For written notes on this lecture, please read chapter 3 of *The Practical Bioinformatician*

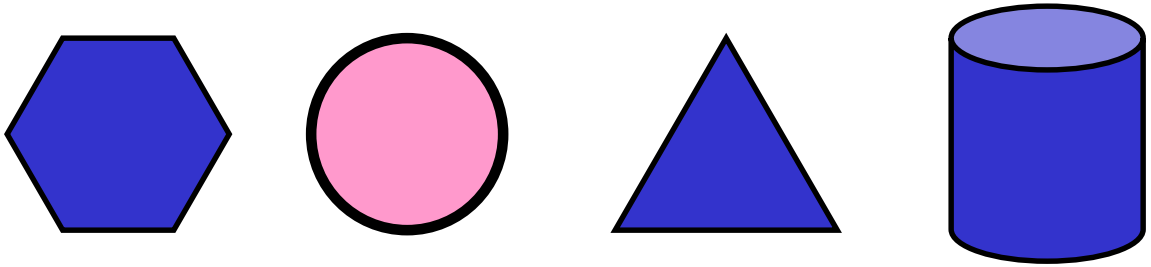
CS2220: Introduction to Computational Biology

Unit 1a: Essence of Knowledge Discovery

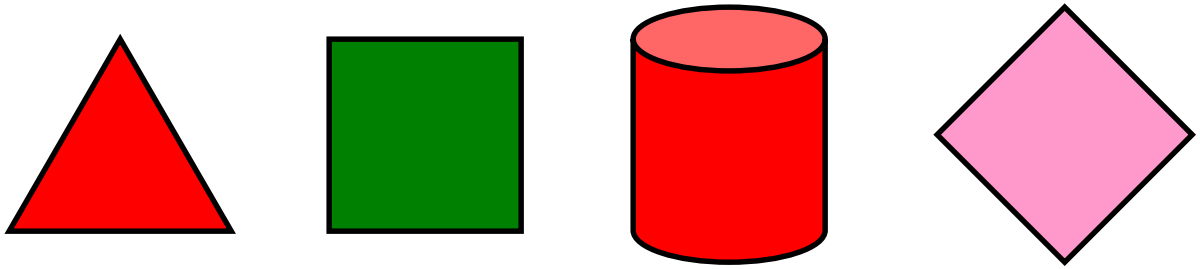
Wong Limsoon

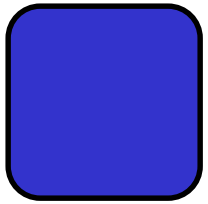
What is knowledge discovery?

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

- **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, ...

Some classifiers / machine learning methods

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 (Balanced population)

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

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?

High accuracy is meaningless if population is unbalanced

What is sensitivity (aka recall)?

	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**

Exercise: Write down the formula for specificity

What is precision?

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

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

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

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%		
C	50	0	150	0	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?**

Exercise #1

Abstract model of a classifier

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

t is the decision threshold of the classifier

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

Example

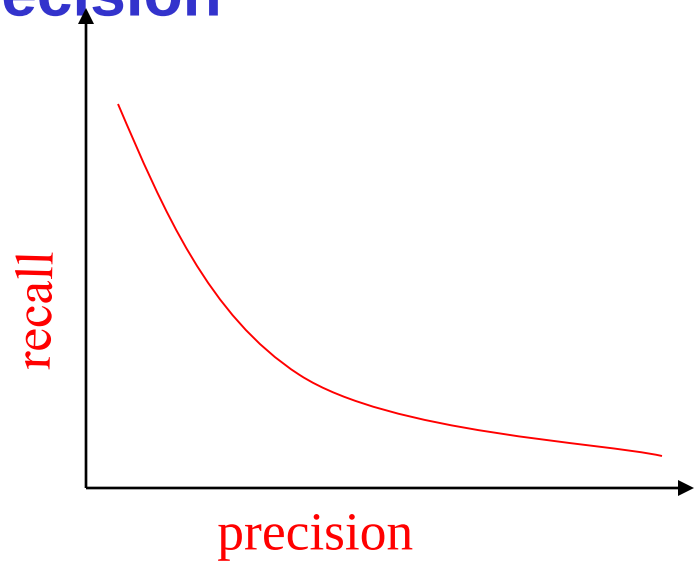
S	P(S)	N(S)	Actual Class	Predicted Class @ t = 3	Predicted Class @ t = 2
2	0.961252	0.038748	P	P	P
3	0.435302	0.564698	N	N	N
6	0.691596	0.308404	P	N	P
7	0.180885	0.819115	N	N	N
8	0.814909	0.185091	P	P	P
10	0.887220	0.112780	P	P	P
			accuracy	5 / 6	6/6
			recall	3 / 4	4/4
			precision	3 / 3	4/4

Recall that …

- **Predict *S* as negative if *p(S) / n(S)* < *t***
- **Predict *S* as positive if *p(S) / n(S)* ≥ *t***

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 apps, once you reach satisfactory precision, you optimize for recall**
- **In some apps, once you reach satisfactory recall, you optimize for precision**

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?

F-measure (Used in info extraction)

- **Take the harmonic mean of recall and precision**

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

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:
C predicts everything as +ve, but still rated better than A

Adjusted accuracy

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

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

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%

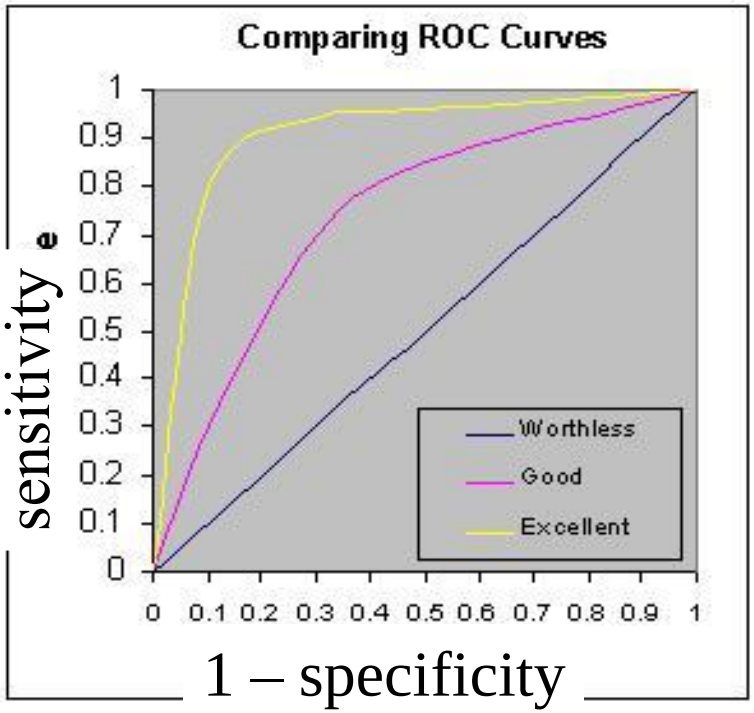
Geometric mean of sensitivity & specificity

Geometric mean = sqrt (Sensitivity * Specificity)

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

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



Food for thought

- **You have a classifier. On a test set having 20% +ve and 80% -ve cases, the classifier's recall and precision are both 80%**
- **Suppose you test it on a new test set having 80% +ve and 20% -ve cases. What do you expect its accuracy to be?**

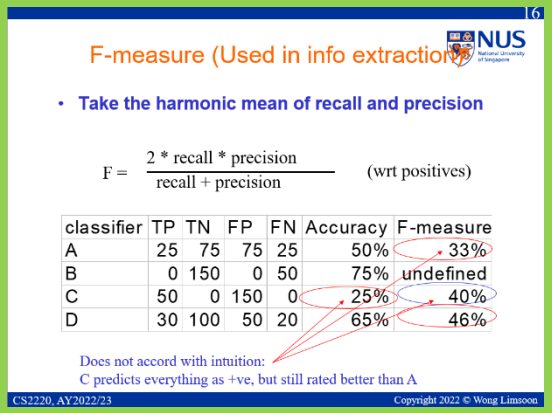
- You may assume that the +ve (resp. –ve) cases in both test sets are equally sufficiently representative of the +ve (resp. –ve) real-world population

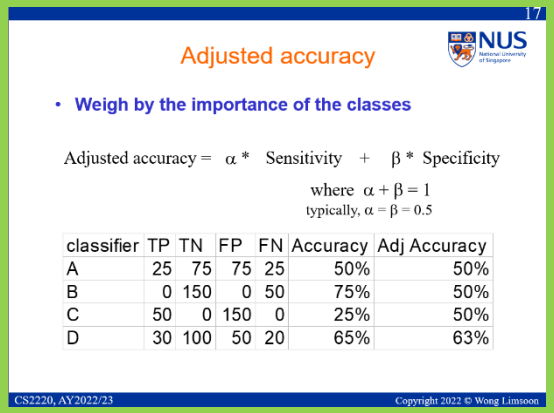
- **What lesson have you learned?**

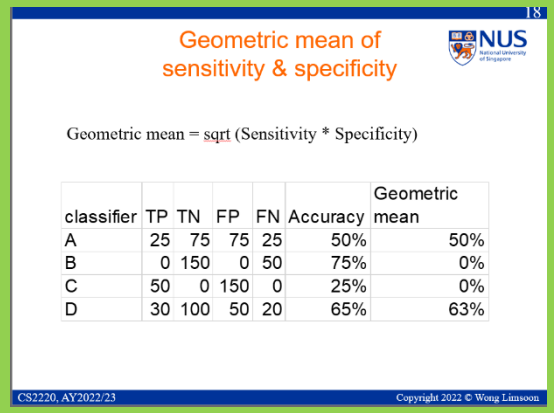
Exercise #2

Food for thought

- **Consider F-measure, adjusted accuracy, and geometric mean of sensitivity & specificity**







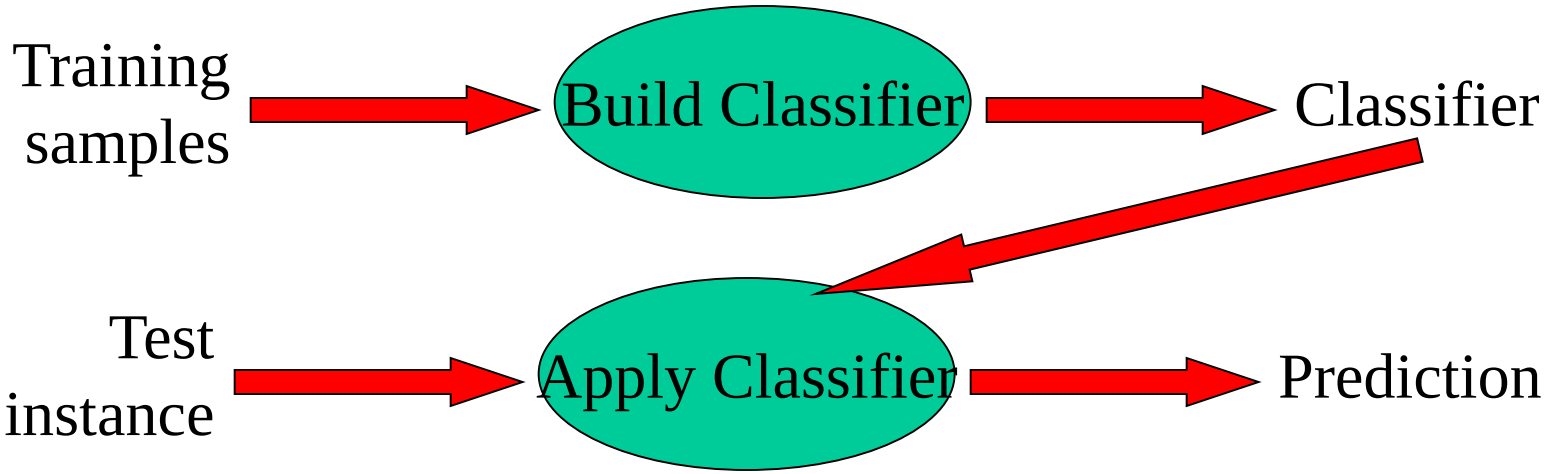
- **Which among these is likely the most robust measure of a classifier's performance?**

- **Why?**

Exercise #3

What is cross validation?

Construction of a classifier



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) / n(S) < t**
- **Predict S as positive if p(S) / n(S) ≥ t**

t is the decision threshold of the classifier

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 take majority vote of 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

Illustration of kNN (k=8)

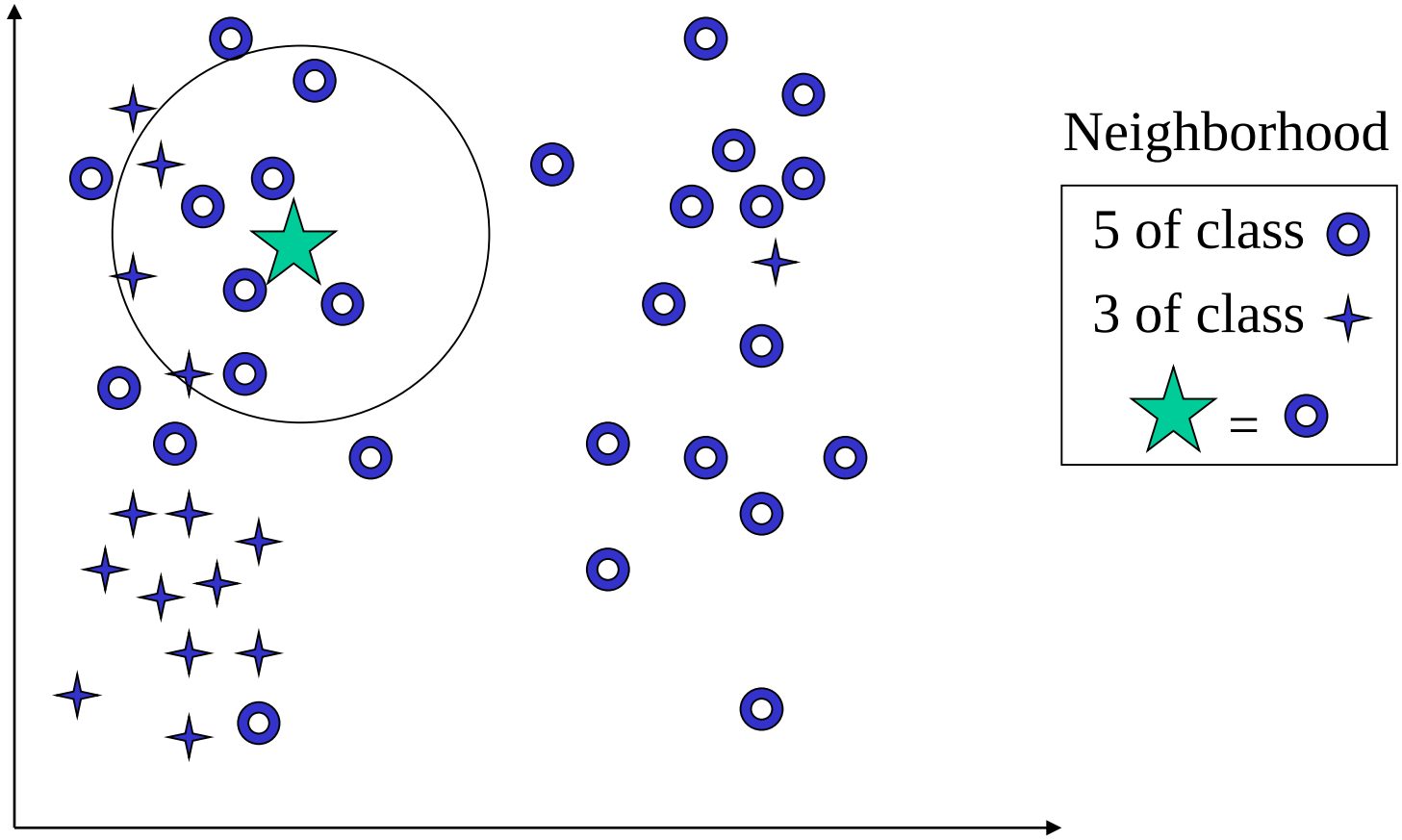
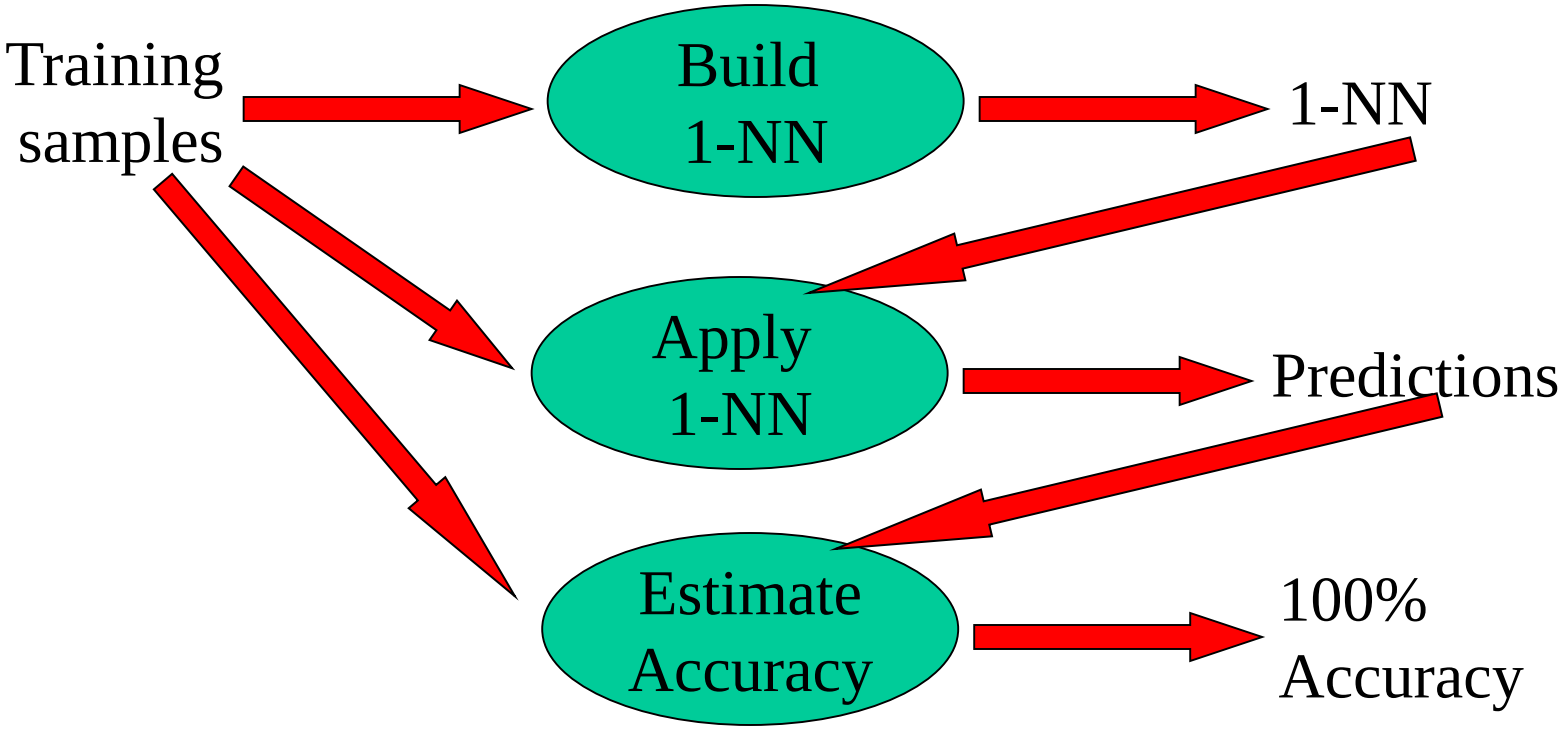


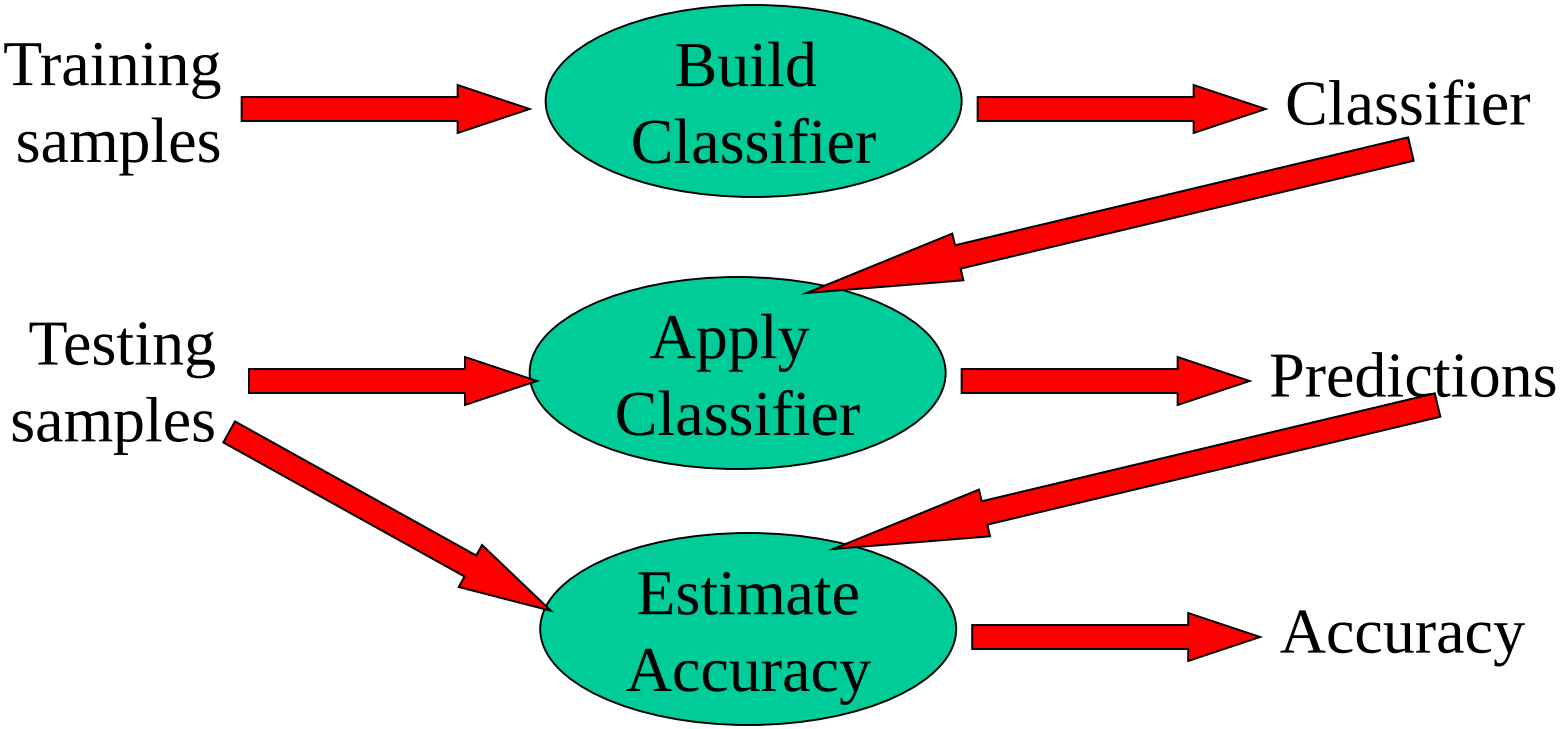
Image credit: Zaki

Estimate accuracy: Wrong way



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

Estimate accuracy: Right way



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

How many training and testing samples?

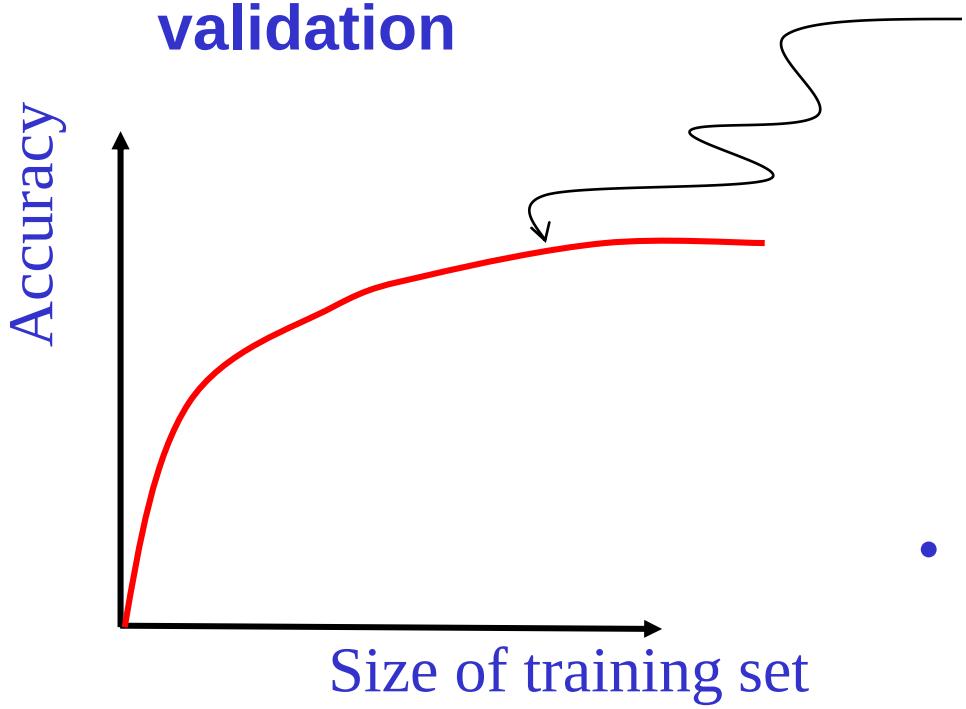
- **No fixed ratio between training and testing samples; but typically 2:1 ratio**
- **Proportion of instances of different classes in testing samples should be similar to proportion in the real world, and preferably also to proportion in the training samples**
- **What if there are insufficient samples to reserve 1/3 for testing?**

Cross validation

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

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**

Food for thought

- **What is the logical basis of cross validation?**
- Hint: Central limit theorem

- **What / whose accuracy does it really estimate?**

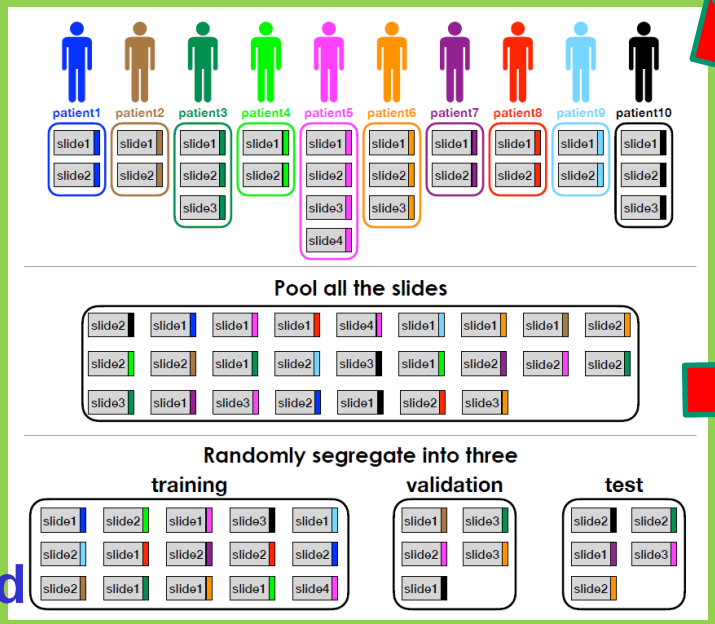
Exercise #4

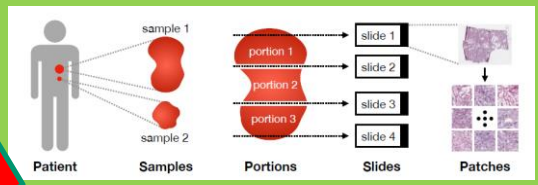
Food for thought: Deep learning from histopath images for lung cancer diagnosis

Coudray et al. report that common mutations in lung cancers can be predicted from histopath images using deep learning

Is this claim sound based purely on their results?

Coudrayet al, *Nat Med* 24:1559-1567, 2018





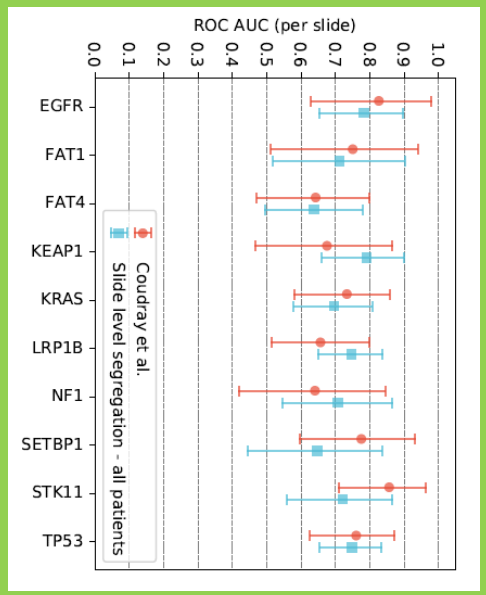


Image credit: Mustafa Umit Oner

Exercise #5

Curse of dimensionality

Recall kNN …

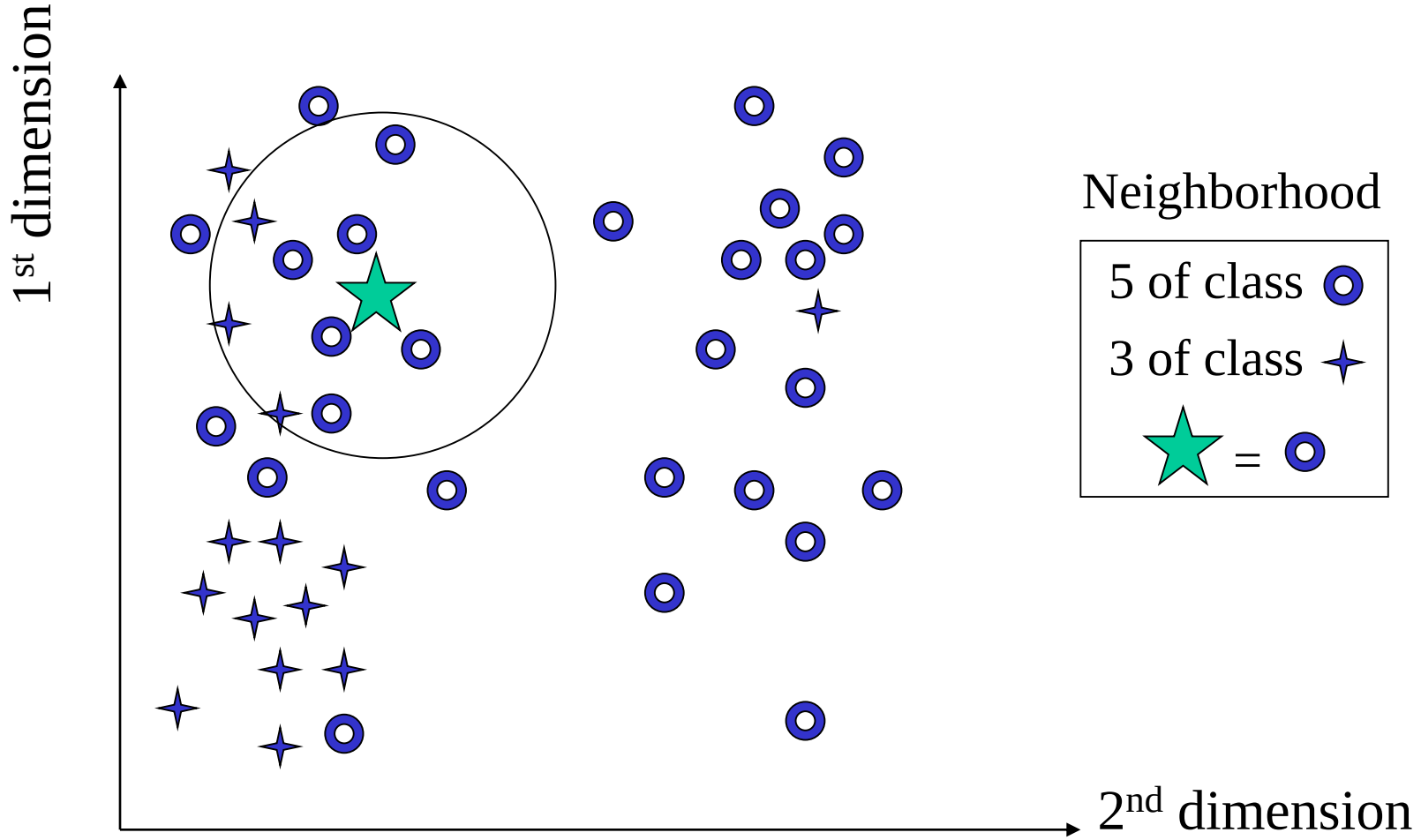
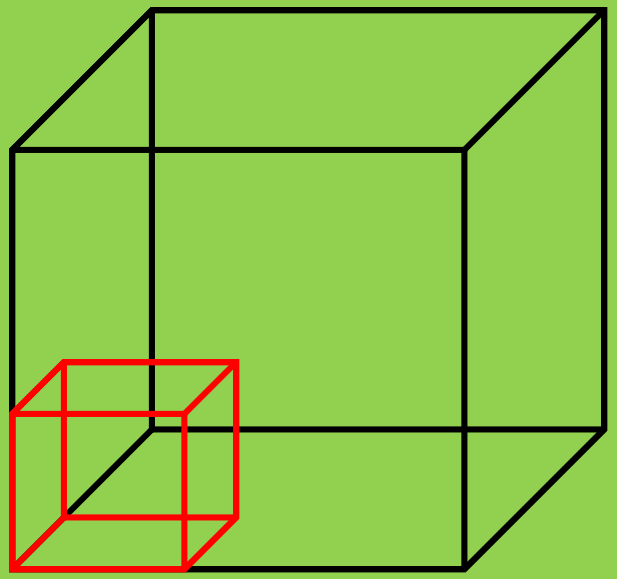


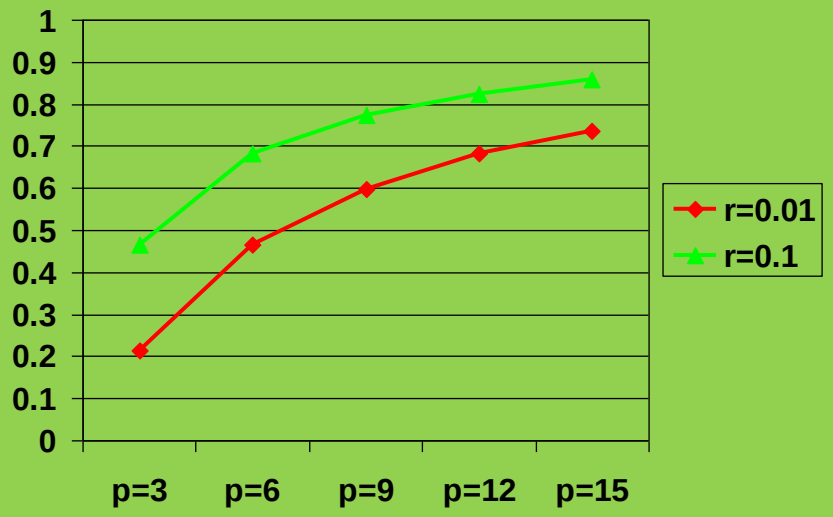
Image credit: Zaki

Curse of dimensionality

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



- **Calculate by $e_p(r) = r^{1/p}$. Why?**
- **So, to cover 10% of a 15-D space, need 85% of each dimension!**



Exercise #6

Consequence of the curse

- **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**

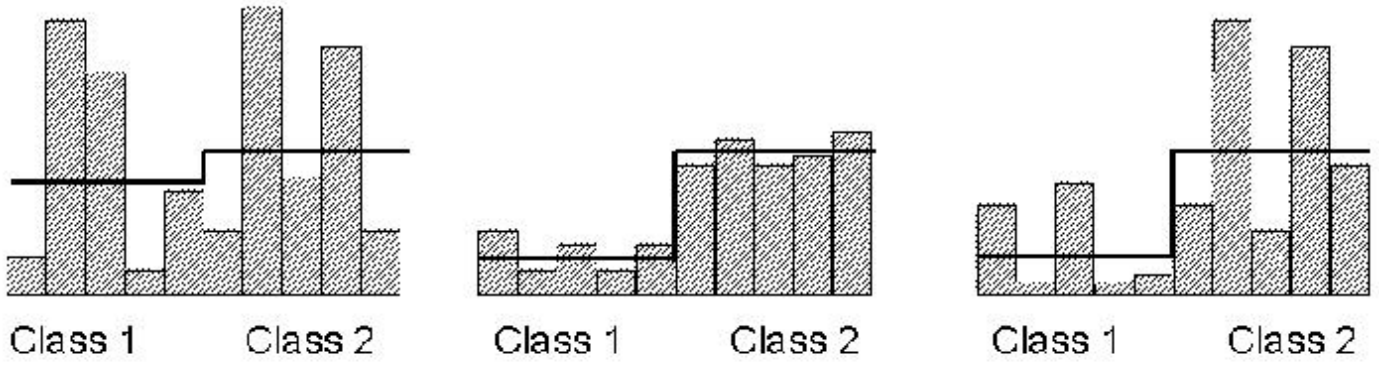
What is feature selection?

Tackling the curse

- **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)**

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.

Food for thought

- **How is the t-statistic typically used?**
- **What are the assumptions required for this way of using the t-statistic?**

Exercise #7

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 the 20 selected features
- The resulting accuracy can be ~90%
- But the true accuracy should be 50%, as the data were derived randomly

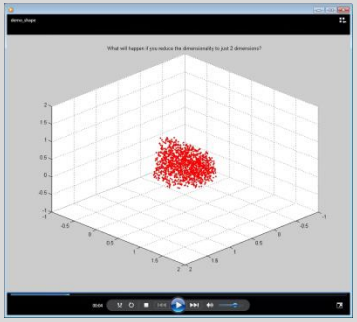
What went wrong?

- **The 20 features were selected from 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**

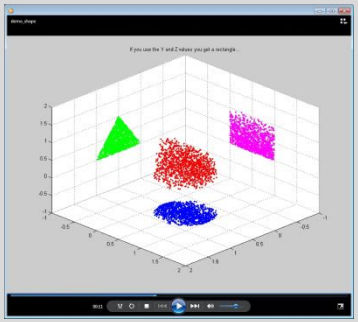
While **dimensionality reduction** is an important tool in machine learning/data mining, we must always be aware that it can distort the data in misleading ways.

Above is a two dimensional projection of an intrinsically three dimensional world….

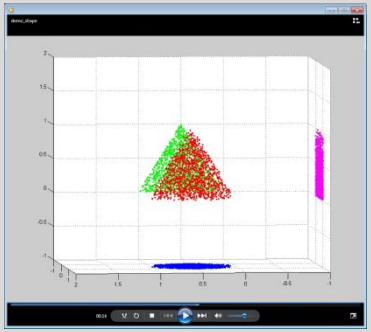
A cloud of points in 3D



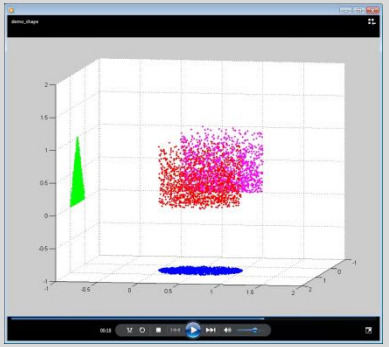
Can be projected into 2D
XY or XZ or YZ



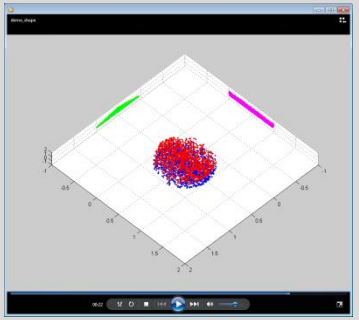
In 2D XZ we see
a triangle



In 2D YZ we see
a square



In 2D XY we see
a circle



Screen dumps of a short video from
www.cs.gmu.edu/~jessica/DimReducDanger.htm

Concluding remarks

NUS
National University
of Singapore

What have we learned?

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

Any questions?

Acknowledgements

- **The first two slides were shown to WLS 20+ years ago by Tan Ah Hwee**

- **The three slides on the dangers of dimensionality reduction were created by Eamonn Keogh**

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