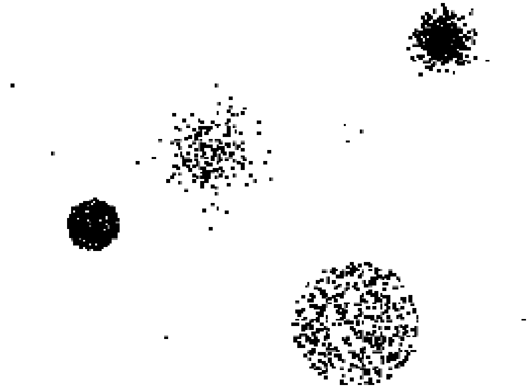
Anomaly Detection

Lecture Notes for Chapter 9

Introduction to Data Mining

Anomaly/Outlier Detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data



- Natural implication is that anomali rare
  - One in a thousand occurs often if you have lots of data
  - Context is important, e.g., freezing temps in July

- Can be important or a nuisance
  - 10 foot tall 2 year old
  - Unusually high blood pressure

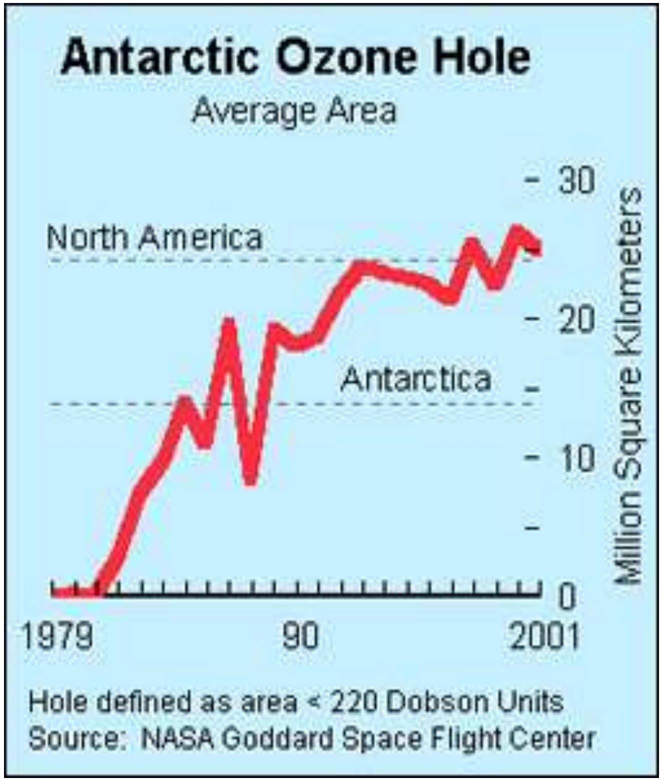
Importance of Anomaly Detection

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels

- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?

- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

Causes of Anomalies

- Data from different classes
  - Measuring the weights of oranges, but a few grapefruit are mixed in
- Natural variation
  - Unusually tall people
- Data errors
  - 200 pound 2 year old

Distinction Between Noise and Anomalies

- Noise is erroneous, perhaps random, values or contaminating objects
  - Weight recorded incorrectly
  - Grapefruit mixed in with the oranges

- Noise doesn't necessarily produce unusual values or objects

- Noise is not interesting

- Anomalies may be interesting if they are not a result of noise

- Noise and anomalies are related but distinct concepts

General Issues: Number of Attributes

- Many anomalies are defined in terms of a single attribute
  - Height
  - Shape
  - Color

- Can be hard to find an anomaly using all attributes
  - Noisy or irrelevant attributes
  - Object is only anomalous with respect to some attributes

- However, an object may not be anomalous in any one attribute

General Issues: Anomaly Scoring

- Many anomaly detection techniques provide only a binary categorization
  - An object is an anomaly or it isn't
  - This is especially true of classification-based approaches
- Other approaches assign a score to all points
  - This score measures the degree to which an object is an anomaly
  - This allows objects to be ranked
- In the end, you often need a binary decision
  - Should this credit card transaction be flagged?
  - Still useful to have a score
- How many anomalies are there?

Other Issues for Anomaly Detection

- Find all anomalies at once or one at a time
  - Swamping
  - Masking
- Evaluation
  - How do you measure performance?
  - Supervised vs. unsupervised situations
- Efficiency
- Context
  - Professional basketball team

Variants of Anomaly Detection Problems

- Given a data set D, find all data points **x** $\in$ D with anomaly scores greater than some threshold t

- Given a data set D, find all data points **x** $\in$ D having the top-n largest anomaly scores

- Given a data set D, containing mostly normal (but unlabeled) data points, and a test point **x**, compute the anomaly score of **x** with respect to D

Model-Based Anomaly Detection

- Build a model for the data and see
  - Unsupervised
    - Anomalies are those points that don’t fit well
    - Anomalies are those points that distort the model
    - Examples:
      - Statistical distribution
      - Clusters
      - Regression
      - Geometric
      - Graph
  - Supervised
    - Anomalies are regarded as a rare class
    - Need to have training data

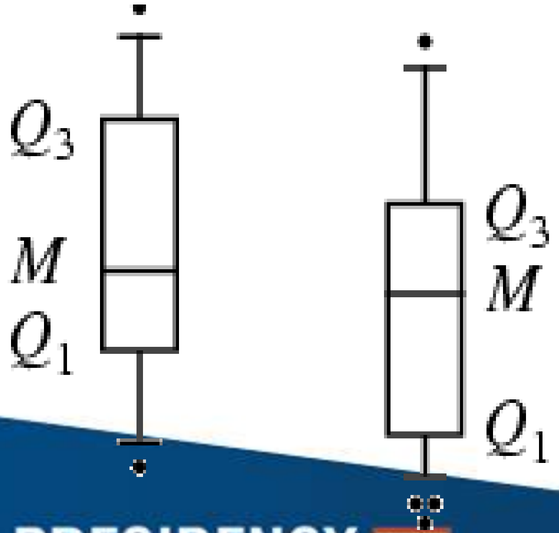
Additional Anomaly Detection Techniques

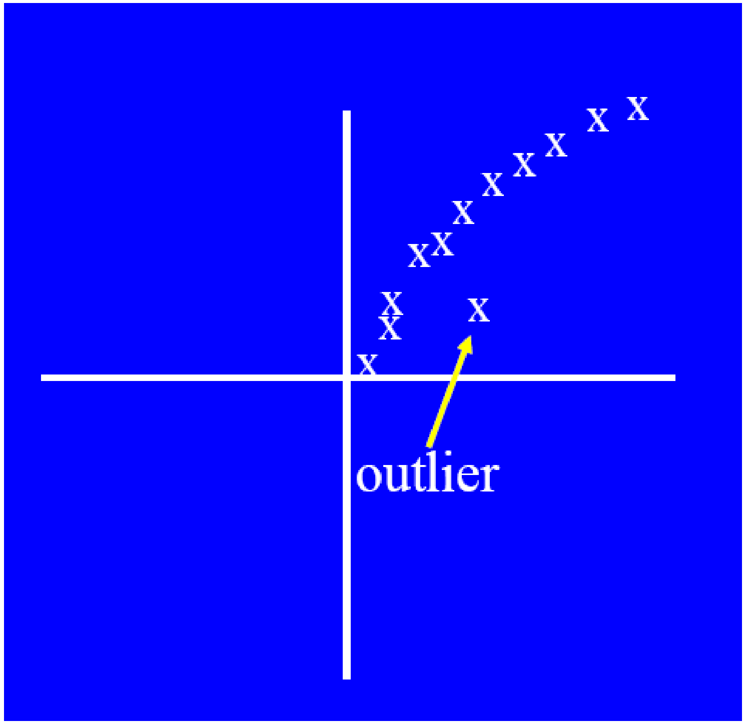
- Proximity-based
  - Anomalies are points far away from other points
  - Can detect this graphically in some cases
- Density-based
  - Low density points are outliers
- Pattern matching
  - Create profiles or templates of atypical but important events or objects
  - Algorithms to detect these patterns are usually simple and efficient

Visual Approaches

- Boxplots or scatter plots

- Limitations
  - Not automatic
  - Subjective



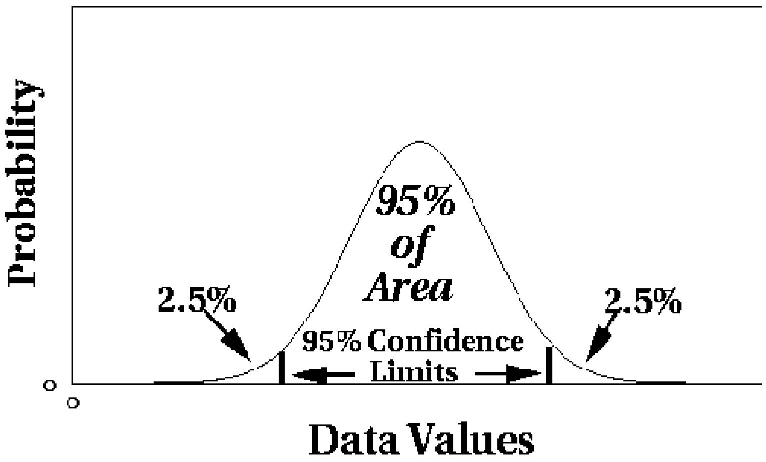


Statistical Approaches

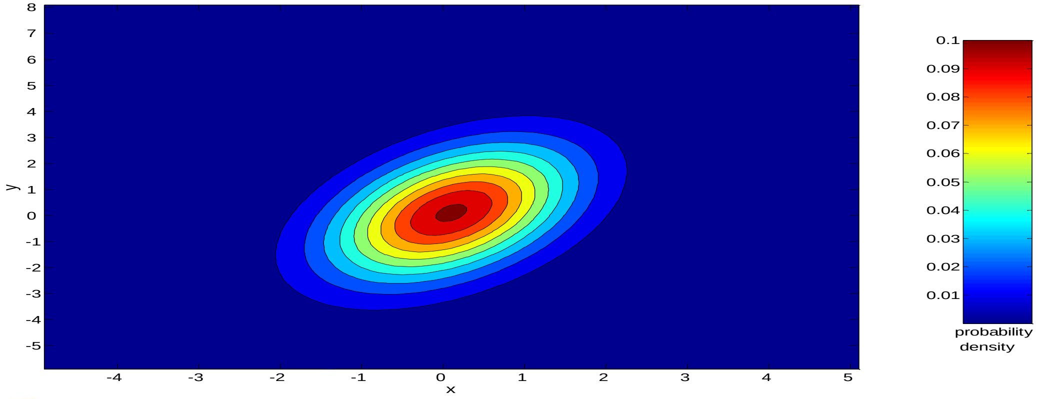
Probabilistic definition of an outlier: An outlier is an object that has a low probability with respect to a probability distribution model of the data.

- Usually assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on
  - Data distribution
  - Parameters of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)
- Issues
  - Identifying the distribution of a data set
    - Heavy tailed distribution
  - Number of attributes
  - Is the data a mixture of distributions?

Normal Distributions



One-dimensional Gaussian



Two-dimensional Gaussian

Grubbs' Test

- Detect outliers in univariate data
- Assume data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat
  - H_0: There is no outlier in data
  - H_A: There is at least one outlier
- Grubbs' test statistic:

$$G = \frac{\max \left| X - \overline{X} \right|}{s}$$

- Reject H_0 if:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N, N-2)}}{N - 2 + t^2_{(\alpha/N, N-2)}}}$$

Statistical-based – Likelihood Approach

- Assume the data set D contains samples from a mixture of two probability distributions:
  - M (majority distribution)
  - A (anomalous distribution)
- General Approach:
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log likelihood of D at time t
  - For each point x_t that belongs to M, move it to A
    - Let L_{t+1} (D) be the new log likelihood.
    - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
    - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Statistical-based – Likelihood Approach

- Data distribution, D = (1 – λ) M + λ A
- M is a probability distribution estimated from data
  - Can be based on any modeling method (naïve Bayes, maximum entropy, etc)
- A is initially assumed to be uniform distribution
- Likelihood at time t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

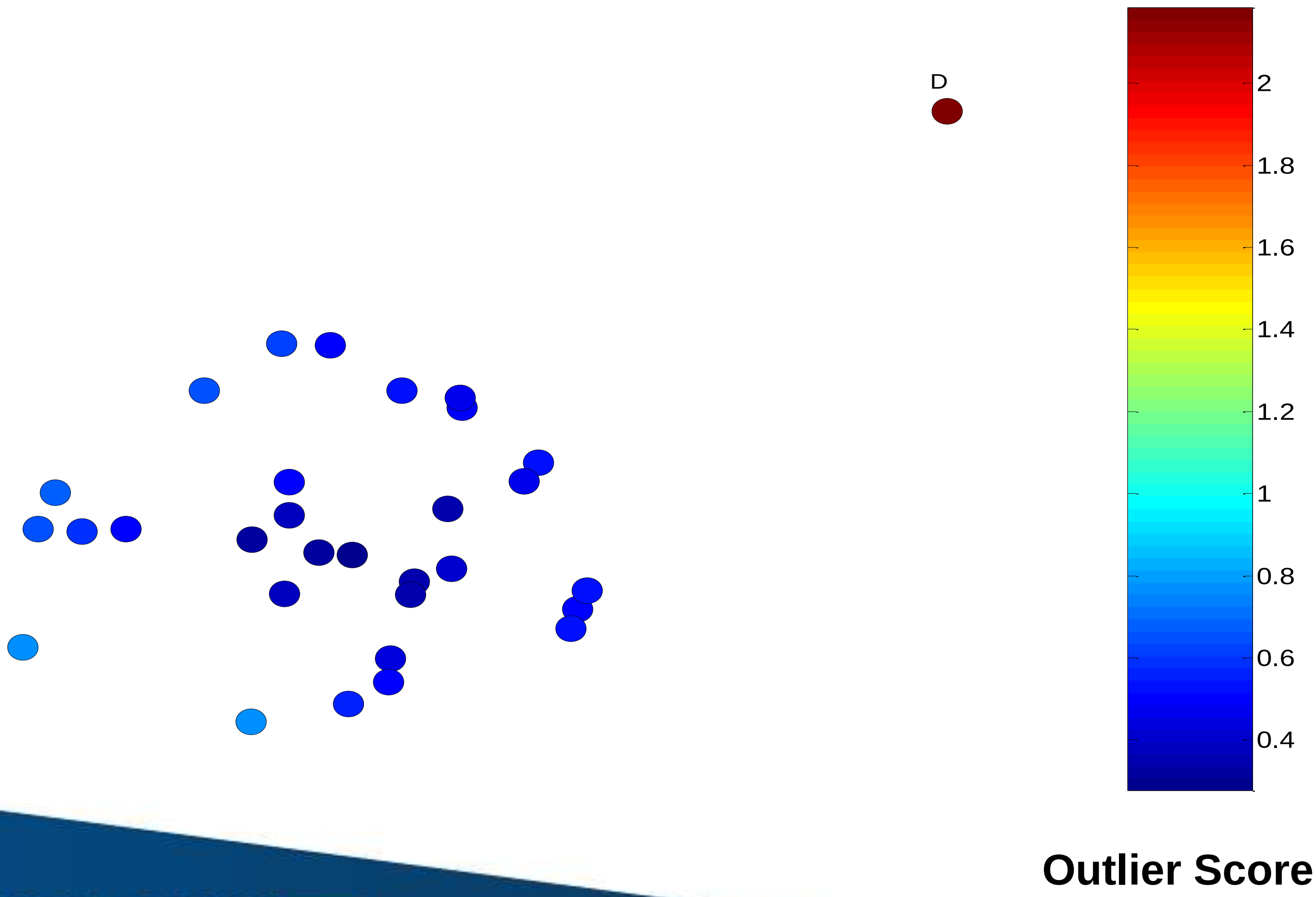
Strengths/Weaknesses of Statistical Approaches

- Firm mathematical foundation
- Can be very efficient
- Good results if distribution is known
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution
- Anomalies can distort the parameters of the distribution

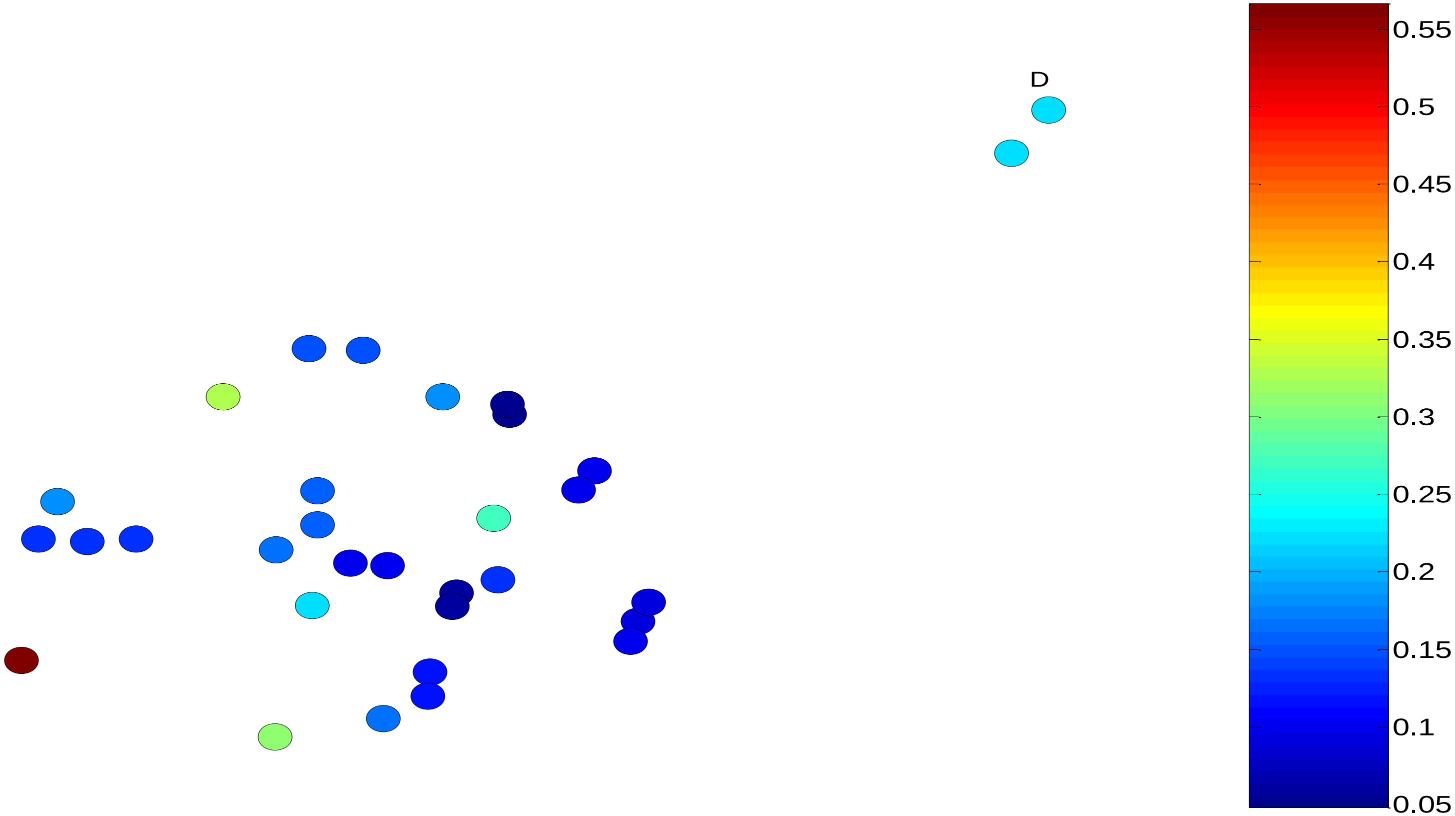
Distance-Based Approaches

- Several different techniques
- An object is an outlier if a specified fraction of the objects is more than a specified distance away (Knorr, Ng 1998)
  - Some statistical definitions are special cases of this
- The outlier score of an object is the distance to its kth nearest neighbor

One Nearest Neighbor - One Outlier

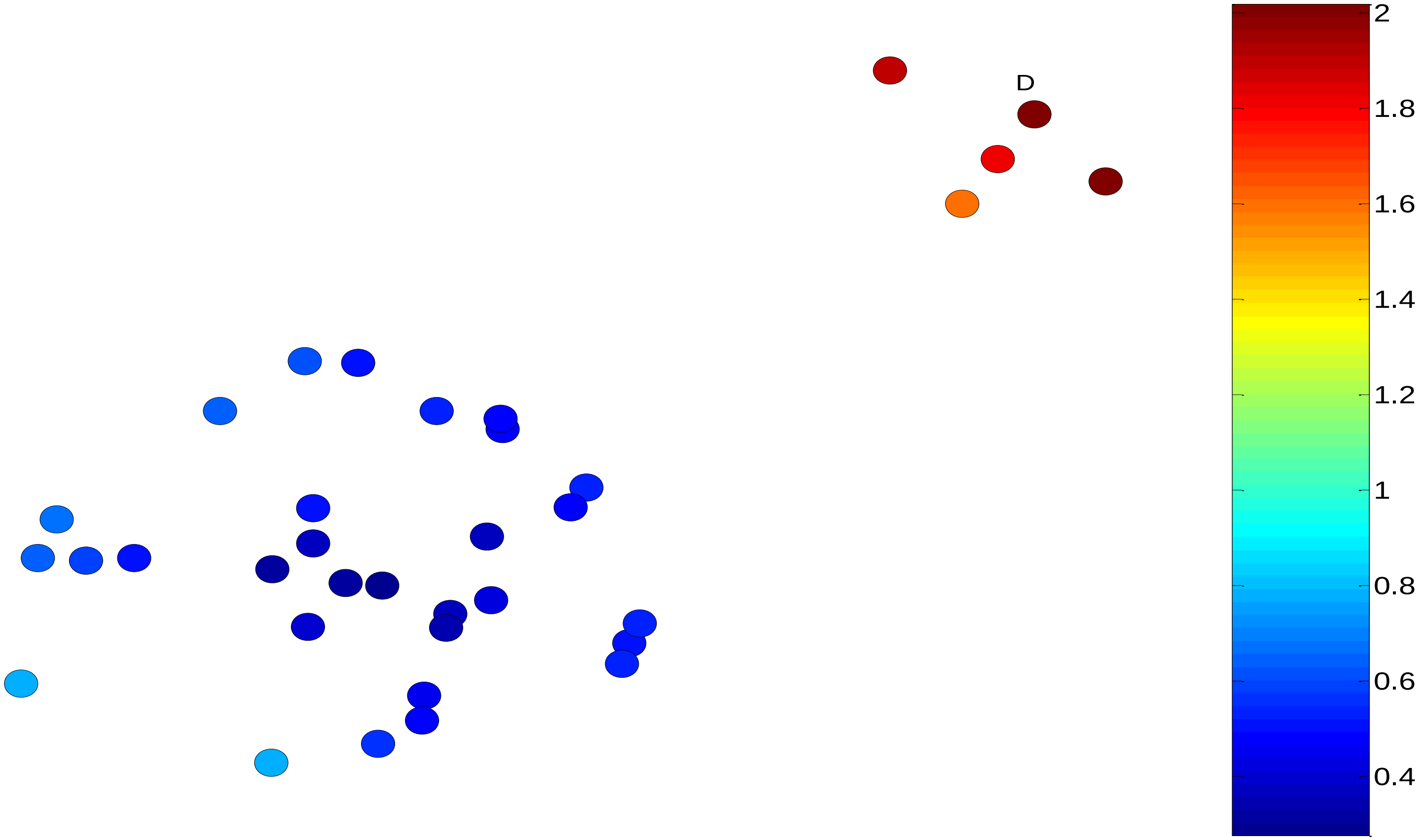


One Nearest Neighbor - Two Outliers



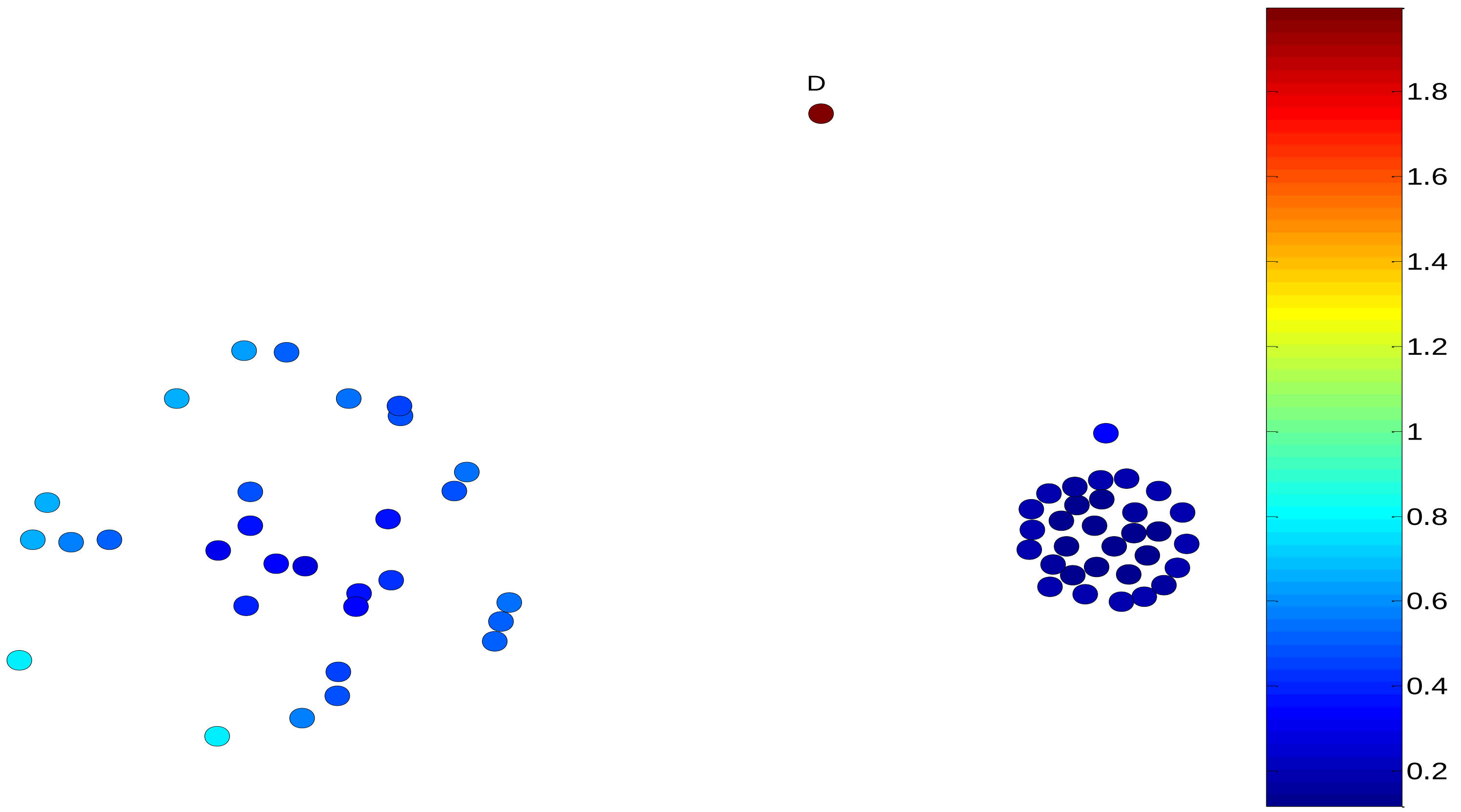
Outlier Score

Five Nearest Neighbors - Small Cluster



Outlier Score

Five Nearest Neighbors - Differing Density



Outlier Score

Strengths/Weaknesses of Distance-Based Approaches

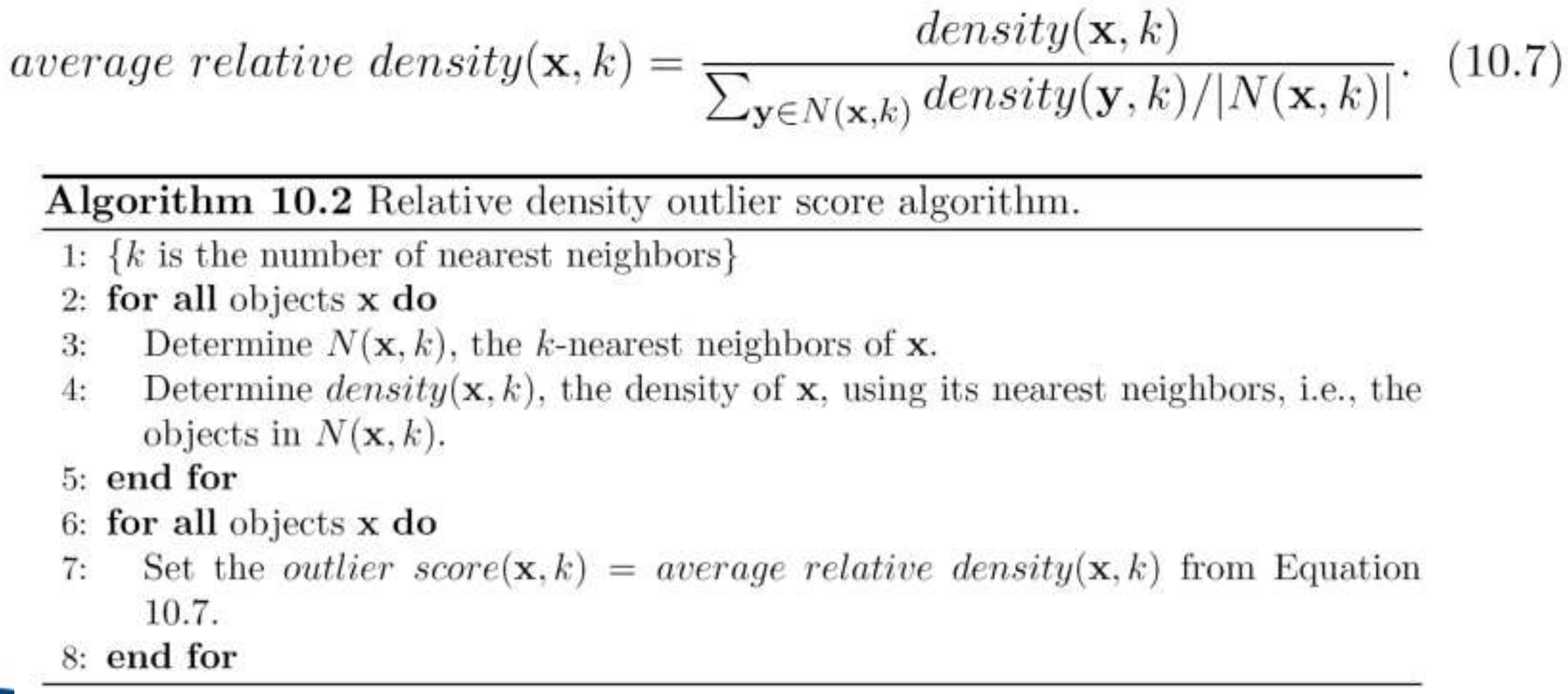
- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Sensitive to variations in density
- Distance becomes less meaningful in high-dimensional space

Density-Based Approaches

- **Density-based Outlier:** The outlier score of an object is the inverse of the density around the object.
  - Can be defined in terms of the k nearest neighbors
  - One definition: Inverse of distance to kth neighbor
  - Another definition: Inverse of the average distance to k neighbors
  - DBSCAN definition
- If there are regions of different density, this approach can have problems

Relative Density

- Consider the density of a point relative to that of its k nearest neighbors

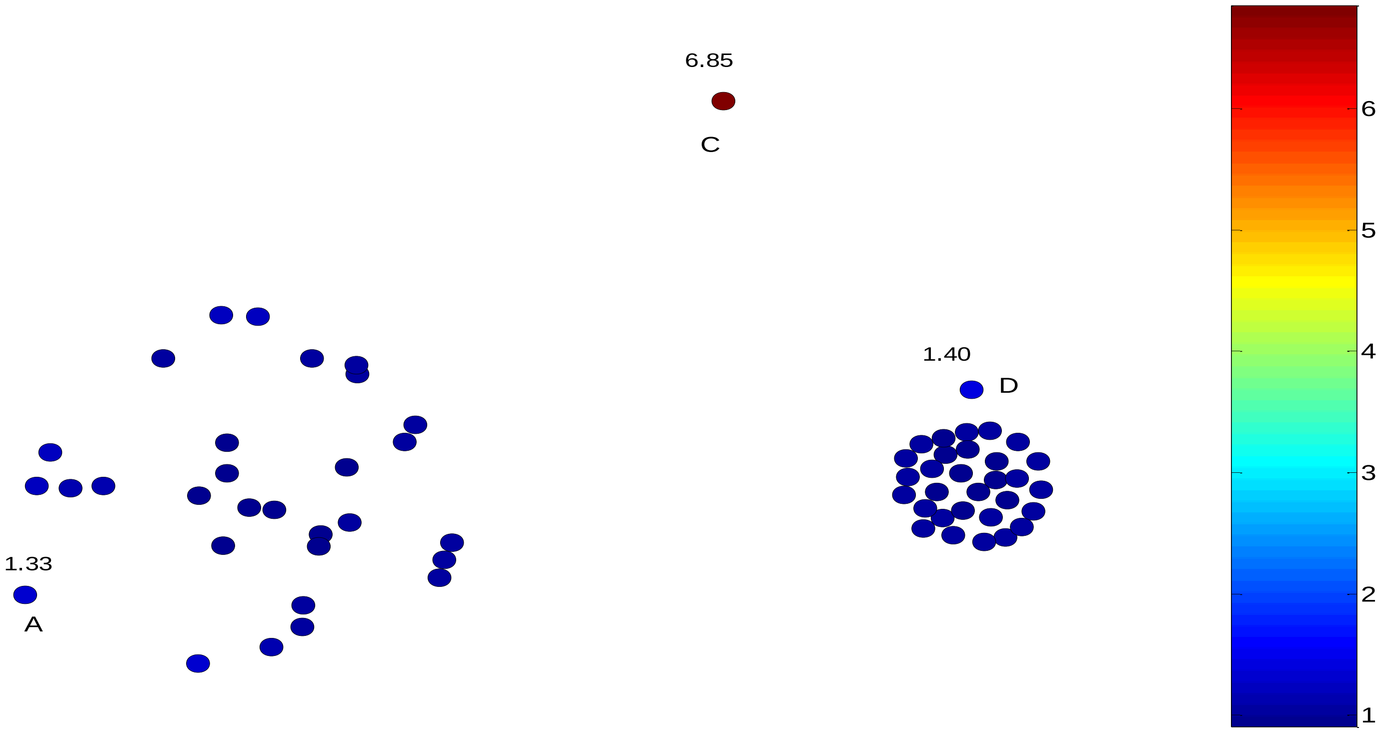


$$average\ relative\ density(\mathbf{x}, k) = \frac{density(\mathbf{x}, k)}{\sum_{\mathbf{y} \in N(\mathbf{x},k)} density(\mathbf{y}, k)/|N(\mathbf{x}, k)|}. \quad (10.7)$$

Algorithm 10.2 Relative density outlier score algorithm.

```
{$k$ is the number of nearest neighbors}
for all objects $\mathbf{x}$ do
    Determine $N(\mathbf{x}, k)$, the $k$-nearest neighbors of $\mathbf{x}$.
    Determine $density(\mathbf{x}, k)$, the density of $\mathbf{x}$, using its nearest neighbors, i.e., the
    objects in $N(\mathbf{x}, k)$.
end for
for all objects $\mathbf{x}$ do
    Set the $outlier\ score(\mathbf{x}, k) = average\ relative\ density(\mathbf{x}, k)$ from Equation
    10.7.
end for
```

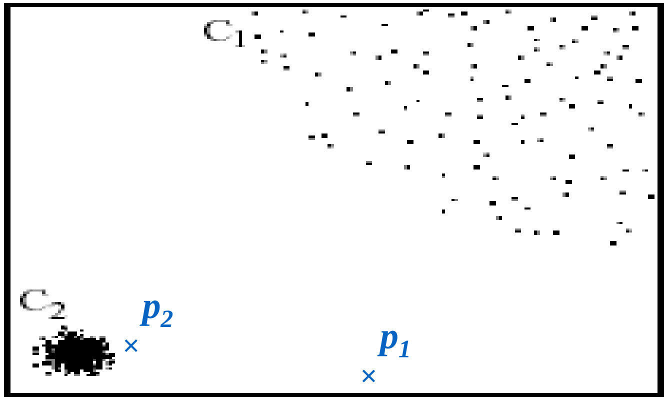
Relative Density Outlier Scores



Outlier Score

Density-based: LOF approach

- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors
- Outliers are points with largest LOF value



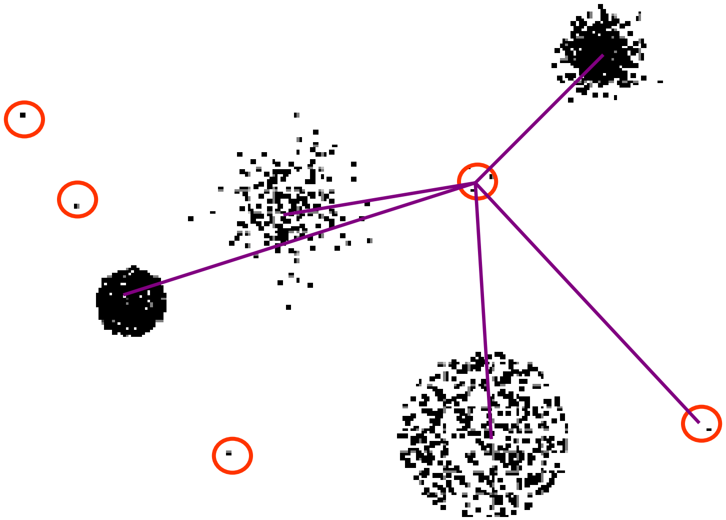
In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Strengths/Weaknesses of Density-Based Approaches

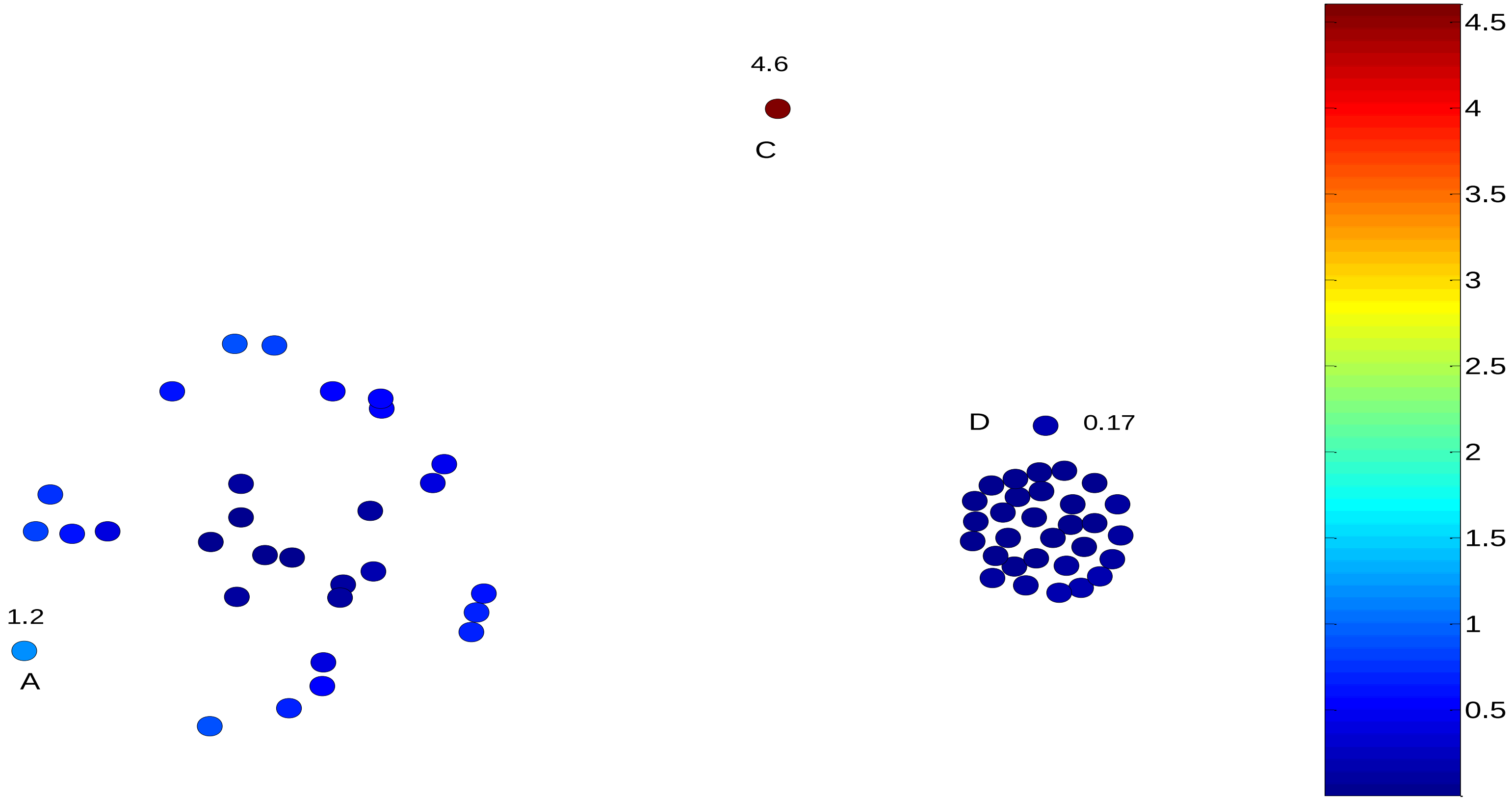
- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Density becomes less meaningful in high-dimensional space

Clustering-Based Approaches

- **Clustering-based Outlier:** An object is a cluster-based outlier if it does not strongly belong to any cluster
  - For prototype-based clusters, an object is an outlier if it is not close enough to a cluster center
  - For density-based clusters, an object is an outlier if its density is too low
  - For graph-based clusters, an object is an outlier if it is not well connected
- Other issues include the impact of outliers on the clusters and the number of clusters

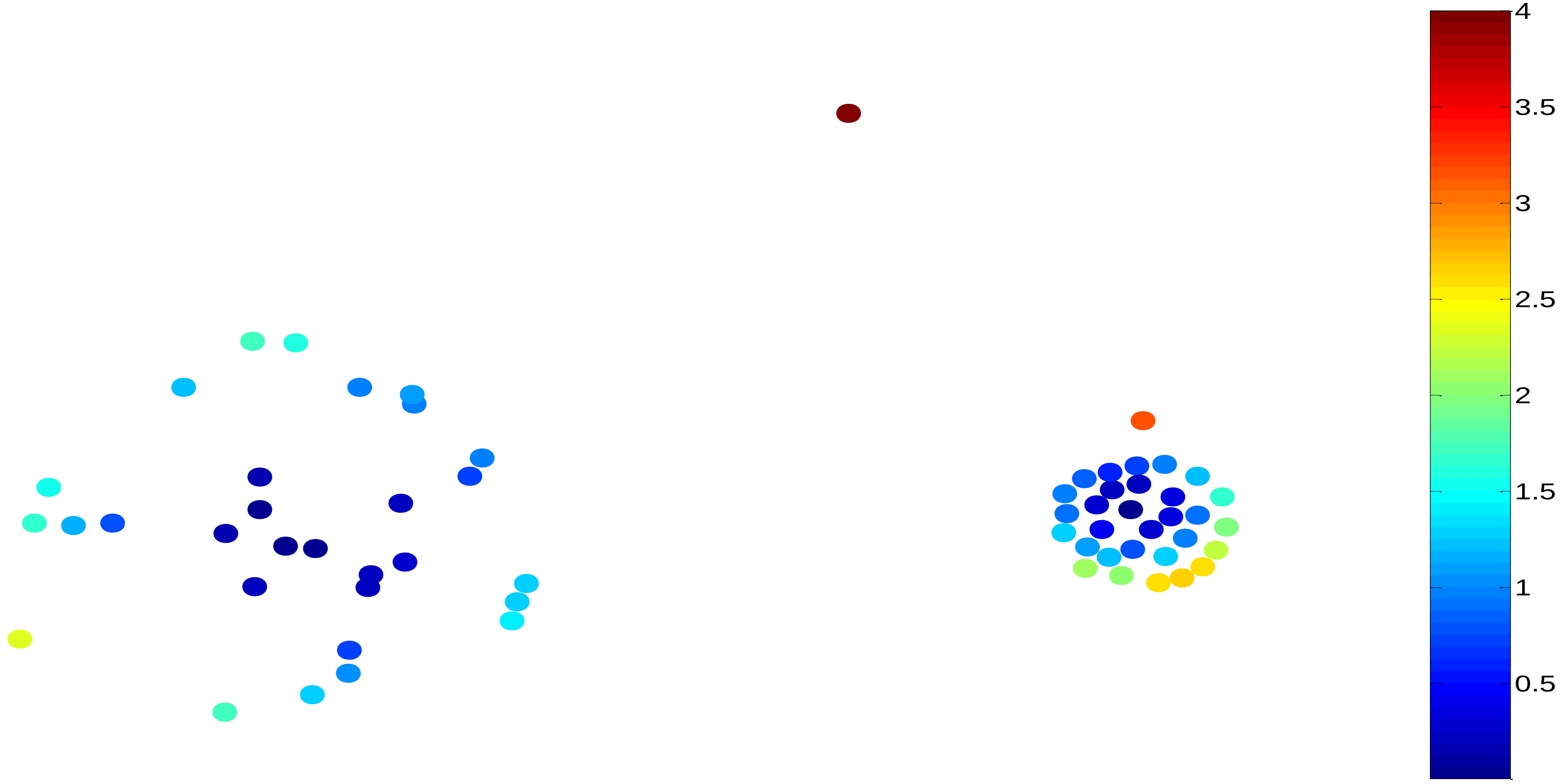


Distance of Points from Closest Centroids



Outlier Score

Relative Distance of Points from Closest Centroid



Outlier Score

Strengths/Weaknesses of Distance-Based Approaches

- Simple
- Many clustering techniques can be used
- Can be difficult to decide on a clustering technique
- Can be difficult to decide on number of clusters
- Outliers can distort the clusters

Thank You

Avoiding False Discoveries

Lecture Notes for Chapter 10

Introduction to Data Mining

Outline

- Statistical Background
- Significance Testing
- Hypothesis Testing
- Multiple Hypothesis Testing

Motivation

- An algorithm applied to a set of data will usually produce some result(s)
  - There have been claims that the results reported in more than 50% of published papers are false. (Ioannidis)
- Results may be a result of random variation
  - Any particular data set is a finite sample from a larger population
  - Often significant variation among instances in a data set or heterogeneity in the population
  - Unusual events or coincidences do happen, especially when looking at lots of events
  - For this and other reasons, results may not replicate, i.e., generalize to other samples of data
- Results may not have domain significance
  - Finding a difference that makes no difference

- Data scientists need to help ensure that results of data analysis are not false discoveries, i.e., not meaningful or reproducible

Statistical Testing

- Statistical approaches are used to help avoid many of these problems
- Statistics has well-developed procedures for evaluating the results of data analysis
  - Significance testing
  - Hypothesis testing
- Domain knowledge, careful data collection and preprocessing, and proper methodology are also important
  - Bias and poor quality data
  - Fishing for good results
  - Reporting how analysis was done
- Ultimate verification lies in the real world

Probability and Distributions

- Variables are characterized by a set of possible values
  - Called the domain of the variable
  - Examples:
    - True or False for binary variables
    - Subset of integers for variables that are counts, such as number of students in a class
    - Range of real numbers for variables such as weight or height
- A **probability distribution function** describes the relative frequency with which the values are observed
- Call a variable with a distribution a **random variable**

Probability and Distributions ..

- For a discrete variable we define a probability distribution by the relative frequency with which each value occurs
  - Let X be a variable that records the outcome flipping a fair coin: heads (1) or tails (0)
  - $P(X=1) = P(X=0) = 0.5$ (P stands for "probability")
  - If f is the distribution of X, $f(1) = f(0) = 0.5$
- Probability distribution function has the following properties
  - Minimum value 0, maximum value 1
  - Sums to 1, i.e., $\sum_{all\ values\ of\ X} f(X) = 1$

Binomial Distribution

- Number of heads in a sequence of n coin flips
  - Let R be the number of heads
  - R has a binomial distribution
  - $P(R = k) = \binom{n}{k} P(X = 1)^{k} P(X = 0)^{n-k}$
  - What is $P(R = k)$ given $n = 10$ and $P(X = 1) = 0.5$?

k	$P(R=k)$
0	0.001
1	0.01
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.01
10	0.001

Probability and Distributions ..

- For a continuous variable we define a probability distribution by using density function
  - Probability of any specific value is 0
  - Only intervals of values have non-zero probability
    - Examples: $P(X > 3)$, $P(X < -3)$, $P(-1 < X < 1)$
    - If f is the distribution of X, $P(X > 3) = \int_3^\infty f(X)\,dx$

- Probability density has the following properties
  - Minimum value 0
  - Integrates to 1, i.e., $\int_{-\infty}^{\infty} f(X) = 1$

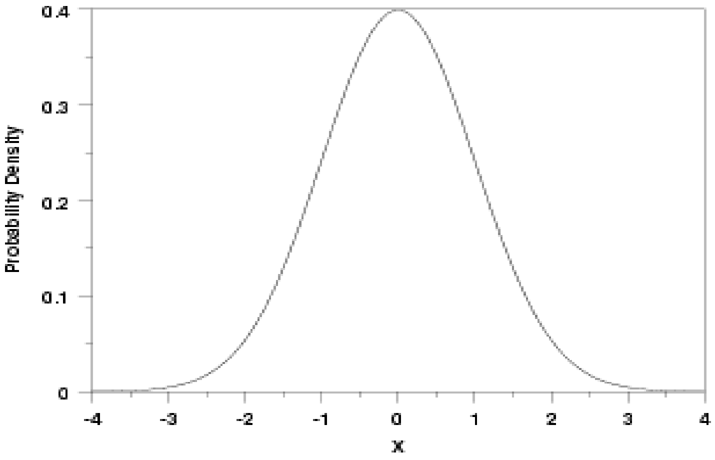
Gaussian Distribution

- The Gaussian (normal) distribution is the most commonly used
  - $f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - Where μ and σ are the mean and standard distribution of the distribution
  - $\mu = \int_{-\infty}^{\infty} X f(X) dx$ and $\sigma = \int_{-\infty}^{\infty} (X-\mu)^2 f(X) dx$

$\mu = 0$ and $\sigma = 1$, i.e., $\mathcal{N}(0,1)$

http://www.itl.nist.gov/div898/handbook/eda/section3/eda3661.htm

http://www.itl.nist.gov/div898/handbook/index.htm



Statistical Testing …

- Make inferences (decisions) about that validity of a result
- For statistical inference (testing), we need two thing:
  - A statement that we want to disprove
    - Called the **null hypothesis (H_0)**
    - The null hypothesis is typically a statement that the result is merely due to random variation
    - It is typically the opposite of what we would like to show
  - A random variable, *R*, called a **test statistic**, for which we know or can determine a distribution if H_0 is true.
    - The distribution of *R* under H0 is called the **null distribution**
    - The value of *R* is obtained from the result and is typically numeric

Examples of Null Hypotheses

- A coin or a die is a fair coin.
- The difference between the means of two samples is 0
- The purchase of a particular item in a store is unrelated to the purchase of a second item, e.g., the purchase of bread and milk are unconnected
- The accuracy of a classifier is no better than random

Significance Testing

- Significance testing was devised by the statistician Fisher
- Only interested in whether null hypothesis is true
- Significance testing was intended only for exploratory analyses of the null hypothesis in the preliminary stages of a study
  - For example, to refine the null hypothesis or modify future experiments
- For many years, significance testing has been a key approach for justifying the validity of scientific results
- Introduced the concept of p-value, which is widely used and misused

How Significance Testing Works

- Analyze the data to obtain a result
  - For example, data could be from flipping a coin 10 times to test its fairness
- The result is expressed as a value of the test statistic, R
  - For example, let R be the number of heads in 10 flips
- Compute the probability of seeing the current value of R or something more extreme
  - This probability is known as the **p-value** of the test statistic

How Significance Testing Works …

- If the p-value is sufficiently small, we reject the null hypothesis, H_0 and say that the result is statistically significant
    - We say we reject the null hypothesis, H_0
    - A threshold on the p-value is called the **significance level, α**
        - Often the significance level is 0.01 or 0.05
- If the p-value is **not** sufficiently small, we say that we fail to reject the null hypothesis
    - Sometimes we say that we accept the null hypothesis but a high p-value does not necessarily imply the null hypothesis is true

Example: Testing a coin for fairness

- H_0: $P(X=1) = P(X=0) = 0.5$
- Define the test statistic R to be the number of heads in 10 flips
- Set the significance level α to be 0.05
- The number of heads R has a binomial distribution
- For which values of R would you reject H_0?

k	$P(S=k)$
0	0.001
1	0.01
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.01
10	0.001

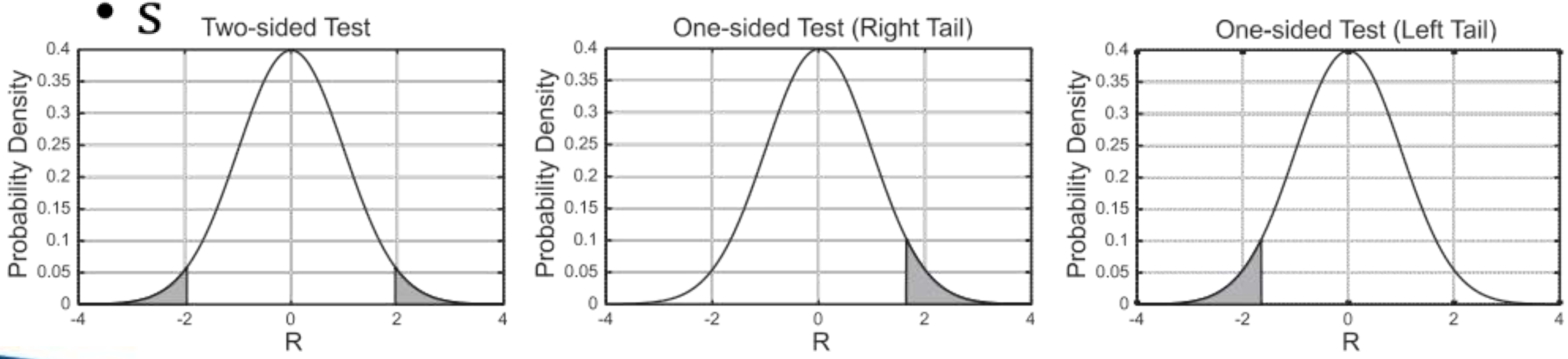
One-sided and Two-sided Tests

- More extreme can be interpreted in different ways
- For example, an observed value of the test statistic, R_{obs}, can be considered extreme if
  - it is greater than or equal to a certain value, R_H ,
  - smaller than or equal to a certain value, R_L, or
  - outside a specified interval, $[R_H, R_L]$.
- The first two cases are "one-sided tests" (right-tailed and left-tailed, respectively),

One-sided and Two-sided Tests …

- Example of one-tailed and two tailed tests for a test statistic *R* that is normally distributed for a roughly 5% significance level.

- *s*



Neyman-Pearson Hypothesis Testing

- Devised by statisticians Neyman and Pearson in response to perceived shortcomings in significance testing
  - Explicitly specifies an **alternative hypothesis, H_1**
  - Significance testing cannot quantify how an observed results supports H_1
  - Define an **alternative distribution** which is the distribution of the test statistic if H_1 is true
  - We define a **critical region** for the test statistic *R*
    - If the value of *R* falls in the critical region, we reject H_0
    - We may or may not accept H_1 if H_0 is rejected
  - The **significance level, α,** is the probability of the critical region under H_0

Hypothesis Testing …

- **Type I Error (α)**: Error of incorrectly rejecting the null hypothesis for a result.
  - It is equal to the probability of the critical region under H_0, i.e., is the same as the significance level, $\boldsymbol{\alpha}$.
  - Formally, $\alpha = P(R$ ⍰ Critical Region / $H_0)$

- **Type II Error (β)**: Error of falsely calling a result as not significant when the alternative hypothesis is true.
  - It is equal to the probability of observing test statistic values outside the critical region under H_1
  - Formally, $\beta = P(R \notin$ Critical Region / $H_1)$.

Hypothesis Testing …

- Power: which is the probability of the critical region under H_1, i.e., $1-\beta$.
    - Power indicates how effective a test will be at correctly rejecting the null hypothesis.
    - Low power means that many results that actually show the desired pattern or phenomenon will not be considered significant and thus will be missed.
    - Thus, if the power of a test is low, then it may not be appropriate to ignore results that fall outside the critical region.

Example: Classifying Medical Results

- The value of a blood test is used as the test statistic, R, to identify whether a patient has a particular disease or not.
  - H_0: For patients **not** having the disease, R has distribution $\mathcal{N}(40, 5)$
  - H_1: For patients having the disease, R has distribution $\mathcal{N}(60, 5)$

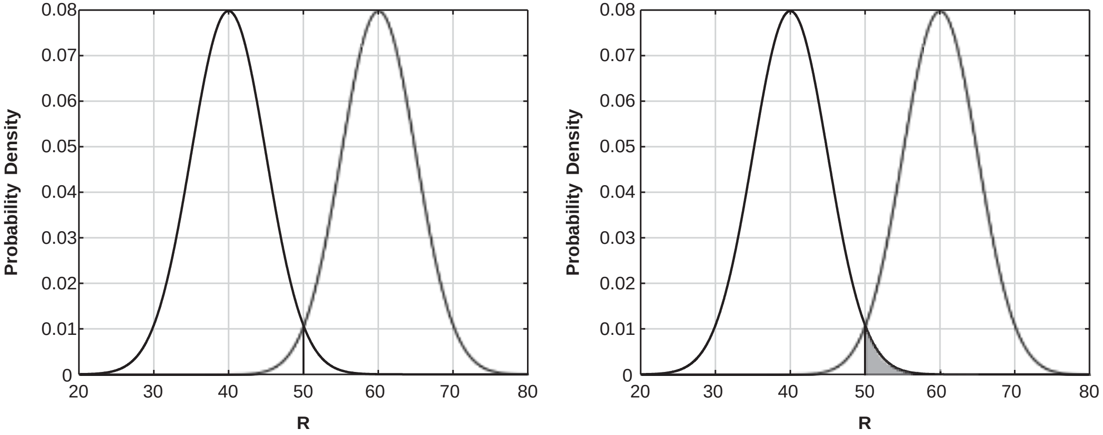
$$\alpha = \int_{50}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(R-u)^2}{2\sigma^2}}\, dR = \int_{50}^{\infty} \frac{1}{\sqrt{50\pi}} e^{-\frac{(R-40)^2}{50}}\, dR = 0.023,\ \mu = 40,\ \sigma = 5$$

$$\beta = \int_{-\infty}^{50} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(R-u)^2}{2\sigma^2}}\, dR = \int_{-\infty}^{50} \frac{1}{\sqrt{50\pi}} e^{-\frac{(R-60)^2}{50}}\, dR = 0.023,\ \mu = 60,\ \sigma = 5$$

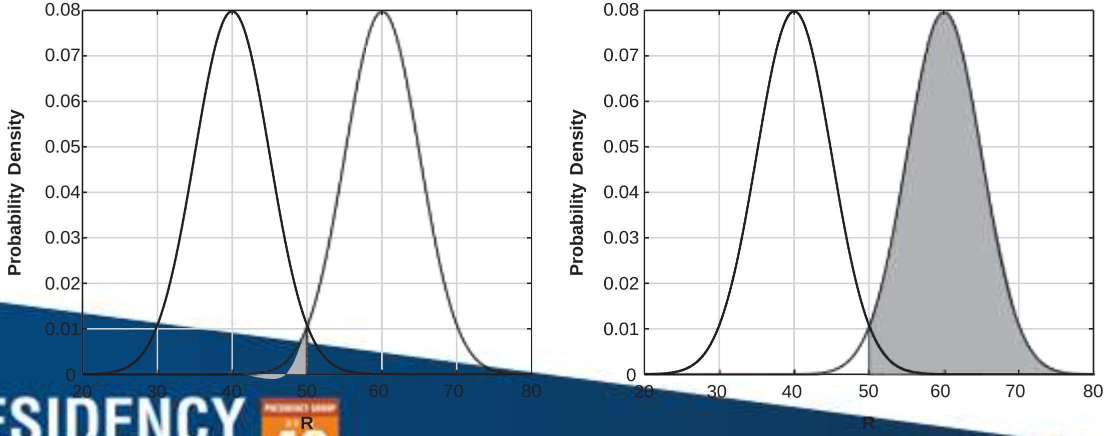
Power = 1 - β = 0.977

- See figures on the next page

α, β and Power for Medical Testing Exam



Distribution of test statistic for the alternative hypothesis (rightmost density curve) and null hypothesis (leftmost density curve). Shaded region in right subfigure is α.



Shaded region in left subfigure is β and shaded region in right subfigure is power.

Hypothesis Testing: Effect Size

- Many times we can find a result that is statistically significant but not significant from a domain point of view
  - A drug that lowers blood pressure by one percent
- **Effect size** measures the magnitude of the effect or characteristic being evaluated, and is often the magnitude of the test statistic.
  - Brings in domain considerations
- The desired effect size impacts the choice of the critical region, and thus the significance level and power of the test

Effect Size: Example Problem

- Consider several new treatments for a rare disease that have a particular probability of success. If we only have a sample size of 10 patients, what effect size will be needed to clearly distinguish a new treatment from the baseline which has is 60 % effective?

R/p(X=1)	0.60	0.70	0.80	0.90
0	0.0001	0.0000	0.0000	0.0000
1	0.0016	0.0001	0.0000	0.0000
2	0.0106	0.0014	0.0001	0.0000
3	0.0425	0.0090	0.0008	0.0000
4	0.1115	0.0368	0.0055	0.0001
5	0.2007	0.1029	0.0264	0.0015
6	0.2508	0.2001	0.0881	0.0112
7	0.2150	0.2668	0.2013	0.0574
8	0.1209	0.2335	0.3020	0.1937
9	0.0403	0.1211	0.2684	0.3874
10	0.0060	0.0282	0.1074	0.3487

Multiple Hypothesis Testing

- Arises when multiple results are produced and multiple statistical tests are performed
- The tests studied so far are for assessing the evidence for the null (and perhaps alternative) hypothesis for a single result
- A regular statistical test does not suffice
  - For example, getting 10 heads in a row for a fair coin is unlikely for one such experiment
    - probability $= \left(\frac{1}{2}\right)^{10} = 0.001$

Summarizing the Results of Multiple Tests

- The following confusion table defines how results of multiple tests are summarized
  - We assume the results fall into two classes, **+** and **–**, which, follow the alternative and null hypotheses, respectively.
  - The focus is typically on the number of false positives (FP), i.e., the results that belong to the null distribution (– class) but are declared significant (+ class).

Confusion table for summarizing multiple hypothesis testing results.

	Declared significant (+ prediction)	**Declared not significant (– prediction)**	**Total**
H_1 **True (actual +)**	True Positive (TP)	False Negative (FN) type II error	Positives (m_1)
H_0 **True (actual –)**	False Positive (FP) type I error	True Negative (TN)	Negatives (m_0)
	Positive Predictions (Ppred)	Negative Predictions (Npred)	m

Family-wise Error Rate

- By family, we mean a collection of related tests
- **family-wise error rate** (FWER) is the probability of observing even a single false positive (type I error) in an entire set of *m* results.
    - FWER = P(FP > 0).
- Suppose your significance level is 0.05 for a single test
    - Probability of no error for one test is $1 - 0.05 = 0.95$.
    - Probability of no error for *m* tests is 0.95^m
    - FWER = P(FP > 0) = 1 – 0.95^m
    - If *m* = 10, FWER = 0.60

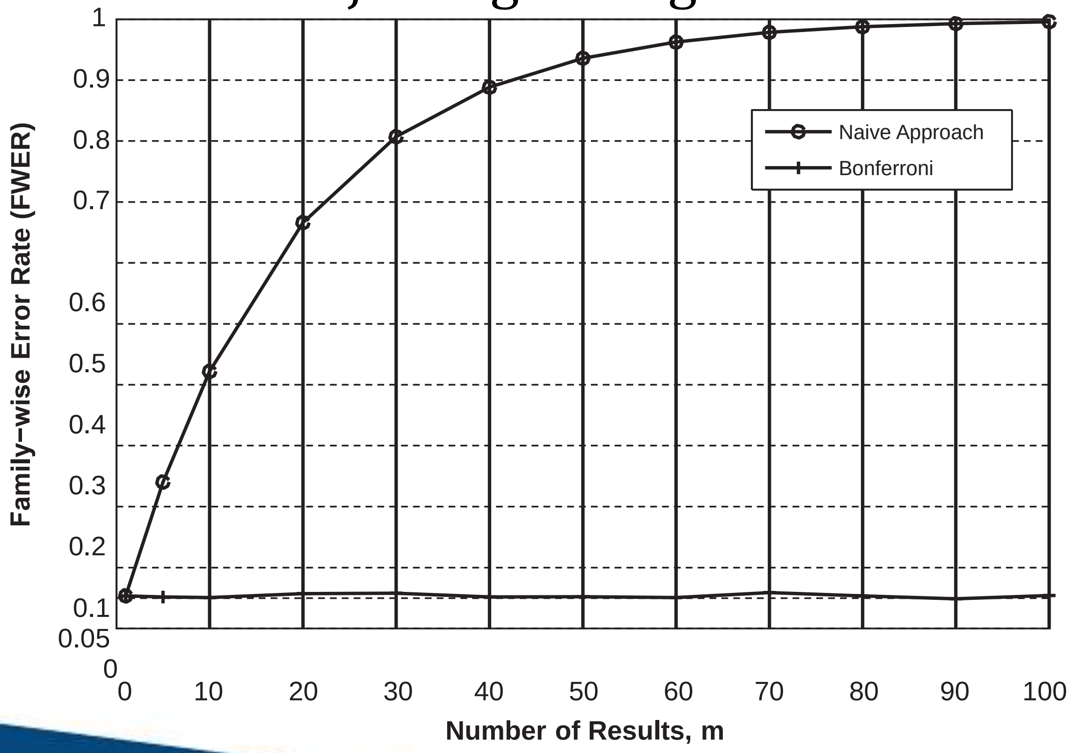
Bonferroni Procedure

- Goal of FWER is to ensure that FWER < α, where α is often 0.05
- Bonferroni Procedure:
  - m results are to be tested
  - Require FWER < α
  - set the significance level, α^* for every test to be $\alpha^* = \alpha/m$.

- If m = 10 and $\alpha = 0.05$ then $\alpha^* = 0.05/10{=}0.005$

Example: Bonferroni versus Naïve approach

- Naïve approach is to evaluate statistical significance for each result without adjusting the significance level.



The family wise error rate (FWER) curves for the naïve approach and the Bonferroni procedure as a function of the number of

False Discovery Rate

- FWER controlling procedures seek a low probability for obtaining any false positives
  - Not the appropriate tool when the goal is to allow some false positives in order to get more true positives
- The **false discovery rate (FDR)** measures the rate of false positives, which are also called false discoveries

$$Q = \frac{FP}{Ppred} = \frac{FP}{TP + FP} \text{ if } Ppred > 0$$

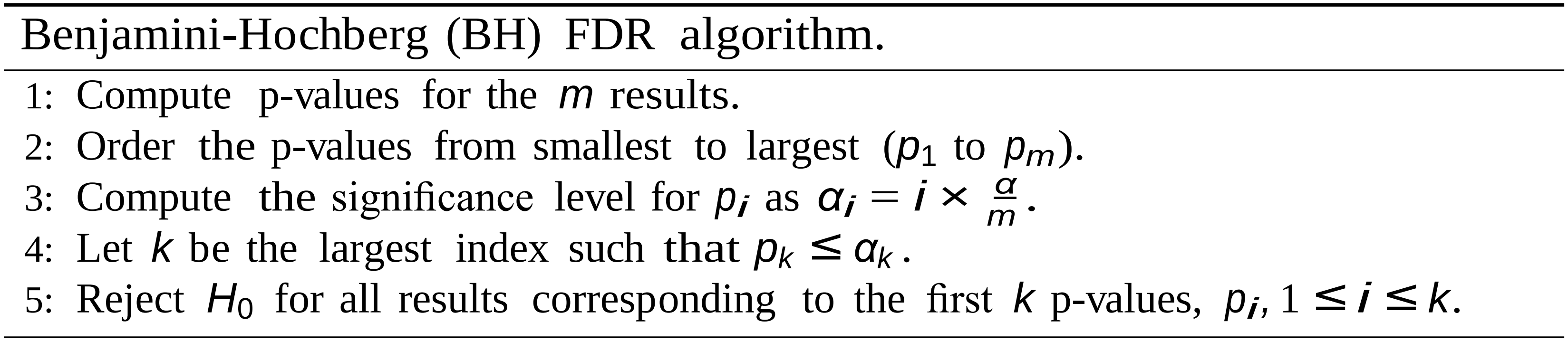
$$= 0 \text{ if } Ppred = 0,$$

where $Ppred$ is the number of predicted positives

- If we know FP, the number of actual false positives, then FDR = FP.
  - Typically we don't know FP in a testing situation
- Thus, FDR = Q P($Ppred$>0) = E(Q), the expected value of Q.

Benjamini-Hochberg Procedure

- An algorithm to control the false discovery rate (FDR)



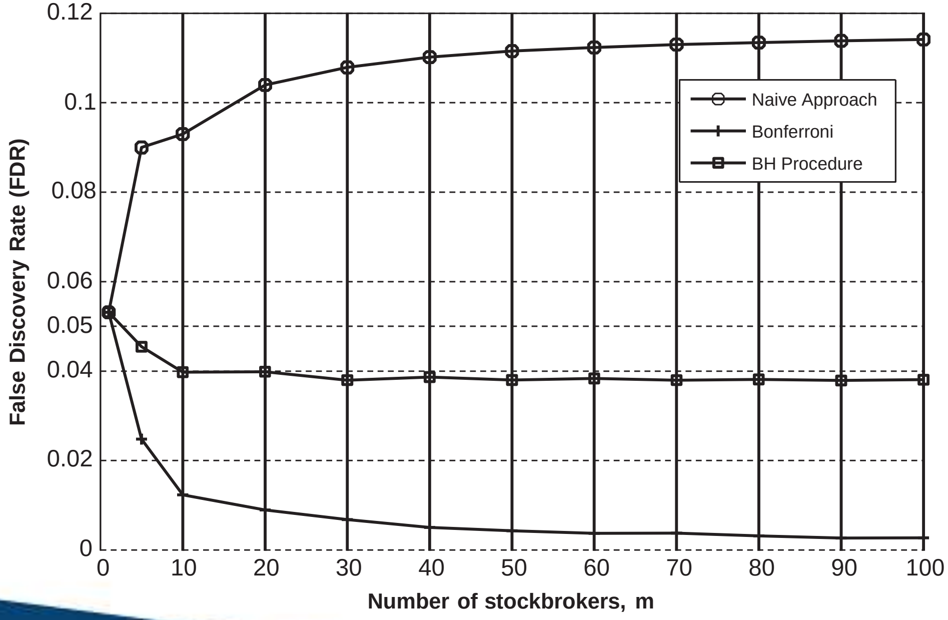
Benjamini-Hochberg (BH) FDR algorithm.

```
1: Compute p-values for the $m$ results.
2: Order the p-values from smallest to largest ($p_1$ to $p_m$).
3: Compute the significance level for $p_i$ as $\alpha_i = i \times \frac{\alpha}{m}$.
4: Let $k$ be the largest index such that $p_k \leq \alpha_k$.
5: Reject $H_0$ for all results corresponding to the first $k$ p-values, $p_i, 1 \leq i \leq k$.
```

- This procedure first orders the p-values from smallest to largest
- Then it uses a separate significance level for each test
  - $\alpha_i = i \times \frac{\alpha}{i}$

FDR Example: Picking a stockbroker

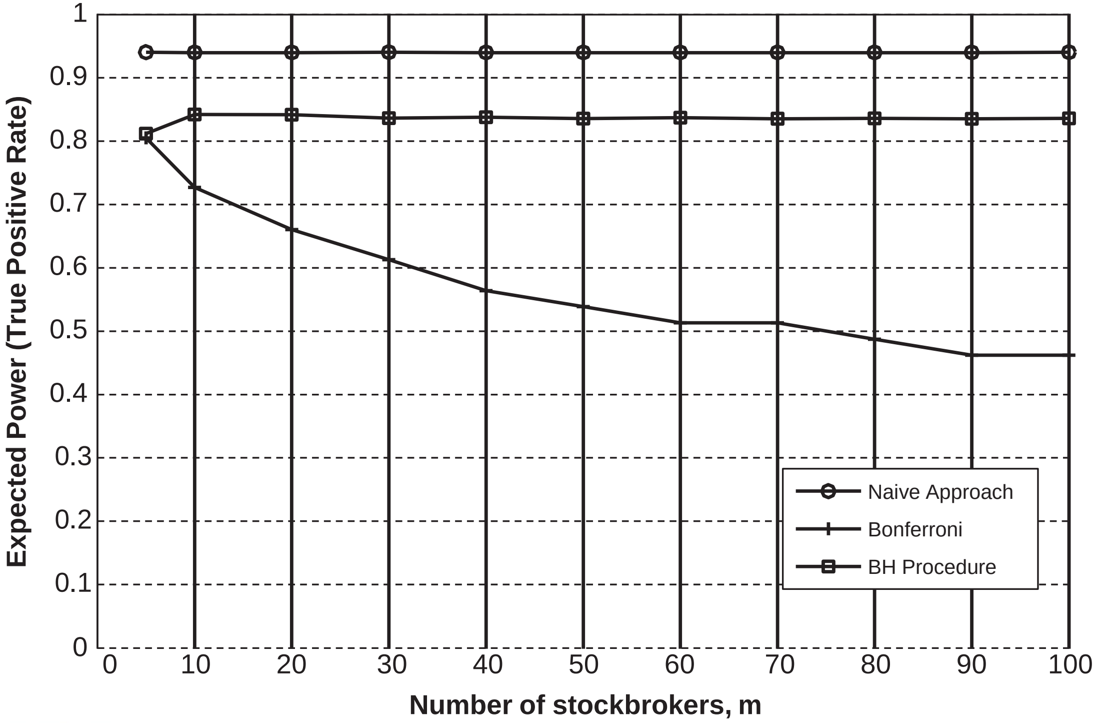
- Suppose we have a test for determining whether a stockbroker makes profitable stock picks. This test, applied to an individual stockbroker, has a significance level, $\alpha = 0.05$. We use the same value for our desired false discovery rate.
  - Normally, we set the desired FDR rate higher, e.g., 10% or 20%
- The following figure compares the naïve approach, Bonferroni, and the BH FDR procedure with respect to the false discovery rate for various numbers of tests, m. 1/3 of the sample were from the alternative distribution.



False Discovery Rate as a function of m.

FDR Example: Picking a stockbroker …

- The following figure compares the naïve approach, Bonferroni, and the BH FDR procedure with respect to the power for various numbers of tests, *m*. 1/3 of the sample were from the alternative distribution.



Expected Power as function of *m*.

Comparison of FWER and FDR

- FWER is appropriate when it is important to avoid any error.
  - But an FWER procedure such as Bonferroni makes many Type II errors and thus, has poor power.
  - An FWER approach has very a very false discovery rate
- FDR is appropriate when it is important to identity positive results, i.e., those belonging to the alternative distribution.
  - By construction, the false discovery rate is good for an FDR procedure such as the BH approach
  - An FDR approach also has good power

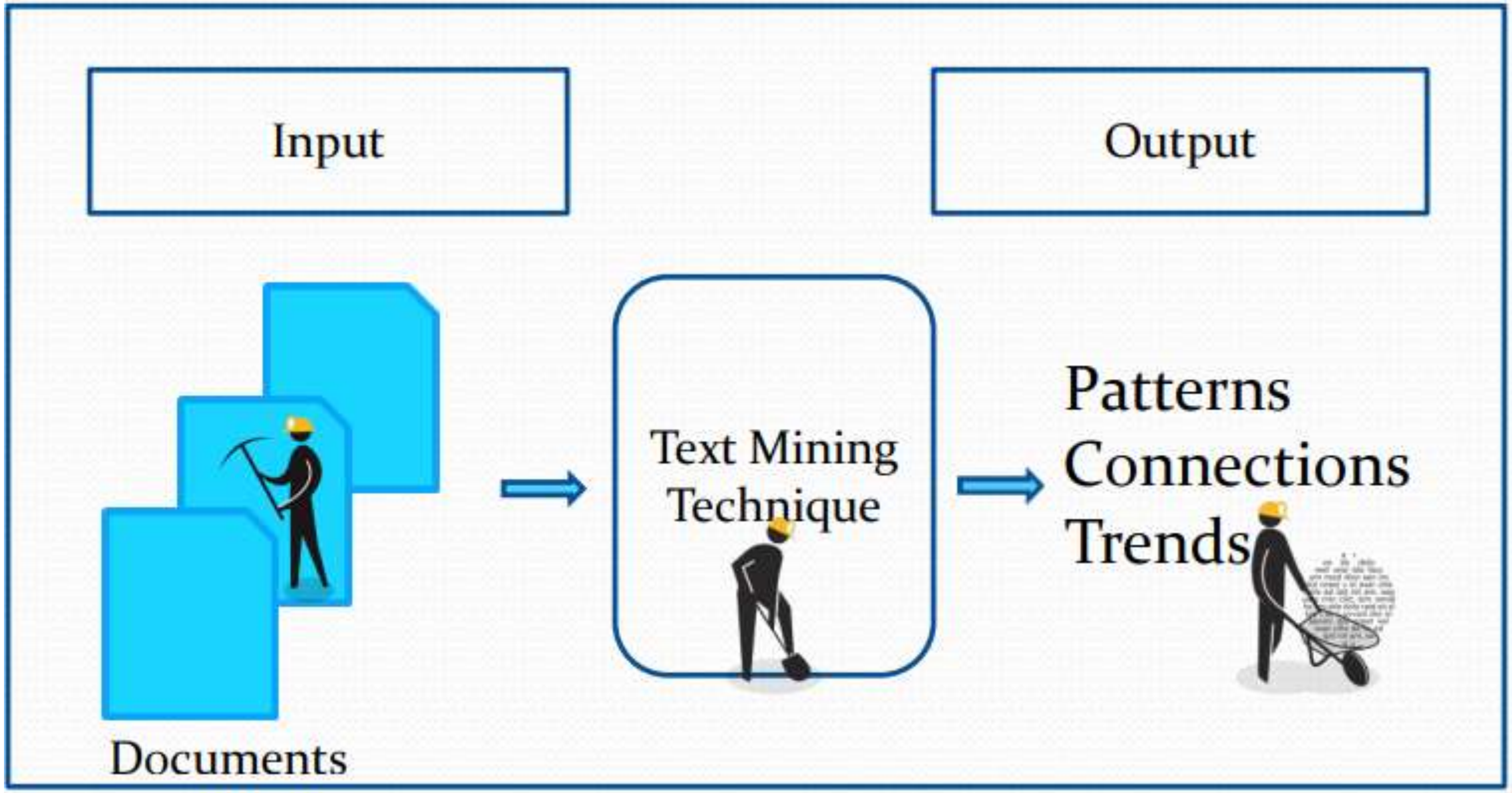
Thank You

Introduction to Text Mining

- Text Mining is a Discovery
- Text Mining is also referred as Text Data Mining (TDM) and Knowledge Discovery in Textual Database (KDT).
- Text Mining is used to extract relevant information or knowledge or pattern from different sources that are in unstructured or semi-structured form.

- Extract and discover knowledge hidden in text automatically
- Helps domain experts by automatically:
  - identifying concepts
  - extracting facts/relations
  - discovering implicit links
  - generating hypotheses

Input-Output Model for Text Mining



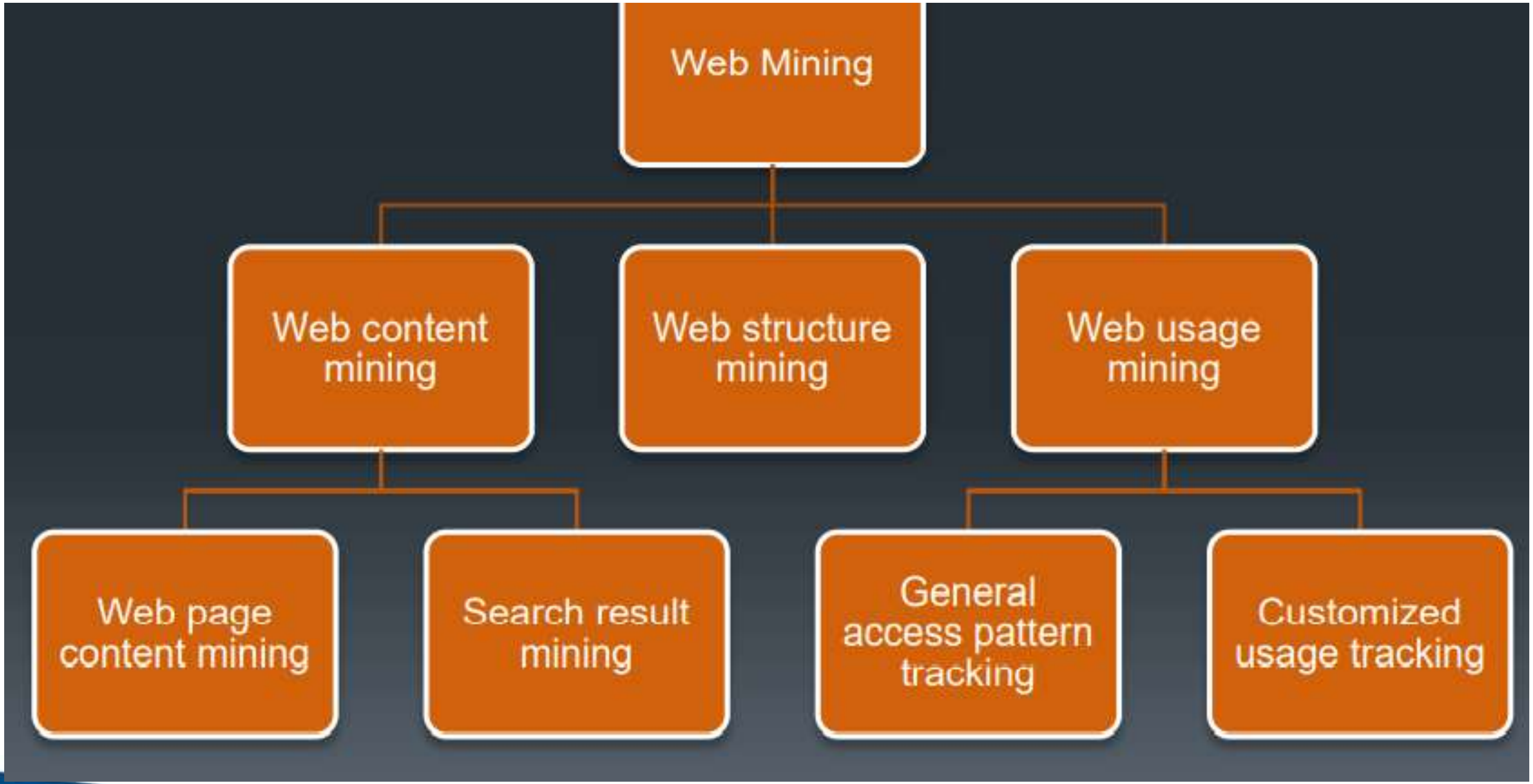
- Steps for Text Mining
  - Pre-Processing the Text
  - Applying Text Mining Techniques
    - Summarization
    - Classification
    - Clustering
    - Visualization
    - Information Extraction
  - Analyzing the Text

WEB MINING

- WEB Mining is the use of the data mining techniques to automatically discover and extract information from web documents/services Discovering useful information from the World-Wide Web and its usage patterns
- Using data mining techniques to make the web more useful and more profitable (for some) and to increase the efficiency of our interaction with the web

- Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet.
  Some users might be looking at only textual data, whereas some others might be interested in multimedia data.
- Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications.

WEB Mining TAXONOMY



Web Mining

Web content mining

Web structure mining

Web usage mining

Web page content mining

Search result mining

General access pattern tracking

Customized usage tracking