Data Mining

Data Mining Concepts/Definitions , Why Data Mining?

- More intense competition at the global scale.
- Recognition of the value in data sources.
- Availability of quality data on customers, vendors, transactions, Web, etc.
- Consolidation and integration of data repositories into data warehouses.
- The exponential increase in data processing and storage capabilities; and decrease in cost.
- Movement toward conversion of information resources into nonphysical form.

تعاريف/مفاهيم التنقيب في البيانات ، لماذا التنقيب في البيانات ؟

- منافسة أكثر حدة على المستوى العالمي.
- التعرف على القيمة في مصادر البيانات.
- توافر بيانات عالية الجودة على العملاء والبائعين والمعاملات والويب وما إلى ذلك.
- توحيد وتكامل مستودعات البيانات في مستودعات او مخازن البيانات.
- الزيادة الأسية في قدرات معالجة البيانات وتخزينها ؛ وانخفاض في التكلفة.
- التحرك نحو تحويل موارد المعلومات إلى صيغة غير مادية او غير ملموسه.

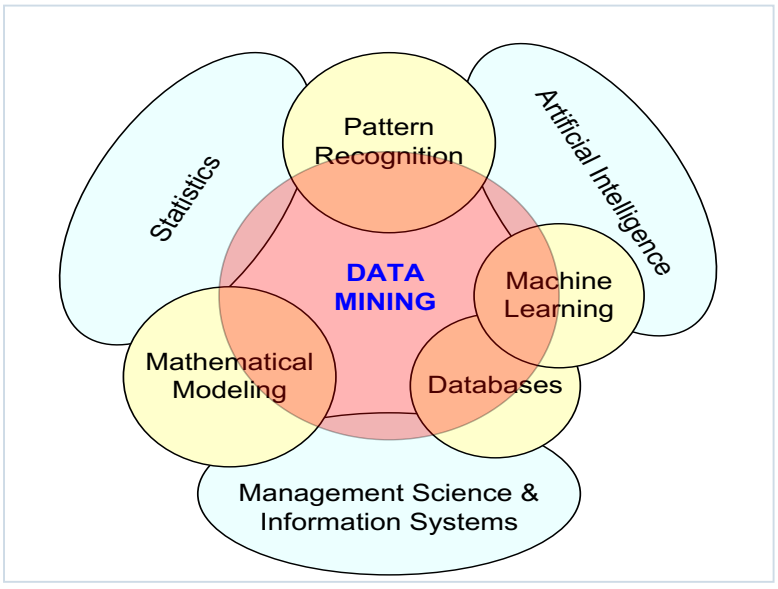
Definition of Data Mining

- The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases. - *Fayyad et al., (1996)*
- Keywords in this definition: Process, nontrivial, valid, novel, potentially useful, understandable.
- Data mining: a misnomer?
- Other names: knowledge extraction, pattern analysis, knowledge discovery, information harvesting, pattern searching, data dredging,…

تعريف التنقيب في البيانات

- عملية غير بديهية لتحديد أنماط صالحة ، جديدة ، قد تكون مفيدة ، ومفهومة في نهاية المطاف في البيانات المخزنة في قواعد البيانات المنظمة. - *Fayyad et al., (1996)*
- الكلمات الرئيسية في هذا التعريف: عملية ، غير بديهية ، صالحة ، جديدة ، يمكن أن تكون مفيدة ومفهومة.
- استخراج البيانات: تسمية خاطئة؟
- الأسماء الأخرى: استخراج المعرفة ، تحليل النمط ، اكتشاف المعرفة ، حصاد المعلومات ، البحث عن الأنماط ، تجريف البيانات ، ...

Data Mining is at the Intersection of Many Disciplines → التنقيب عن البيانات هو في تقاطع العديد من التخصصات



Data Mining Characteristics/Objectives

- Source of data for DM is often a consolidated data warehouse (not always!).
- DM environment is usually a client-server or a Web-based information systems architecture.
- Data is the most critical ingredient for DM which may include soft/unstructured data.
- The miner is often an end user
- Striking it rich requires creative thinking
- Data mining tools' capabilities and ease of use are essential (Web, Parallel processing, etc.)

خصائص و اهداف التنقيب في البيانات

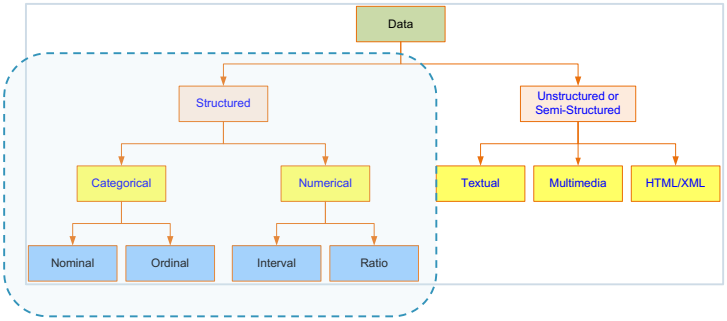
- مصدر البيانات لـ DM (Data Mining) هو مستودع بيانات مدمج (ليس دائمًا!).
- بيئة DM هي عادة خادم العميل أو بنية نظم المعلومات على شبكة الإنترنت.
- البيانات هي العنصر الأكثر أهمية ل DM والتي قد تتضمن بيانات هشة او ناقصة / غير منظمة.
- غالباً ما يكون عامل التنقيب مستخدمًا نهائيًا
- الحصول عليها غني يتطلب التفكير الإبداعي
- قدرة أدوات التنقيب عن البيانات وسهولة الاستخدام ضرورية (الويب ، المعالجة الموازية ، إلخ)

Data in Data Mining

- Data: a collection of facts usually obtained as the result of experiences, observations, or experiments.
- Data may consist of numbers, words, images, …
- Data: lowest level of abstraction (from which information and knowledge are derived).

البيانات في التنقيب عن البيانات

- البيانات: مجموعة من الحقائق التي يتم الحصول عليها عادة كنتيجة للتجارب أو الملاحظات أو التجارب.
- قد تتكون البيانات من أرقام ، كلمات ، صور ، ...
- البيانات: أدنى مستوى من التجريد (الذي تستمد منه المعلومات والمعارف).



What Does DM Do? How Does it Work?

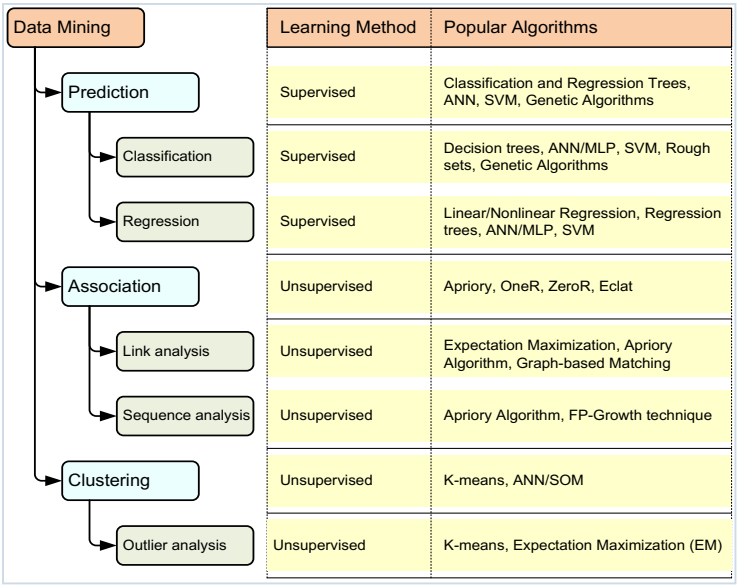
- DM extract patterns from data
  - Pattern? A mathematical (numeric and/or symbolic) relationship among data items
- Types of patterns
  - Association
  - Prediction
  - Cluster (segmentation)
  - Sequential (or time series) relationships

ماذا يعمل التنقيب عن البيانات و كيف يعمل ؟

- أنماط استخراج DM من البيانات
  - نمط؟ علاقة رياضية (رقمية و / أو رمزية) بين عناصر البيانات

- أنواع الأنماط
  - جمعية
  - تنبؤ
  - المجموعة (التجزئة)
  - العلاقات التسلسلية (أو السلاسل الزمنية)

A Taxonomy for Data Mining Tasks → تصنيف لمهام التنقيب عن البيانات



Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

Data Mining Tasks (cont.)

- Time-series forecasting
  - Part of sequence or link analysis?
- Visualization
  - Another data mining task?
- Types of DM
  - Hypothesis-driven data mining
  - Discovery-driven data mining

مهام التنقيب عن البيانات

- التنبؤ بالسلسلة الزمنية
  - جزء من تسلسل أو تحليل الارتباط؟
- تصور
  - مهمة أخرى لتنقيب عن البيانات؟
- أنواع التنقيب عن البيانات
  - التنقيب عن البيانات منقاد بالفرضية
  - التنقيب عن البيانات منقاد بالاكتشاف

Data Mining Applications → تطبيقات التنقيب عن البيانات

- Customer Relationship Management
  - Maximize return on marketing campaigns
  - Improve customer retention (churn analysis)
  - Maximize customer value (cross-, up-selling)
  - Identify and treat most valued customers
- Banking & Other Financial
  - Automate the loan application process
  - Detecting fraudulent transactions
  - Maximize customer value (cross-, up-selling)
  - Optimizing cash reserves with forecasting

- إدارة علاقات العملاء
  - تعظيم العائد على الحملات التسويقية
  - تحسين احتفاظ العملاء (churn analysis)
  - تعظيم قيمة العملاء (cross-, up-selling)
  - تميز وتعامل مع معظم العملاء الكرام
- المصرفية وغيرها من المالية
  - أتمتة عملية طلب القرض
  - كشف المعاملات الاحتيالية
  - تعظيم قيمة العملاء (cross-, up-selling)
  - تحسين الاحتياطي النقدي مع التنبؤ

- Retailing and Logistics → البيع بالتجزئة و الخدمات اللوجستية (التخطيط و التنفيذ)
  - Optimize inventory levels at different locations → تحسين مستويات المخزون في مواقع مختلفة
  - Improve the store layout and sales promotions → تحسين تخطيط المتاجر و العروض الترويجية للمبيعات
  - Optimize logistics by predicting seasonal effects → تحسين اللوجستية عن طريق التنبؤ بالتأثيرات الموسمية
  - Minimize losses due to limited shelf life → تقليل الخسائر نظراً لمحدودية مدة الصلاحية
- Manufacturing and Maintenance → التصنيع و الصيانة
  - Predict/prevent machinery failures
  - Identify anomalies in production systems to optimize the use manufacturing capacity
  - Discover novel patterns to improve product quality
    - تنبأ / منع فشل الآلات
    - التعرف على الحالات الشاذة في أنظمة الإنتاج لتحسين استخدام القدرة التصنيعية
    - اكتشف أنماطًا جديدة لتحسين جودة المنتج

Data Mining Applications (cont.)

- Brokerage and Securities Trading
  - Predict changes on certain bond prices
  - Forecast the direction of stock fluctuations
  - Assess the effect of events on market movements
  - Identify and prevent fraudulent activities in trading
- Insurance
  - Forecast claim costs for better business planning
  - Determine optimal rate plans
  - Optimize marketing to specific customers
  - Identify and prevent fraudulent claim activities

- تداول الوساطة والأوراق المالية
  - تنبأ التغييرات على أسعار سندات معينة
  - توقع اتجاه تقلبات الأسهم
  - تقييم تأثير الأحداث على تحركات السوق
  - تحديد ومنع الأنشطة الاحتيالية في التداول

- تأمين
  - توقعات طلب التكاليف لتحسين تخطيط الأعمال
  - تحديد خطط المعدل الأمثل
  - تحسين التسويق لزبائن محددين و خاصيين
  - تحديد ومنع أنشطة المطالبة الاحتيالية

  - Computer hardware and software
  - Science and engineering
  - Government and defense
  - Homeland security and law enforcement
  - Travel industry
  - Healthcare
  - Medicine
  - Entertainment industry
  - Sports
  - Etc.

} Increasingly more popular application areas for data mining

  - أجهزة الكمبيوتر والبرمجيات
  - العلوم والهندسة
  - الحكومة والدفاع
  - الأمن الداخلي و تطبيق القانون
  - قطاع السفر
  - الرعاية الصحية
  - دواء
  - صناعه او مجال التسلية والترفيه
  - الأنشطة الرياضية

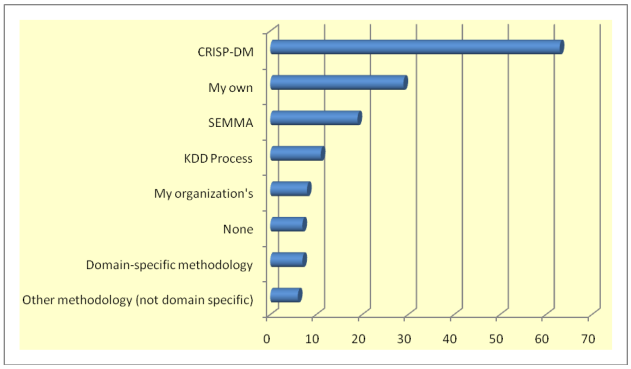
} مجالات التطبيق الأكثر انتشارًا لتنقيب عن البيانات

Data Mining Process → عملية التنقيب عن البيانات

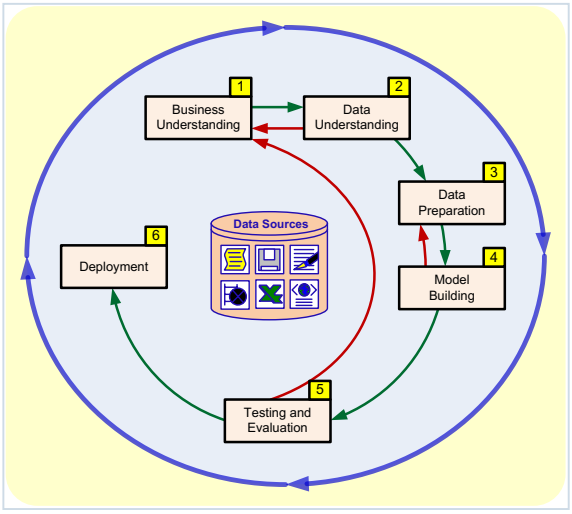
- A manifestation of best practices → مظهر من مظاهر أفضل الممارسات
- A systematic way to conduct DM projects → طريقة منظمة لإجراء مشاريع التنقيب عن البيانات
- Different groups has different versions → المجموعات المختلفة لديها إصدارات مختلفة
- Most common standard processes:
  - CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)
  - KDD (Knowledge Discovery in Databases)

- العمليات القياسية الأكثر شيوعًا:
  - CRISP-DM (عملية قياسية عبر الصناعة لتنقيب عن البيانات)
  - SEMMA (عينه ، استكشاف ، تعديل ، نموذج ، وتقييم)
  - KDD (اكتشاف المعرفة في قواعد البيانات)

Data Mining Process



Data Mining Process: CRISP-DM



Step 1: Business Understanding
Step 2: Data Understanding
Step 3: Data Preparation (!)
Step 4: Model Building
Step 5: Testing and Evaluation
Step 6: Deployment

Accounts for ~85% of total project time

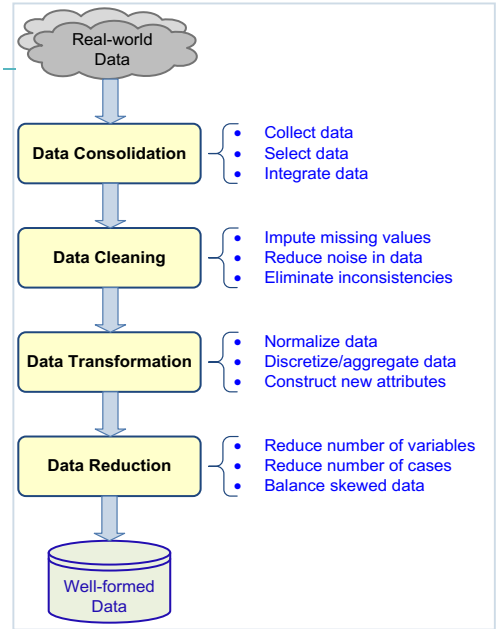
- The process is highly repetitive and experimental (DM: art versus science?)

الخطوة 1: فهم الأعمال
الخطوة 2: فهم البيانات
الخطوة 3: إعداد البيانات (!)
الخطوة 4: بناء النماذج
الخطوة 5: الاختبار والتقييم
الخطوة 6: النشر

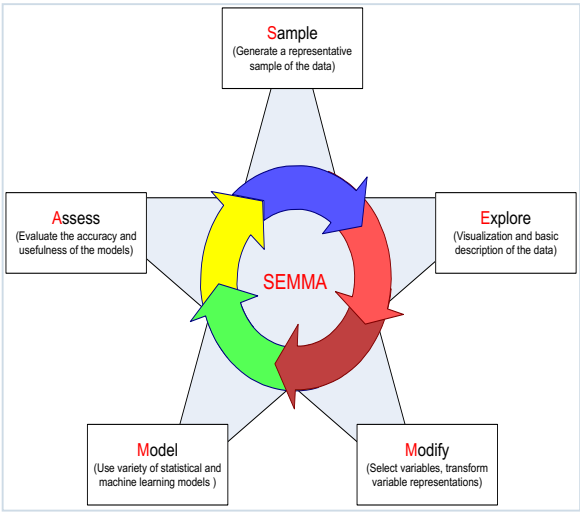
حسابات ~ 85٪ من إجمالي وقت المشروع

- العملية متكررة للغاية والتجريبية (DM: الفن مقابل العلم؟)

Data Preparation – A Critical DM Task



Data Mining Process: SEMMA



Data Mining Methods: Classification

- Most frequently used DM method
- Part of the machine-learning family
- Employ supervised learning
- Learn from past data, classify new data
- The output variable is categorical (nominal or ordinal) in nature
- Classification versus regression?
- Classification versus clustering?

طرق التنقيب عن البيانات : التصنيف

- طريقة DM الأكثر استخداما
- جزء من عائلة التعلم الآلي
- استخدام التعليم تحت الإشراف
- التعلم من البيانات السابقة ، وتصنيف البيانات الجديدة
- المتغير الناتج هو فئوي (اسمي أو ترتيبي) في الطبيعة - "او بمعنى خاص بفئة"
- التصنيف مقابل الانحدار او التراجع؟
- التصنيف مقابل التجمع؟

Assessment Methods for Classification

- Predictive accuracy
  - Hit rate
- Speed
  - Model building; predicting
- Robustness
- Scalability
- Interpretability
  - Transparency, explainability

طرق تقييم التصنيف

- دقة تنبؤية
  - معدل إصابة (الهدف)
- سرعة
  - بناء نموذج؛ توقعي
- متانة او صلابة
- قابلية التوسع
- للتفسير
  - الشفافية ، شرح

Accuracy of Classification Models → دقة نماذج التصنيف

- In classification problems, the primary source for accuracy estimation is the **confusion matrix**
- في مشاكل التصنيف ، يكون المصدر الأساسي لتقدير الدقة هو مصفوفة الارتباك او الحيرة

		True Class: Positive	True Class: Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
Predicted Class	Negative	False Negative Count (FN)	True Negative Count (TN)

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

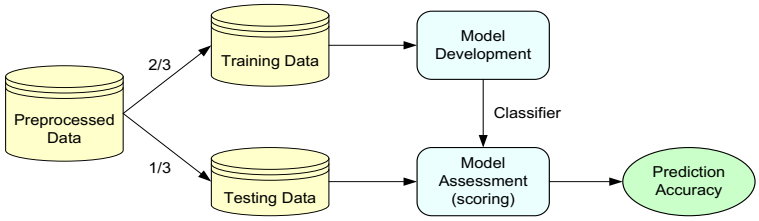
$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Estimation Methodologies for Classification

- **Simple split** (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets training (~70%) and testing (30%)



  - For ANN, the data is split into three sub-sets (training [~60%], validation [~20%], testing [~20%])

منهجيات تقدير التصنيف

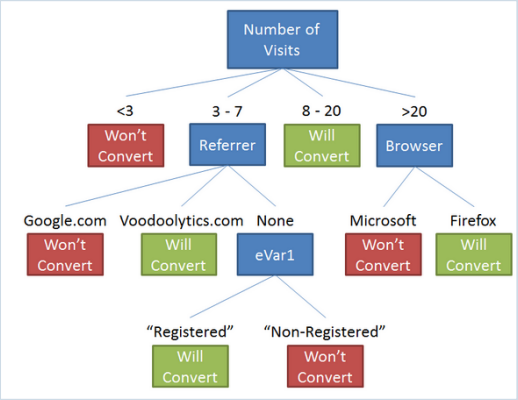
- انقسام بسيط (أو إجراء تقييم للعقد أو اختبار العينة)
  - قسّم البيانات إلى مجموعتين من الدورات التدريبية المتبادلة (~ 70٪) والاختبار (30٪)
  - بالنسبة إلى ANN ، يتم تقسيم البيانات إلى ثلاث مجموعات فرعية (تدريب [~ 60٪] ، التحقق من الصحة [~ 20٪] ، اختبار [~ 20٪])
- ***k*-Fold Cross** Validation (rotation estimation)
  - Split the data into *k* mutually exclusive subsets
  - Use each subset as testing while using the rest of the subsets as training
  - Repeat the experimentation for *k* times
  - Aggregate the test results for true estimation of prediction accuracy training
- Other estimation methodologies → منهجيات تقدير أخرى
  - **Leave-one-out, bootstrapping, jackknifing**
  - **Area under the ROC curve (**graph slide 33**)**

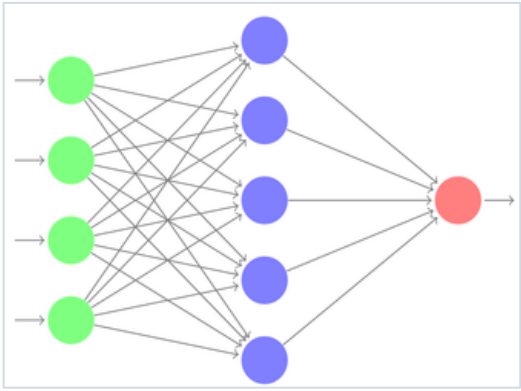
- فعالية k-Fold Cross (تقدير الدوران)
  - قسّم البيانات إلى *k* مجموعات فرعية حصرية متبادلة
  - استخدم كل مجموعة فرعية كاختبار أثناء استخدام بقية المجموعات الفرعية كتدريب
  - كرر تجربة k مرات
  - تجميع نتائج الاختبار لتقدير الصحيح لتدريب دقة التنبؤ

Classification Techniques

- Decision tree analysis
- Statistical analysis
- Neural networks
- Support vector machines
- Case-based reasoning
- Bayesian classifiers
- Genetic algorithms
- Rough sets

- تقنيات التصنيف
- تحليل شجرة القرار
- تحليل احصائي
- الشبكات العصبية
- دعم ناقلات الآلات
- الاستدلال المبني على حالة
- Bayesian classifiers
- الخوارزميات الجينية
- المجموعات قاسية





Decision Trees

- Employs the divide and conquer method
- Recursively divides a training set until each division consists of examples from one class

A general algorithm for decision tree building

1. Create a root node and assign all of the training data to it.
2. Select the best splitting attribute.
3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive subsets along the lines of the specific split.
4. Repeat the steps 2 and 3 for each and every leaf node until the stopping criteria is reached.

شجرة القرارات

- تستخدم طريقة الانقسام والانتزاع
- يقسم مجموعة التدريب بشكل متكرر حتى يتكون كل قسم من أمثلة و من فصل واحد

خوارزمية عامة لبناء شجرة القرارات

1. إنشاء عقدة جذرية وتعيين جميع بيانات التدريب إليها.
2. حدد أفضل سمة التجزئة.
3. إضافة فرع إلى عقدة الجذر لكل قيمة من الانقسام. قسّم البيانات إلى مجموعات فرعية خاصة بالتبادل على طول خطوط التقسيم المحدد.
4. كرر الخطوتين 2 و 3 لكل عقدة كل ورقة حتى الوصول إلى معايير التوقف.

- DT algorithms mainly differ on
  1. Splitting criteria
     - Which variable, what value, etc.
  2. Stopping criteria
     - When to stop building the tree
  3. Pruning (generalization method)
     - Pre-pruning versus post-pruning
- Most popular DT algorithms include → الأكثر شعبية DT و تشمل خوارزميات
  - ID3, C4.5, C5; CART; CHAID; M5

- خوارزميات DT تختلف أساسا على
  1. معايير التقسيم
     - أي متغير وأي قيمة وما إلى ذلك
  2. وقف المعايير
     - متى تتوقف عن بناء الشجرة
  3. التقليم (طريقة التعميم)
     - قبل التقليم Vs ما بعد التقليم

- Alternative splitting criteria
  - Gini index determines the purity of a specific class as a result of a decision to branch along a particular attribute/value
    - Used in CART
  - Information gain uses entropy to measure the extent of uncertainty or randomness of a particular attribute/value split
    - Used in ID3, C4.5, C5
  - Chi-square statistics (used in CHAID)

- معايير تقسيم بديلة
  - يحدد مؤشر Gini نقاوة فئة معينة كنتيجة لقرار التفرع على طول سمة / قيمة معينة
    - تستخدم في CART
  - يستخدم اكتساب المعلومات entropy (هو مقاييس للعشوائية في النظام) لقياس مدى عدم اليقين أو العشوائية لتقسيم سمة / قيمة معينة
    - يستخدم في ID3, C4.5, C5

Cluster Analysis for Data Mining

- Used for automatic identification of natural groupings of things
- Part of the machine-learning family
- Employ unsupervised learning
- Learns the clusters of things from past data, then assigns new instances
- There is not an output variable
- Also known as segmentation

التحليل العنقودي (التجمعي) للتنقيب في البيانات

- تُستخدم للتعرف التلقائي على التجمعات الطبيعية للأشياء
- جزء من عائلة التعلم الآلي
- تستخدم التعلم بدون إشراف
- يتعلم مجموعات من الأشياء من البيانات السابقة ، ثم يعين حالات جديدة
- لا يوجد متغير النتيجة
- يُعرف أيضًا باسم التجزئة

- Clustering results may be used to
  - Identify natural groupings of customers
  - Identify rules for assigning new cases to classes for targeting/diagnostic purposes
  - Provide characterization, definition, labeling of populations
  - Decrease the size and complexity of problems for other data mining methods
  - Identify outliers in a specific domain (e.g., rare-event detection)

- نتائج المجموعات يمكن استخدامها ل
  - تحديد المجموعات الطبيعية للعملاء
  - حدد قواعد لتعيين حالات جديدة لفئات لأغراض الاستهداف / التشخيص
  - توفير توصيف وتعريف وتوسيم او تميز السكان
  - تقليل حجم وتعقيد المشاكل لطرق استخراج البيانات الأخرى
  - تحديد القيم المتطرفة في نطاق معين (على سبيل المثال ، اكتشاف الأحداث النادرة)

- Analysis methods
  - Statistical methods (including both hierarchical and nonhierarchical), such as *k*-means, *k*-modes, and so on.
  - Neural networks (adaptive resonance theory [ART], self-organizing map [SOM])
  - Fuzzy logic (e.g., fuzzy c-means algorithm)
  - Genetic algorithms

- طرق التحليل
  - الأساليب الإحصائية (بما في ذلك التسلسل الهرمي وغير الهرمي) ، مثل k-means و k-modes وما إلى ذلك.
  - الشبكات العصبية (نظرية الرنين التكيفي [ART]، خريطة التنظيم الذاتي [SOM])
  - المنطق الضبابي " يعني الغير واضح" (على سبيل المثال ، fuzzy c-means algorithm)
  - الخوارزميات الجينية

- How many clusters?
  - There is not a “truly optimal” way to calculate it
  - Heuristics are often used
- Most cluster analysis methods involve the use of a distance measure to calculate the closeness between pairs of items.
  - Euclidian versus Manhattan/Rectilinear distance

- كم عدد العناقيد؟
  - لا توجد طريقة "مثالية حقًا" لحسابها
  - وغالبا ما تستخدم الاستدلال
- تتضمن معظم طرق التحليل العنقودية استخدام مقياس المسافة لحساب التقارب بين أزواج العناصر.
  - Euclidian مقابل مسافة Manhattan/Rectilinear

Cluster Analysis for Data Mining (Cont.)

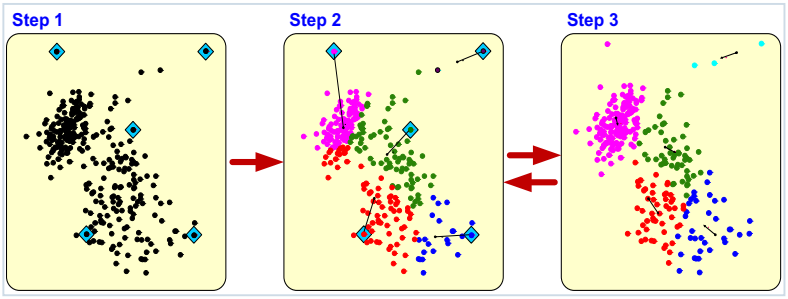
- *k*-Means Clustering Algorithm
  - ***k*** **:** pre-determined number of clusters
  - Algorithm (**Step 0:** determine value of *k*)

Step 1: Randomly generate *k* random points as initial cluster centers.
Step 2: Assign each point to the nearest cluster center.
Step 3: Re-compute the new cluster centers.
Repetition step: Repeat steps 3 and 4 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable).

- تجميع خوارزمية k-Means
  - ***k:*** عدد محدد مسبقًا من العناقيد
  - الخوارزمية **الخطوة 0:** تحديد قيمة (k)

الخطوة 1: إنشاء نقاط عشوائية k كمراكز عنقودية أولية.
الخطوة 2: قم بتعيين كل نقطة إلى مركز العنقود الأقرب.
الخطوة 3: إعادة حساب مراكز العنقود الجديدة.
خطوة التكرار: كرر الخطوتين 3 و 4 حتى يتم استيفاء معيار تقارب (عادة ما يصبح تعيين النقاط إلى التجمعات او العناقيد مستقراً).

Cluster Analysis for Data Mining - *k*-Means Clustering Algorithm



Association Rule Mining

- A very popular DM method in business
- Finds interesting relationships (affinities) between variables (items or events)
- Part of machine learning family
- Employs unsupervised learning
- There is no output variable
- Also known as market basket analysis
- Often used as an example to describe DM to ordinary people, such as the famous "relationship between diapers and beers!"

تنظيم قاعدة التنقيب

- طريقة DM شعبية جدا في مجال الأعمال التجارية
- يجد العلاقات المثيرة للاهتمام (الانتماءات) بين المتغيرات (العناصر أو الأحداث)
- جزء من عائلة التعلم الآلي
- تستخدم التعلم بدون إشراف
- لا يوجد متغير النتيجة
- المعروف أيضا باسم تحليل سلة السوق
- غالبًا ما يستخدم كمثال لوصف DM للأشخاص العاديين .

- **Input:** the simple point-of-sale transaction data
- **Output:** Most frequent affinities among items
- Example: according to the transaction data…
  "Customer who bought a lap-top computer and a virus protection software, also bought extended service plan 70 percent of the time."
- How do you use such a pattern/knowledge?
  - Put the items next to each other
  - Promote the items as a package
  - Place items far apart from each other!

- **الإدخال:** بيانات معاملات نقطة البيع البسيطة
- **الإخراج:** أكثر الانتماءات المتكررة بين العناصر
- مثال: وفقًا لبيانات المعاملة ...

"العميل الذي اشترى جهاز كمبيوتر محمولًا وبرنامج حماية من الفيروسات ، اشترى أيضًا خطة خدمة موسعة بنسبة 70 بالمائة من الوقت".

- كيف تستخدم مثل هذا النمط / المعرفة؟
  - ضع العناصر بجانب بعضها البعض
  - تعزيز العناصر كحزمة
  - ضع العناصر بعيدًا عن بعضها البعض!

Association Rule Mining (Cont.)

- A representative applications of association rule mining include
  - **In business:** cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration
  - **In medicine:** relationships between symptoms and illnesses; diagnosis and patient characteristics and treatments (to be used in medical DSS); and genes and their functions (to be used in genomics projects)

- وتشمل التطبيقات التمثيلية للتنقيب قاعدة الارتباط
  - **في مجال الأعمال التجارية:** التسويق المتقاطع ، البيع المتقاطع ، تصميم المتجر ، تصميم الكتالوج ، تصميم موقع التجارة الإلكترونية ، تحسين الإعلان عبر الإنترنت ، تسعير المنتج ، وتكوين المبيعات / الترويج
  - **في الطب:** العلاقات بين الأعراض والأمراض. التشخيص وخصائص وعلاجات المريض (لاستخدامها في DSS الطبية) ؛ والجينات ووظائفها (لاستخدامها في مشاريع الجينوميات)

- Are all association rules interesting and useful? → هل كل قواعد الجمعيات مثيرة للاهتمام ومفيدة؟
  A Generic Rule: X ⇒ Y [S%, C%]
  X, Y: products and/or services → الخدمات او المنتجات
  X: Left-hand-side (LHS)
  Y: Right-hand-side (RHS)
  S: Support: how often X and Y go together → الدعم
  C: Confidence: how often Y go together with the X → الثقه
  Example: {Laptop Computer, Antivirus Software} ⇒ {Extended Service Plan} [30%, 70%]

- Algorithms are available for generating association rules
  - Apriori
  - Eclat
  - FP-Growth
  - + Derivatives and hybrids of the three
- The algorithms help identify the frequent item sets, which are, then converted to association rules

- الخوارزميات متاحة لتوليد قواعد الإتباط
  - Apriori
  - Eclat
  - FP-Growth
  - + مشتقه وهجين من الثلاثة
- تساعد الخوارزميات على تحديد مجموعات العناصر المتكررة ، والتي يتم تحويلها بعد ذلك إلى قواعد الارتباط

Apriori Algorithm

- Finds subsets that are common to at least a minimum number of the itemsets
- Uses a bottom-up approach
  - frequent subsets are extended one item at a time (the size of frequent subsets increases from one-item subsets to two-item subsets, then three-item subsets, and so on), and
  - groups of candidates at each level are tested against the data for minimum support.

(*see the figure*) → --

- يجد مجموعات فرعية شائعة على الأقل العدد الاصغر لعدد العناصر
- يستخدم نهجًا من القاعدة إلى القمة
  - يتم توسيع مجموعات فرعية متكررة عنصر واحد في كل مرة (يزيد حجم المجموعات الفرعية بشكل متكرر من مجموعات فرعية عنصر واحد إلى مجموعات فرعية اثنين ، ثم مجموعات فرعية ثلاثة عناصر ، وهكذا) ، و
  - يتم اختبار مجموعات المرشحين على كل مستوى مقابل البيانات للحصول على الحد الأدنى من الدعم.

Raw Transaction Data

Transaction No	SKUs (Item No)
1	1, 2, 3, 4
1	2, 3, 4
1	2, 3
1	1, 2, 4
1	1, 2, 3, 4
1	2, 4

One-item Itemsets

Itemset (SKUs)	Support
1	3
2	6
3	4
4	5

Two-item Itemsets

Itemset (SKUs)	Support
1, 2	3
1, 3	2
1, 4	3
2, 3	4
2, 4	5
3, 4	3

Three-item Itemsets

Itemset (SKUs)	Support
1, 2, 4	3
2, 3, 4	3

Data Mining Software → هذي البرمجيات في الرسمه سلايد ٥٢

- Commercial → التجاري
  - IBM SPSS Modeler (formerly Clementine)
  - IBM - Intelligent Miner
  - … many more
  - SAS - Enterprise Miner
  - StatSoft – Statistica Data Miner
- Free and/or Open Source → مجاني او مفتوح المصدر
  - R
  - RapidMiner
  - Weka…

Data Mining Myths → اساطير التنقيب عن البيانات

- Data mining …
  - provides instant solutions/predictions
  - is not yet viable for business applications
  - requires a separate, dedicated database
  - can only be done by those with advanced degrees
  - is only for large firms that have lots of customer data
  - is another name for the good-old statistics

- التنقيب عن البيانات ....
  - يوفر حلول / توقعات فورية
  - ليست قابلة للتطبيق بعد لتطبيقات الأعمال
  - يتطلب قاعدة بيانات منفصلة ومخصصة
  - يمكن القيام به فقط من قبل أولئك الذين حصلوا على درجات علمية متقدمة
  - هو فقط للشركات الكبيرة التي لديها الكثير من بيانات العملاء
  - هو اسم آخر للإحصاءات القديمة

Common Data Mining Blunders → الأخطاء العامة للتنقيب عن البيانات

1. Selecting the wrong problem for data mining
2. Ignoring what your sponsor thinks data mining is and what it really can/cannot do
3. Not leaving insufficient time for data acquisition, selection and preparation
4. Looking only at aggregated results and not at individual records/predictions
5. Being sloppy about keeping track of the data mining procedure and results
6. …more in the book

1. اختيار مشكلة خاطئة لتعدين البيانات
2. تجاهل ما يعتقده الكفيل بالبيانات وماذا يمكن / لا يمكن فعله
3. عدم ترك الوقت غير الكافي لاكتساب البيانات ، والاختيار والتحضير
4. النظر فقط في النتائج المجمعة وليس في السجلات / التوقعات الفردية
5. كونها غير واضح بشأن تتبع إجراءات ونتائج استخراج البيانات