Lecture 5:
Statistical Measurements X Dimension Reduction X Feature Selection

Speaker: Hong-Han Shuai

ECE, NCTU

Thanks to the slides made by Prof. Hung-Yi Lee and Jen-pei Liu from NTU.

Statistical Measurements

Significance testing

- Say I have two classifiers.

- A = 50% accuracy
- B = 75% accuracy

- B is better, right?

Significance Testing

- Say I have another two classifiers

- A = 50% accuracy
- B = 50.5% accuracy

- Is B better?

Basic Evaluation

- Training data – used to identify model parameters
- Testing data – used for evaluation
- Optionally: Development / tuning data – used to identify model hyperparameters.

- Difficult to get significance or confidence values

Some criticisms of cross-validation

- While the test data is independently sampled, there are a lot of overlap samples in training data.
  - The model performance may be correlated.
  - **Underestimation** of variance.
  - **Overestimation** of significant differences.
- One proposed solution is to repeat **2-fold** cross-validation 5 times rather than **10-fold** cross validation

Significance Testing

- Is the performance of two classifiers different with statistical significance?
- Means testing
  - If we have two samples of classifier performance (accuracy), we want to determine if they are drawn from the same distribution (no difference) or two different distributions.

steps in doing significance test:

- Statement of the problem.
- Formulation of hypothesis, null and alternate
- Decide upon a significant level, α of the test.
- Choose a test statistic, t, z
- Compare test-statistics with relevant tabulated value
- Make statistical decision
- Conclude.

- **Research hypothesis:** It is the conjecture or supposition that motivates the research.
- **Statistical hypothesis:** are hypothesis that are stated in such a way that they may be evaluated by appropriate statistical techniques.

問題範例

犯人有罪與否? (Yes or No?)

藥品是否有效? (Yes or No?)

食品防腐劑是否超出政府所訂的標準? (Yes or No?)

乳品中是否含三聚氰胺? (Yes or No?)

建構式數學是否提升小學生數學能力? (Yes or No?)

產品是否符合規格? (Yes or No?)

實證科學 (Evidence-based Science)

- 以數據的經驗證據(Empirical evidence)做出資訊決策(Informed Decision)。
- 決策的方式只有兩種:是或否(Yes or No?)。
- 數據的經驗證據來自樣本。
- 決策是推論至整個母體。

統計假說檢定(Statistical Hypothesis Testing)

統計方法進行決策的過程(Decision-Making Process)，將探討的問題二分為兩種假說：

虛無假說(Null Hypothesis，H_o)

對立假說(Alternative Hypothesis，H_a)

對立假說：吾人欲證明的事件(所感興趣)

虛擬假說：對立假說之補事件(不感興趣)

例：若法官對審判的目的為證明嫌犯有罪

H_o：無罪

vs.

H_a：有罪

若藥廠要證明所研發的新藥有療效

H_o：無療效

vs.

H_a：有療效

自今天生產的奶粉罐隨機取樣36罐奶粉，其樣品平均值為485g，若族群標準偏差σ=30g，是否有足夠證據證明奶粉罐平均重量不足500公克？

vs. H_o：新藥不具療效
H_a：新藥具有療效

事實(Truth)		
決策(Decision)	H_o：新藥不具療效為真	H_a：新藥具有療效為真
無法拒絕H_o Not reject H_o	決策正確	型Ⅱ錯誤
拒絕H_o Reject H_o	型Ⅰ錯誤	決策正確

型Ⅰ錯誤(Type Ⅰ Error)

拒絕虛無假說|當H_0為真時

決策判定新藥有療效|事實上新藥無療效

reject H_0|H_0 is true

消費者的風險(Consumer's Risk)

型Ⅱ錯誤(TypeⅡError)

無法拒絕虛無假說|當H_a為真時

決策判定新藥無療效|事實上新藥具有療效

無法拒絕H_o|H_a is true

生產者的風險(Producer's Risk)

診斷結果		
決策(診斷)	事實(Truth)	
	Ho：無病為真	Ha：有病為真
無法拒絕Ho	決策正確	型Ⅱ錯誤
拒絕Ho	型Ⅰ錯誤	決策正確

型Ⅰ錯誤：診斷有病|事實上無病
拒絕Ho| Ho為真
偽陽性(False Positive)

型Ⅱ錯誤：診斷無病|事實上有病
無法拒絕Ho| Ha為真
偽陰性(False Negative)

統計假說檢定之邏輯基礎：反證法

目的：證明H_a為真

方法：利用資料證明H_o不成立
=>間接地證明H_a為真

結論：二種可能性

1. 拒絕H_o => 證明H_a
2. 無法拒絕H_o
   不代表證明H_o
   僅說明資料無法提供足夠證據推翻H_o

目的：反證法證明H_a為真
必須先控制型Ⅰ錯誤
($拒絕H_o \mid H_o$為真)

顯著水準α=P[型Ⅰ錯誤]
=P[$拒絕H_o \mid H_o$為真]
=P[偽陽性]

β=P[型Ⅱ錯誤]
=P[無法拒絕$H_o \mid H_a$為真]
=P[偽陰性]

檢定力=1－β
=P[$拒絕H_o \mid H_a$為真]

統計假說檢定之步驟

1. 設立虛無假說(H_o)及對立假說(H_a)

   應將欲證明之假說放於H_a

   其補集合放於H_o

   消費者基金會：奶粉重量不足500公克

   H_o：$\mu \geq 500g$ vs. H_a： $\mu < 500g$

   H_o：$\mu \geq \mu_o$ vs. H_a：μ $\mu_o < \mu_o$=500g

2. 設定顯著水準:通常α=0.05或α=0.01

3. 選擇適當的檢定統計量(Test Statistic)

$$\text{Z} = \frac{\text{分子}}{\text{分母}}$$

分子：樣品估算值－虛無假設所定族群母數(μ_o)

$$\bar{x} - 500$$

分母：樣品估算值的抽樣誤差： $\frac{\sigma}{\sqrt{n}}$

$$Z = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}} = \frac{\bar{x} - 500}{\sigma/\sqrt{n}}$$

4. 決定棄卻域(Rejection Region)或決策規則(Decision Rule)

H_o：$\mu \geq 500g$ vs. H_a：$\mu < 500g$

H_a只考慮奶粉平均重量小於500公克

=>單尾檢定

標準常態分佈之$(1-\alpha)\%$百分位 $z_{1-\alpha}$

如果 $Z = \dfrac{\bar{x} - \mu_o}{\sigma/\sqrt{n}} < -z_{1-\alpha}$

=>拒絕H_o

$\alpha=0.05$， $z_{0.95}=1.645$

表示 $\bar{x}-\mu_o$ 的差異

無法以抽樣誤差解釋=>

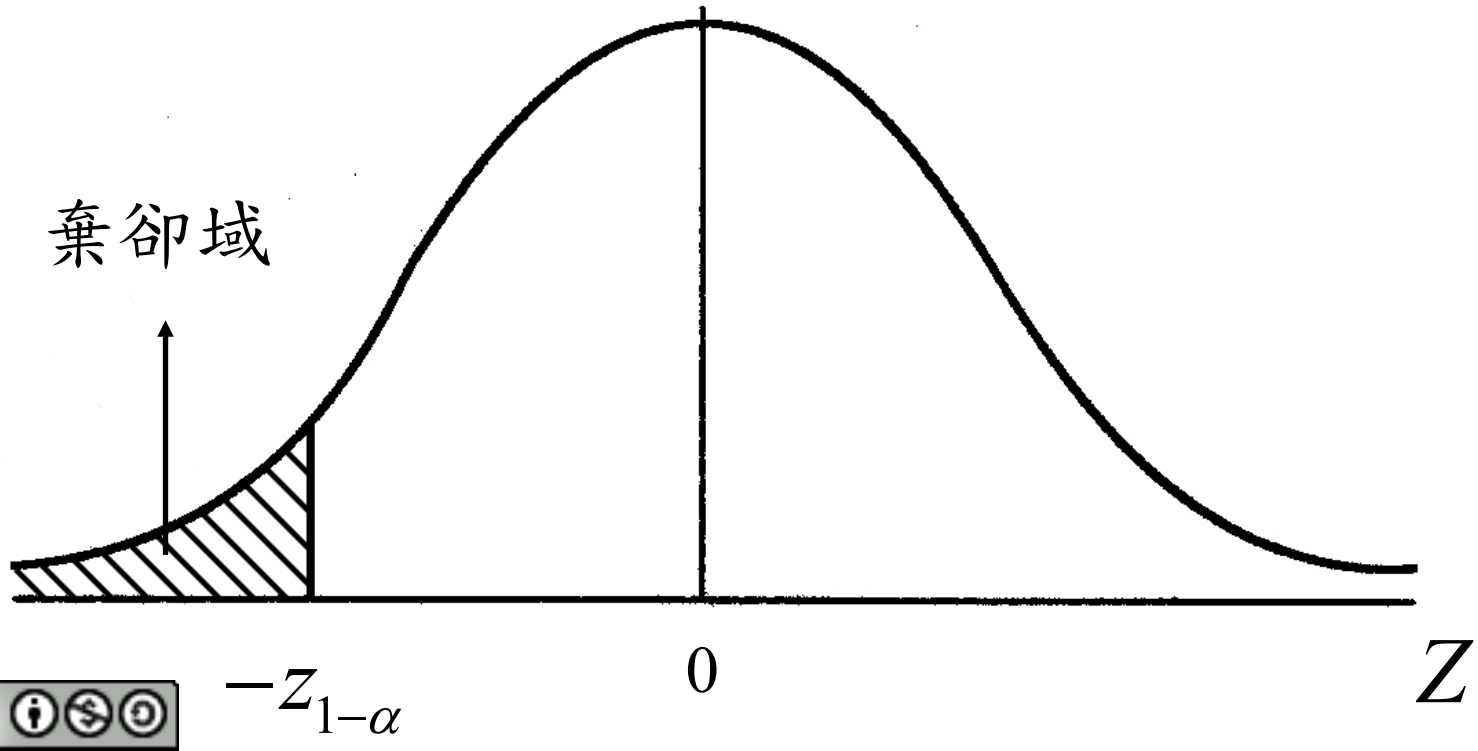
$\bar{x}-\mu_o$ 的差異可能是其他原因

如自動化裝罐機問題，人工操作的原因

如果 $Z=\dfrac{\bar{x}-\mu_o}{\sigma/\sqrt{n}}>-z_{1-\alpha}$

=>無法拒絕H_o

=> $\bar{x}-\mu_o$ 的差異未超過抽樣誤差



5. 進行實驗或取樣取得樣品計算
樣品統計量及檢定統計量Z

根據步驟4判定拒絕Ho或無法拒絕Ho
本日隨機取樣36罐奶粉

$\bar{x} = 485\text{g}$　　若σ=30g

$$Z = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}} = \frac{485 - 500}{30/\sqrt{36}} = -3.0$$

若 $\alpha = 0.05$　　$-z_{0.95} = -1.645$

$Z = -3.0 < -1.645$　=>拒絕Ho

=>本日奶粉罐重量小於500公克
=>此決定是有5%的型Ⅰ錯誤機率
=>重複進行20次假設檢定後的決策有一次是錯誤

單尾檢定：對立假說是單一方向

H_o：$\mu \geq \mu_o(500g)$ vs. H_a：$\mu < \mu_o(500g)$

若 $Z = \dfrac{\bar{x} - \mu_o}{\sigma/\sqrt{n}} < -z_{1-\alpha}$ =>拒絕H_o

H_o：$\mu \leq \mu_o(1000kg/ha)$ vs. H_a：$\mu > \mu_o(1000kg/ha)$

若 $Z = \dfrac{\bar{x} - \mu_o}{\sigma/\sqrt{n}} > z_{1-\alpha}$ =>拒絕H_o

雙尾檢定：對立假說是兩個方向

H_o：$\mu=500g\ (\mu_o)$ vs. H_a：$\mu\neq 500g\ (\mu_o)$

若 $Z=\dfrac{\bar{x}-\mu_o}{\sigma/\sqrt{n}}<-z_{1-\alpha/2}$ 或

$Z=\dfrac{\bar{x}-\mu_o}{\sigma/\sqrt{n}}>z_{1-\alpha/2}$ =>拒絕H_o

$|Z|=\left|\dfrac{\bar{x}-\mu_o}{\sigma/\sqrt{n}}\right|>z_{1-\alpha/2}$ =>拒絕H_o

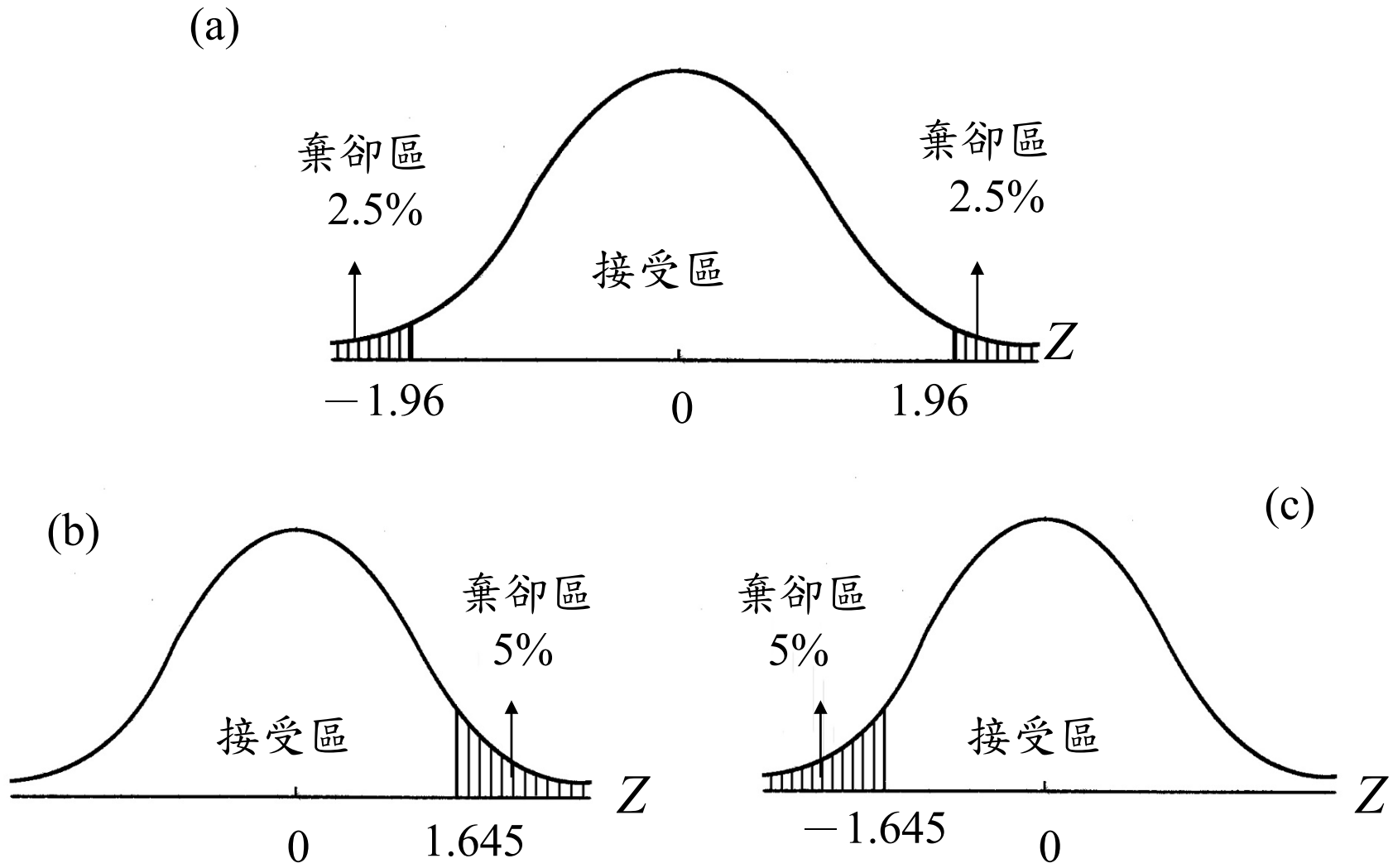


圖7.3 雙尾與單尾檢定圖

例子：正常成人血中平均膽固醇為180mg/dL
標準偏差為50mg/dL.今調查某地區16位
成人平均膽固醇為200mg/dL問此地區
平均膽固醇是否與180mg/dL有差異？

1. Ho：μ=180mg/dL vs. Ha： μ≠180mg/dL
2. α=0.05
3. $Z = \dfrac{\bar{x} - 180}{\sigma / \sqrt{n}}$
4. 若 $|Z| > z_{1-\alpha/2} = z_{0.975} = 1.96$ =>拒絕Ho
5. $Z = \dfrac{200 - 180}{50 / \sqrt{16}} = 1.6$

因 $|Z| = 1.6 < z_{0.975} = 1.96$ =>無法拒絕Ho

Null Hypothesis

. Null Hypothesis (symbolized as H_0) can be defined as the statistical hypothesis of no difference.

. H_0 is an artificial ‘ straw man” that provides a reference for examining the departure of data actually obtained from the data that would be expected under the null hypothesis.

Alternate hypothesis(H_a) : Is any other hypothesis which we are willing to accept when the H_0 is rejected.

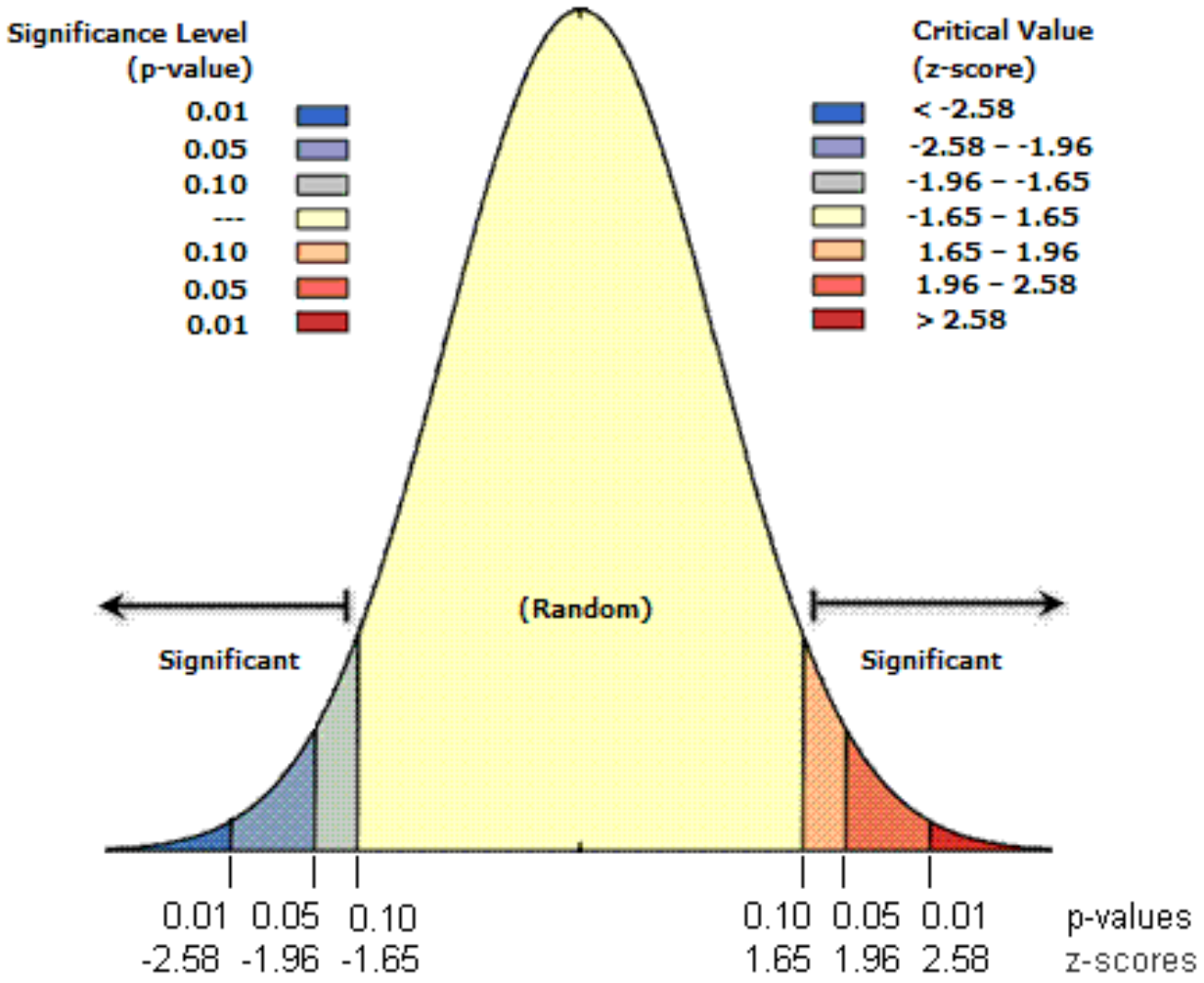
How to write a null hypothesis.

- Alternative hypothesis (H_a) guides the writing of the null hypothesis(H_0).
- So consider the form of the alternative hypothesis first.
- remember H_a is the reflection of your Research hypothesis.
- The Research hypothesis is usually written in narrative form while the Ha is written in algebraic form of inequality.

- **level of Significance**
- In hypothesis testing, the null hypothesis is either accepted or rejected, depending on whether the p value is above or below a predetermined cut-off point,known as the Significance level of the test, usually it is taken as 5% level.

Calculated value:

- tabulated value: for a certain degree of freedom highest value of test statistics obtainable by chance corresponding to probability of .05 or .01.


Significance Level
(p-value)
0.01
0.05
0.10

0.10
0.05
0.01
Critical Value
(z-score)
< -2.58
-2.58 – -1.96
-1.96 – -1.65
-1.65 – 1.65
1.65 – 1.96
1.96 – 2.58
> 2.58
(Random)
Significant
Significant
0.01 0.05 0.10
-2.58 -1.96 -1.65
0.10 0.05 0.01
1.65 1.96 2.58
p-values
z-scores

Dimension Reduction

Speaker: Hong-Han Shuai

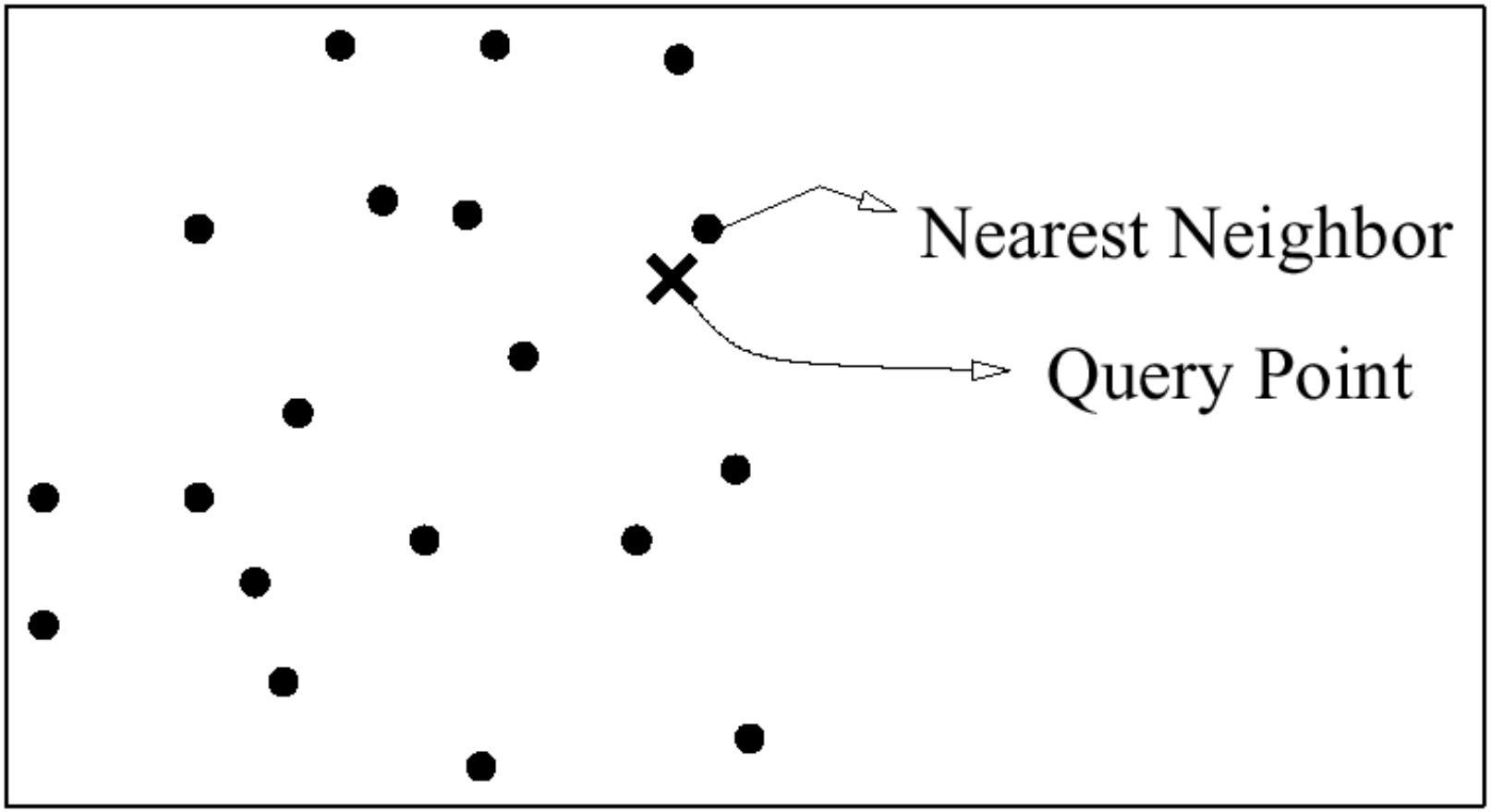
ECE, NCTU

Thanks to the slides made by Prof. Hung-Yi Lee from NTU.

Preliminaries : Nearest Neighbor Search

- Given a collection of data points and a query point in m-dimensional metric space, find the data point that is closest to the query point
- Variation: k-nearest neighbor
- Relevant to clustering and similarity search
- Applications: Geographical Information Systems, similarity search in multimedia databases

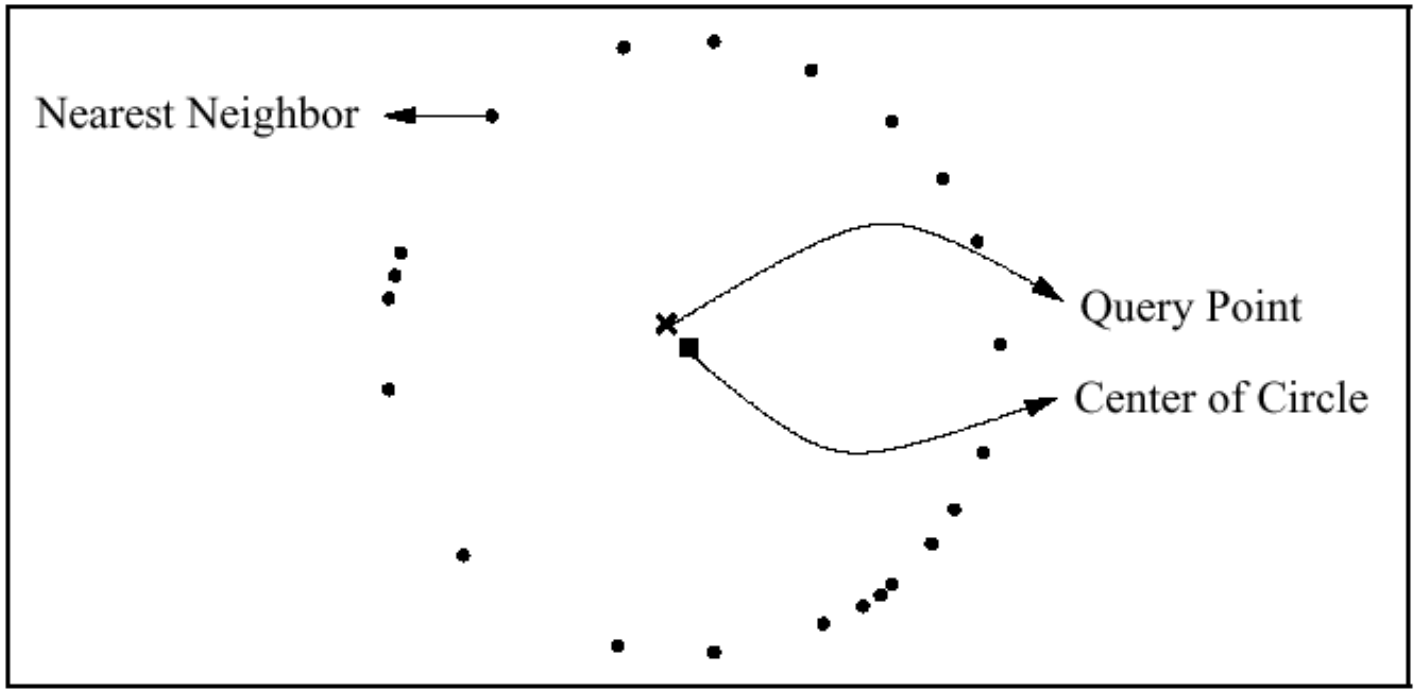
NN Search Con’t



Source: [2]

Problems with High Dimensional Data

- A point's nearest neighbor (NN) loses meaning



Source: [2]

Problems (Con't)

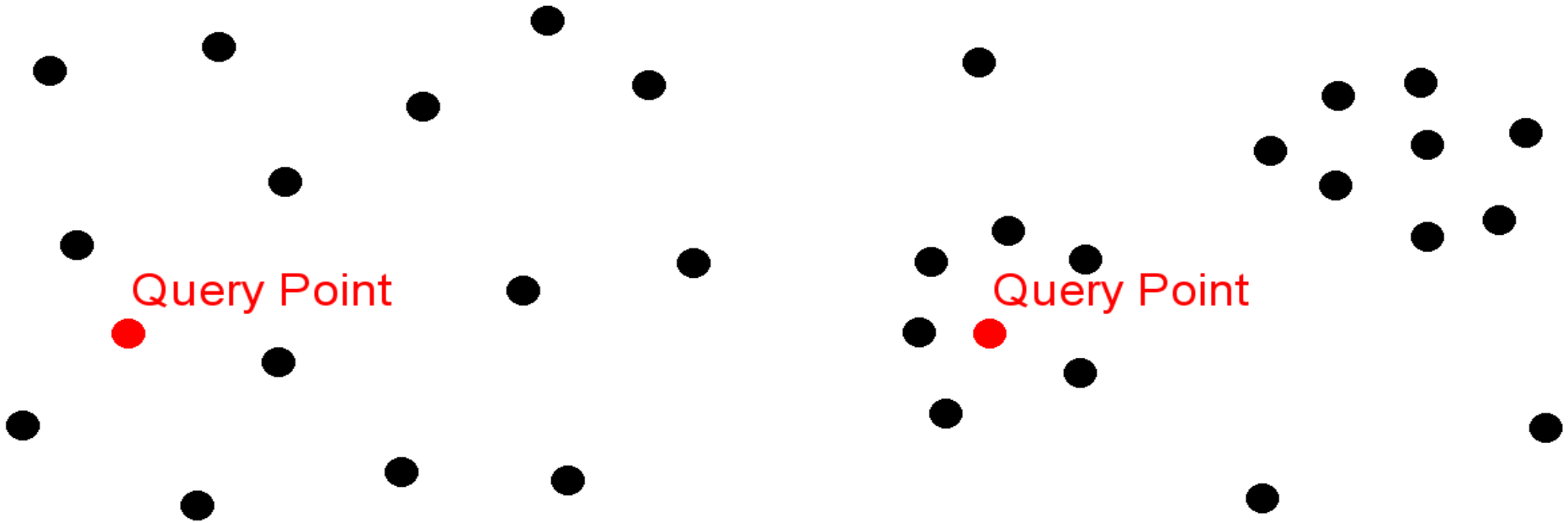
- NN query cost degrades – more strong candidates to compare with
- In as few as 10 dimensions, linear scan outperforms some multidimensional indexing structures (e.g. KD-tree, R* tree, SR tree)
- Biology and genomic data can have dimensions in the 1000's.

Problems (Con't)

- The presence of irrelevant attributes decreases the tendency for clusters to form
- Points in high dimensional space have high degree of freedom; they could be so scattered that they appear uniformly distributed

Problems Con't

- In which cluster does the query point fall?



The Curse

- Refers to the decrease in performance of query processing when the dimensionality increases
- In particular, under certain conditions, the distance between the nearest point and the query point equals the distance between the farthest and query point as dimensionality approaches infinity

Curse (Con't)

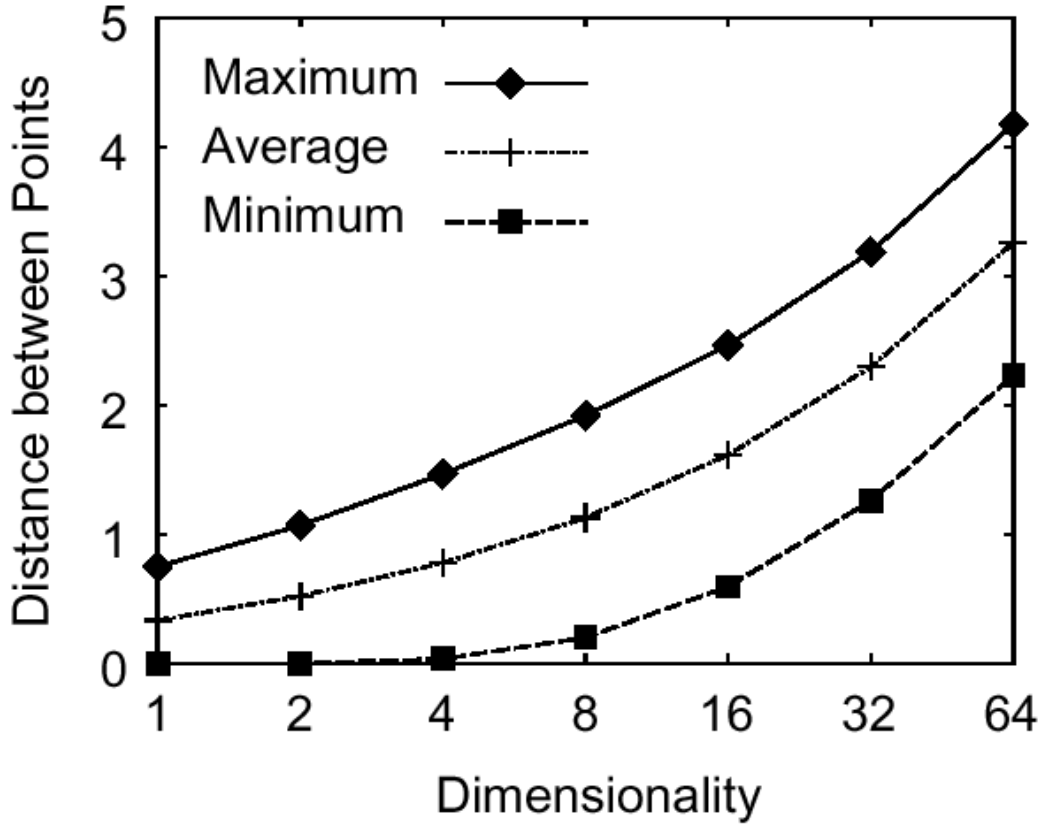


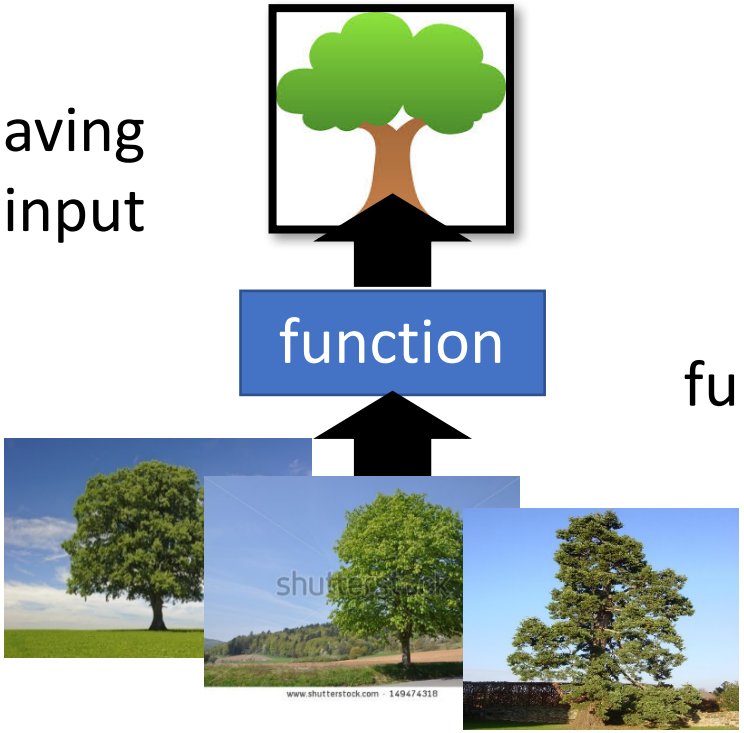
Figure 1. Distances among 100k points generated at random in a unit hypercube

Source: N. Katayama, S. Satoh. Distinctiveness Sensitive Nearest Neighbor Search for Efficient Similarity Retrieval of Multimedia Information. ICDE Conference, 2001.

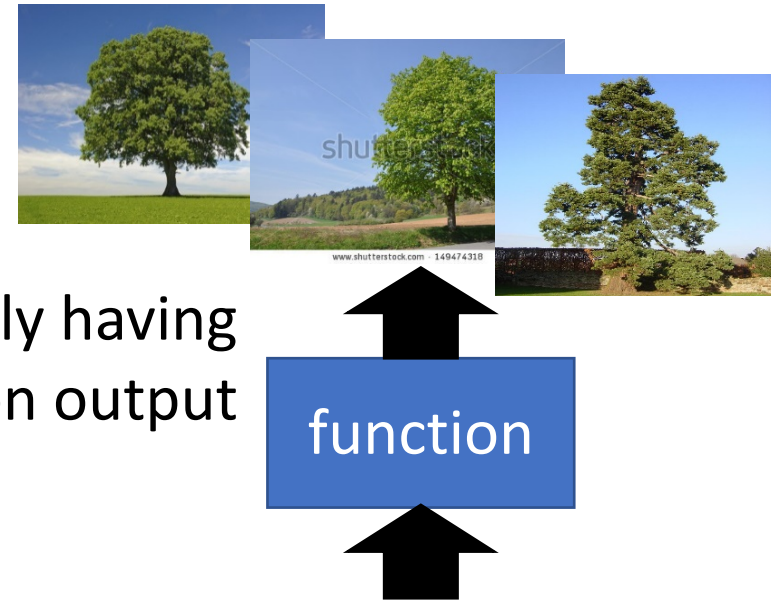
Unsupervised Learning

- Clustering & Dimension Reduction (化繁為簡)
- Generation (無中生有)

only having function input



only having function output

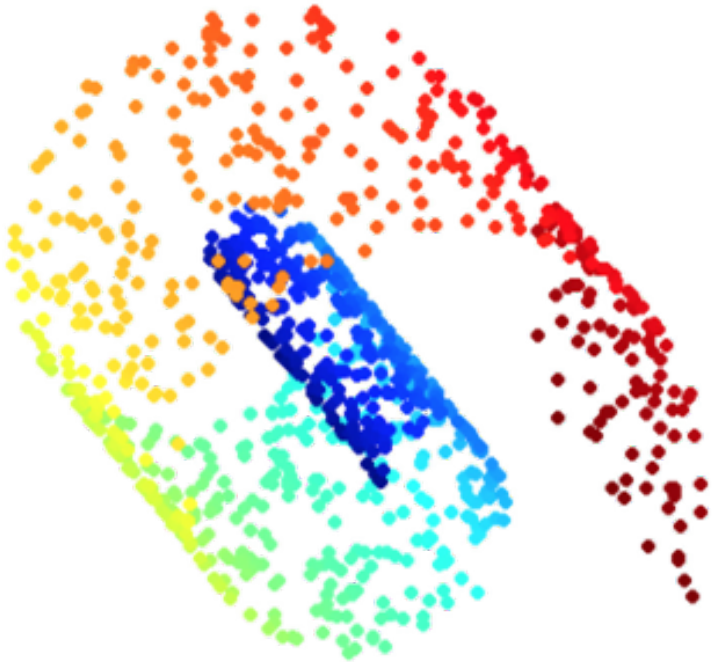


Clustering & Dimension Reduction in these slides

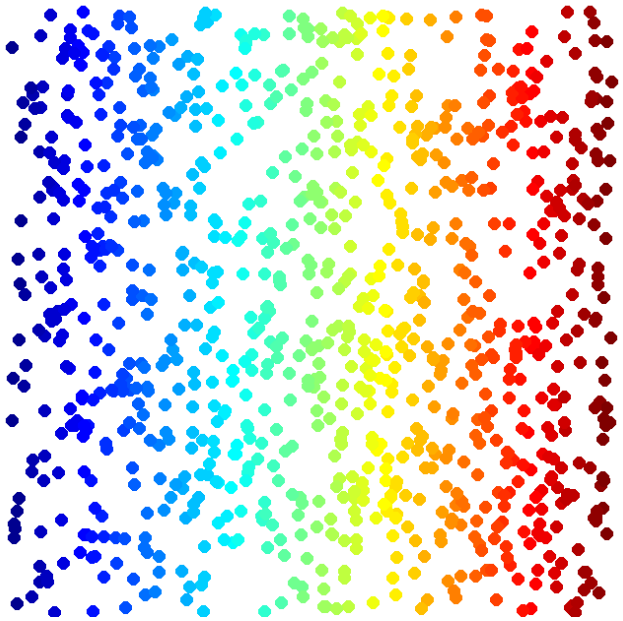
Clustering

- K-means
  - Clustering $X = \{x^1, \cdots, x^n, \cdots, x^N\}$ into K clusters
  - Initialize cluster center c^i, i=1,2, … K (K random x^n from X)
  - Repeat
    - For all x^n in X: $b_i^n \begin{cases} 1 & x^n \text{ is most "}\textbf{\textit{close}}\text{" to } c^i \\ 0 & \text{Otherwise} \end{cases}$
    - Updating all c^i: $c^i = \sum_{x^n} b_i^n x^n \Big/ \sum_{x^n} b_i^n$

Dimension Reduction



Looks like 3-D



Actually, 2-D

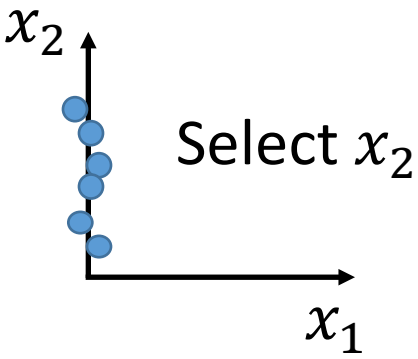
http://reuter.mit.edu/blue/images/research/manifold.png

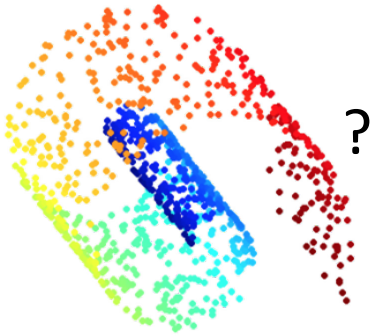
http://archive.cnx.org/resources/51a9b2052ae167db310fda5600b89badea85eae5/isomapCNXtrue1.png

Dimension Reduction

The dimension of z would be smaller than x

- Feature selection





- Principle component analysis (PCA)
  [Bishop, Chapter 12]

$$z = Wx$$

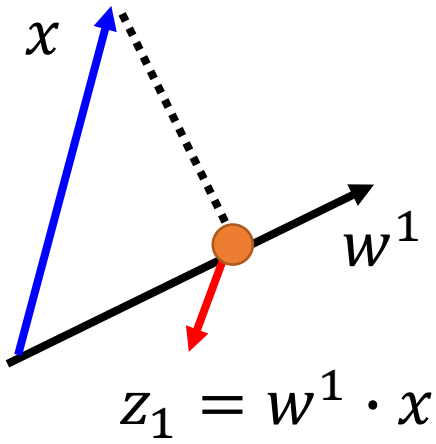
Principle Component Analysis (PCA)

PCA

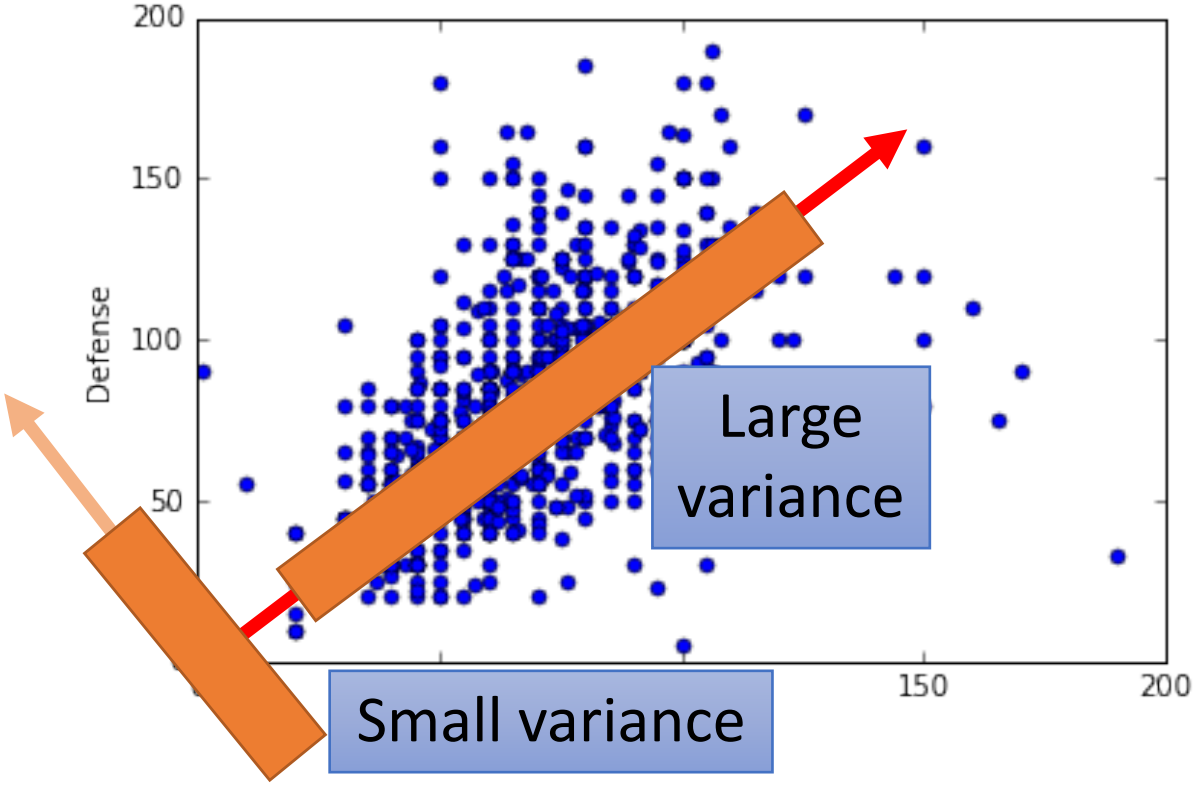
$z = Wx$

Reduce to 1-D:

$z_1 = w^1 \cdot x$



$z_1 = w^1 \cdot x$



Project all the data points x onto w^1, and obtain a set of z_1

We want the variance of z_1 as large as possible

$$Var(z_1) = \sum_{z_1} (z_1 - \bar{z_1})^2 \quad \|w^1\|_2 = 1$$

PCA

$$z = Wx$$

Reduce to 1-D:

$$z_1 = w^1 \cdot x$$

$$z_2 = w^2 \cdot x$$

$$W = \begin{bmatrix} (w^1)^T \\ (w^2)^T \\ \vdots \end{bmatrix}$$

Orthogonal matrix

Project all the data points x onto w^1, and obtain a set of z_1

We want the variance of z_1 as large as possible

$$Var(z_1) = \sum_{z_1} (z_1 - \bar{z_1})^2 \quad \|w^1\|_2 = 1$$

We want the variance of z_2 as large as possible

$$Var(z_2) = \sum_{z_2} (z_2 - \bar{z_2})^2 \quad \|w^2\|_2 = 1$$

$$w^1 \cdot w^2 = 0$$

Warning of Math

PCA

$$z_1 = w^1 \cdot x$$

$$\bar{z_1} = \frac{1}{N}\sum z_1 = \frac{1}{N}\sum w^1 \cdot x = w^1 \cdot \frac{1}{N}\sum x = w^1 \cdot \bar{x}$$

$$Var(z_1) = \frac{1}{N}\sum_{z_1}(z_1 - \bar{z_1})^2$$

$$= \frac{1}{N}\sum_{x}(w^1 \cdot x - w^1 \cdot \bar{x})^2$$

$$= \frac{1}{N}\sum\left(w^1 \cdot (x - \bar{x})\right)^2$$

$$= \frac{1}{N}\sum (w^1)^T (x - \bar{x})(x - \bar{x})^T w^1$$

$$= (w^1)^T \boxed{\frac{1}{N}\sum (x - \bar{x})(x - \bar{x})^T} w^1$$

$$= (w^1)^T Cov(x) w^1$$

$S = Cov(x)$

$$(a \cdot b)^2 = (a^T b)^2 = a^T b a^T b$$

$$= a^T b (a^T b)^T = a^T b b^T a$$

Find w^1 maximizing

$$(w^1)^T S w^1$$

$$\|w^1\|_2 = (w^1)^T w^1 = 1$$

Find w^1 maximizing $(w^1)^T S w^1$ $\quad (w^1)^T w^1 = 1$

$S = Cov(x)$ Symmetric positive-semidefinite
(non-negative eigenvalues)

Using Lagrange multiplier [Bishop, Appendix E]

$$g(w^1) = (w^1)^T S w^1 - \alpha\left((w^1)^T w^1 - 1\right)$$

$$\partial g(w^1)/\partial w_1^1 = 0$$
$$\partial g(w^1)/\partial w_2^1 = 0$$
$$\vdots$$

$$Sw^1 - \alpha w^1 = 0$$
$$Sw^1 = \alpha w^1$$ w^1 : eigenvector

$$(w^1)^T S w^1 = \alpha (w^1)^T w^1$$
$$= \alpha$$ Choose the maximum one

w^1 is the eigenvector of the covariance matrix S
Corresponding to the largest eigenvalue λ_1

Find w^2 maximizing $(w^2)^T S w^2 \quad (w^2)^T w^2 = 1 \quad (w^2)^T w^1 = 0$

$$g(w^2) = (w^2)^T S w^2 - \alpha\left((w^2)^T w^2 - 1\right) - \beta\left((w^2)^T w^1 - 0\right)$$

$\partial g(w^2)/\partial w_1^2 = 0$

$\partial g(w^2)/\partial w_2^2 = 0$

$\vdots$

$$Sw^2 - \alpha w^2 - \beta w^1 = 0$$

$$0 - \alpha\, 0 - \beta\, 1 = 0$$

$$= \left((w^1)^T S w^2\right)^T = (w^2)^T S^T w^1$$

$$= (w^2)^T S w^1 = \lambda_1 (w^2)^T w^1 = 0$$

$$Sw^1 = \lambda_1 w^1$$

$\beta = 0$: $\quad Sw^2 - \alpha w^2 = 0 \quad Sw^2 = \alpha w^2$

w^2 is the eigenvector of the covariance matrix S

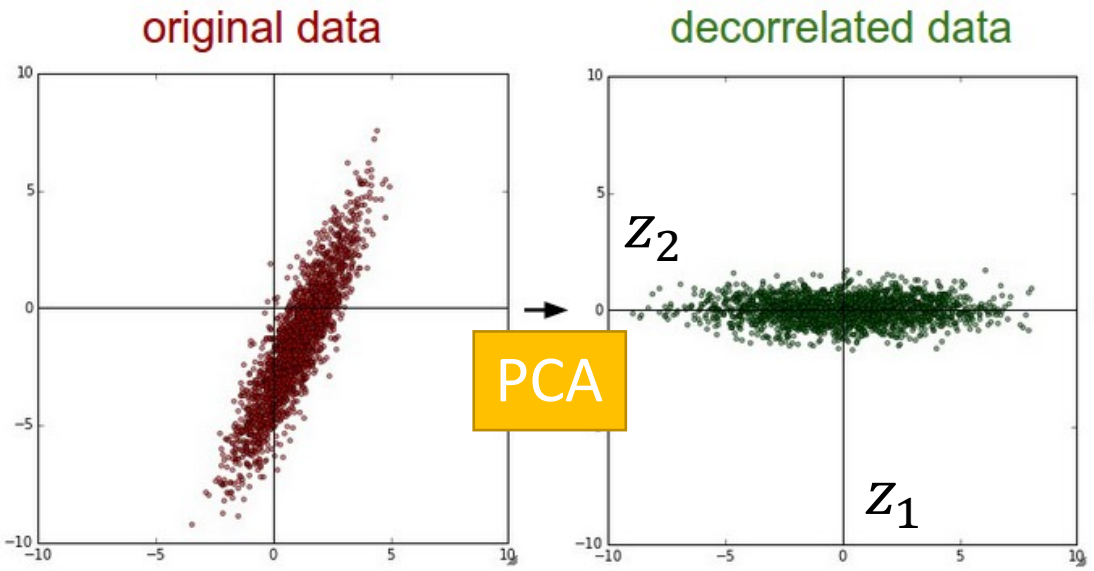
Corresponding to the 2nd largest eigenvalue λ_2

PCA - decorrelation

$$z = Wx$$

$$Cov(z) = D$$

Diagonal matrix



$$Cov(z) = \frac{1}{N}\sum (z - \bar{z})(z - \bar{z})^T = WSW^T \qquad S = Cov(x)$$

$$= WS[w^1 \quad \cdots \quad w^K] = W[Sw^1 \quad \cdots \quad Sw^K]$$

$$= W[\lambda_1 w^1 \quad \cdots \quad \lambda_K w^K] = [\lambda_1 Ww^1 \quad \cdots \quad \lambda_K Ww^K]$$

$$= [\lambda_1 e_1 \quad \cdots \quad \lambda_K e_K] = D$$

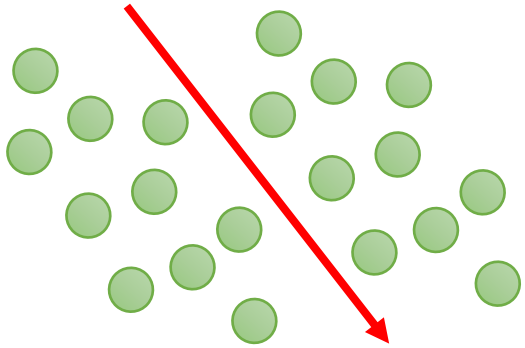
Diagonal matrix

End of Warning

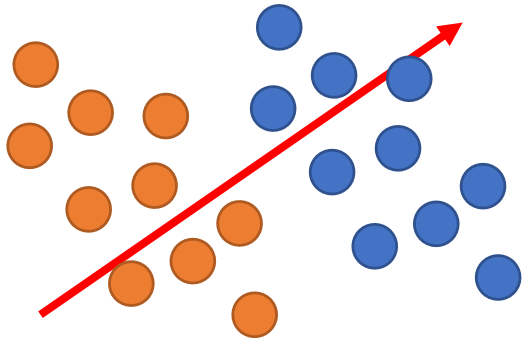
Weakness of PCA

- Unsupervised

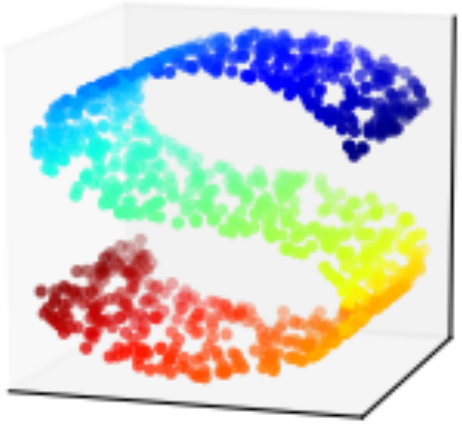
PCA

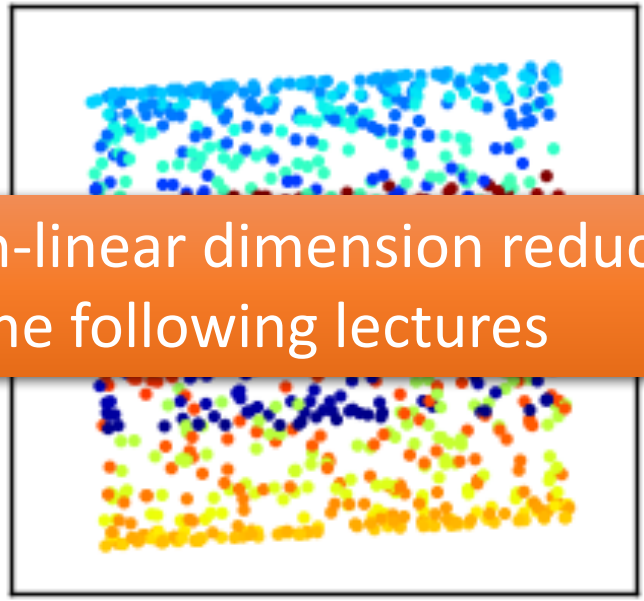


LDA



- Linear





Non-linear dimension reduction in the following lectures

http://www.astroml.org/book_figures/chapter7/fig_S_manifold_PCA.html

Matrix Factorization

Matrix Factorization

Number in table: number of figures a person has

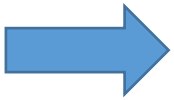
A
B
C
D
E

There are some common *factors* behind customers and characters.

http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/

More about Matrix Factorization

- Considering the individual characteristics

$r^A \cdot r^1 \approx 5$  $r^A \cdot r^1 + b_A + b_1 \approx 5$

b_A: how customer A likes to buy

b_1: how popular character 1 is

Minimizing $L = \sum_{(i,j)} \left(r^i \cdot r^j + b_i + b_j - n_{ij}\right)^2$

Find r^i, r^j, b_i, b_j by gradient descent (can add regularization)

- Ref: Matrix Factorization Techniques For Recommender Systems

Matrix Factorization for Topic analysis

character→document,
customer→word

- Latent semantic analysis (LSA)

	Doc 1	Doc 2	Doc 3	Doc 4
投資	5	3	0	1
股票	4	0	0	1
總統	1	1	0	5
選舉	1	0	0	4
立委	0	1	5	4

Number in Table:

Term frequency (weighted by inverse document frequency)

Latent factors are topics (財經、政治 ……)

- Probability latent semantic analysis (PLSA)
  - Thomas Hofmann, Probabilistic Latent Semantic Indexing, SIGIR, 1999
- latent Dirichlet allocation (LDA)
  - Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5): pp. 993–1022.

More Related Approaches Not Introduced

- Multidimensional Scaling (MDS) [Alpaydin, Chapter 6.7]
  - Only need distance between objects
- Probabilistic PCA [Bishop, Chapter 12.2]
- Kernel PCA [Bishop, Chapter 12.3]
  - non-linear version of PCA
- Canonical Correlation Analysis (CCA) [Alpaydin, Chapter 6.9]
- Independent Component Analysis (ICA)
  - Ref: http://cis.legacy.ics.tkk.fi/aapo/papers/IJCNN99_tutorialweb/
- Linear Discriminant Analysis (LDA) [Alpaydin, Chapter 6.8]
  - Supervised

Feature Selection
-A Data Perspective

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys*, *50*(6).

https://arxiv.org/abs/1601.07996

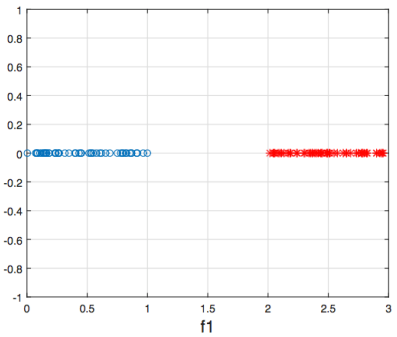
Intro

- The problems when handling **high dimensional data**
  - Curse of dimensionality: data becomes sparser in high dimensional space
  - Overfitting: cause performance degradation on unseen data

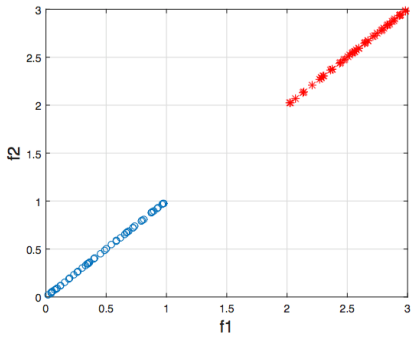
- Dimensionality reduction
  - Feature extraction
    - Projects original high dimensional feature space to a new one with low dimensionality
    - Principle Component Analysis, Linear Discriminant Analysis, etc.
  - **Feature selection**
    - Selects a subset of relevant features for the use model construction.
    - Lasso, Information Gain, Relif, Laplacian, etc.

Intro

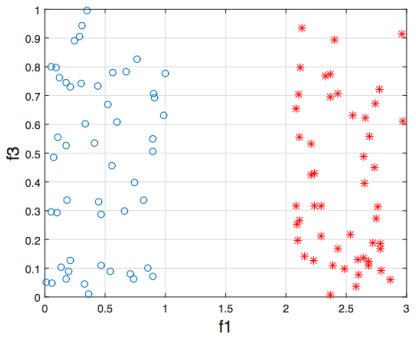
- Why feature selection?
  - Feature extraction builds a set of new features, further analysis is problematic as we cannot get the physical meaning of these feauture in the transformed space.
  - Reducing storage and computational cost while avoiding significant loss of information or negative degradation of learning performance.



(a) relevant feature f_1



(b) redundant feature f_2



(c) irrelevant feature f_3

Intro

- Traditional Categorizations of Feature Selection Algorithms
  - Label Perspective

Feature Selection	Supervised	Unsupervised	Semi-Supervised
Availability of label information	when sufficient	do not require any label information	In many real-world applications, we usually have a small number of labeled samples and a large number of unlabeled samples.
Designed for type of problems	Classification or regression problem	Clustering problem	

Intro

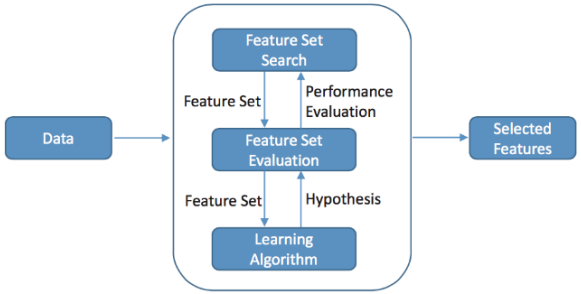


Figure 6: A General Framework of Wrapper Feature Selection Methods.

- Traditional Categorizations of Feature Selection Algorithms
  - Search Strategy Perspective
    - Wrapper Method
      - Rely on the predictive performance of a predefined learning algorithm to evaluate the quality of selected features.
      - Steps: (1)Search for a subset of features and (2)Evaluate selected features.
      - Repeats (1) and (2) until some stopping criteria are satisfied or the desired learning performance is obtained.
      - Disadvantage: The search space for d features is 2^d, which makes the exhaustive search impractical when d is large.

Intro

- Traditional Categorizations of Feature Selection Algorithms
  - Search Strategy Perspective
    - Filter Method
      - Independent of any learning algorithms.
      - Typically more efficient than wrapper methods.
      - Rely on certain characteristics of data to assess the importance of features.
      - (1) Feature importance is ranked by a feature score according to some feature evaluation criteria.
      - (2) Low ranking features are filtered out and the remaining features are selected.
      - Disadvantage: Due to the lack of a specific learning algorithm guiding the feature selection phase, the selected features may not be optimal for the target learning algorithms.

Intro

- Traditional Categorizations of Feature Selection Algorithms
  - Search Strategy Perspective
    - Embedded Method
      - A trade-off solution between filter and wrapper methods which embed the feature selection with the model learning.
      - Include the interactions with the learning algorithm.
      - Far more efficient than the wrapper methods since there is no need to evaluate feature sets iteratively.
      - The most widely used embedded methods are the regularization models which targets to fit a learning model by minimizing the fitting errors and forcing the feature coefficients to be small (or exact zero) simultaneously.

Intro - Feature Selection Algorithms from A Data Perspective

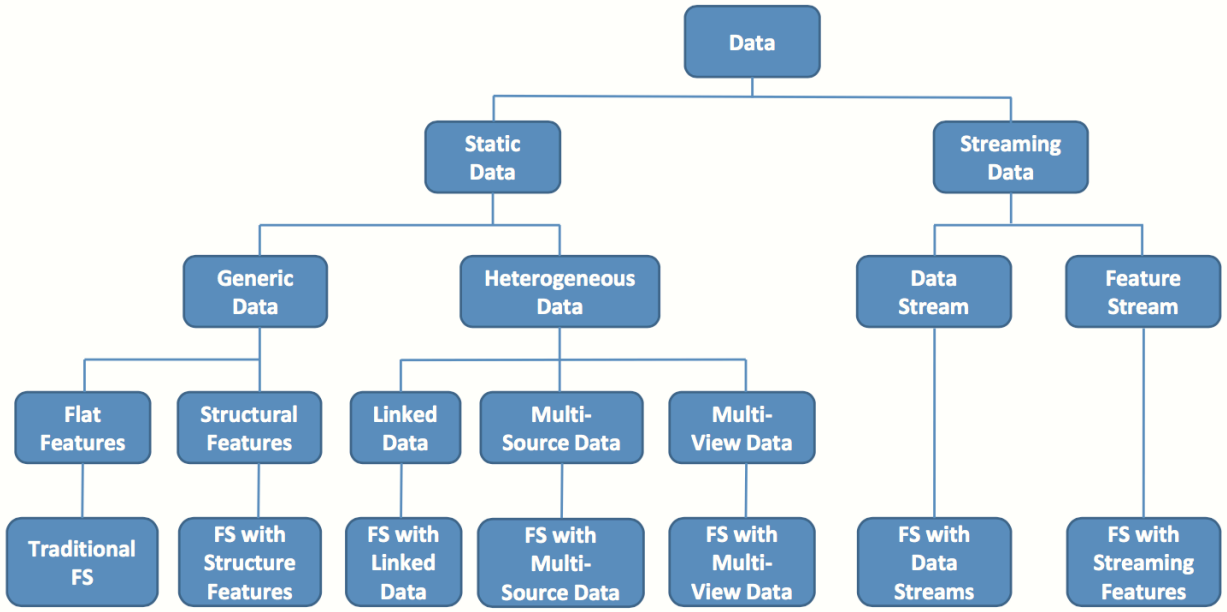


Figure 7: Feature Selection Algorithms from the Data Perspective.

Intro

- Feature Selection Algorithms from A Data Perspective
  - Streaming Data and Features
    - For example in Twitter, new data like posts and new features like slang words are continuously being generated. It is impractical to apply traditional batch-mode feature selection algorithms to find relevant features at each round when new data or new feature arrives.
  - Heterogeneous Data
    - Most existing algorithms of feature selection assume that the data is independent and identically distributed (i.i.d.). However, multi-source data is quite prevalent in many domains.
    - For instance, with the existence of link information, the widely adopted i.i.d. assumption in most machine learning algorithms does not hold. How to appropriately utilize link information for feature selection is still a challenging problem.

Intro

- Feature Selection Algorithms from A Data Perspective
  - Structures Between Features
    - Some well-known structures among features are group structure, tree structure, graph structure, etc.
    - Incorporating the prior knowledge of feature structures can possibly help select relevant features to greatly improve the learning performance.

Intro

- A feature selection repository in Python named ***scikit-feast*** which is built upon the widely used machine learning package ***scikit- learn*** and two scientific computing packages Numpy and Scipy.
- Includes more than 40 representative feature selection algorithms.
- The website of the repository: http://featureselection.asu.edu/scikit-feast/.

Organization of the Survey

1. Feature Selection with Generic Data (Section 2)
   a) Similarity based Feature Selection Methods
   b) Information Theoretical based Feature Selection Methods
   c) Sparse Learning based Feature Selection Methods
   d) Statistical based Feature Selection Methods
2. Feature Selection with Structure Features (Section 3)
3. Feature Selection with Heterogeneous Data (Section 4)
4. Feature Selection with Streaming Data (Section 5)
5. Performance Evaluation (Section 6)
6. Open Problems and Challenges (Section 7)
7. Summary of the Survey (Section 8)

Section 2. Feature Selection on Generic Data

- Only involves filter methods and embedded methods while the wrapper methods are excluded.

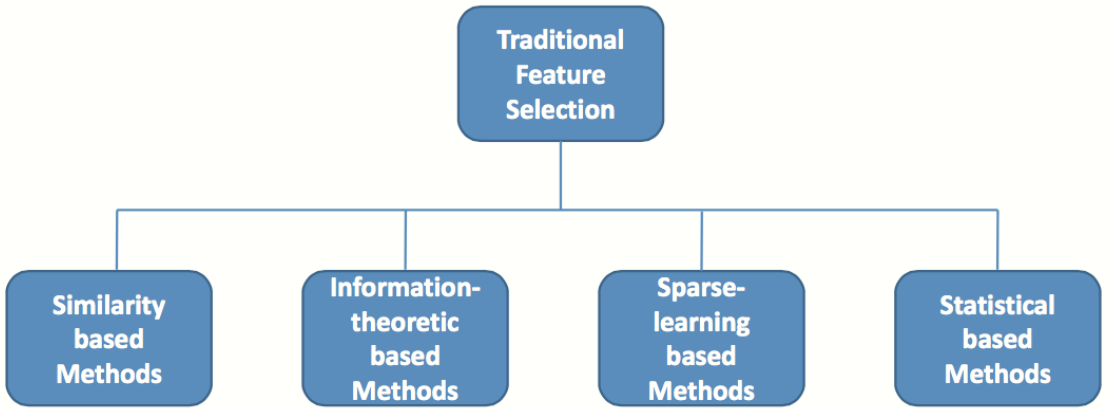


Figure 8: Categorization of Traditional Feature Selection Algorithms According to The Adopted Techniques.

2.1 Similarity based Methods

- Different feature selection algorithms exploit various types of criteria to define the relevance of features
  - such as distance, separability, information, correlation, dependency, and reconstruction error.
- Among them, there is a family of methods assessing the importance of features by its ability to preserve data similarity.

Notation

Notations	Definitions or Descriptions
n	number of instances in the data
d	number of features in the data
k	number of selected features
c	number of classes (if exist)
$\mathcal{F}$	original feature set which contains d features
$\mathcal{S}$	selected feature set which contains k selected features
$\{i_1, i_2, ..., i_k\}$	index of k selected features in $\mathcal{S}$
$f_1, f_2, ..., f_d$	d features
$f_{i_1}, f_{i_2}, ..., f_{i_k}$	k selected features
$x_1, x_2, ..., x_n$	n data instances
$\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_d$	d feature vectors corresponding to $f_1, f_2, ..., f_d$
$\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, ..., \mathbf{f}_{i_k}$	k feature vectors corresponding to $f_{i_1}, f_{i_2}, ..., f_{i_k}$
$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$	n data vectors corresponding to $x_1, x_2, ..., x_n$
$y_1, y_2, ..., y_n$	class labels of all n instances (if exist)
$\mathbf{X} \in \mathbb{R}^{n \times d}$	data matrix with n instances and d features
$\mathbf{X}_{\mathcal{F}} \in \mathbb{R}^{n \times k}$	data matrix on the selected k features
$\mathbf{y} \in \mathbb{R}^n$	class label vector for all n instances (if exist)

Table 1: Symbols.

2.1 Similarity based Methods

- Given a dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$ with n instances and d features, the pairwise similarity among instances can be encoded in an affinity matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$.
- The affinity matrix $\mathbf{S}$ is symmetric and its (i, j)-th entry indicates the similarity between the i-th instance x_i and the j-th instance x_j, the larger the value of $\mathbf{S}_{i,j}$ is, the more similarity x_i and x_j share.

2.1 Similarity based Methods

- Suppose we want to select k most relevant features from $\mathcal{F}$, then the utility of these k features is maximized as
  - $\max_{\mathcal{F}} \sum_{f \in \mathcal{F}} SC(f) = \max_{\mathcal{F}} \sum_{\mathbf{f} \in \mathcal{F}} \widehat{\mathbf{f}}' \hat{\mathbf{S}} \hat{\mathbf{f}}$

  where SC is a function that measures the utility of feature $\mathbf{f}$

  $\widehat{\mathbf{f}}'$ are the normalized feature

  $\hat{\mathbf{S}}$ are refined affinity matrix obtained from $\mathbf{f}$ and $\mathbf{S}$
- We would select a subset of features from $\mathcal{F}$ such that they can well preserve the data similarity structures defined in $\hat{\mathbf{S}}$.
- Usually solved by greedily selecting the top k features that maximize their individual utility $\widehat{\mathbf{f}}' \hat{\mathbf{S}} \hat{\mathbf{f}}$. Methods in this category vary in the way the similarity matrix $\mathbf{S}$ is designed.

2.1.1 Laplacian Score (He et al., 2005) (Unsupervised)

- An unsupervised feature selection algorithm which selects features that can best preserve the data manifold structure.
- Three phases:
  - First, it constructs a nearest neighbor graph $\mathcal{G}$ with n nodes where the i-th node corresponds to x_i. If x_i is among the p nearest neighbors of x_j or x_j is among the p nearest neighbors of x_i, nodes i and j are connected in $\mathcal{G}$ (p is a predefined number).
  - Second, if nodes i and j are connected, the entry in the
  - affinity matrix $\mathbf{S}_{\mathrm{ij}}$ is $\mathbf{S}(i,j) = e^{-\frac{\left\|x_i - x_j\right\|^2}{t}}$, where t is a constant, otherwise $\mathbf{S}(i,j) = 0$.

2.1.1 Laplacian Score (He et al., 2005) (Unsupervised)

- The diagonal matrix $\mathbf{D}$ is defined as $\mathbf{D}(i,i) = \sum_{j=1}^{n} \mathbf{S}(i,j)$
- The laplacian matrix $\mathbf{L}$ is $\mathbf{L} = \mathbf{D} - \mathbf{S}$
- the Laplacian Score of each feature f_i
  - $laplacian\ score(f_i) = \frac{\widetilde{\mathbf{f}}_i' \mathbf{L} \widetilde{\mathbf{f}}_i}{\widetilde{\mathbf{f}}_i' \mathbf{D} \widetilde{\mathbf{f}}_i}$, where $\widetilde{\mathbf{f}}_i = \mathbf{f}_i - \frac{\mathbf{f}_i' \mathbf{D} \mathbf{1}}{\mathbf{1}' \mathbf{D} \mathbf{1}} \mathbf{1}$

- Laplacian Score evaluates the importance of each feature individually, the task of selecting the k features can be solved by greedily picking the top k features with the **smallest** Laplacian Scores.

Justification

- A ”good” feature should the one on which two data points are close to each other if and only if there is an edge between these two points.

$$L_r = \frac{\sum_{ij}(f_{ri} - f_{rj})^2 S_{ij}}{Var(\mathbf{f}_r)}$$

- Var($\mathbf{f}_r$) is the estimated variance of the r-th feature

$$\sum_{ij}(f_{ri} - f_{rj})^2 S_{ij} = \sum_{ij}\left(f_{ri}^2 + f_{rj}^2 - 2f_{ri}f_{rj}\right) S_{ij}$$

$$= \quad 2\sum_{ij} f_{ri}^2 S_{ij} - 2\sum_{ij} f_{ri} S_{ij} f_{rj} = 2\mathbf{f}_r^T D\mathbf{f}_r - 2\mathbf{f}_r^T S\mathbf{f}_r = 2\mathbf{f}_r^T L\mathbf{f}_r$$

$$Var(\mathbf{f}_r) = \sum_i (\mathbf{f}_{ri} - \mu_r)^2 D_{ii}$$

$$\mu_r = \sum_i \left(\mathbf{f}_{ri} \frac{D_{ii}}{\sum_i D_{ii}}\right) = \frac{1}{\left(\sum_i D_{ii}\right)} \left(\sum_i \mathbf{f}_{ri} D_{ii}\right) = \frac{\mathbf{f}_r^T D\mathbf{1}}{\mathbf{1}^T D\mathbf{1}}$$

2.1.1 Laplacian Score (He et al., 2005) (Unsupervised)

- Reformulate
  - $laplacian\ score(f_i) = \frac{\widetilde{\mathbf{f}}_i' \mathbf{L} \widetilde{\mathbf{f}}_i}{\widetilde{\mathbf{f}}_i' \mathbf{D} \widetilde{\mathbf{f}}_i} = \frac{\widetilde{\mathbf{f}}_i' (\mathbf{D}-\mathbf{S}) \widetilde{\mathbf{f}}_i}{\widetilde{\mathbf{f}}_i' \mathbf{D} \widetilde{\mathbf{f}}_i} = 1 - \frac{\widetilde{\mathbf{f}}_i' \mathbf{S} \widetilde{\mathbf{f}}_i}{\widetilde{\mathbf{f}}_i' \mathbf{D} \widetilde{\mathbf{f}}_i}$

$$= 1 - \left(\frac{\widetilde{\mathbf{f}}_i}{\left\| \mathbf{D}^{\frac{1}{2}} \widetilde{\mathbf{f}}_i \right\|} \right)' \mathbf{S} \left(\frac{\widetilde{\mathbf{f}}_i}{\left\| \mathbf{D}^{\frac{1}{2}} \widetilde{\mathbf{f}}_i \right\|} \right)$$

  - $\widetilde{\mathbf{f}}_i' \mathbf{D} \widetilde{\mathbf{f}}_i$ is the weighted data variance of feature f_i (denoted as σ_i^2)
  - $\left\| \mathbf{D}^{\frac{1}{2}} \widetilde{\mathbf{f}}_i \right\|$ is the standard data variance (denoted as σ_i)
  - $\frac{\widetilde{\mathbf{f}}_i}{\left\| \mathbf{D}^{\frac{1}{2}} \widetilde{\mathbf{f}}_i \right\|}$ is interpreted as a normalized feature vector $\widehat{\mathbf{f}}_i = \frac{\mathbf{f}_i - \mu_i \mathbf{1}}{\sigma}$

- Therefore, Laplacian Score feature selection can be reformulated by maximizing the following:

$$\max_{\mathcal{F}} \sum_{\mathbf{f} \in \mathcal{F}} \widehat{\mathbf{f}}' \hat{\mathbf{S}} \hat{\mathbf{f}}$$

2.1.2 SPEC (Zhao and Liu, 2007) (Unsupervised and Supervised)

- An extension of Laplacian Score that work for both supervised and unsupervised scenarios.
- In the **unsupervised** scenario, without label information, the data similarity is measured by the RBF kernel function:

$$\mathbf{S}(i,j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma^2}}$$

- In the **supervised** scenario, using label information, data similarity can be defined by:

$$\mathbf{S}(i,j) = \begin{cases} \frac{1}{n_l}, & \text{if } y_i = y_j = 1 \\ 0, & \text{otherwise} \end{cases}$$

where n_l is the number of data samples in the class l

Other methods

- Fisher score
- Trace Ratio Criterion
- ReliefF
- information gain

Q & A

Thanks.