



Learning from data

د.أبي صندوق

2020\7\8

محتوى مجاني غير مخصص للبيع التجاري

RB

Informatics;

data engineering

تحدثنا في المحاضرة السابقة عن طرق التفكير واليوم سوف نتحدث عن الخوارزميات فسوف نستنتج أن لكل شيء له أداة معينة والأداة يمكن أن تعطيني أكثر من معلومة ممكن أن تكون معلومة توصيفية ويمكن أن تكون ربط بين العناصر ولأن هذا هدفي فيلزماني هذه الأداة (الأدوات المتاحة لي هذه وهذه و...) وأختار أداة معينة فسوف نتحدث عن الأدوات فإن الـ *KDD (Knowledge discovery and data mining)* وهذا يعبر عن مصطلح الـ *data mining* أن لدي داتا استخرج منها معلومة حيث هذه المعلومة إحصائية أي إذا أخرجت معلومة فليست واثق منها وهذه المعلومة تتكرر

مثلا شخص يقوم بشراء غرض فهو يفعل ذلك بنسبة مثلا 70% فهذا النوع من المعلومات تخرجه بشكل إحصائي وهذا النوع من المعلومات الذي ينتج معنا يدعى بالـ *pattern* وهو الشيء الذي يتكرر هذه الحالة تتكرر مرة ومرتين وهكذا حتى 70% من الداتا تطبق عليهم هذا الـ *pattern* فهذا الـ *pattern* الذي أحاول البحث عنه ويكون من نمط الذي واصفته *x* هكذا قيمته والذي واصفته *y* هكذا قيمته والذي *x, y* يمثلو قيمة معينة *z*، هذه قيمتها لأننا لا نعرف لذلك نبحث عنه في نمط الـ *pattern* ففي البداية ننظف الداتا ومن ثم نبحث مثلا الـ *x* تعبر عن الطلاب الذين حصلوا على العلامات كذا في المادة كذا ينتموا للطلاب التي علامتهم عالية بالرياضيات مثلا يعني هذا النمط الذي أبحث عنه فأعبر عن المعلومات بشكل إحصائي فحتى أحصل على الـ *pattern* ممكن أن أصنف العينة أو أخرج الترابطات بين الـ *features* بشكل إحصائي مثلا الـ *support* لها 70% أي موجودة بالداتا بنسبة 70% (أي أنا واثقة من هذه القاعدة بنسبة 70%).

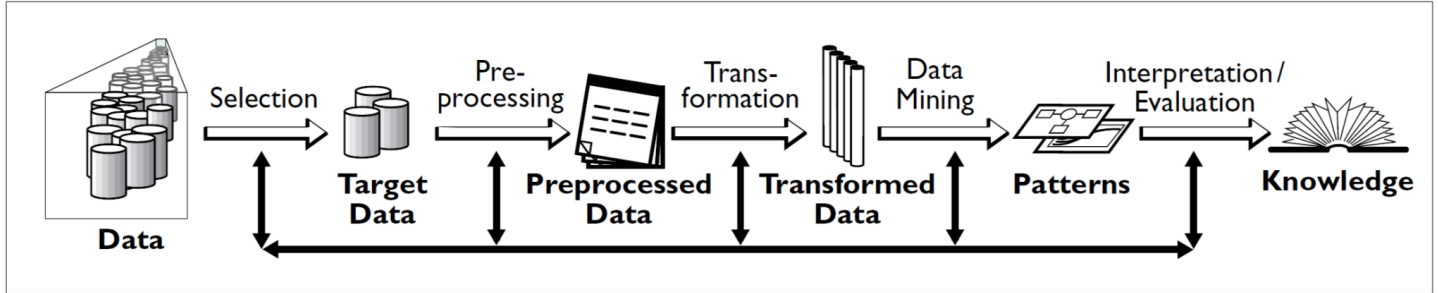
تعريف الـ *KDD*:

" the nontrivial process of identifying valid , novel , potentially, useful , and ultimately understandable pattern in data"

أي هو عملية غير بسيطة لاكتشاف وتحديد الـ *pattern* (الذي يتكرر بالداتا) وهي صحيحة وممكن أن يكون لها فائدة بالعمل والهدف النهائي أن تستطيع قراءتها وتحدد الاسباب وممكن أن نحصل على *pattern* قابلة للاستخدام ولكن غير قابلة للتفسير .

❖ حيث أنه يمر بهذه المراحل:

Figure 1. Overview of the steps constituting the KDD process



عندي مثلاً داتا موزعة جزء منها على الإنترنت وجزء منها على النظام وأجت داتا موزعة بداتا بيز كثيرة بتعمل عليها *selection* للداتا الي بدك تعمل عليها تحليل وأحياناً بتكون الداتا محدودة (جاهزة) فبتكون قطعت عملية ال *selection* بعدين بتنظف الداتا (*preprocessing*) مثلاً إذا فيها تعارضات نخلص منها وإذا فيها شيء غير دقيق حتى تصل لعينة تقدر أن تتعلم منها (فيها كل معلوماتها لحد ما صحيحة) أي تحقق معايير ال *Quality* نقوم بعمل لها *transformed* أي شغلات قابلة للقياس حيث تكون جاهزة للتحليل (نقوم بعمل عليهم عمليات ال *data mining*) حيث سوف يخرج منه *pattern* نقوم بعمل عليه *evaluation* يصبح *knowledge* حيث هذا ال *knowledge* يمكن أن يؤثر على جميع قراراتك السابقة أو على شو الداتا التي سوف تستخدمها أو يغيرك الداتا يمكن يغيرك قراراتك على ال *preprocess* الي ممكن شي أهملته وبس وصلت للأخير لقيته شيء مفيد وممكن يغيرك قراراتك بال *transformed* وممكن يغير رأيك بال *pattern* الي طلعت).

❖ حيث سوف نطبق الخوارزميات التالية :

❖ Association

❖ Clustering

❖ Classification

❖ Regression

❖ Point of interest detection

❖ Information filtering

❖ Visualization

❖ Summarization

Association Analysis

سوف نتحدث في البداية عن خوارزميات ال Association تحدثنا سابقا *collaration* والتي تعني لدينا متحولين في ارتباط بينهم يعني إذا واحد تغير يتغير الثاني وعنا نوعين من الارتباط فالارتباط الإيجابي: إذا المتحول الأول زاد فالمتحول الثاني يزيد أما الارتباط السلبي يعني إذا المتحول الأول زاد فالثاني ينقص .

وهذه الحالة ممكن أن تدرس على أي *2 features* ونقيس ال *collaration* مثلا ويطلع رقم بين $[-1, +1]$ حيث أن ال 0 تعني أنه لا يوجد ارتباط وال 1 تعني أن الارتباط إيجابي بنسبة 100% وال -1 تعني أن الارتباط سلبي بنسبة 100% وأي رقم بينهما يعتبر درجة من الارتباط .

هذا بالنسبة لل *2 features* أما بالنسبة ل *3 features* كيف نقول أنه يوجد ارتباط بين الثلاثة ؟؟

إحصائيا فيك تقول في ارتباط بين الأولى والثانية أو الثانية والثالثة أو الأولى والثالثة وفيما كمان نقول أنو إذا كانت الأولى لها القيمة كذا والثانية لها القيمة كذا فإن الثالثة لها قيمة هكذا أي نعني الارتباط بين الأولى والثانية أثر على الثالثة وكل مازاد عدد ال *features* كلما تعقد الأمر.

مثلا العلاقة بين *3 features* أن الثلاثة تحققوا سوية بنسبة 70% .

لدينا المثال التالي موجود في السلايد فلدينا *dataset* فلاسطر تعبر عن ال *transaction* والأعمدة عن ال *features* وليكن مثلا القاعدة $\{Milk\} \Rightarrow \{Butter, Bread\}$.

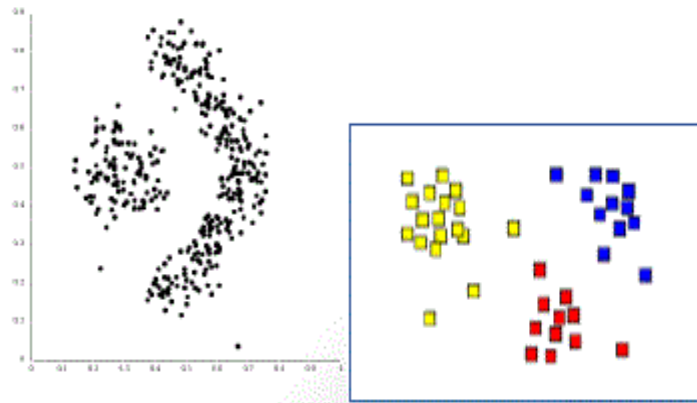
كيف بدنا نقيس أن القاعدة محققة أم لا ففي البداية سوف نرى كم سطر فيه الثلاثة سوا بغض النظر عن الترتيب فنلاحظ أنه يوجد 1 من أصل 5 وهذا يدعى *support* والتي هي 20% فكلمة *support* تعني قدي أنت متأكد للرأي الذي تحطه فهنا عندي الداتا 20% تدعم هذا الرأي ونلاحظ أن ال *support* نقيس مدى جودتها حسب الداتا فإذا كانت الداتا نوعا ما ليست كبيرة فيكون في حال كانت بنسبة 10% لحد ال 60% فهي ممتازة لأننا نقيسها على مستوى وجود جميع ال *features* ال *rules* أما ال *confidence* فتعني كم مرة جاء القسم اليساري في حال جاء القسم اليميني وهنا قيمتها 100% وحتى تكون جيدة يجب أن تكون قيمتها عالية أي فوق ال 60% أما ال *lift* فهي اختصار للمعيارين السابقين وربط بين ال *support* وال *confidence* والعلاقة بينهما تناسب طردي فإذا كانوا المفهومين السابقين عاليين فال *lift* يكون عالي وتحسب ال *lift* من خلال مجموع ال *support, confidence* على جداءهم فيكون 50%.

القواعد التي تخلق من ال *dataset* تطبق عليها خوارزمية *aprior algorithm*.

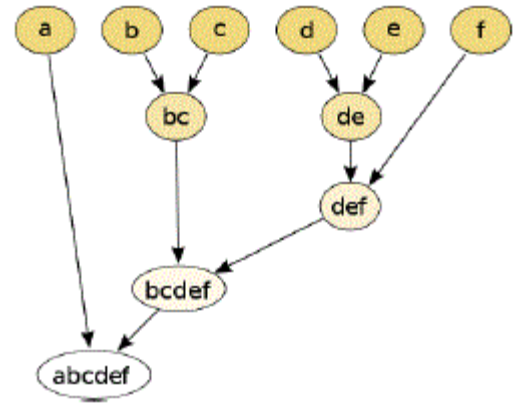
Clustering

وهو النوع الثاني من أنواع الخوارزميات وهنا يهملنا الأسطر التي تشبه بعضها والأمثلة التي تشبه بعضها (القريبين من بعضهما البعض) وليس بالضرورة أن تكون قيمة ال *cluster* واحدة أي ممكن يكون العشرين مثال متشابهين وممكن الأربعين مثال متشابهين وهكذا وليكن لدينا علامات طلاب وأريد تقسيمهم على العلامات المتقاربة من بعضها (علامات الطلاب القريبة من بعضها بغض النظر هل هم متقاربين عاليين أم متقاربين ضعاف أو.....) فهذه الخوارزمية يهملها المتقاربين من بعض

مثلا لدينا صف ينقسم إلى 15 مجموعة كل مجموعة متقاربة من بعضها على المستوى الأكاديمي , ومثلا بالمسألة لدي بدي اعرف الزبائن التي تشبه بعضها فنقسم الزبائن لمجموعات (من يأتي كثيرا أو من يأتي نادرا وهكذا ...) فنقول أن هؤلاء ال *cluster* يشبهون بعض (يأتون كثيرا) وهؤلاء ال *cluster* يشبهون بعض (يأتون نادرا) وكل *cluster* تدعى بـ *schema* فنحن سوف نقوم بتصميم تابع هو من يقوم بقياس التشابه بين القيم ليعطي ال *cluster* وبناء على تصميمي لهذا التابع يسهل أو يصعب ال *clustering* (العملية سهلة جدا في حال معرفتنا بتابع التشابه) فمثلا لدينا الشكلى :



فيني أحسب المسافة الاقليدية بين أي نقطة و *cluster* لنعرف إلى أي *cluster* تنتمي , ففي الرسمة الثانية كما نلاحظ لدينا *2cluster* فلو استخدمنا أي خوارزمية لحساب مسافة اقليدس سوف يخرج لدينا *2cluster* بحساب النقاط القريبة وهذه الطريقة نسميها بـ *global clustering* لأننا نرى المسألة ككتلة واحدة , ويوجد نوع ثاني يسمى *hierchical* فهذا ينظر إلى المسألة شغلة شغلة ليس ككل ونرى تدريجيا النقط القريبة من بعضها البعض ونرى مثلا كل نقطتين قريبين من بعض نضعهم مع بعض ونستبدلهم بنقطة وحدة ونمشي تدريجي كلشي نقاط قريبة عليه ضيفن .



لدينا نوعين من الclustering:

- *hard* : كل نقطة تنتمي لcluster 1.
- *soft* : كل نقطة تنتمي ل many clustering ولكن بنسب معينة.

Classification

خوارزميات classification هي حالة خاصة من التنبؤ , فيكون لدينا مجموعة من features يتم استعمالها لايجاد قيمة feature أخرى.

حيث تكون مجموعة features هي الدخل والنتيجة هي label

كما ذكرنا سابقا انه يمكن لمسائل prediction ان تكون للمستقبل او للحاضر , ويمكن أيضا استعمال label الناتجة حاليا كدخل لحالة تنبؤ أخرى (مثل استخدام العمر للتنبؤ بالطول , او العكس استعمال الطول للتنبؤ بالعمر وفي الحالتين هي عملية classification و prediction)

والهدف من هذه العملية الوصول الى هذا التابع :

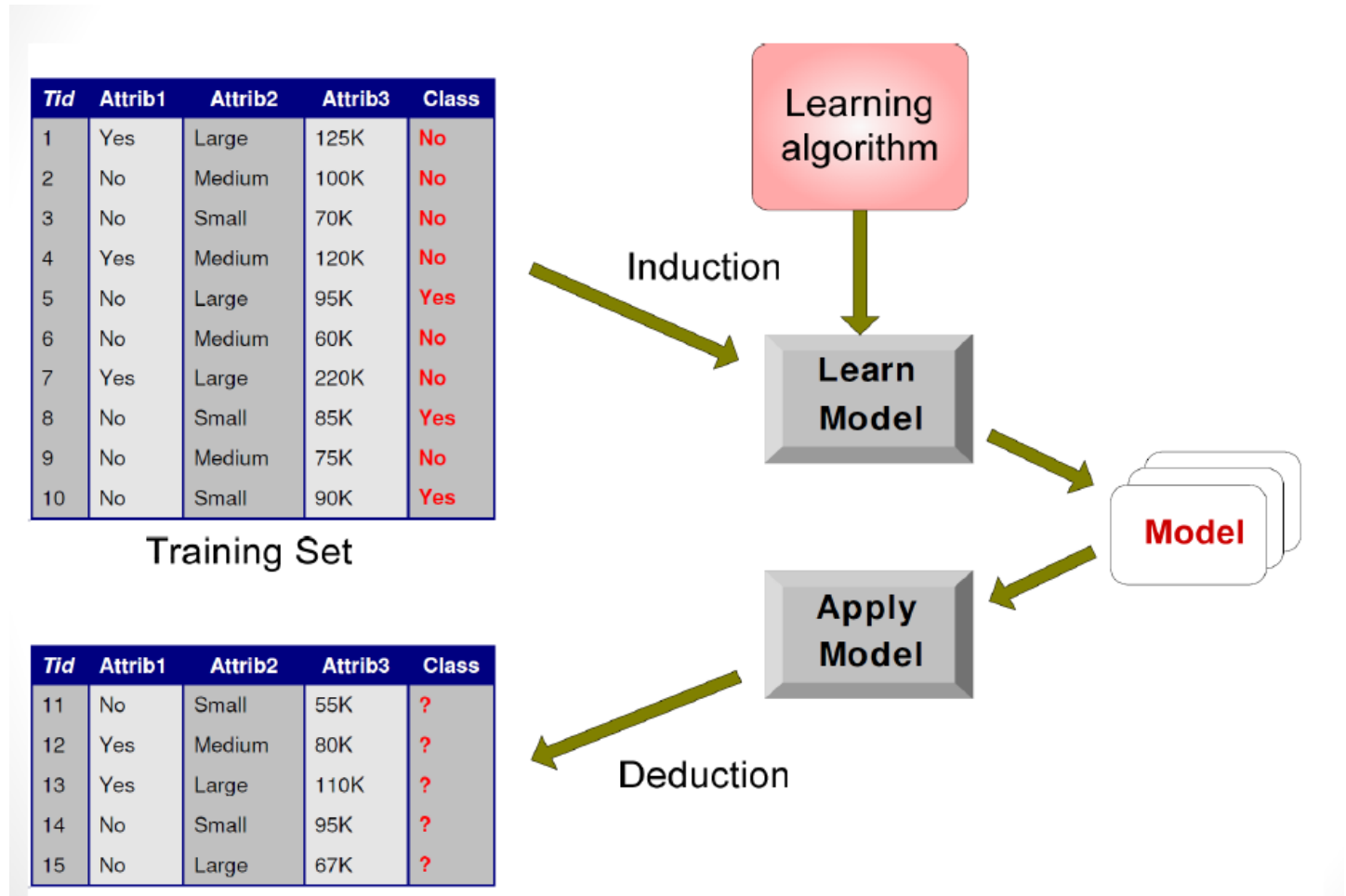
$$f(\text{features: parameters}) \rightarrow \text{label}$$

عند اعطائه feature يعطي label بأقرب ما يمكن ل label الحقيقية ,

مثلا : لدينا label ان الطالب سينجح في مادة هندسة نظم معلومات ام لا فيكون الدخل عبارة عن معلومات عن زملاء الطالب في نفس الدفعة وماذا كان وسطي علاماتهم في مواد معينة (برمجة , خوارزميات ...) وبعدها بناء التابع على هذا الدخل

يتم الاستفادة من هذا العمل في المستقبل حيث يمكن التنبؤ بنتيجة طالب (النجاح او الرسوب وهي LABEL) في مادة هندسة النظم بناء على علاماته السابقة في مواد أخرى

إذا يجب محاولة الوصول الى model مشابه للتالي :



بفرض لدينا عينة تحوي past experience فيها 3 features و label فالمسألة هي استعمال هذه الداتا من اجل التعلم عن طريقها وانشاء model وفي المستقبل عند الحصول على داتا جديدة نستطيع تطبيق function الناتج عن التعلم السابق للوصول الى label الخاص بها , أي نحاول دائما العمل على قيم جديدة ل features وذلك من اجل التعميم (أي إيجاد model)

بفرض حصل تقاطع بين الداتا القديمة وداتا جديدة (أي حصلنا على نفس ال record) فيجب تذكر هذه الداتا وإعطاء النتيجة دون الحاجة الى تمريرها على function ولكن ليس هذا الغرض من classification لذلك يجب العمل دائما مع داتا جديدة

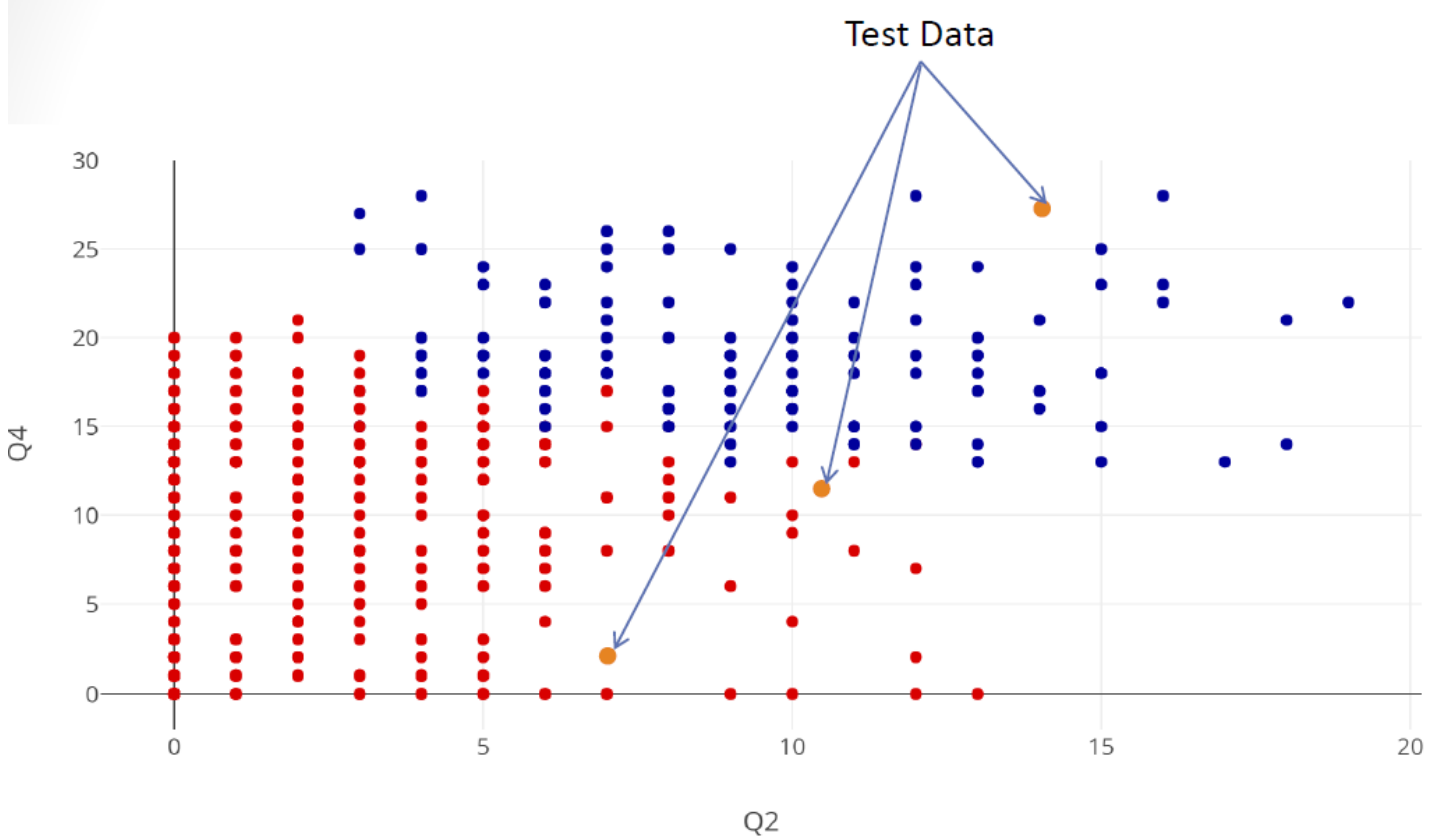
مصطلح التعميم generalizability مصطلح أساسي

نأخذ من الداتا ونعمم لداتا أخرى (أي نأخذ كل features و labels) ونقوم بادخالهم الى خوارزمية learning model تنتج عنها model وهو ال function ونقوم بتثبيت ال parameters الخاصة به , أي لعملية learning لا نعلم parameters ولكنها تنتج عنها عن طريق خوارزمية ما ,

يمكن ان يحوي ال model على نسبة خطأ أي عند ظهور النتيجة تكون مع معامل ثقة معين .

Predictive Data Analysis

بفرض هذه العينة :



النقاط الحمراء والزرقاء تعبر عن العينة اذا لدينا 2 features (عمودين من الداتا , وعدد الاسطر مساوي لعدد النقاط)

والنتائج معبر عنها بالحالتين اما ازرق او احمر

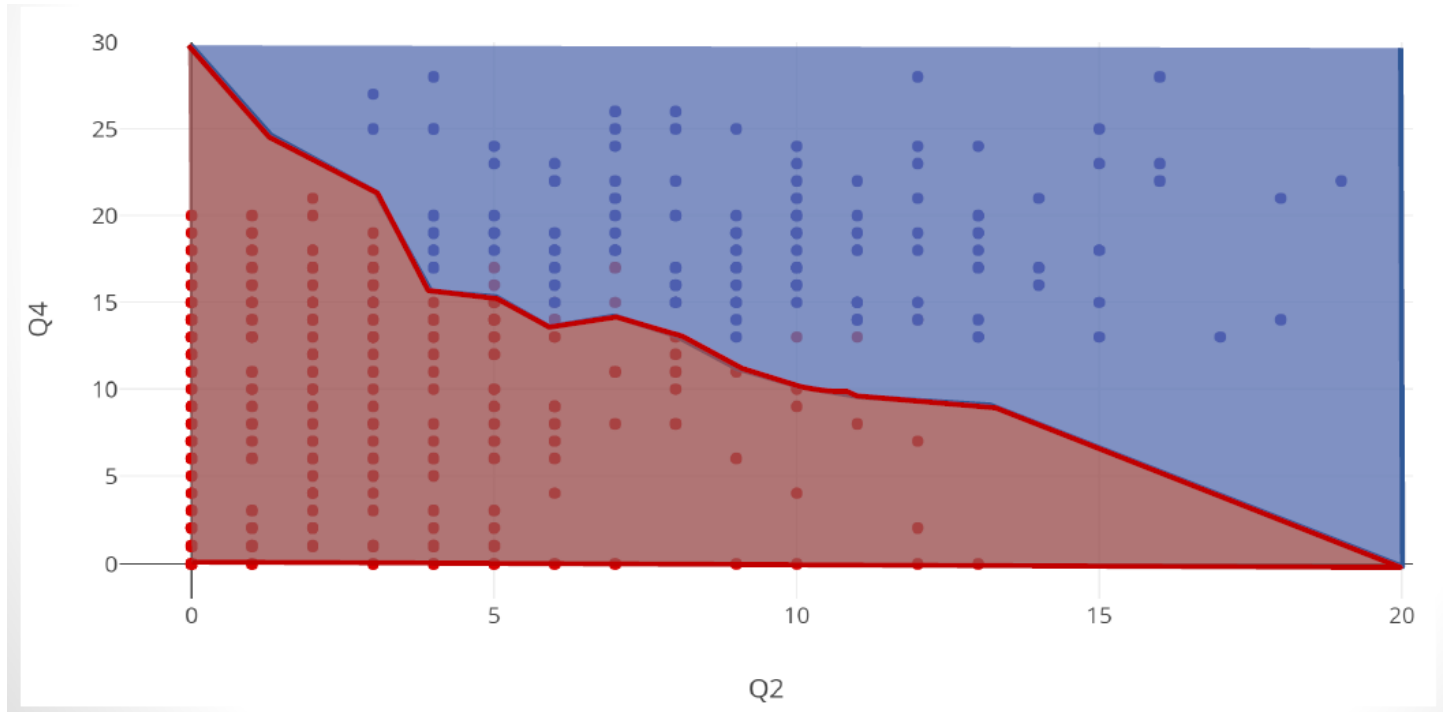
النقاط البرتقالية هي الداتا الجديدة التي لا يوجد عليها معرفة مسبقة ويتم إدخالها الى function لايجاد label الخاص بها

كيف نقوم بذلك ؟

احدى الخوارزميات البسيطة المساعدة على ذلك هي :

k-nn

تفترض الخوارزمية ان كل نقطة في الفضاء ما هي اقرب النقاط اليها , يتم التلوين بنفس لونها فينتج الشكل التالي :



بتلوين الفضاء للعينة السابقة (بما انه بعدين فهو سهل التلوين) فلاحظ انقسامه الى قسمين , الحد الفاصل بين اللونين هو ما نبحث عنه وهو حد القرار decision boundry وجميع خوارزميات prediction تسعى للوصول اليه

كلما ابتعدنا عن حد القرار نكون متأكدين اكثر اما بقربه فنسبة الثقة ضعيفة

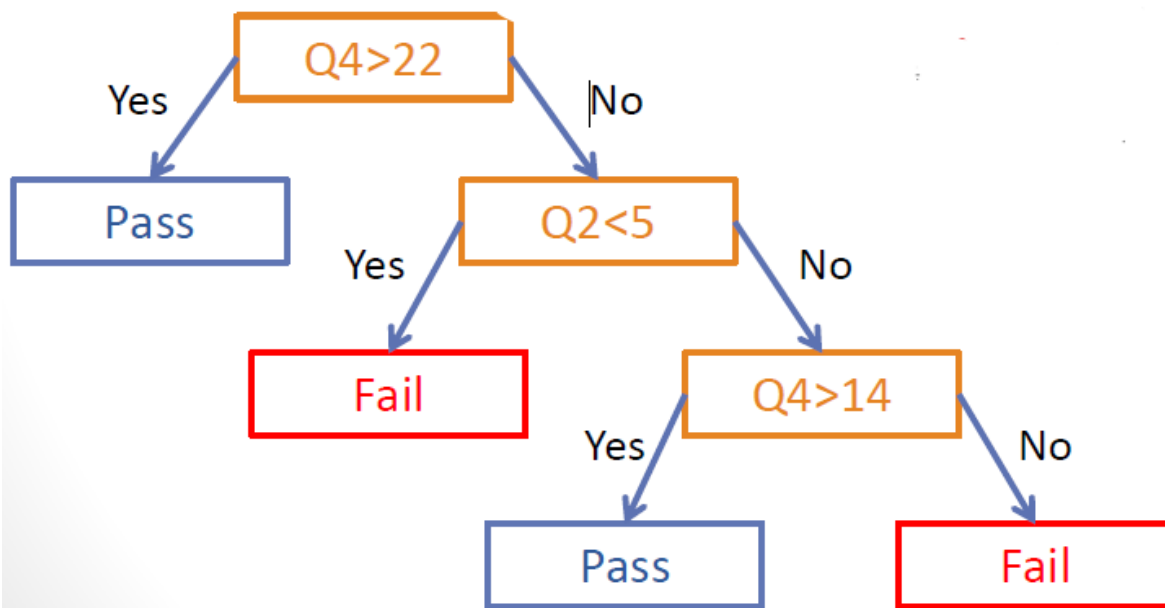
في clustering الداتا هي من تقرر ال label اما predictive نحن من نقرر label

النمط الثاني من خوارزميات classification هي Decision Tree

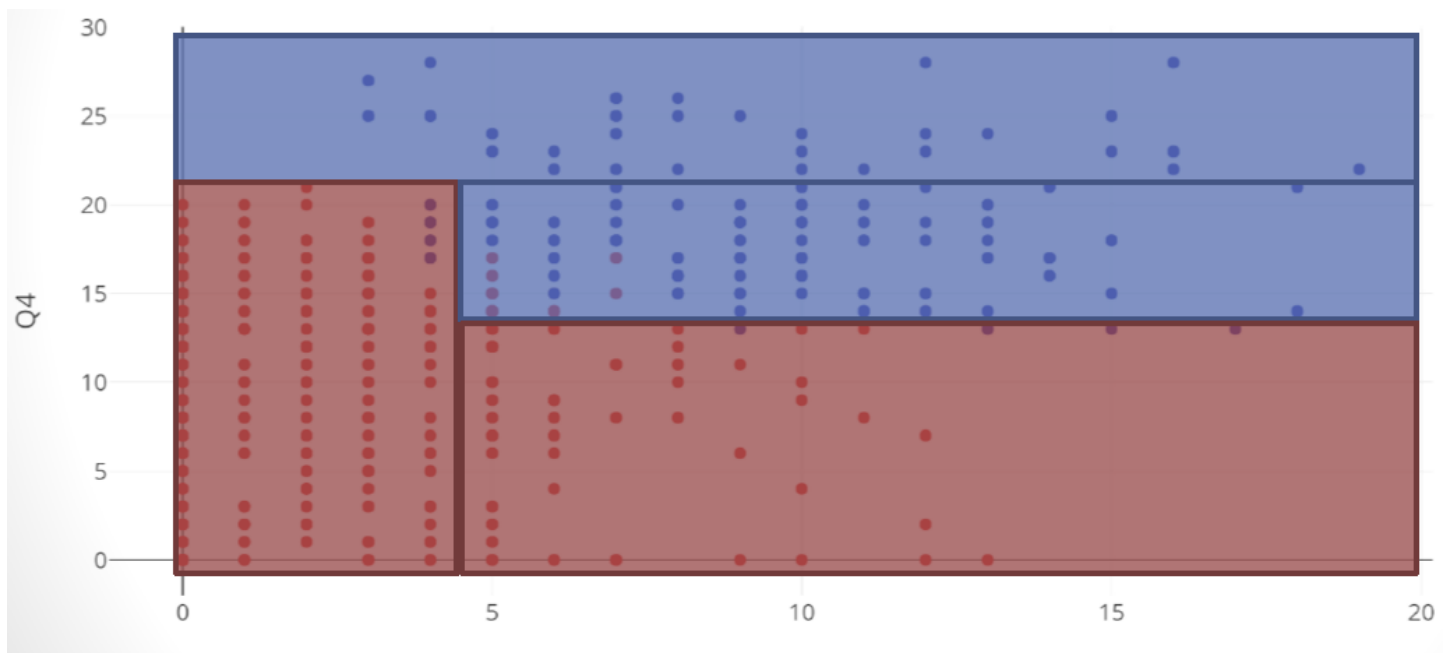


Decision Tree

وهي تقوم بتقسيم الفضاء الى أجزاء :



تفترض ان اذا كان الشرط محقق نفذ a والا نفذ b فيتم تقسيم الفضاء الى قطع , كما يبين الشكل التالي :

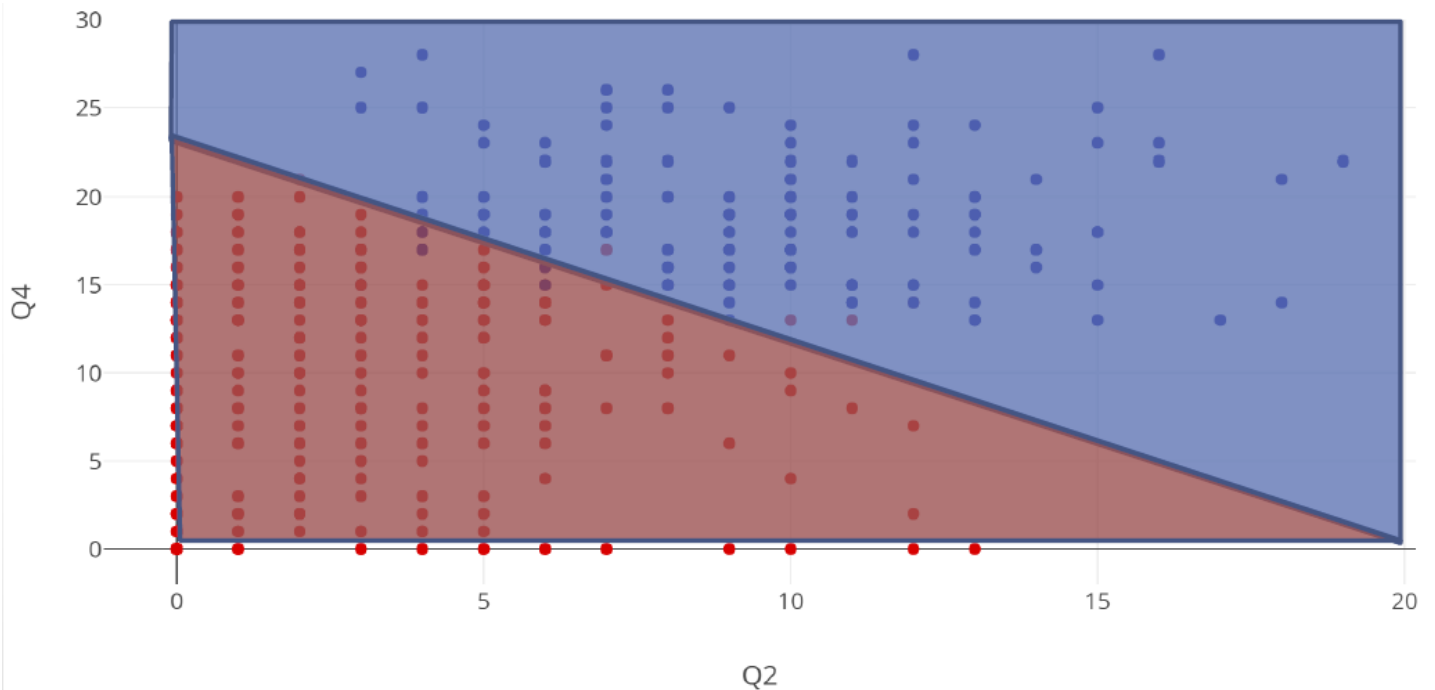


وال parameters في هذه الحالة هي حدود القرار , (في مثال العلامات السابق : 22 و 25 ... هي ما تتعلمه خلال training وبعمق الشجرة الناتجة يتم اخذ القرارات , فاذا كان عمقها 3 هذا يعني باننا سناخذ 3 قرارات , في الشكل الحد الفاصل بين الأزرق والاحمر هو حد القرار

الطريقة الثالثة لايجاد حد القرار هي :

Perceptron

وهي تعتمد على إيجاد خط واحد (حد القرار) بمعادلة خطية



في هذه الحالة اذا كانت لدينا نقاط حمراء في الخانة الزرقاء تصبح تابعة لها وذلك لان حد القرار خطي مثال :

باعتقاد على dataset علامات الخوارزميات السؤالين 2 و 4 هما features نرمز لهما q_2 , q_4 وال parameters هي w_1, w_2 والثابت فبحل المعادلتين :

$$f(Q_1, Q_2; w_1, w_2, c) = w_1 * Q_2 + w_2 * Q_4 + c \leq 0$$

فالتابع من الشكل : $ax_1 + bx_2 + c$

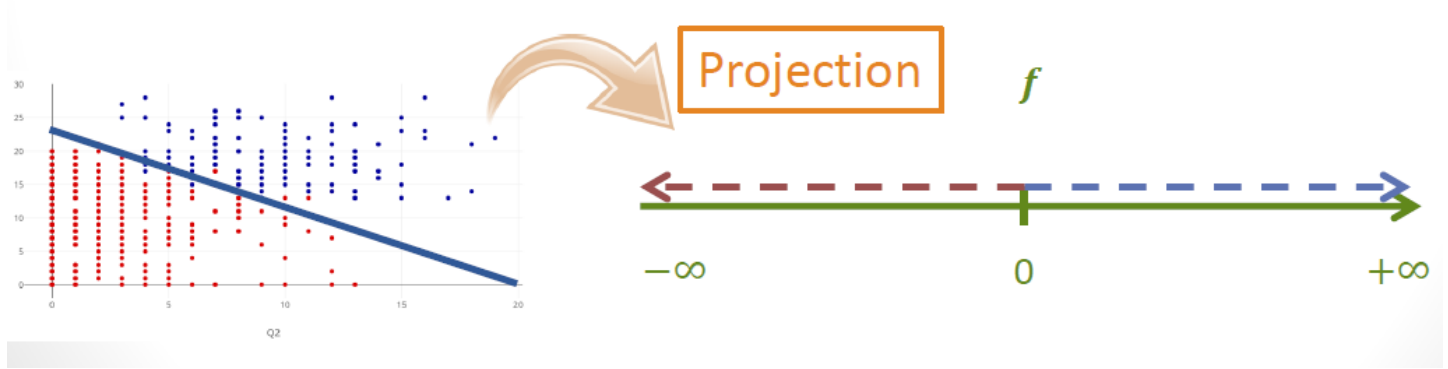
الثوابت a, b, c نحصل عليها خلال التعلم , أي باختيار قيم q_2 , q_4 كبيرة مثل 15 و 20 فيجب ان تكون النتيجة موجبة :

$$f(15, 20; w_1, w_2, c) = w_1 * 15 + w_2 * 20 + c > 0$$

اما اذا كانت القيم ل q_2, q_4 اقل فالنتيجة سالبة :

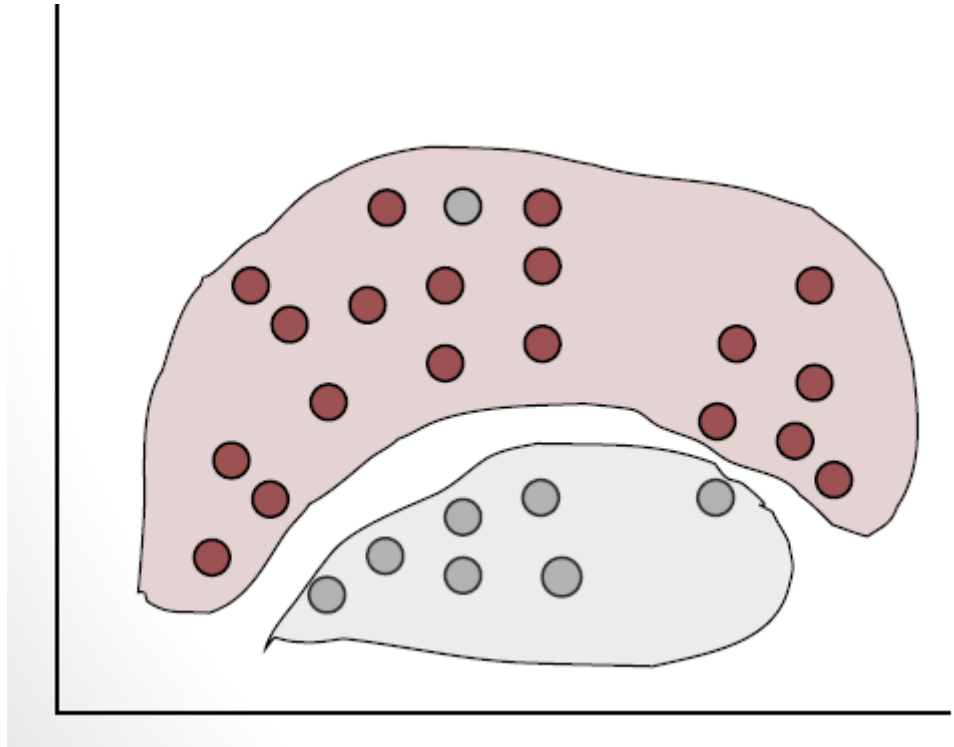
$$f(5,10;w_1,w_2,c)=w_1*5+w_2*10+c<0$$

فنحن نبحت عن قيم ل w_1, w_2, c تحقق كل النقاط من اجل تحويل المسألة الى مسألة خطية يمكن تمثيلها على محور الاعداد وكلما ابتعدنا عن الصفر اصبح معامل الثقة اكبر

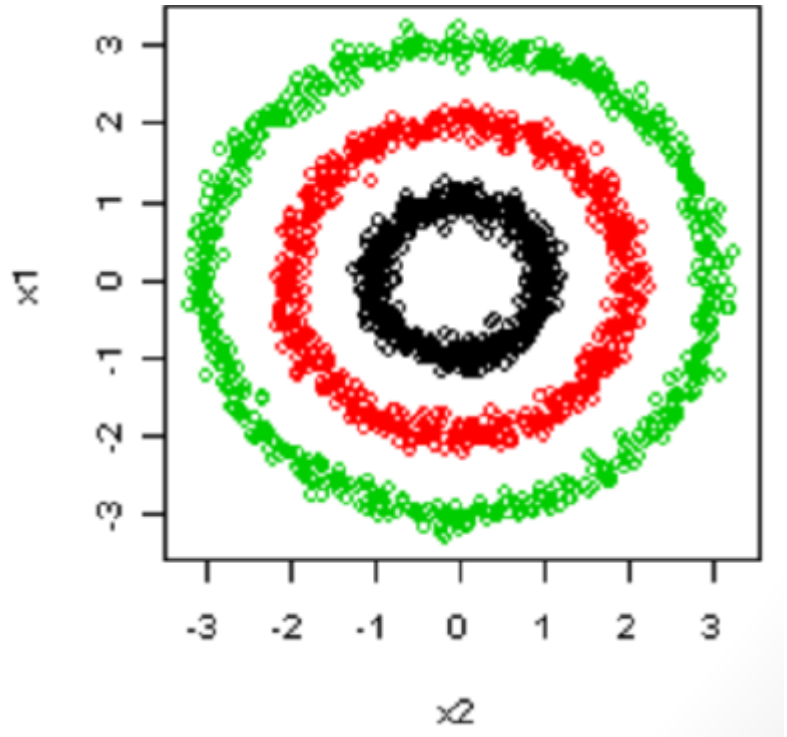


Projections

كل مسائل prediction يجب تخيلها بالشكل السابق ولكن قد يكون فضاءها معقد مثل الفضاء التالي :



او بهذا الشكل :



بهذه الحالة لم يعد بالإمكان حلها خطيا , ويصبح حد القرار معقدا ,

مثلا : في الشكل السابق يوجد 3 labels ويكون شكل حد القرار : الأول بين الأسود والاحمر والثاني بين الأحمر والاخضر فعندها النقاط الأقرب للمركز تعتبر اسود , النقاط في المسافة بين الأحمر والأسود احمر , النقاط في المسافة بين الأحمر والاخضر اخضر

إذا يمكن تخيل ان المسألة يمكن ارجاع شكلها الى مستقيم والمطلوب معرفة البعد عن نقطة معينة فيها من اجل تحديد الثقة في القرار

: Projections may be non-linear

Kernel SVM, neural networks, etc... : يمكن ان تحل مثل هذه المسائل (يستطيع إيجاد model يحدد

حد القرار)

The concept is the same

Require huge amounts of data to tune non-linear models

Non-linear models have higher capacity to model the real world.

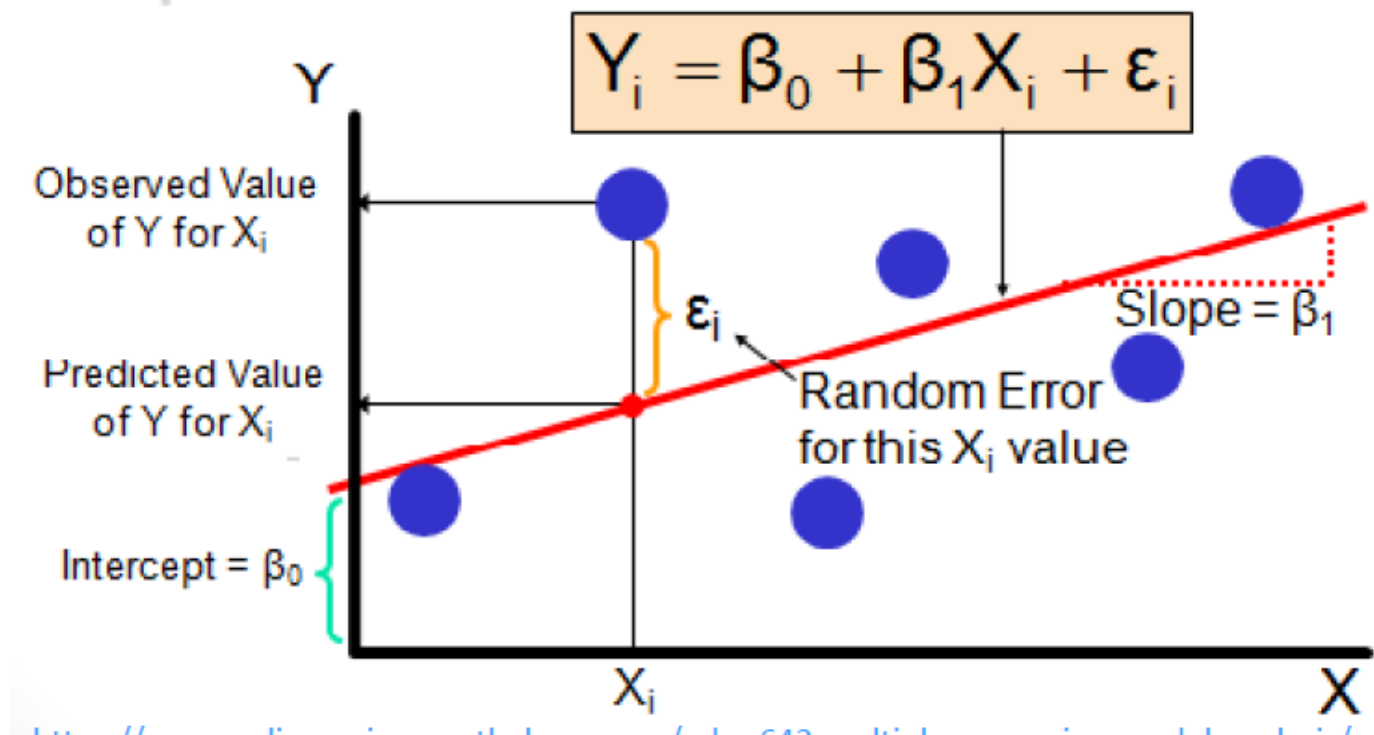
Regression

وهي مسألة تعميم لمسألة classification ولكن label فيها هو رقم وليس class

أي بدلا من محاولة معرفة اذا كان الطالب سينجح او يرسب فهي تبحث في ما هي علامته

اذا أسندت المسألة الى ranges عادت الى classification , اما اذا اردنا الرقم الدقيق تصبح regression او curve fitting وهي تدعي بهذا الاسم لأننا نبحث عن معادلة دخلها ارقام وخرجها ارقام , أي تمر في مجموعة النقاط التي لدينا

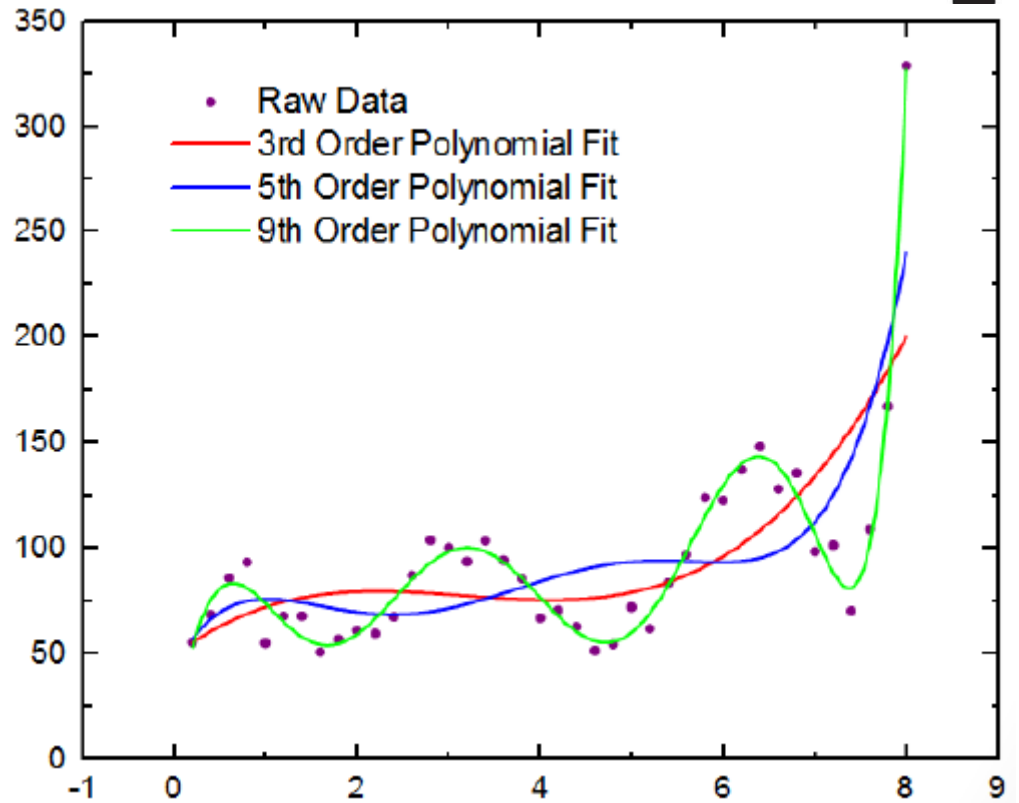
قد لا يتحقق وجود هكذا تابع ليمر على كل النقاط المطلوبة , لذا يجب إيجاد الطريقة المناسبة لجعل المستقيم يمر اقرب ما يمكن الى كل النقاط
فينتج لدينا تابع من الشكل :



يمكن ان يكون Ridge Regression, LASSO Regression, Polynomial Regression, etc

في بعض الحالات يمكن لاكثر من curve ان يكون fit :

في الحالة التالية لدينا عدد كبير من النقاط , فيمكن القيام بعمل fit بتابع اسي وكلما تم رفع الاس زاد تعقيد التابع وقام بعمل fit افضل لكنه يكون موجه لداتا محددة وبالتالي يصبح تعميمه سيء



الداتا التالية اذا تم عمل fit لها بتابع خطي فسوف يكون سيء

ملخص :

المسألة التي يجب ان نتعامل معها هي تعلم تابع عندما نقوم بتمرير features له وبالاكتفاء على parameters التي لديه يعطيني target ويقبل وجود هامش للخطأ

$$f(\text{features}; \text{parameters}) \rightarrow \text{Target} + \epsilon$$

Point of Interest Detection

مجموعة خوارزميات لكشف لأمثلة خارج القاعدة ويتطلب ذلك تعريف القاعدة

مرتبطة الى حد كبير بنظرية القرار ومستويات الثقة

عند إجراء معاملة في بنك نقل من حساب لحساب , يظهر البنك أنه نشاط مشبوه suspicious activity ويتم توفير الحساب حيث عادة عند الدخول الى الحساب فإن العمليات المتعارف عليها تمشي على رتم معين pattern وإذا لم يتم المرور بهذا ال pattern يتنبه لوجود خطأ او مشكلة ما

كيف عرف ذلك ؟

تم تعليم الداتا لتعرف شيء اسمه norm (normal behavior) وكل شيء خارج هذا الحد هو ليس norm
 حلول هذه المسألة ممكن أن تكون إحصائية ويتم تعريف حد القرار والوسط يدويا
 وممكن أن يكون حلها عن طريق ال classification أمثلة عادية وأمثلة suspicious وبالحالتين نصل لنفس
 النتيجة
 أمثلة عن استخدام هذه الخوارزميات :

Fraud Detection;
 Scene Change;
 Speaker Change;

Information Filtering


























مجموعة خوارزميات لفلتر العناصر التي تكون مفيدة أكثر من غيرها وضمن اهتمام المستخدم باستخدام السلوك
 السابق

تحليل الداتا والتفاعل معها

ممكن أن تكون عناصر النموذج والصلة بينها , المستخدمين والصلة بينهم والتفاعل بين العنصر والمستخدم
 ممكن أن تتم رؤيته ك نمط خاص من تقييم توقع نماذج الانحدار لعنصر معطى من قبل مستخدم ما
 ال Recommender systems تنتج recommendations باستخدام :

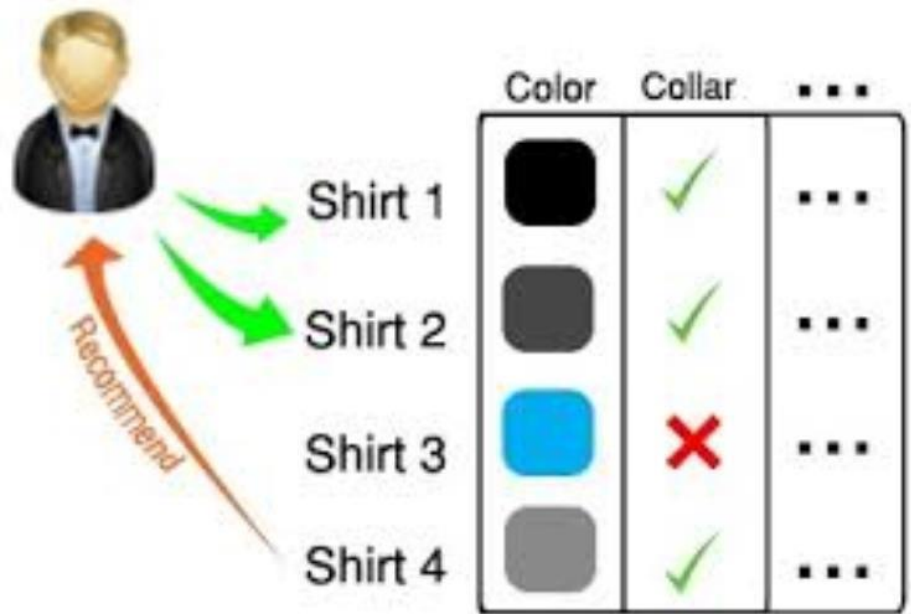
Collaborative filtering:

بناء نموذج من سلوك سابق لمستخدم (العناصر التي تم شراؤها او اختيارها او التقييمات المعطاة لهذه العناصر)
 بالإضافة الى قرار مشابه تم اتخاذه من قبل مستخدم اخر حيث تقوم الفلتر عن طريق التشابه على أعمدة معينة

Content-Based Filtering:

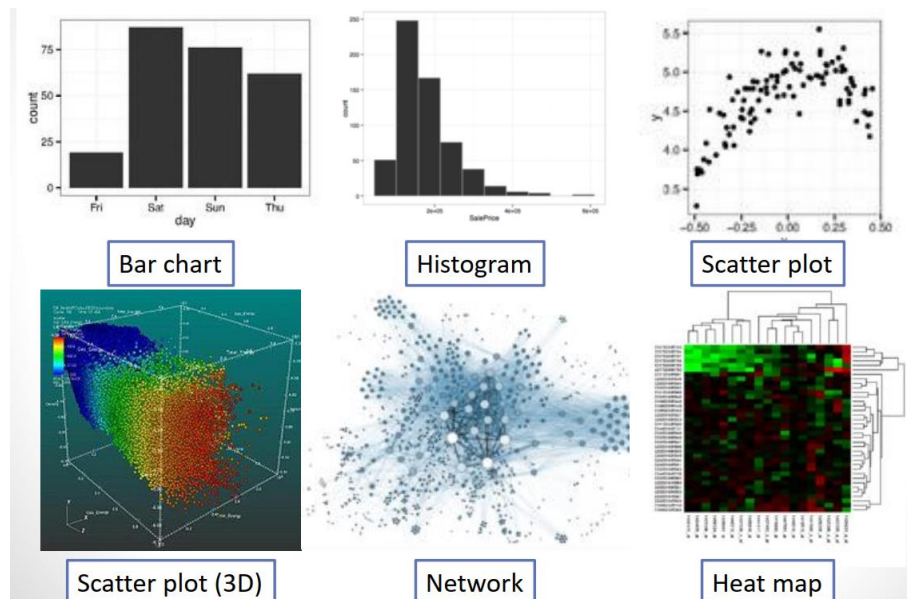
الاستفادة من سلسلة من الخصائص المنفصلة الخاصة بعنصر ما من اجل التوصية بعناصر إضافية مماثلة بالخصائص , حيث يتم قياس المسافة بين الاعمدة التي تشبه بعضها بناء على داتا خارجية , اذا الفلتر قائمة بناء على الواصفات



ويوجد نوع هجين بين النوعين السابقين بنتيجة وسطية

Visualization

سبب تعقيدها هو اخراج chart بديهي للفهم حيث نحتاج لدعم ال process التي تدعم القرار وبذلك محم بحاجة ل visualization مناسب ينظر له متخذ القرار

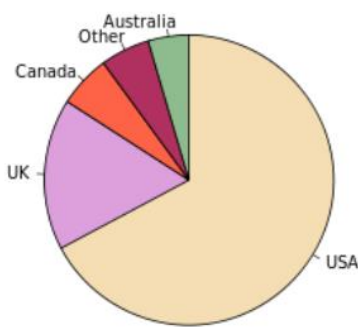


ال bar chart يدرس الكميات و يقيس العلاقة بين شيئين ونحن نحدد على أي أساس يتم التقسيم
 ال Histogram هو تجميعية لل bar chart ويعرض التوزيع الحقيقي والطبيعي للداتا والمحور مقسم لأنماط اذا هو
 يدرس التكرار

ال scatter plot يظهر العلاقة بين بعدين ودرجة الارتباط
 ننتقل للأكثر تعقيدا مثل ال network التي تبين اين الترابط الأساسي (الخط السميك للوصلات الأساسية والرفيع
 للوصلات الضعيفة)

ال scatter plot 3d يقيس المسافة بين كل مثال ومثال
 ال heat map مسافات بين n عنصر و n عنصر وتعطي معلومات عن العناصر القريبة من بعضها

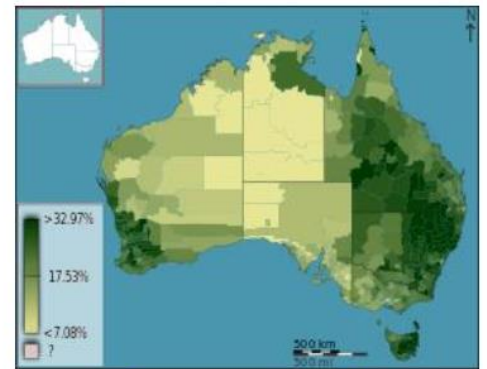
Descriptive Analysis – Vis.



Pie chart



Donut chart



Map chart

ال pie chart توضح نسب مئوية في حين ال donut chart , مثلا عدد المبيعات لكل حي وكل حي تابع لمدينة وكل
 مدينة تابعة لبلد حيث ال pie chart يمثل الدول ويتم التفصيل من الداخل الى الخارج (المقطع الداخلي ينقسم
 للخارج) ولا يشترط تواجد كل المدن او كل الاحياء .
 ال map chart يوضح توزع البيانات على الخريطة

Summerization

وصف مضغوط لمجموعة فرعية من ال data
مثلا الانحرافات المعيارية لكل المجالات وغالبا ما يستخدم التلخيص في الاستكشاف التفاعلي وتحليل البيانات وتوليد التقارير المؤتممة
التلخيص التلقائي: عملية تقصير نص او مستند او مقطع فيديو مع المحافظة على نقاط الاهتمام الاصلية وإعطاء نتيجة بملخص متماسك ومتجانس
وبشكل عام هي تقنيات تعتمد على الاستخدام والتجريد وهي مسألة information extraction وتستخدم بكثافة في أنظمة المستوى التنفيذي
الدكتور لم يتطرق لل Types of model

Structuring Data

تقنيات لترميز الأمثلة الفردية ولتحسين تمثيلهم لأدوات استخراج البيانات

Text (or Text Mining)
Images (or Computer Vision)
Audio
Video
Categorised data
Missing values
Feature normalisation
Representation Learning : كيف نحول مسألتنا لمسألة رقمية

-انتهت المحاضرة-

