0. Definitions of Data Mining

- **The discovery of new information in terms of patterns or rules from vast amounts of data.**
- **The process of finding interesting structure in data.**
- **The process of employing one or more computer learning techniques to automatically analyze and extract knowledge from data.**

- اكتشاف معلومات جديده من حيث الأنماط أو القواعد من كميات هائله من البيانات.
- عمليه العثور علي بنيه مثيره للاهتمام في البيانات.
- عمليه استخدام تقنيه واحده أو أكثر من تقنيات التعلم بالكمبيوتر لتحليل المعلومات واستخراجها تلقائيا من البيانات.

1. Data Warehousing

- **The data warehouse is a historical database designed for decision support.**
- **Data mining can be applied to the data in a warehouse to help with certain types of decisions.**
- **Proper construction of a data warehouse is fundamental to the successful use of data mining**

- ومستودع البيانات هو قاعده بيانات تاريخيه مصممه لدعم القرارات.
- ويمكن تطبيق استخراج البيانات علي البيانات في المستودع للمساعدة في أنواع معينه من القرارات.
- البناء السليم لمستودع البيانات أمر أساسي للنجاح في استخدام البيانات التعدين

2. Knowledge Discovery in Databases (KDD)

- **Data mining is actually one step of a larger process known as knowledge discovery in databases (KDD).**
- **The KDD process model comprises six phases**
  - ✓ **Data selection**
  - ✓ **Data cleansing**
  - ✓ **Enrichment**
  - ✓ **Data transformation or encoding**

- البيانات التعدين هو في الواقع خطوه واحده من عمليه أكبر المعروفة باسم اكتشاف المعرفة في قواعد البيانات (KDD.)
- ويتالف نموذج عمليه KDDمن ست مراحل
  - ✓ اختيار البيانات
  - ✓ تنقيه البيانات
  - ✓ تخصيب
  - ✓ تحويل البيانات أو ترميزها

3. Goals of Data Mining and Knowledge Discovery (PICO)

- Prediction:
  - Determine how certain attributes will behave in the future.
- Identification:
  - Identify the existence of an item, event, or activity.
- Classification:
  - Partition data into classes or categories.
- Optimization:
  - Optimize the use of limited resources.

- تنبؤ:
- تحديد كيفيه سلوك سمات معينه في المستقبل.
- تحديد:
- تحديد وجود عنصر أو حدث أو نشاط.
- تصنيف:
- تقسيم البيانات إلى فئات أو فئات.
- أمثل:
- تحسين استخدام الموارد المحدودة.

4. Types of Discovered Knowledge

- **Association Rules**
- **Classification Hierarchies**
- **Sequential Patterns**
- **Patterns Within Time Series**
- **Clustering**

- قواعد الانتساب
- التسلسل الهرمي للتصنيف
- أنماط متسلسلة
- أنماط داخل السلسلة الزمنيه
- تجميع

5. Association Rules

- **Association rules are frequently used to generate rules from market-basket data.**
- **A market basket corresponds to the sets of items a consumer purchases during one visit to a supermarket.**
- **The set of items purchased by customers is known as an itemset.**
- **An association rule is of the form X=>Y, where X ={x1, x2, …., xn }, and Y = {y1,y2, …., yn} are sets of items, with xi and yi being distinct items for all i and all j.**
- **For an association rule to be of interest, it must satisfy a minimum support and confidence.**

- يتم استخدام قواعد الاقتران بشكل متكرر لإنشاء قواعد من بيانات سله السوق.
- وتتطابق سله السوق مع مجموعات الأصناف التي يشتريها المستهلك خلال زيارة واحده إلى سوبرماركت.
- تعرف مجموعه الأصناف التي تم شراؤها من قبل العملاء بمجموعه العناصر.
- قاعده اقتران هي من النموذج X = > Y، حيث X = {x1، x2، ....، س} ، وصاد = { y1، y2، ....، yn}هي مجموعات من البنود ، مع الحادي عشر والياء ويجري البنود المميزة للجميع انا وجميع j.
- ولكي تكون قاعده الجمعيات ذات اهميه ، يجب ان تلبي الحد الأدنى من الدعم والثقة.

6. Association Rules Confidence and Support

- **Support:**
- **The minimum percentage of instances in the database that contain all items listed in a given association rule.**
- **Support is the percentage of transactions that contain all of the items in the itemset, LHS U RHS.**
- **Confidence:**
- **Given a rule of the form A=>B, rule confidence is the conditional probability that B is true when A is known to be true.**
- **Confidence can be computed as**
- **support(LHS U RHS) / support(LHS)**

- دعم:
- النسبة المئوية الدنيا للحالات في قاعده البيانات التي تحتوي علي كافة العناصر المسرودة في قاعده اقتران معطي.
- الدعم هو النسبة المئوية من الحركات التي تحتوي علي كافة العناصر الواردة في مجموعه البنود ، والنظام المنسق الخاص.
- ثقة:
- النظر إلى قاعده النموذج a = > b، تكون ثقة القاعدة هي الاحتمالية الشرطية التي تكون Bصحيحه عندما يكون من المعروف انها صحيحه.
- ويمكن احتساب الثقة
- الدعم (النظام المنسق الإقليمي)/الدعم ((hs

7. Generating Association Rules

- **The general algorithm for generating association rules is a two-step process.**
- **Generate all itemsets that have a support exceeding the given threshold. Itemsets with this property are called large or frequent itemsets.**
- **Generate rules for each itemset as follows:**
- **For itemset X and Y a subset of X, let Z = X – Y;**
- **If support(X)/Support(Z) > minimum confidence, the rule Z=>Y is a valid rule.**

- الخوارزميه العامة لإنشاء قواعد الاقتران هي عمليه من خطوتين.
- إنشاء كافة مجموعات البنود التي لها دعم يتجاوز العتبة المعطية. تسمي مجموعات البنود مع هذه الخاصية مجموعات البنود الكبيرة أو المتكررة.
- إنشاء قواعد لكل مجموعه البنود كما يلي:
- لمجموعه البنود xو yمجموعه فرعيه من x، دع Z = x-y;
- إذا كان الدعم (س)/الدعم (< zالحد الأدنى من الثقة ، والقاعدة X = > Zهو قاعده صالحه.

8. Reducing Association Rule Complexity

- **Two properties are used to reduce the search space for association rule generation.**
  - **Downward Closure**
    - **A subset of a large itemset must also be large**
  - **Anti-monotonicity**
    - **A superset of a small itemset is also small. This implies that the itemset does not have sufficient support to be considered for rule generation.**

- يتم استخدام خاصيتين لتقليل مساحة البحث لإنشاء قاعده الاقتران.
- الإغلاق التنازلي
- يجب ان تكون مجموعه فرعيه من مجموعه بنود كبيره أيضا كبيره
- المضادة لرتيب
- مجموعه من مجموعه البنود الصغيرة هي أيضا صغيره. وهذا يعني ان مجموعه البنود لا يتوفر لها الدعم الكافي للنظر فيها من أجل توليد القواعد.

9. Generating Association Rules: The Apriori Algorithm

- **The Apriori algorithm was the first algorithm used to generate association rules.**
  - **The Apriori algorithm uses the general algorithm for creating association rules together with downward closure and anti-monotonicity.**

- وكانت الخوارزميه الخاصة بالبروتوكول الأول الخوارزميه المستخدمة لإنشاء قواعد الاقتران.
- الخوارزميه التي تستخدم الخوارزميه العامة لإنشاء قواعد الاقتران مع الإغلاق الهبوطي والمضادة للرتابة.

10. Generating Association Rules: The Sampling Algorithm

- **The sampling algorithm selects samples from the database of transactions that individually fit into memory. Frequent itemsets are then formed for each sample.**
  - **If the frequent itemsets form a superset of the frequent itemsets for the entire database, then the real frequent itemsets can be obtained by scanning the remainder of the database.**
  - **In some rare cases, a second scan of the database is required to find all frequent itemsets.**

- تقوم خوارزميه أخذ العينات بتحديد العينات من قاعده بيانات الحركات التي تلائم الذاكرة بشكل فردي. ويتم بعد ذلك تشكيل مجموعات البنود المتكررة لكل عينه.
- إذا كانت مجموعات البنود المتكررة تشكل مجموعه فائقه من مجموعات البنود المتكررة لقاعده البيانات بأكملها ، فانه يمكن الحصول علي مجموعات البنود المتكررة الحقيقية عن طريق مسح ما تبقي من قاعده البيانات.
- وفي بعض الحالات النادرة ، يلزم اجراء فحص ثان لقاعده البيانات للعثور علي كافة مجموعات البنود المتكررة.

---

11. Generating Association Rules: Frequent-Pattern Tree Algorithm

- The Frequent-Pattern Tree Algorithm reduces the total number of candidate itemsets by producing a compressed version of the database in terms of an FP-tree.
- The FP-tree stores relevant information and allows for the efficient discovery of frequent itemsets.
- The algorithm consists of two steps:
  - Step 1 builds the FP-tree.
  - Step 2 uses the tree to find frequent itemsets.

- وتخفض خوارزميه الأشجار المتكررة النمط العدد الإجمالي لمجموعات البنود المرشحة عن طريق إنتاج نسخه مضغوطه من قاعده البيانات من حيث شجره FP.
- وتخزن شجره FP المعلومات ذات الصلة وتتيح الاكتشاف الفعال لمجموعات البنود المتكررة.
- الخوارزميه يتكون من خطوتين:
- الخطوة 1 يبني شجره.FP
- الخطوة 2 يستخدم الشجرة للعثور علي مجموعات البنود المتكررة.

12. Step 1: Building the FP-Tree

- First, frequent 1-itemsets along with the count of transactions containing each item are computed.
- The 1-itemsets are sorted in non-increasing order.
- The root of the FP-tree is created with a “null” label.
- For each transaction T in the database, place the frequent 1-itemsets in T in sorted order. Designate T as consisting of a head and the remaining items, the tail.
- Insert itemset information recursively into the FP-tree as follows:
  - if the current node, N, of the FP-tree has a child with an item name = head, increment the count associated with N by 1 else create a new node, N, with a count of 1, link N to its parent and link N with the item header table.
  - if tail is nonempty, repeat the above step using only the tail, i.e., the old head is removed and the new head is the first item from the tail and the remaining items become the new tail.

- أولا ، يتم احتساب مجموعات البنود المتكررة بالاضافه إلى عدد المعاملات التي تحتوي علي كل صنف.
- يتم فرز مجموعات البنود 1 في ترتيب غير متزايد.
- يتم إنشاء الجذر من شجره-FP مع تسميه."null"
- بالنسبة لكل معامله T في قاعده البيانات ، ضع مجموعات البنود المتكررة في t في ترتيب مفروز. تعيين T باعتبارها تتكون من الراس والعناصر المتبقية ، والذيل.
- ادراج معلومات مجموعه البنود بشكل متكرر في شجره FP كما يلي:
- إذا كانت العقدة الحالية ، n، من شجره FP لها طفل باسم عنصر = الراس ، قم بزيادة العدد المقترن ب n بواسطة 1 آخر إنشاء عقده جديده ، n، مع عدد 1 ، الارتباط n إلى الأصل والارتباط n بجدول راس العنصر.
- إذا كان الذيل غير فارغ ، كرر الخطوة أعلاه باستخدام الذيل فقط ، اي الراس القديم هو أزاله والراس الجديد هو العنصر الأول من الذيل والعناصر المتبقية تصبح الذيل الجديد.

13. Step 2: The FP-growth Algorithm For Finding Frequent Itemsets

```
Input: Fp-tree and minimum support, mins
Output: frequent patterns (itemsets)
procedure FP-growth (tree, alpha);
Begin
 if tree contains a single path P then
  for each combination, beta of the nodes in the path
   generate pattern (beta U alpha)
   with support = minimum support of nodes in beta
 else
  for each item, i, in the header of the tree do
   begin
    generate pattern beta = (i U alpha) with support = i.support;
    construct beta's conditional pattern base;
    construct beta's conditional FP-tree, beta_tree;
    if beta_tree is not empty then
     FP-growth(beta_tree, beta);
    end;
End;
```

14. Generating Association Rules:
The Partition Algorithm

- Divide the database into non-overlapping subsets.
- Treat each subset as a separate database where each subset fits entirely into main memory.
- Apply the Apriori algorithm to each partition.
- Take the union of all frequent itemsets from each partition.
- These itemsets form the global candidate frequent itemsets for the entire database.
- Verify the global set of itemsets by having their actual support measured for the entire database.

- تقسيم قاعده البيانات إلى مجموعات فرعيه غير متراكبه.
- معالجه كل مجموعه فرعيه كقاعده بيانات منفصلة حيث كل مجموعه فرعيه تناسب تماما في الذاكرة الرئيسية.
- تطبيق الخوارزميه الخاصة بكل قسم.
- اتخاذ الاتحاد من جميع البنود المتكررة من كل قسم.
- وتشكل مجموعات البنود هذه مجموعات البنود المرشحة العالمية المتكررة لقاعده البيانات بأكملها.
- تحقق من المجموعة العمومية من مجموعات البنود بواسطة الحصول علي الدعم الفعلي الخاص بهم تقاس لقاعده البيانات بأكملها.

---

15. Complications seen with Association Rules

- The cardinality of itemsets in most situations is extremely large.
- Association rule mining is more difficult when transactions show variability in factors such as geographic location and seasons.
- Item classifications exist along multiple dimensions.
- Data quality is variable; data may be missing, erroneous, conflicting, as well as redundant.

- العلاقة السببية لمجموعات البنود في معظم الحالات كبيره للغاية.
- الرابطة القاعدة التعدين أكثر صعوبة عندما تظهر المعاملات التفاوت في عوامل مثل الموقع الجغرافي والمواسم.
- تصنيفات الأصناف موجودة علي طول ابعاد متعددة.
- جوده البيانات متغيرة; قد تكون البيانات مفقوده ، خاطئه ، متعارضة ، وكذلك زائده عن الحاجة.

16. Classification

- Classification is the process of learning a model that is able to describe different classes of data.
- Learning is supervised as the classes to be learned are predetermined.
- Learning is accomplished by using a training set of pre-classified data.
- The model produced is usually in the form of a decision tree or a set of rules.

- التصنيف هو عمليه تعلم نموذج قادر علي وصف فئات مختلفه من البيانات.
- ويتم الاشراف علي التعلم لان الفصول المقرر تعلمها محدده سلفا.
- ويتم التعلم باستخدام مجموعه من البيانات السابقة لتصنيف التدريب.
- وعاده ما يكون النموذج المنتج في شكل شجره قرار أو مجموعه من القواعد.

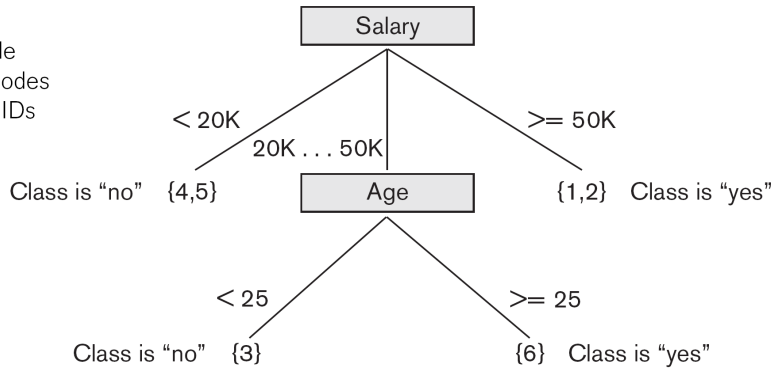

Figure 28.7
Decision tree based on sample training data where the leaf nodes are represented by a set of RIDs of the partitioned records.

17. An Example Rule

- Here is one of the rules extracted from the decision tree of Figure 28.7.

IF 50K > salary >= 20K
AND age >=25
THEN class is "yes"

18. Clustering

- Unsupervised learning or clustering builds models from data without predefined classes.
- The goal is to place records into groups where the records in a group are highly similar to each other and dissimilar to records in other groups.
- The k-Means algorithm is a simple yet effective clustering technique.

- التعلم أو التجميع غير الخاضع للاشراف يبني نماذج من البيانات دون فئات معرفه مسبقا.
- والهدف هو وضع السجلات في مجموعات تكون فيها السجلات الموجودة في مجموعه مشابهه لبعضها البعض بشكل كبير وتختلف عن السجلات الموجودة في مجموعات أخرى.
- و-k الوسائل الخوارزميه هي تقنيه بسيطه ولكنها فعاله التكتلات.

19. Additional Data Mining Methods

- QSequential pattern analysis
- Time Series Analysis
- Regression
- Neural Networks
- Genetic Algorithms

- تحليل الأنماط المتسلسلة
- تحليل السلاسل الزمنيه
- انحدار
- الشبكات العصبية
- خوارزميات وراثيه

---

20. Sequential Pattern Analysis

- Transactions ordered by time of purchase form a sequence of itemsets.
- The problem is to find all subsequences from a given set of sequences that have a minimum support.
- The sequence S_1, S_2, S_3, .. is a predictor of the fact that a customer purchasing itemset S_1 is likely to buy S_2 , and then S_3, and so on.

- المعاملات المطلوبة حسب وقت الشراء نموذج تسلسل من مجموعات البنود.
- المشكلة هي العثور علي كافة مسار من مجموعه معينه من التسلسلات التي لديها الحد الأدنى من الدعم.
- تسلسلS1 ، S2 ، S3 ..هو التنبؤ من حقيقة ان العملاء شراء مجموعه البنود S1 من المرجح ان شراءS2 ، ومن ثمS3 ، وهلم تم.

21. Time Series Analysis

- Time series are sequences of events. For example, the closing price of a stock is an event that occurs each day of the week.
- Time series analysis can be used to identify the price trends of a stock or mutual fund.

Time series analysis is an extended functionality of temporal data management

- سلاسل الوقت هي تسلسل من الاحداث. علي سبيل المثال ، يعتبر سعر إغلاق المخزون حدثا يحدث كل يوم من أيام الأسبوع.
- ويمكن استخدام تحليل السلاسل الزمنيه لتحديد اتجاات الأسعار في الأسهم أو الصناديق المشتركة.
- تحليل السلاسل الزمنيه هو وظيفة موسعه من أداره البيانات الزمنيه

22. Regression Analysis

- A regression equation estimates a dependent variable using a set of independent variables and a set of constants.
- The independent variables as well as the dependent variable are numeric.
- A regression equation can be written in the form $Y=f(x_1,x_2,\ldots,x_n)$ where Y is the dependent variable.
- If f is linear in the domain variables x_i, the equation is call a linear regression equation.

- وتقدر معادله الانحدار متغيرا تابعا باستخدام مجموعه من المتغيرات المستقلة ومجموعه من الثوابت.
- المتغيرات المستقلة وكذلك المتغير التابعة هي رقميه.
- يمكن كتابه معادله الانحدار في النموذج y = f (x1 ، x2 ، ...، xcln)حيث y هو المتغير التابع.
- إذا كانت f خطيه في متغيرات المجال الحادي عشر ، فان المعادلة تسمي معادله انحدار خطي.

23. Neural Networks

- A neural network is a set of interconnected nodes designed to imitate the functioning of the brain.
- Node connections have weights which are modified during the learning process.
- Neural networks can be used for supervised learning and unsupervised clustering.
- The output of a neural network is quantitative and not easily understood.

- الشبكة العصبية هي مجموعه من العقد المترابطة المصممة لتقليد أداء الدماغ.
- تكون لاتصالات العقدة أوزان يتم تعديلها اثناء عمليه التعلم.
- ويمكن استخدام الشبكات العصبية للتعلم تحت اشراف والتكتلات غير الخاضعة للاشراف.
- إنتاج الشبكة العصبية الكمية وليس من السهل فهمها.

24. Genetic Learning

- Genetic learning is based on the theory of evolution.
- An initial population of several candidate solutions is provided to the learning model.
- A fitness function defines which solutions survive from one generation to the next.
- Crossover, mutation and selection are used to create new population elements.

- ويستند التعلم الوراثي علي نظرية التطور.
- ويقدم لنموذج التعلم عدد اولي من الحلول للمرشحين.
- وظيفة اللياقة البدنية يحدد الحلول التي البقاء علي قيد الحياة من جيل إلى آخر.
- وتستخدم التحويلات والطفرات والاختيارات لإنشاء عناصر سكانية جديده.

25. Data Mining Applications

- Marketing
  - Marketing strategies and consumer behavior
- Finance
  - Fraud detection, creditworthiness and investment analysis
- Manufacturing
  - Resource optimization
- Health
  - Image analysis, side effects of drug, and treatment effectiveness

- تسويق
  - استراتيجيات التسويق وسلوك المستهلك
- مالي
  - الكشف عن الغش والجدارة الائتمانية وتحليل الاستثمار
- تصنيع
  - تحسين الموارد
- صحة
  - تحليل الصور ، والآثار الجانبية للمخدرات ، وفعالية العلاج
  -

26. Recap

- Data Mining
- Data Warehousing
- Knowledge Discovery in Databases (KDD)
- Goals of Data Mining and Knowledge Discovery
- Association Rules
- Additional Data Mining Algorithms
  - Sequential pattern analysis
  - Time Series Analysis
  - Regression
  - Neural Networks
  - Genetic Algorithms