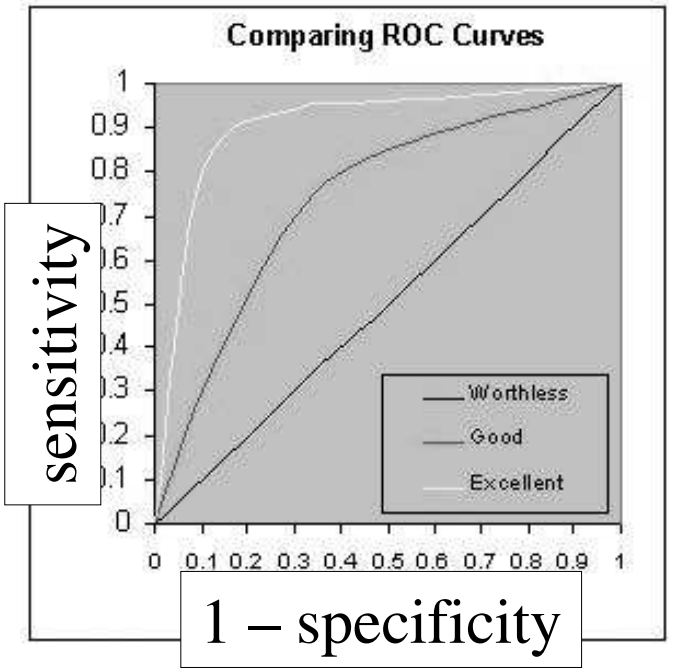
classifier	TP	TN	FP	FN	Accuracy	Adj Accuracy
A	25	75	75	25	50%	50%
B	0	150	0	50	75%	50%
C	50	0	150	0	25%	50%
D	30	100	50	20	65%	63%

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

Copyright ©2003 limsoon wong

ROC Curves

- By changing *t*, we get a range of sensitivities and specificities of a classifier
- A predicts better than B if A has better sensitivities than B at most specificities
- Leads to ROC curve that plots sensitivity vs. (1 – specificity)
- Then the larger the area under the ROC curve, the better



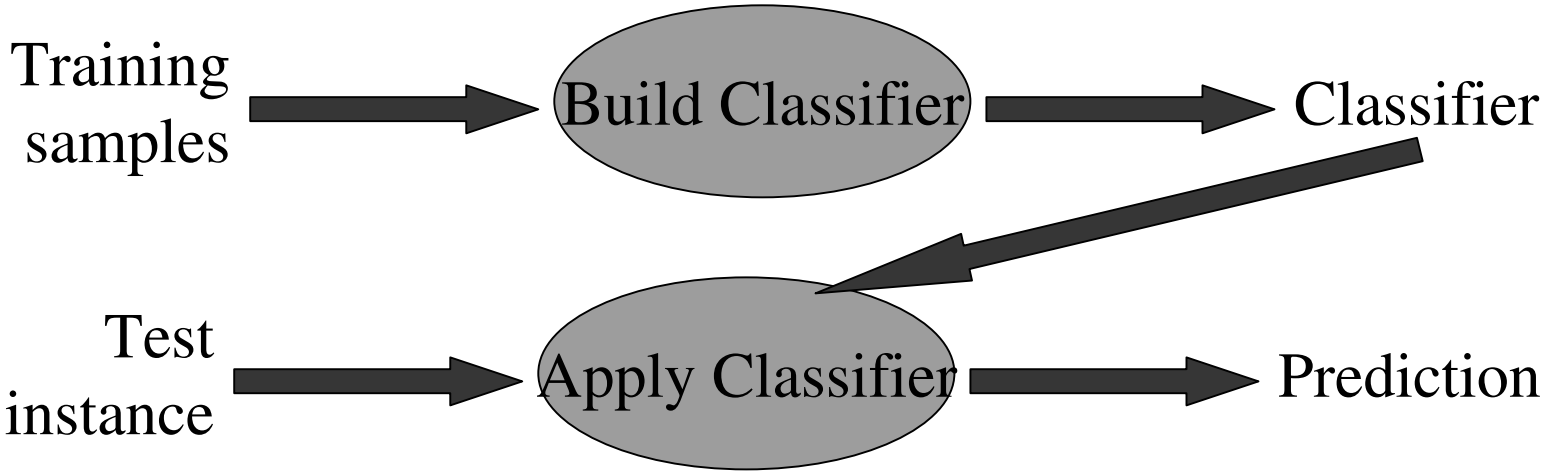
Copyright ©2003 limsoon wong

Institute for Infocomm Research
A*STAR

What is Cross Validation?

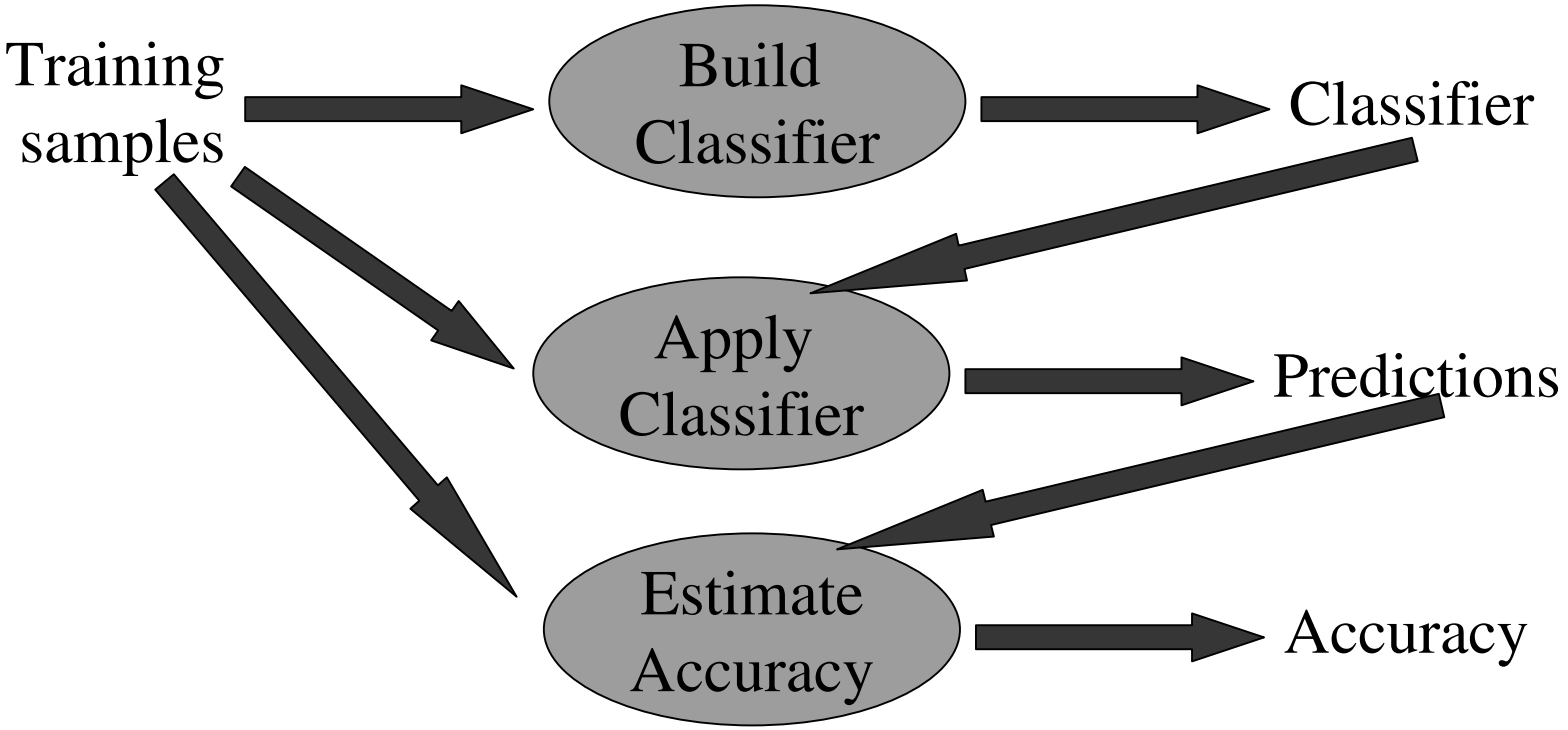
Copyright ©2003 limsoon wong

Construction of a Classifier



Copyright ©2003 limsoon wong

Estimate Accuracy: Wrong Way



Why is this way of estimating accuracy wrong?

Copyright ©2003 limsoon wong

Recall ...

…the abstract model of a classifier

- Given a test sample *S*
- Compute scores *p(S), n(S)*
- Predict *S* as negative if $p(S) < t * n(s)$
- Predict *S* as positive if $p(S) \geq t * n(s)$

t is the decision threshold of the classifier

Copyright ©2003 limsoon wong

K-Nearest Neighbour Classifier (k-NN)

- Given a sample S, find the k observations S_i in the known data that are "closest" to it, and average their responses.
- Assume S is well approximated by its neighbours

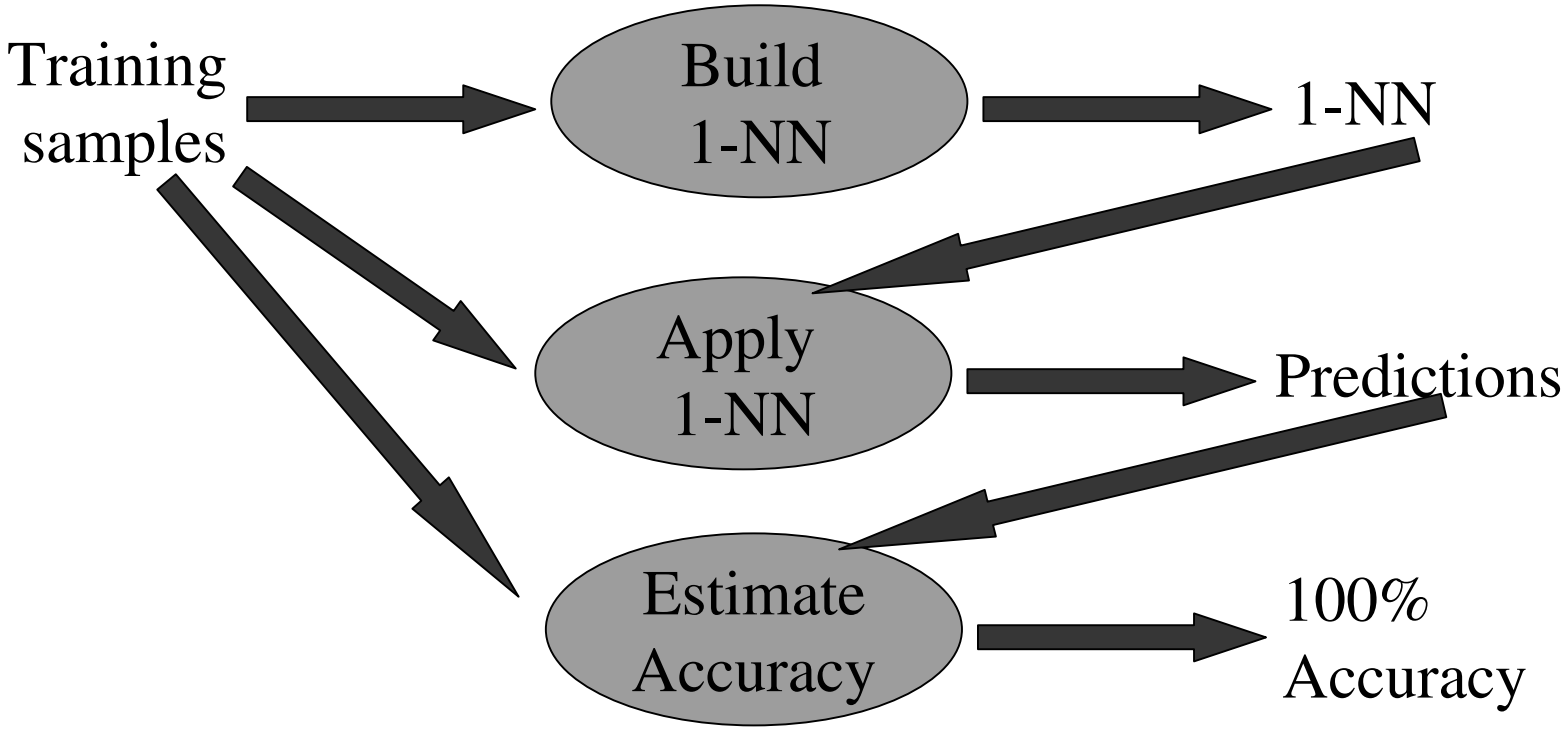
$$p(S) = \sum_{S_i \in N_k(S) \cap D^P} 1 \qquad n(S) = \sum_{S_i \in N_k(S) \cap D^N} 1$$

where $N_k(S)$ is the neighbourhood of S defined by the k nearest samples to it.

Assume distance between samples is Euclidean distance for now

Copyright ©2003 limsoon wong

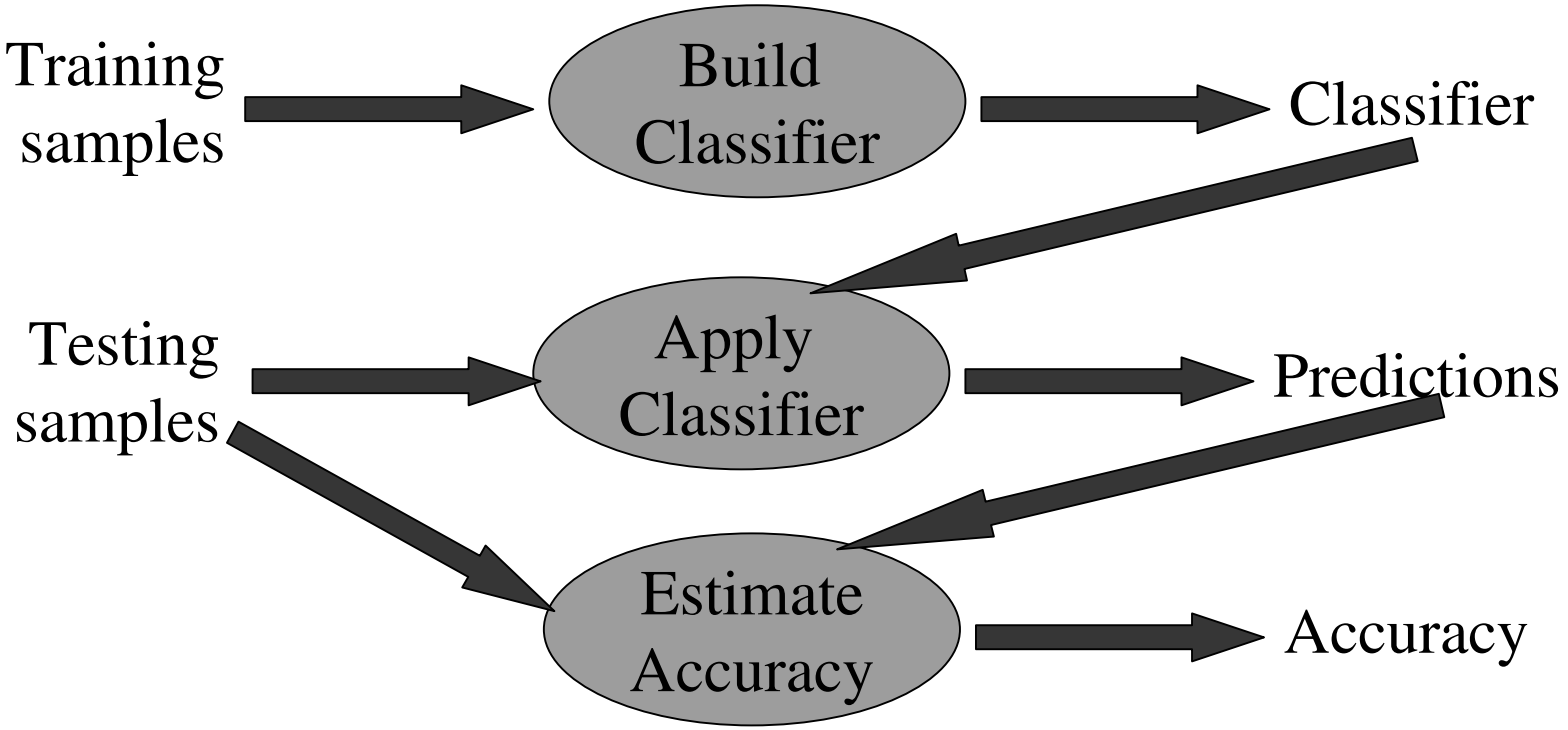
Estimate Accuracy: Wrong Way



For sure k-NN (k = 1) has 100% accuracy in the "accuracy estimation" procedure above. But does this accuracy generalize to new test instances?

Copyright ©2003 limsoon wong

Estimate Accuracy: Right Way



Testing samples are NOT to be used during "Build Classifier"

Copyright ©2003 limsoon wong

How Many Training and Testing Samples?

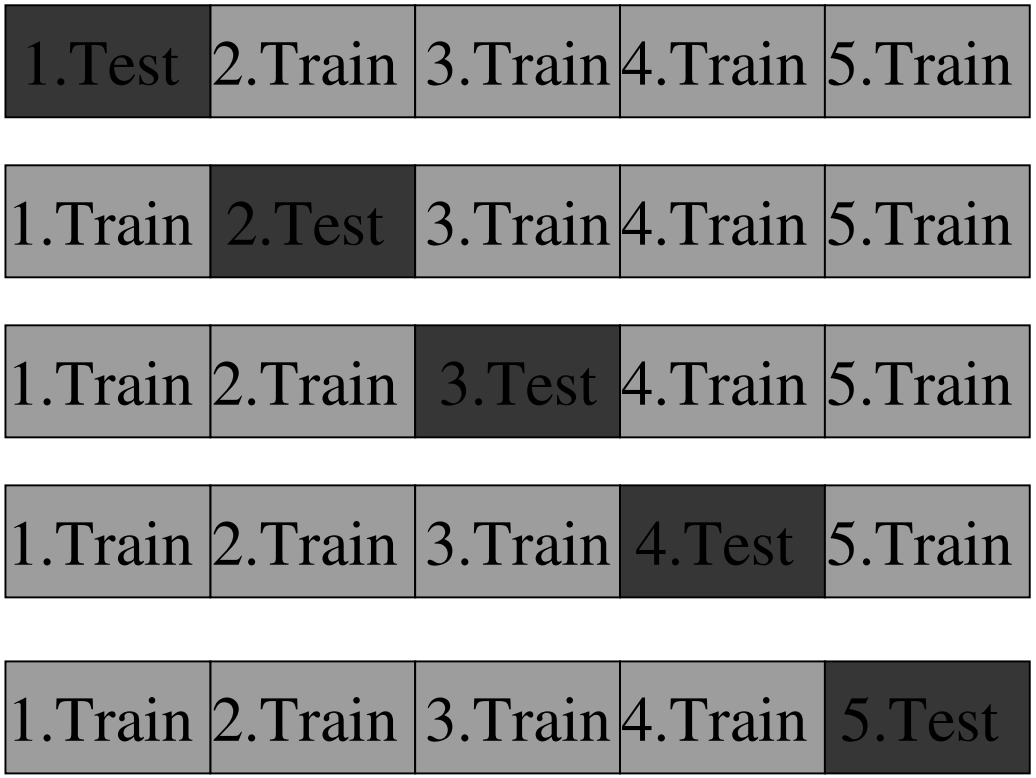
Institute for Infocomm Research
A*STAR

- No fixed ratio between training and testing samples; but typically 2:1 ratio
- Proportion of instances of different classes in testing samples should be similar to proportion in training samples

- What if there are insufficient samples to reserve 1/3 for testing?
- Ans: Cross validation

Copyright ©2003 limsoon wong

Cross Validation

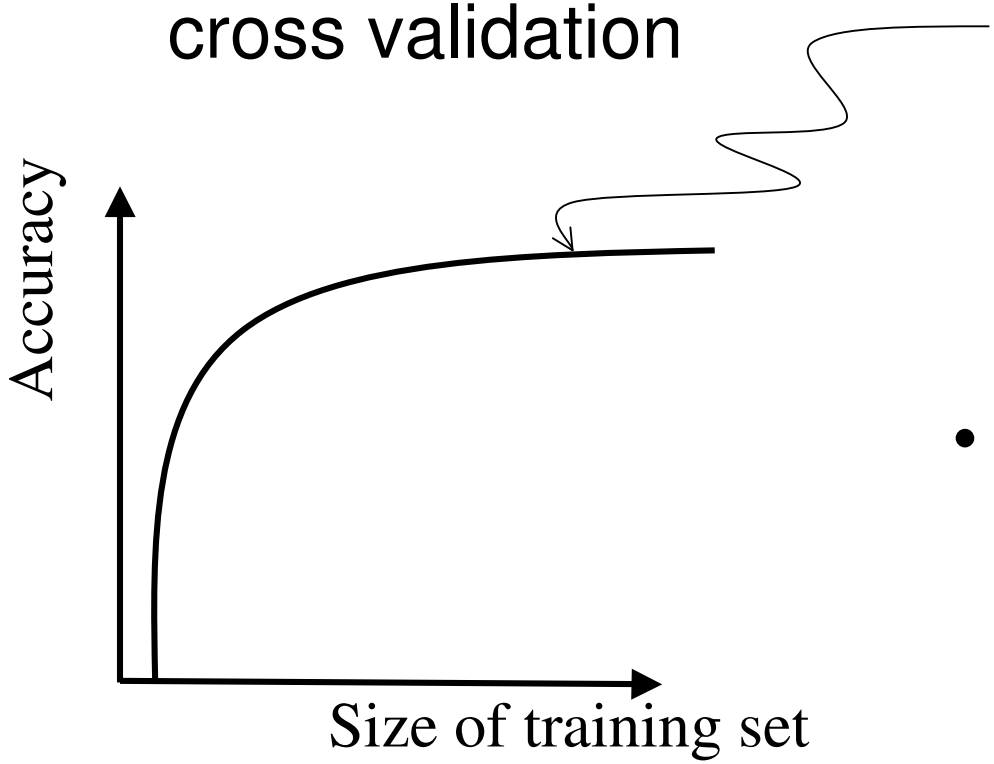


- Divide samples into k roughly equal parts
- Each part has similar proportion of samples from different classes
- Use each part to testing other parts
- Total up accuracy

Copyright ©2003 limsoon wong

How Many Fold?

- If samples are divided into k parts, we call this k-fold cross validation



- Choose k so that
  - the k-fold cross validation accuracy does not change much from k-1 fold
  - each part within the k-fold cross validation has similar accuracy
- k = 5 or 10 are popular choices for k.

Copyright ©2003 limsoon wong

Bias-Variance Decomposition

Institute for Infocomm Research
A*STAR

- Suppose classifiers C_j and C_k were trained on different sets S_j and S_k of 1000 samples each
- Then C_j and C_k might have different accuracy
- What is the expected accuracy of a classifier C trained this way?

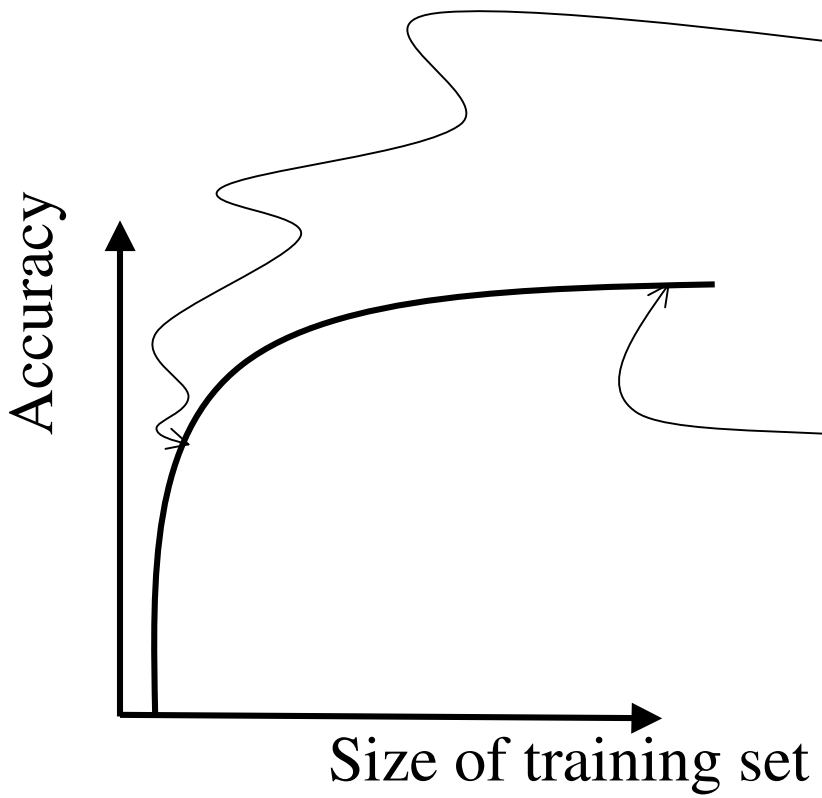
- Let Y = f(X) be what C is trying to predict
- The expected squared error at a test instance x, averaging over all such training samples, is

$$E[f(x) - C(x)]^2$$
$$= E[C(x) - E[C(x)]]^2 \quad \text{(variance)}$$
$$+ [E[C(x)] - f(x)]^2 \quad \text{(bias)}$$

Copyright ©2003 limsoon wong

Bias-Variance Trade-Off

Institute for Infocomm Research
A*STAR



- In k-fold cross validation,
  - small k tends to under estimate accuracy (i.e., large bias downwards)
  - Large k has smaller bias, but can have high variance

Copyright ©2003 limsoon wong

Institute for Infocomm Research
A*STAR

Curse of Dimensionality

Copyright ©2003 limsoon wong

Recall ...

…the abstract model of a classifier

- Given a test sample *S*
- Compute scores *p(S), n(S)*
- Predict *S* as negative if $p(S) < t * n(s)$
- Predict *S* as positive if $p(S) \geq t * n(s)$

t is the decision threshold of the classifier

Copyright ©2003 limsoon wong

K-Nearest Neighbour Classifier (k-NN)

- Given a sample S, find the k observations S_i in the known data that are "closest" to it, and average their responses.
- Assume S is well approximated by its neighbours

$$p(S) = \sum_{S_i \in N_k(S) \cap D^P} 1 \qquad n(S) = \sum_{S_i \in N_k(S) \cap D^N} 1$$

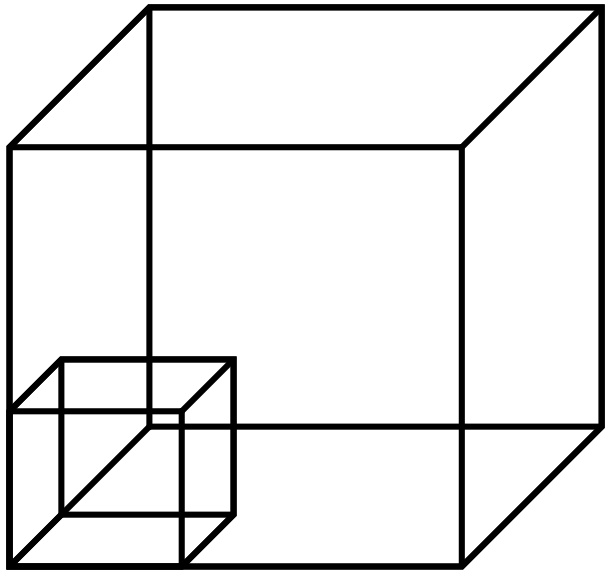
where $N_k(S)$ is the neighbourhood of S defined by the k nearest samples to it.

Assume distance between samples is Euclidean distance for now

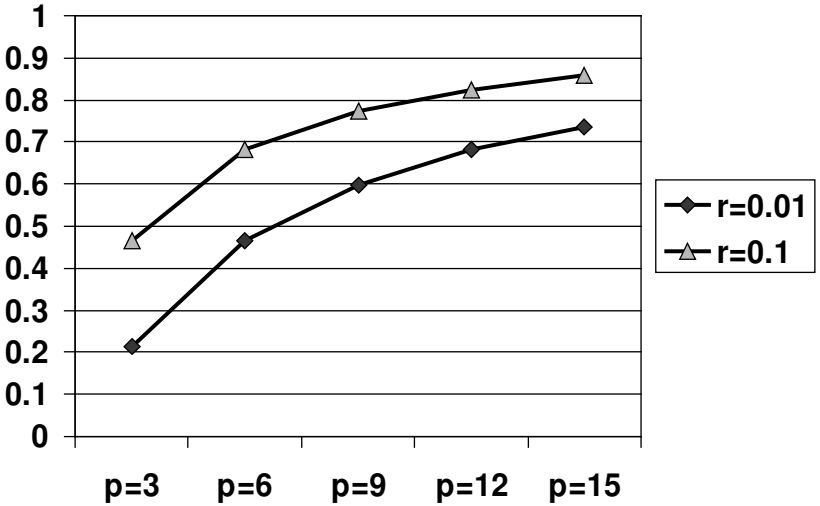
Copyright ©2003 limsoon wong

Curse of Dimensionality

- How much of each dimension is needed to cover a proportion *r* of total sample space?



- Calculate by $e_p(r) = r^{1/p}$
- So, to cover 1% of a 15-D space, need 85% of each dimension!



Copyright ©2003 limsoon wong

Consequence of the Curse

Institute for Infocomm Research
A*STAR

- Suppose the number of samples given to us in the total sample space is fixed
- Let the dimension increase
- Then the distance of the k nearest neighbours of any point increases
- Then the k nearest neighbours are less and less useful for prediction, and can confuse the k-NN classifier

Copyright ©2003 limsoon wong

Institute for Infocomm Research
A*STAR

What is Feature Selection?

Copyright ©2003 limsoon wong

Tackling the Curse

Institute for Infocomm Research
A*STAR

- Given a sample space of p dimensions
- It is possible that some dimensions are irrelevant
- Need to find ways to separate those dimensions (aka features) that are relevant (aka signals) from those that are irrelevant (aka noise)

Copyright ©2003 limsoon wong

Signal Selection (Basic Idea)

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

Copyright ©2003 limsoon wong

Signal Selection (eg., 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.

Copyright ©2003 limsoon wong

Self-fulfilling Oracle

Institute for Infocomm Research
A*STAR

- 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

Copyright ©2003 limsoon wong

What Went Wrong?

Institute for Infocomm Research
A*STAR

- The 20 features were selected from the whole dataset
- Information in the held-out testing samples has thus been “leaked” to the training process

- The correct way is to re-select the 20 features at each fold; better still, use a totally new set of samples for testing

Copyright ©2003 limsoon wong

Notes

Copyright ©2003 limsoon wong

Exercises

Institute for Infocomm Research
A*STAR

- Define specificity, write down its formula
- Draw typical curve for sensitivity vs. specificity
- What area under an ROC curve indicates a classifier is useless? good? great?
- Why does 1-NN have 100% accuracy if the same samples are used for both training and testing?
- Summarise key points of this lecture

Copyright ©2003 limsoon wong

References

- John A. Swets, “Measuring the accuracy of diagnostic systems”, *Science* 240:1285--1293, June 1988
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, *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, November 2002

Copyright ©2003 limsoon wong