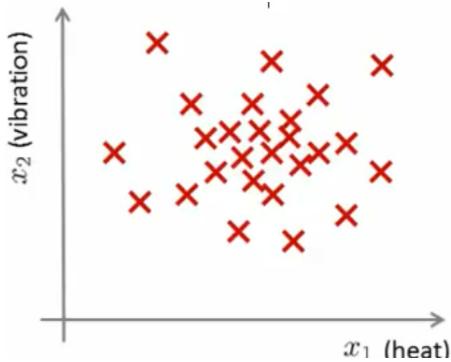


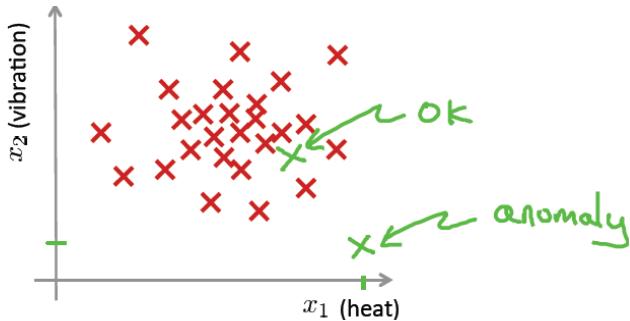
الأسبوع التاسع

Anomaly Detection

- و ترجمتها : اكتشاف القيم الشاذة
- و معناها , اني سواء في الـ **unsupervised** او الـ **supervised** , مهم اني اشوف اغلب القيم عندي موجودة فين , وده عشان اعرف لو عندي قيم غير طبيعية , وده ممكن يشير الي عطل موجود في الصناعة او فاكهة فيها مشكلة , او تعاملات بنكية مريبة
- فمثلا , لو عندي فحص لمحركات معينة , وكان عندي كذا **feature** بافحصه , اهمهم هو درجة حرارته و مقدار اهتزازه وقت التشغيل
- ولما جيت ارسم الاثنين **features** دول , لقيتهم بالشكل ده

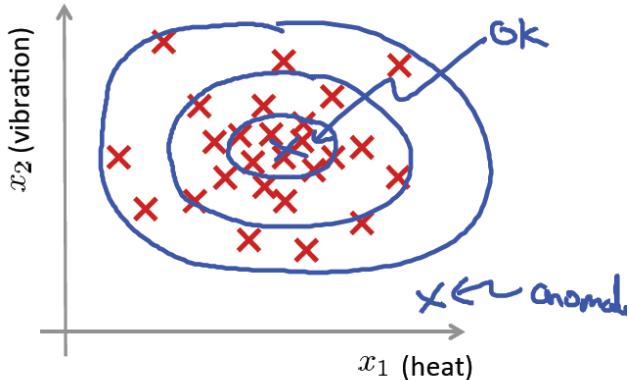


- كدة بقى عندي تصور عام عن القيم التقريبية لدرجة الحرارة و مقدار الاهتزاز و علاقاتهم بعض بعض
- بعدها جيت افحص عدد من المحركات لقيت قيمهم كدة



- فمن الواضح ان القيمة الخضرا اللي في النص قيمتها طبيعية بالنسبة لاخواتها , بينما القيمة اللي على اليمين شاذة عنهم , فده يخلينا نعيدي فحص المحرك ده , و نعرف هل فيه مشكلة ولا لا
- نفس الموضوع يتم في حالة تعاملات بنكية , او فاكهة غير سلية و هكذا
- و كانه من الآخر , بقياس مدى اقتراب او ابعاد البيانات المختبرة **test data** عن اغلبية القيم اللي عندي

- و الاقراب و الابعد يتم قياسهم بالنسبة لمركز البيانات ، يعني كانها الكتلة اللي في النص ، وبيكون ليها رقم P_x اللي بيمثل الدالة probability ، وبيتروح من 0 لـ 1 حسب اقترابه من المركز
 - فمثلا هنا :

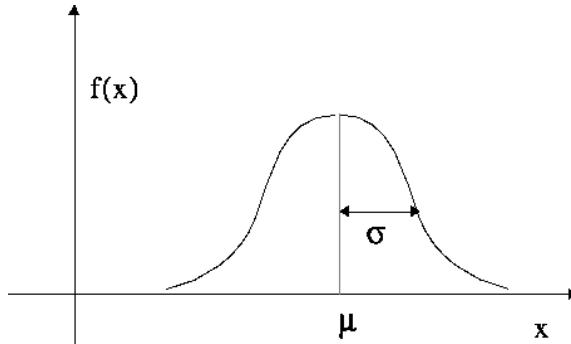


- اي قيمة في الدائرة الصغيرة ، ه تكون P_x بتاعتتها كبيرة ، وكل ما اروح للدواير البعد كل ما القيمة تقل ، لغاية لما اخرج برة اخر دائرة ، تبقى كدة قيمة شاذة anomalous
 - و بيتم تحديد قيمة معينة ابسلون ϵ واللي لو كانت الدالة P_x اكتر منها ، هتبقي القيمة موافق عليها ، ولو اقل منها هتبقي القيمة غير سلية ، وتعتبر قيمة شاذة
-

- لو تخيلنا تطبيق معين لده
 - عندنا موقع الكتروني ، وباعمل رصد لسلوك الزوار ، عدد الزائرين الف (قيمة m) و الدالة features اللي هرصدها 5 حاجات هي :
 - مدة زيارته
 - كام صفحة فتحها
 - كام تعليق عمله
 - كام شير عمله
 - كام فيديو فتحه
 - فهيكون عندي جدول ، فيه الف صف ، و 5 عواميد
 - فممك اعمل موديل للجدول ده، بحيث لما بيجي حد جديد (او حتى من بين الالاف) يعمل حاجة مش طبيعية (عدد صفحات اكتر من اللازم ، او مش بيعمل اي تعليق خالص) ، ساعتها اخذ باللي ان فيه حاجة مش طبيعية
 - مش لازم نرسمها ، خاصة اني مش هاعرف ارسم 5 ابعاد ، لكن الخوارزم نفسه هيحسبها و يطلعلي القيمة اللي الابسلون فيها اقل من الطبيعي ، يعني قيمة شاذة
 - ديه غالبا الطريقة اللي الفيسبوك ، او جوجل ، بيحس فيها ان ده fraud account و يطلب منه بيعت صورة الباسبور ، او يوقفه مدة معينة ، عشان يشوف حواره ايه
-

• توزيع Gaussian Distribution or Normal Distribution

- وهو من الادوات الهامة المستخدمة في شغل الـ **anomaly** و مهم نعرف تفاصيله



- اهم قيمتين في الـ GD هي الميو و السيجما

- الميو بتكون مركز الجرس المرسوم ، والسيجما (الانحراف المعياري SD) بيكون عرض الجرس

$$X \sim N(\mu, \sigma^2)$$

"distribution"

- و خد بالك ان السيجما σ هي الانحراف المعياري SD بينما σ^2 هي التوع variance

- كمان لازم مجموع المساحة الكلية تحت الجرس ده يكون بـ 1 كامل

- و يقال ان الإكس ، يتم توزيعها (تتعمل بالرمز ده ~) عن طريق الـ GD (يتعمله بالرمز ده كل اللي يشبه N مائلة) ، بعدها يتعمل قيمتين ، قيمة ميو ، بعدها سيجما بالشكل ده

$$X \sim N(\mu, \sigma^2)$$

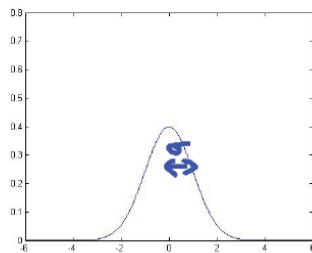
mean = μ
x has normal distribution
variance = σ^2

- أما لو عايز اجيب قيمة الاحتمالية P فهي تكون دالة في إكس ، وميو و سيجما مع بعض بالشكل ده

$$\phi(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

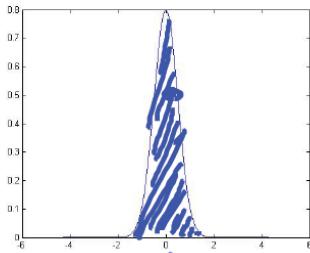
- و هنشوف هنا كذا شكل من اشكال الـ GD

$$\Rightarrow \mu = 0, \sigma = 1$$



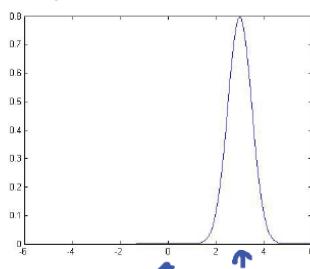
- هنا الميو بصفر و السيجما 1 ، فكان ليها الشكل ده و مركزها في الصفر

$$\rightarrow \mu = 0, \sigma = 0.5$$



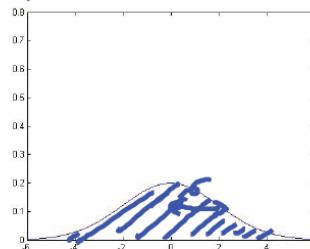
- بينما هنا الميز برضه بصفر ، لكن السيجما بنص فلازم تكون رفيعة شوية ، و لان المساحة الكلية لازم تكون 1 ،
فهيكون الجرس مرفوع لفوق شوية

$$\rightarrow \mu = 3, \sigma = 0.5$$



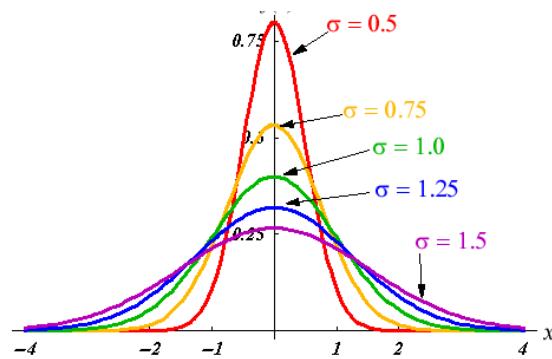
- هنا زي اللي قبلها بالضبط ، لكن هيكون مرکزها عند 3 مش 0

$$\rightarrow \mu = 0, \sigma = 2$$



- هنا لان السيجما تساوي 2 ، هتكون عريضة ، ولازم تكون قصيرة عشان المساحة 1

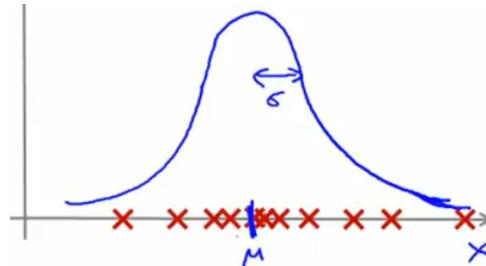
○ و ممكن نلخص الكلام ده في الصورة دي



- فلو كان عندي بيانات معمولة في بعد واحد كدة :



- فممكن ارسم الـ GD بالشكل ده



- لاحظ ان تم اختيار ميو و سيجما بالقيم و الاماكن ديه , بناء على كثافة النقط المتوزعة , فلازم الميو (مركز النقط) تكون في المكان ده

- وممكن احدد القيمتين دول بالارقام مش من الرسم , عن طريق معاذلة الميو (المتوسط) و السيجما (الانحراف المعياري)

$$\bar{\mu} = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

- بينما الـ SD

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

- خ بالك ان الـ n هنا اللي هي الـ m يعني عدد الصفوف

- احيانا بيتم القسمة علي $m-1$ بدل m لكن اصلا الفرق بسيط جدا , لأن اصلا عدد الصفوف بيكون كبير

- تطبيق الـ GD مع البيانات الشاذة

- لو عندي مجموعة من البيانات اللي عايز اشوف الشواذ منها , اول حاجة احسب الميو و السيجما زي ما شفنا فوق بالقوانين بتوعهم , بعدها اجيب احتمالية كل نقطة فيهم و اللي ه تكون مثلا للنقطة الاولى :

$$p(x_1; \mu, \sigma^2)$$

- ولما اجي اعم لباقي النقط , هلاقي ان $P(X$ هتساوي حاصل ضرب الاحتماليات في بعض , يعني :

$$p(x_n; \mu, \sigma^2) \cdot p(x_{n-1}; \mu, \sigma^2) \cdots p(x_3; \mu, \sigma^2) \cdot p(x_2; \mu, \sigma^2) \cdot p(x_1; \mu, \sigma^2)$$

- يعني ممكن اقول انها تساوي

$$p(x) = \prod_{j=1}^n p(x_j; \mu_j, \sigma_j^2)$$

- والي هتساوي ده :

$$\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

○ لاحظ ان الرمز Π يعني حاصل ضرب ارقام في بعض

○ لاحظ كمان ان الاحتمالية دي ممكن تسمى Density estimation

• فالخطوات كالتالي :

○ اولاً : اختيار الـ features المناسبة للتقدير ، زي ما هنقول بعد شوية

○ ثانياً : احسب قيمة ميو لكل feature :

■ يعني لو عندي الف عميل , وكل عميل عنده 50 معلومة , فهجيب متوسط كل معلومة فيهم , عن طريق اني اجمع قيمة المعلومة الاولى (الطول مثلا) لลألف عميل , واقسمها علي الالف , بعدها اجمع قيمة المعلومة الثانية (الوزن) و اقسمها علي الالف وهكذا

$$\underline{\mu}_j = \frac{1}{m} \sum_{i=1}^m \underline{x}_j^{(i)}$$

■ عشان كدة مكتوب ان عشان اجيب ميو 5 مثلا (المعلومة الخامسة او العود الخامس) هجمع كل قيم اكس في العمود الخامس لكل الصفوف , واقسمها علي عدد الصفوف m

○ ثالثاً : احسب قيمة سيجما لكل feature بنفس الطريقة

○ رابعاً : هات الاحتمالية الكلية لمعلومة الاختبار :

■ يعني عندي عميل بنك جديد , فهبدأ اجيب الخمسين معلومة feature اللي عنده , واجيب احتمالية كل معلومة بناء علي الميو و السيجما الخاصة بالمعلومة ديه

■ يعني اطبق القانون ده $p(x_j; \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$ اللي هو ده على مثلا قيمة حسابه الدولاري (feature معين) و استخدم الميو و السيجما اللي حسبتهم من الحسابات الدولارية للاف عميل سابق

■ اكرر خطوة الاحتمالية لباقي الـ features (كل الاعمدة)

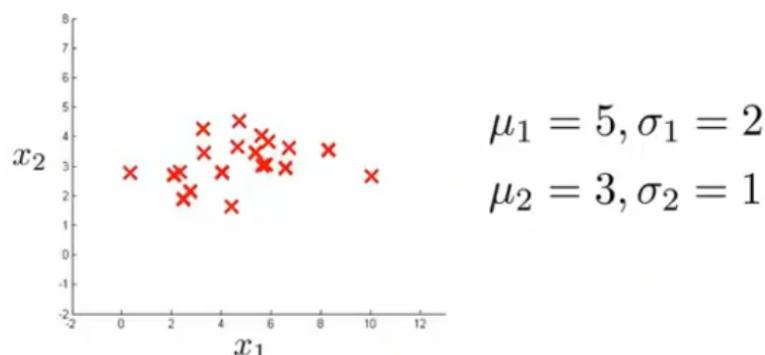
■ اضرب كل الاحتماليات في بعض عن طريق الـ Π

■ اخيرا , لو لقيت ان قيمة P النهائية (الاحتمالية) اقل من اسلون ϵ اللي تم تحديدها , يبقى العنصر ده في مشكلة و ابدا اخذ بالي منه

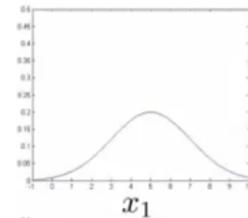
- يعني ايه بقى اختيار الـ **features** المناسبة
 - لما اجي اعمل فحص للعناصر الشاذة ، مهم اني احدد انهي **features** هتعامل معها ، يعني مش لازم اخذ كله
 - و بيتم تحديدها على اساس ، اني اشوف انهي **features** بالتحديد اللي هتحدد هل العنصر شاذ ولا لا
 - يعني لو عندي عدد من الـ **features** الخاصة بعملاء البنك هي :

سنها ■
 طوله ■
 جنسيته ■
 حسابه ■
 ايداعه بالدولار ■
 ايداعه باليورو ■
 وظيفته ■
 سحبه بالدولار ■
 سحبه باليورو ■
 تاريخ سفرياته ■
 دفتر شيكاته ■

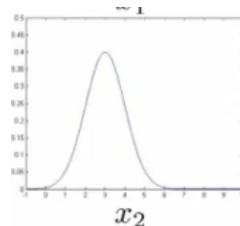
- فلما احدد الـ **fratures** اللي هحطها في الـ **GD** و اجيب الميو و السيجما ليها ، مش هاخد كل حاجة لكن المعمولها بولد بس لأن ديه غالبا اللي بتحدد هو شاذ ولا لا ، لكن ادخال عناصر تانية مش خاصة بالموضوع ده هيأثر سلبا علي دقة البيانات و تمييزها



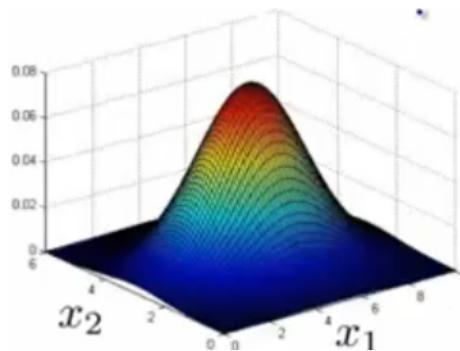
- هناقي ان غالبا متوسط اكس 1 يساوي 5 ، وعرض الجرس تقريبا 2 ، بينما لأكس 2 هتكون بـ 3 و العرض 1
- لاحظ ان عرض جرس اكس 2 اقل من اكس 1 ، و ده منطقي لأن اكس 2 اقل في التوزع
- نشوف جرس اكس 1 هيكون كدة



○ بينما جرس اكس 2



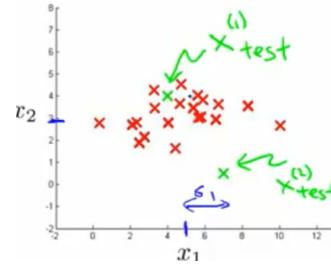
○ عشان نفهمها بشكل كامل ، لازم نشوف الرسم ثلاثي الابعاد ليهم مع بعض اللي هيكون كدة



○ خ بالك ان مسقط الجرس المجسم, من منظور اكس 1 , هو بالضبط رسمه جرس اكس 1 , ومسقطه من منظور اكس 2 هو جرس اكس 2

○ معني كدة ان الجرس المجسم , يكون تخين شوية في محور اكس 1 , ورفع شوية في محور اكس 2

○ دلوقتي لما نقيم نقطة جديدة , هيكون ليها اكس 1 و اكس 2 , فلو كانت النقطة في قلب الزحمة , مثلاً نقطة 1 :



○ هنلاقي ان اساقطها هيكون في قلب الجرس 1 , والجرس 2, يعني في قلب الجرس المجسم , بيقي الـ P اكبر من ϵ بينما لو لقينا نقطة بعيدة شوية زي 2 , هي موجودة في جرس 1 , بس بعيدة عن حدود جرس 2 , ساعتها هتخرج برة الجرس المجسم , يعني الـ P اقل من ϵ

○ معني كدة ان اي نقطة ه تكون تحت سطح الجرس المجسم , تبقى مقبولة , اي نقطة فوق سطحه , تبقى شاذة

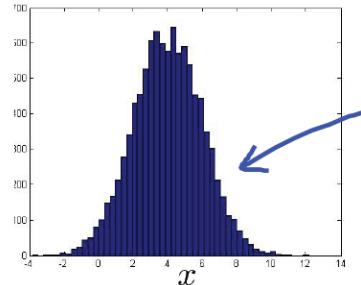
- تقييم نظام كشف البيانات الشاذة
 - عشان اعمل كشف للبيانات الشاذة هنعمل اول خطوة , تقسيم البيانات
 - و التقسيم يكون بالطريقة ديه :
 - او لا هنقول ان البيانات العاديّة (المش شاذة) قيمة الـ $y = 0$ بينما البيانات الشاذة قيمة $y = 1$
 - ها عمل في الـ **training data** اللي البرنامج هيستخرج منها الخوارزم , هخلي فيها بس البيانات العاديّة
 - مش الشاذة , يعني بس اللي $y=0$
 - هقسم البيانات الصحيحة كالتالي :
 - قيمة 60 % منها لـ **training data**
 - قيمة 20 % منها لـ **CV**
 - قيمة 20 % منها لـ **test data**
 - بينما البيانات الشاذة عندي :
 - قيمة 50 % منها لـ **CV**
 - قيمة 50 % منها لـ **test data** (مش هحط منها لـ **CV**)
 - يعني ممكن لو البيانات السليمة 10 الاف , والغلط 20 تكون كدة :
 - Training set: 6000 good engines
 - CV: 2000 good engines ($y = 0$), 10 anomalous ($y = 1$)
 - Test: 2000 good engines ($y = 0$), 10 anomalous ($y = 1$)
 - والخطوات تكون كالتالي :
 - او لا يتم تقسيم البيانات بالشكل الموضح اعلاه
 - ثانيا اتناول بيانات الـ **training** (اللي مفيهاش اي شواد) عشان اخلي الخوارزم , يقدر يحدد قيم ميو و سيجما
 - ثالثا امسك بيانات الـ **CV** او **test** (اللي فيها شواد) و ابدأ اعمل اختبار ليهم , وهبطل حاجة من اربع حاجات
 - اما **True Positive** يعني هي اصلا بايطة , والخوارزم اكتشفها
 - او **false positive** الخوارزم قال انها بايطة بس هي اصلا سليمة
 - او **true negative** يعني هي سليمة والخوارزم قال انها سليمة
 - او **false negative** يعني الخوارزم قال انها سليمة بس هي بايطة
 - رابعا احسب قيم الـ **Precision** , **Recall** , **F1 Score**
 - ولاحظ ان احد الطرق اللي بحدد بيه قيمة ابسلون ϵ هي اني اعمل قيم مختلفة ليها في المثال اللي فات , واشوف انه فيهم بتعملني اعلي **F1 Score**
-

• كشف الشوادم التعليم المشرف :

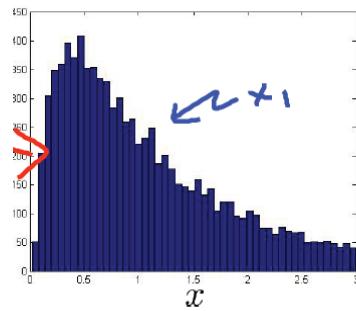
- السؤال يدور , طالما لدى معلومات كاملة , بما فيها قيمة y ليه اساسا عملتها بنظام الشوادم مش بنظام الـ classification و اقسامهم افضل , يعني بالـ supervised مش ه تكون اسهل ?
 - يعني بطريقة تانية , لو عندي نوعين من المواتير , موتور سليم و موتور بايظ , ايه اللي يخليني اقول ان السليم الاصل و البايظ الشاذ , فهمشي علي خوارزم الشوادم , او اني اقول ده نوعين منفصلين , فهمشي بنظام التقسيم
 - الحد الاساسي اللي يخليني اروح هنا او هنا هو عدد القيم الموجبة ($y=1$)
 - لو كان عددا قليل جدا , فهي تعتبر شاذة , وبالتالي مش هاتعتبرها كانها قسم من الاقسام اساسا , بل هي قيمة شاذة , فمش هدخلها في الخوارزم , وبس هخلي قيمتها في التبست
 - لو كان عددا معقول (يعني علي الاقل 20% من البيانات) ببقي ديه مش شاذة لكنها قيمة من القيم , فهروح علي الـ classification
 - كمان حاجة تاني بتخليني اروح هنا او هنا , وهي منطقة اكتر
 - لو كانت القيم الموجبة ($y=1$) ماشية بنمط معين , وبالتالي اقدر استنتجها (الايميل السبام , الورم الخبيث , العميل المتوقع يشتري) , و كمان اقدر استنتاج عليها قيم مستقبلية , فده معناها ان القيم الموجبة هي صنف من الاصناف , فهيكون ليها مساحة خاصة بيها , حتى لو كانت نسبتها قليلة , يعني فيه تشابه نسبي بين القيم الموجبة , ببقي classification
 - بينما لو اكنت القيمة الموجبة ملهاش نمط معين بتمشي عليه , لكن هي فقط , بعيدة عن القيم المعتادة , وبالتالي مفيش اسلوب اقدر استنتاج منه القيم المستقبلية , ببقي ديه قيم شاذة مش صنف من الاصناف , يعني كل قيمة موجبة بعيدة شوية عن باقي اخواتها
 - وبالنالي من تطبيقات القيم الشاذة :
 - المعاملات البنكية المشبوهة
 - اجهزة الكمبيوتر اللي شغالة بشكل غير طبيعي
 - الاعطال في الصناعة
 - الفواكه التالفة
 - بينما من تطبيقات الـ classification
 - الايميل السبام
 - توقع الطقس
 - الفحوص الطبية
 - مع ملاحظة ان اي عنصر فيهم ممكن يتنتقل من هنا لها , في حالة ان توافر فيه الشرطين اللي قلناهم من شوية

- ضبط عناصر خوارزم القيم الشاذة

- في بعض الأحيان ، تكون قيم الاكتسات لدينا جاهزة ، ويتم رسمها ببساطة بشكل جرس جوسبيان زي كدة



- و حتى لو مكانتش الشكل دقيق 100 % ، لكن هو يقاس على جوسبيان ، وبالتالي مش هاعمل فيه حاجة ، وهجيب الميو و السيجما بشكل طبيعي
- لكن ممكن اجي ارسمها الاقيها بشكل مختلف زي ده



- يبقى هنا عايزين نعمل تغيير في معاملات الدالة ، عشان تنطبع ، وده اللي اسمه ضبط عناصر خوارزم القيم الشاذة
- بشكل عام عشان احول من الشكل ده  ، لده  ، تحتاج العب في الدالة نفسها
- يعني ممكن بدل ما هي اكس ، اخليها حاجة من دول مثل :

- $\log x$
- $\log(x + c) == c$ is constant
- \sqrt{x} ===== whatever the root
- X^3 ===== whatever the power

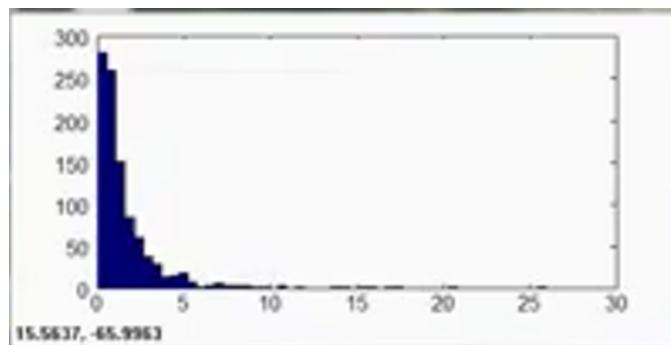
- يعني انا معايا الامكانية بتغيير دالة اكس باي تعامل رياضي ، عشان اوصل لجرس جوسبيان
 - و طبعاً لما اوصلها ، يبقى اي قيمة شاذة عشان اكشف عليها ، هعمل لها نفس التحويل
-

- تعالى نشوف عملي

- في اوكتيف ، هنستخدم امر `hist` عشان نرسم
- فلو عملنا الامر :

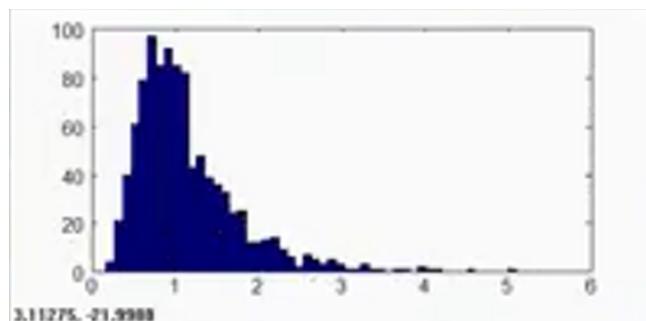
`hist(x,50)`

○ هيكون الرسم كدة



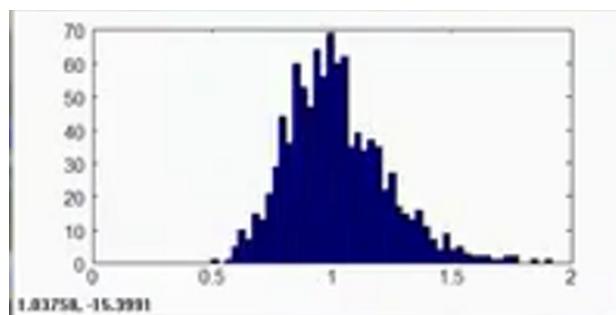
○ بينما لو عملتها بالجذر

`hist(x^0.5,50)`



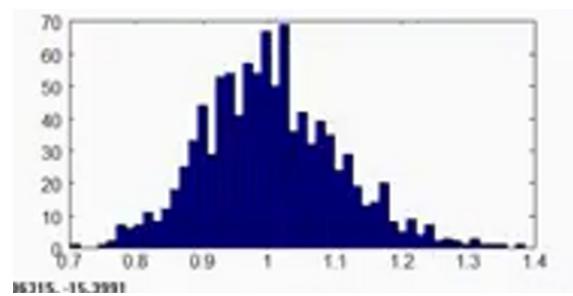
○ نعمل اس 0.2

`hist(x^0.2,50)`



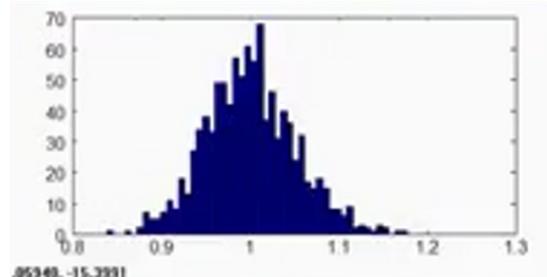
○ نعمل اس 0.1

`hist(x^0.1,50)`



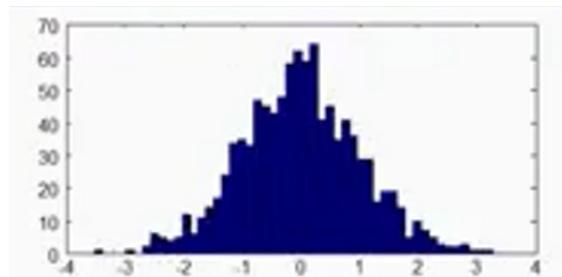
- نعمل اس 0.05

`hist(x^0.05,50)`



- ولو استخدمنا اللوج

`hist(log(x),50)`

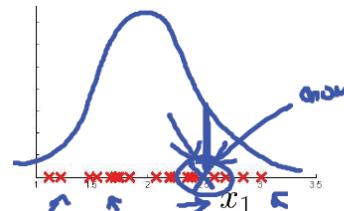


- فهنا لاحظ اننا كل ما نعدل الاس او نغير الدالة ، كل ما الدالة اتحولت لجوسينيان اكتر ، لغاية لما اوصل للشكل المناسب

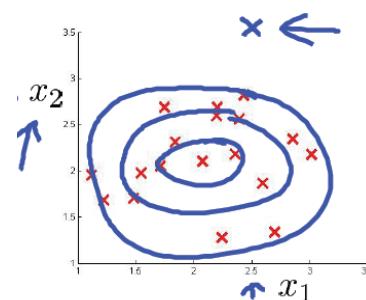
- و هنا فيه حاجة مهمة ، متعلقة باختيار الـ **features**

○ لو انا عايز اكتشف البيانات الشاذة، فلازم اخذ بالي ان ممكن يكون فيه بيانات مش شاذة و بيانات شاذة

○ فمثلا هنا ، النقطة المعروفة X_1 ، موجودة في قلب الزحمة ، فهي مش شاذة ولا حاجة



○ بينما لما جيت اضيف لها X_2 لقيت ان قيمتها بقت شاذة ، وخلتها بعيدة عن الدوائر المحتملة



○ وكان ممكن أحد القيم (وزن التقاحة) يكون طبيعي ، بينما قيمة تانية (لونها) يكون شاذ

● مثال تاني :

- لو عندي وحدة فيها الاف السيرفات اللي شغالة علي حاجة معينة ، ويهمني اعمل مراقبة كاملة ليهم
- فممكن نقول ان فيه عدد من العوامل براقبها هي :

x_1 = memory use of computer

x_2 = number of disk accesses/sec

x_3 = CPU load ↘

x_4 = network traffic ↘

- ساعتها ابدا اسئل المختصين في المجال ، يا ترى انهي عامل فيهم اللي بيشير ان القيمة بتاعته هي اللي ممكن تكون شاذة ، وابدا ارقبها

- بل اني ممكن اخترع قيمة جديدة ، من القيم ديه ، يعني ممكن اقول ان :

$$x_5 = \frac{x_3}{x_4}$$

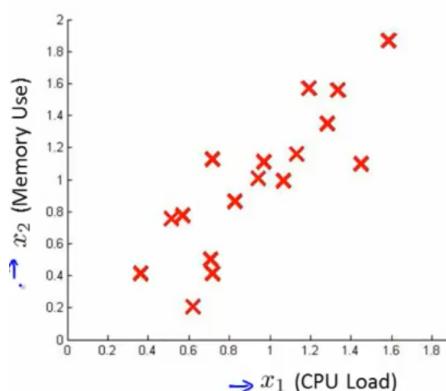
○ او

$$x_5 = \frac{x_3^2}{x_4}$$

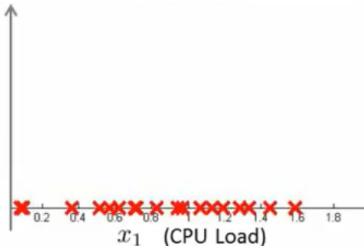
- يعني المجال مفتوح اني اقول قيمة علي قيمة او قيمة في قيمة او تربيع او جذر و هكذا

● الكشف عن شواز باكثر من متغير

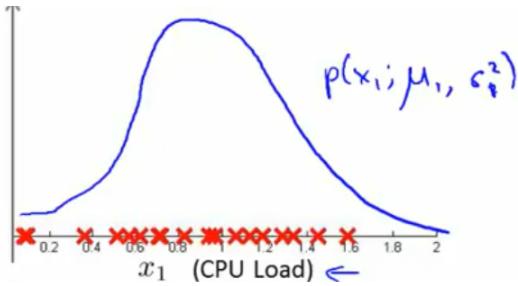
- و هي طريقة لها مميزات و عيوب ، ويمكنها الكشف عن شواز لو تكتشف بالطريقة السابقة
- و عشان نفهم معناها ، تعالى نشوف مشكلة من مشاكل كشف الشواز
- لو عندي بيانات جهاز كومبيوتر ، و العلاقة بين استهلاك الـ CPU و بين الذاكرة ، وكان الرسم كدة



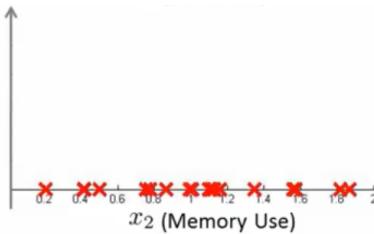
- فممكن احدد بيانات قيمة اكس 1 لوحدة كدة



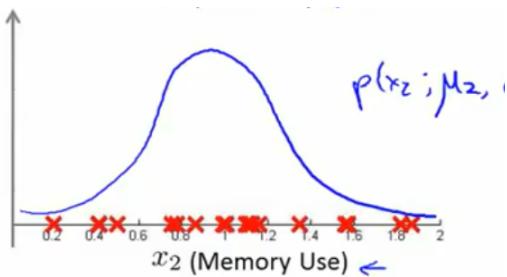
و ارسم جرس جوسیان كدة



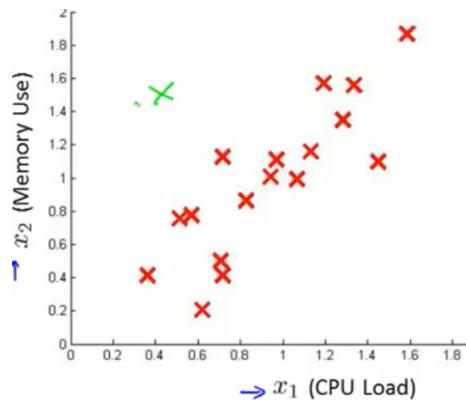
كمان ممكن احدد بیانات اکس 2 لوحدها



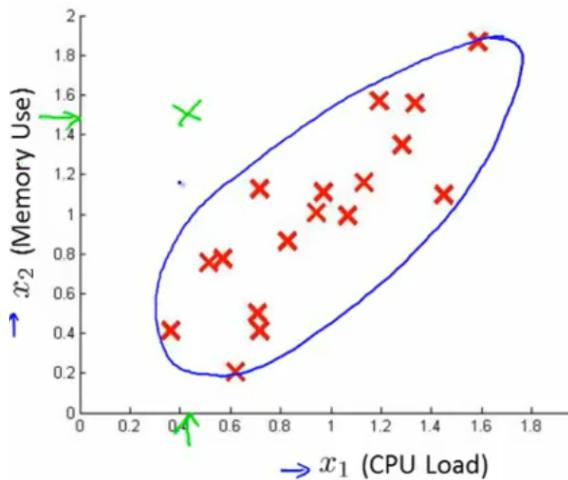
و ارسم الجرس



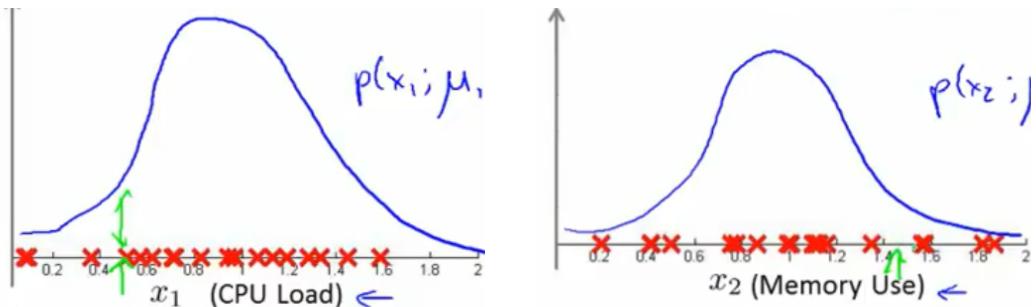
لو فلنا دلوقتي ان عندي قيمة جديدة لونها اخضر ، عايز اعرف هي عاديه ولا شاذة



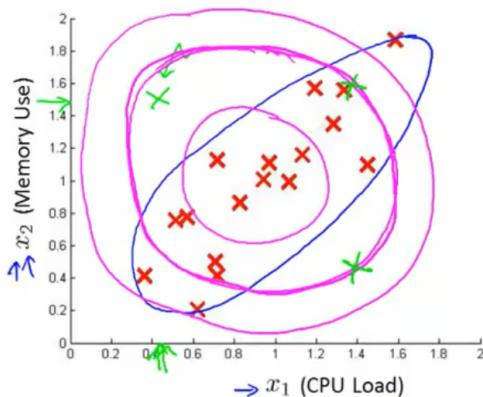
- من الرسم واضح انها شاذة ، لأن استهلاك الـ CPU قليل مع استهلاك ذاكرة كبير ، وحتى بادئه انها بعيدة عن الرسم



- لكن المشكلة ان عرض القيمة ديه لوحدها في اكس 1 او اكس 2 اتنين هيبيان انها طبيعية (النقطة الخضرا هنا و هنا)



- بل حتى لو عملت دواير في الرسم الثاني ، هلاقي انها مش شاذة جدا ، و ان راسها براس النقطتين الخضراء اوتين التانين اللي واحدة فيهم منطقية و الثانية مش منطقية



- و ده يدل علي قصور في قدرة الآلية ديه في تمييز نقطة انا عارف انها شاذة ، بس الجراف مقدرش يميزها
- و غالبا المشكلة جایة من الـ correlation يعني الربط ، بين قيمتين كل قيمة فيهم هي منطقية ، بس المشكلة انها مش هييجو مع بعض ابدا

- زي ما اربط بين سلعة معقولة (عربية) ، وسعر معقول (20 دولار) كل حاجة فيهم سليمة ، بس مينفعش مع بعض

• فيكون الحل ، اننا منحبيش كل P لوحدها ، لكن P المجمعة ليهم

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

◦ القانون الغريب ده ، بيجمع كل تفاصيل الاكسات مع بعض ، وهيجيب الـ P بشكل دقيق

◦ و متنساش ان الـ Σ هي المصفوفة اللي اتكلمنا عنها قبل كدة ، وابعادها $n \times n$

◦ كمان ان $|\Sigma|$ معناها قيمة المصفوفة و بتتعمل بامر \det في اوكتيف او ماتلاب

$$\frac{1}{m} \sum_{i=1}^m x^{(i)}$$

◦ متنساش ان الميو بتكون فيكتور $n \times 1$ و قيمتها تساوي

$$\frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

◦ بينما السيجما بتكون مصفوفة $n \times n$ وقيمتها

◦ و السيجما غالبا بتكون **diagonal matrix** مصفوفة قطرية ، كل عناصرها اصفار ما عدا القطر اللي هو سيجما 1 ، سيجما 2 ، وهكذا (امحرافات معيارية)

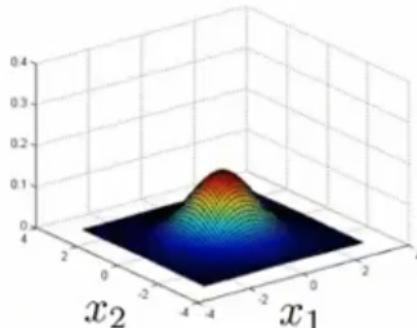
$$\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_n \end{bmatrix}$$

• تعالى نشوف رسوم عملية عشان نفهمها كويis :

◦ لو قلنا ان عندي اكس 1 و اكس 2 ، وان الميو بصفرين ، والسيجما زي هنا

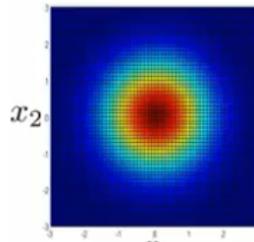
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

◦ هتللاقي ان الرسم هيكون كدة :



◦ وده معناه ان مركز الجرس المجمس هو صفر في اكس 1 و صفر في اكس 2 ، لأن المركز هو الميو ، وان عرض الجرس هو 2 ، لانه ضعف قيمة سيجما (السيجما هي الانحراف المعياري ، نص عرض الجرس)

- و هنلاحظ هنا ان اقصى قيمة للاحتمالية , لما بتكون كلا من الاكس 1 و 2 عند المتوسط ميو , ساعتها الاحتمالية بتساوي 1 , يعني 100%
- وان لو اكس 1 او اكس 2 لو واحدة قلت حتى لو الثانية كبيرة, فالاحتمالية هتقى كتير , و ده منطقى في التعامل مع العنصرين المرتبطين بعض زي اللي فوق, يعني عشان اجيب احتمالية كبيرة لازم الاكتستين مع بعض تكون 0 ولو شفنا المجسم ده , بس من فوق هنلقي الرسم كدة



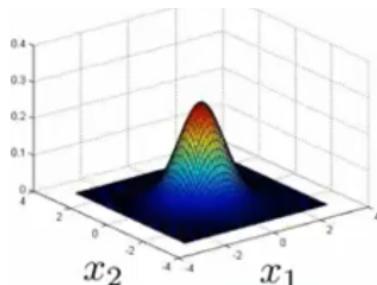
- واللى بيجبيلنا نفس المعنى , بس برسم 2D , مع اعتبار ان القيم الحمرا عالية , والاصفر اقل , والازرق صفر
- فعشان اجيب اعلى قيمة , لازم الاكتستين مع بعض يكونو عاليين

• طيب لو قللنا شوية عرض الجرس (السيجما)

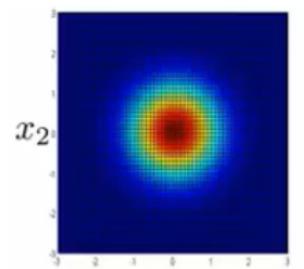
- لو خلينا الاتنين سيجما (للاكس 1 و 2) قيمة اقل من 1

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$$

- طبعي ان العرض هيقل , معنى كدة انها هتطول لان المساحة الكلية لازم تكون 1



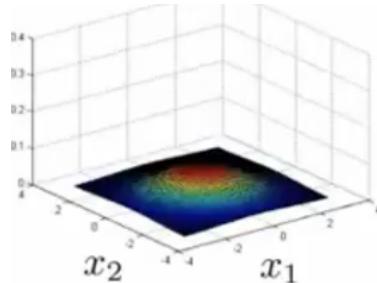
- لاحظ ان المركز نفسه , لكن الجرس طول لفوق و عرضه قل
- وده معناه ان اي ابتعاد لقيم اكس 1 و 2 للبيانات الجديدة , معناه انخفاض رهيب في الاحتمالية
- وهيكون الرسم البعدين كدة , الدواير قلت شوية , وده منطقى



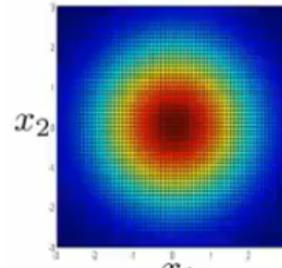
• و بالعكس لو زودنا السيجما

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

○ معني كدة ان العرض هيزيد ، يبقى ه تكون اقصر



○ واللي معناها ان الابتعاد عن قيم اكس 1 او 2 ، مش معناها انهيار فوي للاحتمالية ، لكن لازالت القيم طبيعية
○ و هيكون الرسم البعدين ، والدوایر اعرض

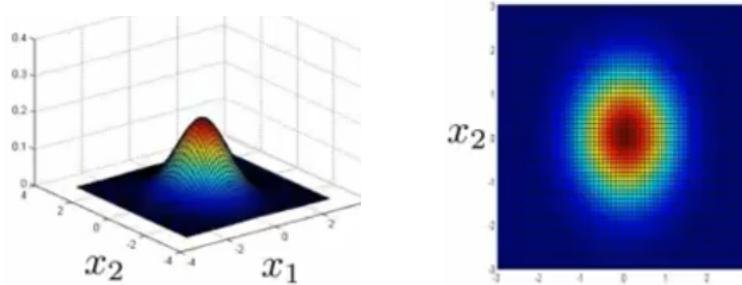


• طيب ماذالو غيرنا قيمة سيجما لواحدة فيهم عن الثانية

○ يعني نخلي ، قيمة سيجما لاكس 1 بقيمة اقل من سيجما اكس 2

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$

○ هنلاقي ان الجرس بقى كدة

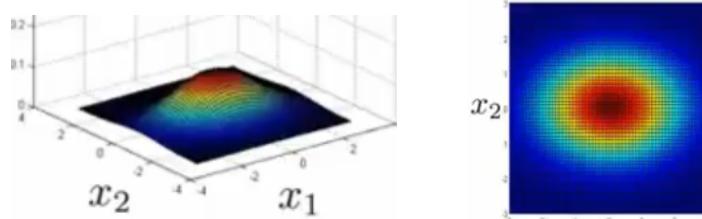


○ فنلاحظ ان عرض الجرس من ناحية اكس 1 بقى ارفع شوية، و عرضه من ناحية اكس 2 زي ما هو ، يعني
مفيش مشكلة اني اقل شوية في البيانات الجديدة من ناحية اكس 2 ، لكن اي تقليل من ناحية اكس 1 ه تكون مشكلة
○ خد بالك ان الارتفاع هيكون اعلى من حالة ان السيجمات 1 (الحالة الاولى) ، واقل من حالة ان السيجات بـ 0.6
(الحالة الثانية) لأن لازم الحجم يظل بـ 1

• ولو هعمل العكس ، اني اثبت اكس 2 ، واخلي اكس 2 اكبر من 1

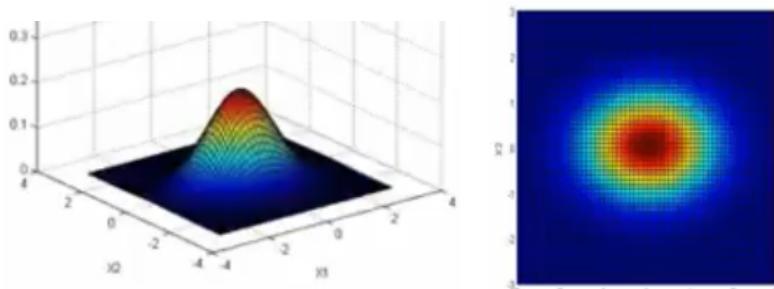
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

○ ساعتها برضه هيثبت عرض الجرس من ناحية اكس 2 ، لكن هيزيد من ناحية اكس 1



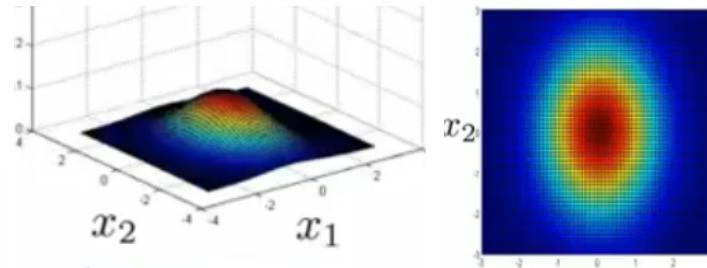
• و العكس ، لو ثبّتنا اكس 1 ، وقلّلنا اكس 2 ، فعرضها هيقل

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0.6 \end{bmatrix}$$



• و برضه لو زودنا اكس 2

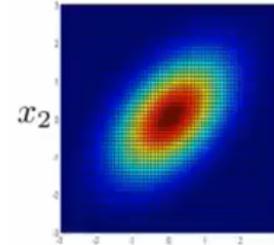
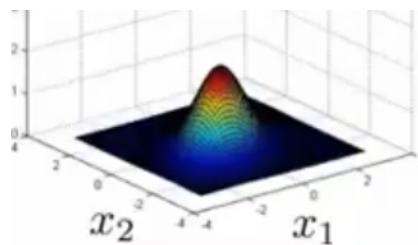
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$



• طيب في حالة اتنا غيرنا القيمتين فوق على اليمين وتحت على الشمال ، بدل $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ خليناها $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ مع ثبات السيجمات هنا و هنا بـ 1

○ ساعتها الدائرة بتبدا تميل شوية بشوية ، لما كانت القيم ديه بصفر ، كانت درجة الميل صفر ، لما بيدها الرقم يزيد من الصفر و يقرب للواحد ، بتبدا الدائرة تنمط عشان تكون شبه الخط المستقيم بـ $slop = 1$

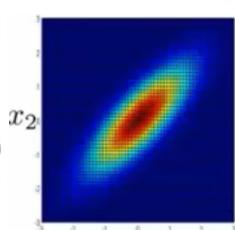
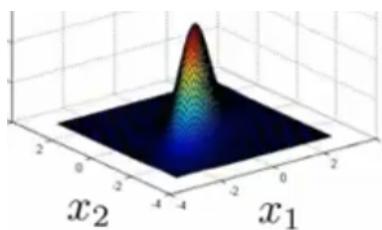
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



○ وده معناها ان الاحتمالية ه تكون اقصى قيمة لما الاكتستين يساوو الميو لكن برضه ه تتطل قيمتهم كويسة لو زادو مع بعض او قلو مع بعض , بينما لو زادت واحدة و قلت الثانية ه تكون مشكلة

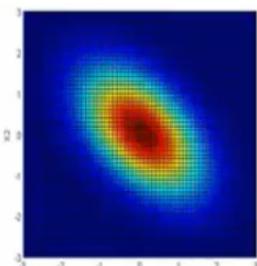
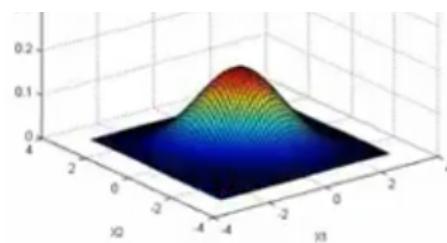
- ولو الرقم خليناه 0.8 بدل 0.5 , الاستطالة ه تزيد و ه تكون شبيهه بالخط المستقيم , فكل ما يزيد الرقم يتتط الشكل

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



- اما لو خلينا الارقام سوالب , ه تكون العكس في الشكل

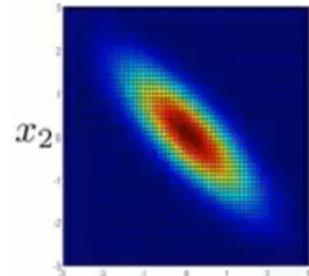
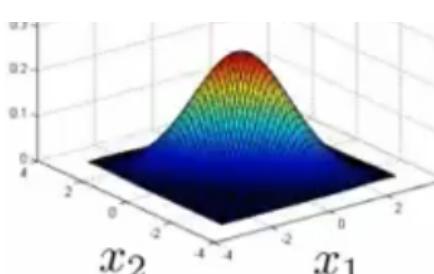
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$



○ المط بقى الاتجاه العكس , يعني يا اما تبقى الاكتستين بقيمة الميو , يا اما واحدى تقل و الثانية تزيد , ساعتها الاحتمالية تزيد , لكن الاثنين يقولو او يزيدو مع بعض تبقى مشكلة

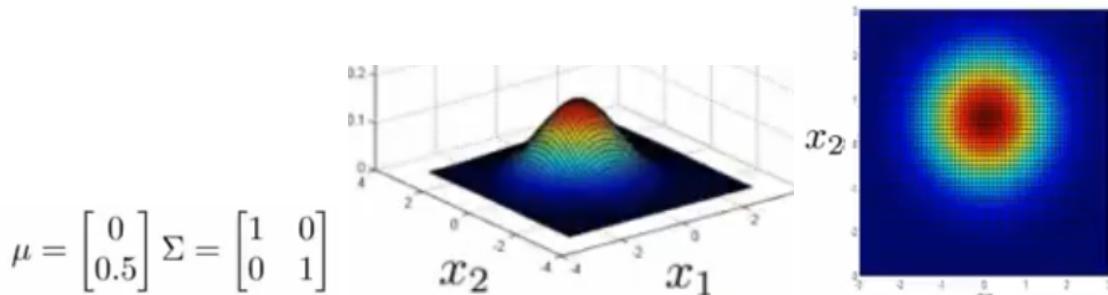
- ولو القيمة بالسالب بقت اقرب لـ -1 - المط هيزيذ

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

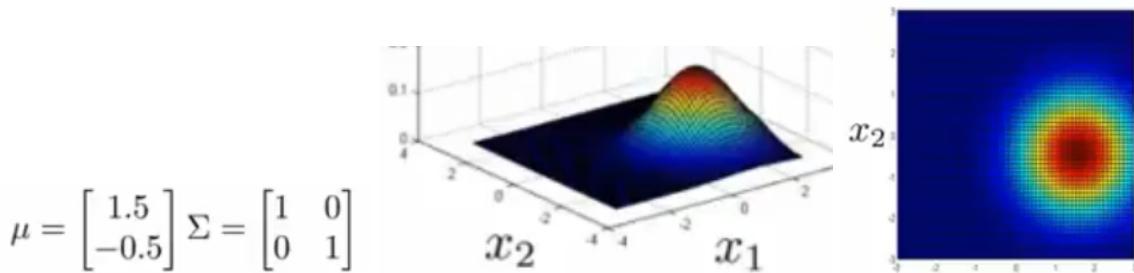


● بينما لو غيرنا قيمة الميو , فكل اللي هيحصل ان مركز الجرس المجسم هيتغير تبعا ليها , و نفس تأثيرات السيجما هي هي

- يعني هنا غيرنا قيمة ميو لاكس 2 :



- و هنا لو غيرنا قيمتي الميو مع بعض



- والأهم من ده كله : انك يكون عندك القدرة على تفسير اي شكل مناسب لاي مشكلة حياتية , ويا ترى في انهي تطبيق في الصناعة او الزراعة هستخدم الشكل الرابع , لأن الاكتستين لازم يكبرو مع بعض عشان الاحتمالية تزيد و هكذا
-

● التطبيق في الكشف عن الشواد :

- هنستخدم ما ذكرناه عن القيم المتعددة في الكشف عن الشواد
- والخطوات هتكون كالتالي :

■ تحديد قيم ميو و سيجما من القوانين الخاصة بيهم

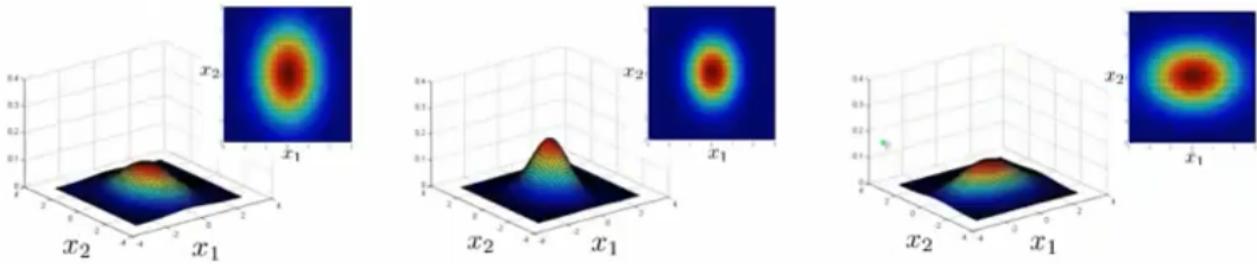
$$\begin{aligned} \bar{\mu} &= \frac{1}{m} \sum_{i=1}^m x^{(i)} \\ \Sigma &= \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \bar{\mu})(x^{(i)} - \bar{\mu})^T \end{aligned}$$

■ ايجاد قيمة P عن طريق القانون الخاص بيها

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (x - \bar{\mu})^T \Sigma^{-1} (x - \bar{\mu}) \right)$$

- في حالة عدم تواجد ابسلون , تحديد قيمتها عبر تكرار فرض قيم لها و اختيار اعلي قيمة لـ F1 Score
- تحديد اي نقطة يكون قيمة P فيها اقل من ابسلون عشان تكون نقطة شادة

■ رسم الموديل المناسب بناء على قيم ميو و سيجما



- متتساش ان في حالة متغير واحد ، القيمة النهائية لـ P ه تكون حاصل ضرب قيم p الصغيرة

$$p(x) = p(x_1; \mu_1, \sigma_1^2) \times p(x_2; \mu_2, \sigma_2^2) \times \cdots \times p(x_n; \mu_n, \sigma_n^2)$$

- بينما في حالة اكتر من متغير

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

- و كمقارنة سريعة بين المتغير الواحد ، وبين متعدد المتغيرات :

- المتغير الواحد :

● مميزاته :

- ارخص في تكلفة العملية
- يعمل بجودة مع العدد الكبير للـ features حتى لو عشر الاف
- يعمل بجودة حتى لو عدد العناصر m قليل

● عيوبه :

- لابد من عمل علاقة بين المتغيرات بشكل يدوي (زي اني اقسم اكس 1 على اكس 2)

○

● متعدد المتغيرات:

● مميزاته :

- الحصول على العلاقة بين المتغيرات تلقائي

● عيوبه :

● مكلف اكتر

● ابطئ

- يستحيل عمله مع عدد كبير من الـ features (لاني باعمل inverse للسيجما)
 - لابد قيمة m تكون اكبر من n والا مش هاعرف اعمل inverse للسيجما (يفضل قيمة m تكون على الاقل عشر اضعاف n)
 - ولو لقيت ان المصفوفة مش بيتعملها inverse يبقى فيه سبب من اتنين
 - اما ان عدد الـ m اقل من الـ n
 - او فيه صفين **redundant** يعني ملوش لازمة , يعني فيه صف من الصفوف مطابق او مضاعف صف ثاني (او عمود) (يعني يكون فيه **feature** يساوي **feature** او يكون مضروب في رقم او فيه عمود يساوي عمودين مجموعين علي بعض)
-

Recommender System

نظام الاقتراحات

• ويقصد به

- هو بناء نظام برمجي , يقوم بعمل ترشيحات و اقتراحات , لمستخدم موقع معين , بحيث تتوافق مع متطلباته , مما يزيد من احتمالية تجاوب المستخدم معه
- وعلى قدر توافق الترشيح المقترن , تكون كفاءة النظام في انتقاء الشئ المناسب
- وهذا الامر على قدر كبير من الأهمية لدى العاملين و المهتمين بمجال الـ ML
- مثال عملی لنفهم الأمر
- نتخيل ان لدينا اربعة مستخدمين Alice , Bob , Carol , Dave سيقومون بالتصويت لخمس افلام (ثلاث رومانسي بلون اخضر , واثنين اكشن بلون ازرق)
- التصويت سيكون بدرجة من صفر الى خمسة



○ اذا تم التصويت بالشكل التالي :

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)
Love at last	5	5	0	0
Romance forever	5	?	?	0
Cute puppies of love	?	4	0	?
Nonstop car chases	0	0	5	4
Swords vs. karate	0	0	5	?

- لاحظ وجود علامات استفهام , لأن المستخدم لم يصوت
- ممكن ببساطة استنتاج ان كل من Alice , Bob يحبون الرومانسي و يكرهون الاكشن , بينما Carol , Dave العكس , يحبون الاكشن و يكرهون الرومانسي
- فممكن نستنتاج القيم الناقصة , تقييم اليـس سيكون 5 و بوب سيكون 4.5 , بينما كارول سفر , و ديف صفر ثم 4
- طبعا ليست ارقام اكيدة لكنها ستكون قريبة منها بالطبع
- استنتاج القيم الغير موجودة معناه ايه : ■
○ معناه ان لما واحد يدخل علي عدد من الافلام , ويعمل تقييم ايجابي للافلام الاكشن , وسلبي للرومانسي ,
فموقع نيتفليكس , بيكون عايز يعرف تقييمه ايه للافلام اللي لسة مشافهاش

- فلما دلـ ML يعمل التقييم و يظهر عنده انه كان ممكن يعمل تقييم كبير لافلام كذا كذا , يعرضها عنده و يخفي الافلام اللي كان هي عمل عليها تقييم سلبي
- مش بس التقييم هو الاساس (احنا واخدinne مثل عشان واضح) ممكن يكون من ضمن العناصر , انه شاف فيلم لآخره , او انه شافه مرتين , او انه بيبحث عن اسم فيلم معين , او بيفتح نوعية افلام و هكذا
- ويجب ان تعرف ان :

 - عدد المستخدمين نسميه N_u
 - عدد المستخدمين نسميه N_m
 - قيمة $r(i, j)$ ستكون بصفر اذا لم يتم التصويت , و 1 اذا تم التصويت
 - قيمة $y^{(i,j)}$ ستكون بقيمة التصويت اذا تم التصويت

- و بناء نظام الاقتراحات له العديد من التحاولات , منها ما يسمى الترشيح بناء على المحتوى Content Based Recommendation

-
- الترشيح بناء على المحتوى CBR
 - نمسك المثال السابق , ونعمل دائرة علي القيم الناقصة
 - اي فيلم بيكون فيه عدد كبير من الـ features , نقول ان عندنا 2 , هما :
 - مقدار الرومانسية في الفيلم X_1
 - مقدار الأكشن في الفيلم X_2
 - نقول ان كل فيلم من الافلام الخمسة , فيه مقدار كذا من X_1 و مقدار كذا من X_2

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)	$\downarrow x_1$ (romance)	$\downarrow x_2$ (action)
→ Love at last	5	5	0	0	0.9	0
Romance forever	5	?	?	0	1.0	0.01
Cute puppies of love	?	4	0	?	0.99	0
Nonstop car chases	0	0	5	4	0.1	1.0
Swords vs. karate	0	0	5	?	0	0.9

- نبدا نجمع الـ features الخاصة بكل فيلم , هنقول ان X^1 (يعني فيتشرز الفيلم الاول) هتكون

$$\begin{bmatrix} 1 \\ 0.9 \\ 0 \end{bmatrix}$$

- لاحظ ان رقم 1 اللي في الاول ده خاص بـ X_0^1 يعني اول feature لإكس واحد ، واللي دائمًا تكون بوحدة زيز ما كنا بنعمل في الـ linear regression ، اما باقي القيمتين 0.9 ، 0 فهما قيمة الرومانسية والاكتشن فيهم
- دلوقي هنعمل linear regression (توقع قيم خطية ، زي توقع اسعار بيوت مش معروفة بناء على بيانات بيوت موجودة)
- و بالتالي كل مستخدم هيكون ليه فيكتور ثيتات ، يعني مصفوفة عمود واحد في عدد من الصفوف ، عدد الصفوف هيكون $n+1$ ، يعني هنا هيكون 3
- قيم θ واحد واثنين و ثلاثة مش معروفين ، هافرضهم في البداية ، و هدخلهم في اللوب الطويلة عشان اجيب قيمهم اللي تتوافق مع اختياراته السابقة
- و هنكون معادلة الـ Hx

$$(\theta^{(j)})^T x^{(i)}$$

- يعني همسك فيكتور الثيتات (3.1) و اعمله ترانزبوس (1.3) و اضربه في فيكتور الاكتسات (3.1) فيعمل قيمة واحدة محددة
- فلو عايزين نجيب قيمة تصويب ليس للفيلم الثالث ، و نفرض ان قيمة ثيتا ليس هي :

$$\theta^{(1)} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}$$

- بينما قيمة X لفيلم الثالث باینة عندنا :

$$X^{(3)} = \begin{bmatrix} 1 \\ 0.99 \\ 0 \end{bmatrix}$$

- هنتطبق القانون ، ونضرب ترانزبوس الثيتا في الاكتس :

$$(\theta^{(1)})^T X^{(3)}$$

- واللي هيساوي 4.95
- وهاعمل نفس نظام الـ linear regression في اني هحدد قيمة ثيتات، عن طريق اني اشوف اقل قيمة لـ J اللي ه تكون مجاميع مربعات الفروق بين القيم المتوقعة لضرب ثيتا في اكتس (Hx) والقيمة الحقيقية y

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

- كل مستخدم بيكون ليه رقم ثيتا معين ، فاللي هي ثيتا 1 ، بوب ثيتا 2 و هكذا
- اذن عدد الـ 2 features (المعلومات المتاحة)، وعدد الداتا هو 5 (عدد الافلام) ، اما عدد المستخدمين فهو مش في العملية ، لانه هيتكرر العملية مع كل مستخدم

○ و هيكون القانون النهائي :

$$\min_{\theta^{(j)}} \frac{1}{2} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{k=1}^n (\theta_k^{(j)})^2$$

○ حيث :

$(\theta^{(j)})$ هي معاملات كل مستخدم ■

$x^{(i)}$ قيم الـ features للفيلم ■

$r(i,j)$ تكون بوحدة لو المستخدم عمل تصويت للفيلم ، وصفر لو معملش ■

$y^{(i,j)}$ قيمة التصويت ■

العداد تحت السيمحا من رقم 1 ، و مرورا بـ n_u بالارقام اللي تم التصويت ليها عشان كدة كتب ■

لاحظ ان تم حذف قيمة m اللي كانت موجودة ، لأنها ثابت في الطرفين ، واحدنا عايزيين نقلل القيمة ■

فاخلاقتها مش هي عمل مشكلة ■

متتساش اننا اضفنا في الآخر قيمة الـ regulation عشان الـ OF ■

○ ولو عايزيين نجيب قيم كل المستخدمين ، هنعمل قانون اعم

$$\min_{\theta^{(1)}, \dots, \theta^{(n_u)}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n (\theta_k^{(j)})^2$$

• ولما اجي اعملها gradient descend عن طريق التقاضل الجزئي ، هتكون كدة

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \sum_{i:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) x_k^{(i)} \text{ (for } k = 0)$$

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \left(\sum_{i:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) x_k^{(i)} + \lambda \theta_k^{(j)} \right) \text{ (for } k \neq 0)$$

• و هتللحظ ان هي نفسها زي الـ linear regression باستثناء عدم وجود الـ m اللي حذفناها

• الطريقة ديه كانت معتمدة على وجود features في المحتوى ، عشان كدة اسمها Content Based

• لأنها معتمدة على وجود ترشيحات وتقدير سابق في المحتوى Recommendation

• لكن أحيانا بيكون فيه حاجات مفيهاش امكانية المحتوى ، زي السلع اللي هتظهر لك في امازون ، فهنشوف تكنولوجيا تاني اسمه

Collaborative Filters المرشحات المشاركة

• المرشحات المترافقية Collaborative Filters

- و ديه اهم ميزة فيها ، انها مش محتاجة لـ features عشان تتعلم منها، هي بتتعلم من نفسها اول باول
- و ديه ميزة لها مش بس في الحاجات اللي مفيهاش features واضحة زي السلع ، ده حتى الأفلام نفسها هيكون مكلف جداً إنك تجيء حد يشوف الأفلام و يعملها تقييم من درجات ، و ممكن كمان متباين دقيقه
- فاحنا هنفرض هنا ان معندينا قيم للإكسات ، اللي زي تقييم الأفلام الرومانسي وال액شن ، لكن عندنا قيم للإكسات ، وكأنها استطلاع لرأي المستخدم ، هو بيحب انهي انواع أفلام

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)	x_1 (romance)	x_2 (action)
Love at last	5	5	0	0	?	?
Romance forever	5	?	?	0	?	?
Cute puppies of love	?	4	0	?	?	?
Nonstop car chases	0	0	5	4	?	?
Swords vs. karate	0	0	5	?	?	?

- وشافين ان قيم x_2 و x_1 مش عارفين القيم بتاعتكم ، فهنجاهم
- بينما استطلاع رأي المستخدمين ، خلنا نحدد الثيتات الخاصة بيهم زي كدة

$$\theta^{(1)} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}, \theta^{(2)} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}, \theta^{(3)} = \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix}, \theta^{(4)} = \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix}$$

- فهنجعل عكس اللي عملناه في الـ CBR اننا هنقول ان اول رقم 5 (تصويت ليس لاول فيلم) هو عبارة عن حاصل ضرب الثيتا ترانسبوز ، في الإكس اللي منعرفهاش

$$(\theta^{(j)})^T x^{(i)}$$

- يعني تبني كدة ، مع مراعاة ان الثيتا بتزيد ، بينما كلها اكس 1 ، طالما ده الفيلم الاول

$$\begin{aligned} & (\theta^{(1)})^T x^{(1)} \\ & (\theta^{(2)})^T x^{(2)} \\ & (\theta^{(3)})^T x^{(3)} \\ & (\theta^{(4)})^T x^{(4)} \end{aligned}$$

- و كدة ممكن نطبق القانون بتاعنا

$$\min_{x^{(i)}} \frac{1}{2} \sum_{j:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{k=1}^n (x_k^{(i)})^2$$

- و خد بالك اننا في العادي (في الا CBR او في linear regression) بيكون عندنا اكسات و بنحسب الثيتا , دلوقتي العكس عندنا الثيتا و بنحسب الاقسات و لو هنعمل لكل الفلام :

$$\min_{x^{(1)}, \dots, x^{(n_m)}} \frac{1}{2} \sum_{i=1}^{n_m} \sum_{j:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^n (x_k^{(i)})^2$$

• تطبيق المرشحات المترافق

- عرفنا من شوية حاجتين , الاولى لو عندي features ازاي اقدر اجيب ثيتا و كان قانونها كدة :

$$\min_{\theta^{(1)}, \dots, \theta^{(n_u)}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n (\theta_k^{(j)})^2$$

- والثانية لو عندي ثيتا ازاي اقدر اجيب features و كان قانونها كدة :

$$\min_{x^{(1)}, \dots, x^{(n_m)}} \frac{1}{2} \sum_{i=1}^{n_m} \sum_{j:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^n (x_k^{(i)})^2$$

- دلوقتي هنحاول ندمجهم مع بعض عشان نطبق المرشحات المترافق وهيكون بالقانون المجمع ده :

$$J(x^{(1)}, \dots, x^{(n_m)}, \theta^{(1)}, \dots, \theta^{(n_u)}) = \frac{1}{2} \sum_{(i,j):r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^n (x_k^{(i)})^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n (\theta_k^{(j)})^2$$

$$\min_{x^{(1)}, \dots, x^{(n_m)}, \theta^{(1)}, \dots, \theta^{(n_u)}} J(x^{(1)}, \dots, x^{(n_m)}, \theta^{(1)}, \dots, \theta^{(n_u)})$$

- و تفسيرها كالتالي :

- هنا دمج بين العمليتين السابقتين معا , بدل ما نعمل بحث عن ثيتا , ثم اكس لتقليل L , باعملهم مع بعض في المعادلتين الاولى والثانية فيه اتنين سميشن , لكن في المجموعة واحدة , وده لان في كلا منها ماشية على اللي قيمة ٢ تساوي ١ , يعني بس اللي تم التصويت , في المجموعة عملتهم مع بعض عن طريق جملة

$$(i,j):r(i,j)=1$$

- القيمتين الفرديتين تم اضافتهم ورا بعض

- كمان خد بالك , في الحالات السابقة , كنا بنخلify فكتور X بيكون طولة $n+1$ لاننا كنا بنعمل X تساوي 1 , لكن هنا هنحذفها , لأن المعادلة دي مش محتاجة اي قيمة اضافية لضبطها , وكمان هنحذف ثيتا صفر , فهبيكون كلا من اكس و ثيتا فيكتور طوله n

$$X \in \mathbb{R}^n$$

$$\Theta \in \mathbb{R}^n$$

• ماذا عن الخطوات :

- اولاً نحدد قيم مبدئية للاكسات و الثيتات , بارقام صغيرة عشوائية , في خطوة مشابهة لما فعلناه في الـ NN
- ثانياً , اتعامل مع المعادلتين دول , لايجاد القى المثلثي للاكس و الثيتا , مع اقل قيمة للـ J , و متتساش ان المعادلتين دول هما التفاضل الجزئي للـ J

$$x_k^{(i)} := x_k^{(i)} - \alpha \left(\sum_{j:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) \theta_k^{(j)} + \lambda x_k^{(i)} \right)$$

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \left(\sum_{i:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) x_k^{(i)} + \lambda \theta_k^{(j)} \right)$$

- دلوقتي مع فيلم بقيم ثيتا , ومستخدم بقيم اكسات , ممكن اتنبأ بالتقدير المتوقع ليه بقيمة $x^T \theta$
-

• طريقة الـ Vectorization

- و هي خاصة بعمل ترشيحات للمستخدم الذي يختار سلعة معينة , بنقوم بترشيح عدد من السلع المرتبطة بها له
- تعالي نفترض تقدير الأفلام
- لو عندي 4 مستخدمين عملو تقدير لخمس افلام , فهياكون الجدول كدة

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)
Love at last	5	5	0	0
Romance forever	5	?	?	0
Cute puppies of love	?	4	0	?
Nonstop car chases	0	0	5	4
Swords vs. karate	0	0	5	?

- فهنجيب القيم ديها , ونحطها في مصفوفة بالترتيب , بما فيها علامات الاستفهام عن القيم اللي لسة متمش اختيارها

$$Y = \begin{bmatrix} 5 & 5 & 0 & 0 \\ 5 & ? & ? & 0 \\ ? & 4 & 0 & ? \\ 0 & 0 & 5 & 4 \\ 0 & 0 & 5 & 0 \end{bmatrix}$$

- ولأن كل قيمة فيها , هي عبارة عن حاصل ضرب الثيتا ترانسيوز في الاكسات ممكن المصفوفة ديها اكتبها كدة

$$\begin{bmatrix} (\theta^{(1)})^T(x^{(1)}) & (\theta^{(2)})^T(x^{(1)}) & \dots & (\theta^{(n_u)})^T(x^{(1)}) \\ (\theta^{(1)})^T(x^{(2)}) & (\theta^{(2)})^T(x^{(2)}) & \dots & (\theta^{(n_u)})^T(x^{(2)}) \\ \vdots & \vdots & \vdots & \vdots \\ (\theta^{(1)})^T(x^{(n_m)}) & (\theta^{(2)})^T(x^{(n_m)}) & \dots & (\theta^{(n_u)})^T(x^{(n_m)}) \end{bmatrix}$$

- متسااش ان قيمة كل ثيتا تعبر عن مستخدم ، بينما التكستات تعبر عن افلام ، فالعمود الاول بيعرض تقييمات ليس ، فهناقي ان كلها ثيتا 1 ، و ثاني عمود ثيتا 2 وهكذا
- بينما كل صف بيعبر عن فيلم من الافلام ، فهناقي ام الصف الاول لакс 1 و الثاني 2 وهكذا
- و ممكن نبسط الماتريكس المخيفه ديه بحاجة ابسط و هي كالتالي :
- لو قلنا ان فيه فيكتور للإكتسات ، عدد صفوفه يساوي عدد الافلام و عمود واحد

$$X = \begin{bmatrix} - (x^{(1)})^T - \\ - (x^{(2)})^T - \\ \vdots \\ - (x^{(n_m)})^T - \end{bmatrix}$$

■ و قلنا ان فيه فيكتور للثيتات ، عدد صفوفه يساوي عدد المستخدمين و عمود واحد

$$\Theta = \begin{bmatrix} - (\theta^{(0)})^T - \\ - (\theta^{(1)})^T - \\ \vdots \\ - (\theta^{(n_u)})^T - \end{bmatrix}$$

■ ساعتها نقول ان المصفوفة الكبيرة اللي فوق هي اكس في ثيتا ترانسيوز $\Theta^T X$.
■ و بيكون اسمها low rank matrix

● أخيرا حاجتين مهمتين :

- لازم تكون ذكي و انت بتختار الـ features المؤثرة ليك اصلا في عوامل تشابه السلع او الافلام ، فمثلا : ممكن يكون نوع الفيلم او البطل او المخرج او سنة الانتاج عوامل مؤثرة ، بينما مدير التصوير ، و تشابه حروف الفيلم ، واسم الشركة المنتجة ، من العوامل الغير مؤثرة فالاحذفها لان وجودها مضلل misleading
- طريقة اختيار فيلم قريب من فيلم ، اني اجيب اقل قيمة للفرق بين الـ features هنا و هنا ، يعني لو الفيلم اللي المستخدم اختاره هو X^5 فهاعمل بحث مع كل الافلام ، واختار اكتر 3 افلام ، بحيث الفرق بين قيمة X لهم و X^5 هي اقل ما يكون

$$\text{small } \|x - (i)x\|$$

• أداة Mean Normalization

○ وهي أداة تستخدم في موضوع الترشيحات ، لتحسين الأداء

○ ويبدا الموضوع من : في حالة وصل مستخدم جديد ، ولم يقم بعمل اي تقييمات بعد للافلام ، وبالتالي كل قيم ثيتا بالنسبة لي مجهولة عنه

○ ولما اجي اطبق في القانون

$$J(x^{(1)}, \dots, x^{(n_m)}, \theta^{(1)}, \dots, \theta^{(n_u)}) = \frac{1}{2} \sum_{(i,j):r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^n (x_k^{(i)})^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n (\theta_k^{(j)})^2$$

○ هافرض في البداية ان قيم ثيتا بالنسبة له اصفار ، ساعتها هلاقي ان كل التقييمات اللي هتلطع ه تكون اصفار

○ وبالتالي مش هقدر اعمل اي ترشيح بالنسبة له ، لأن كل الافلام بالنسبة له تقييمها صفر

○ و ده تناول غير سليم ، فهنهعمل حاجة تاني

• فنهعتمد علي متوسط ترشيحات الناس اللي قبله

○ لو جبنا كل ترشيحات المستخدمين اللي قبله كدة و عملنا عمود فاضي للمستخدم الجديد :

$$Y = \begin{bmatrix} 5 & 5 & 0 & 0 & ? \\ 5 & ? & ? & 0 & ? \\ ? & 4 & 0 & ? & ? \\ 0 & 0 & 5 & 4 & ? \\ 0 & 0 & 5 & 0 & ? \end{bmatrix}$$

○ ساعتها ممكن نحسب قيمة جديدة هي ميو ، اللي ه تكون عبارة عن متوسط ترشيحات المستخدمين قبله

$$\mu = \begin{bmatrix} 2.5 \\ 2.5 \\ 2 \\ 2.25 \\ 1.25 \end{bmatrix}$$

○ الخطوة الجایة اني هاطرح قيم المصفوفة الاصلية ، ناقص الميو ، ومش مشكلة موجب او سالب ، زي ما عملنا قبل كدة

$$Y = \begin{bmatrix} 2.5 & 2.5 & -2.5 & -2.5 & ? \\ 2.5 & ? & ? & -2.5 & ? \\ ? & 2 & -2 & ? & ? \\ -2.25 & -2.25 & 2.75 & 1.75 & ? \\ -1.25 & -1.25 & 3.75 & -1.25 & ? \end{bmatrix}$$

• بعدها نعمل ايه ؟

○ طالما احنا طرحت الميو من القيمة ، بيقى هنضيفها في المعادلة الاخيرة كدة :

$$(y^{(i)} - \mu) + \mu$$

- ساعتها هتلaci ان قيمة اي تقييم بقى زي ما هو ، طرحتنا الاول الميو ، بعدها ضفناها تاني فبقت نفسها
- طب ليه عملنا اللغة دي؟ عشان هتنفعنا جدا لما نيجي نقيم حد مقىمش قبل كدة ، زي العنصر الخامس ، واللي لما بيجي تقييمه يكون بصفر ، هنضيف عليه الميو اجباري ، فيكون برقم معقول

$$\frac{1}{n} + \frac{(1-x)^2}{n}$$

- كدة المستخدم الجديد مبقاش بفر ، لكن برقم ميو ، اللي هو اصلاً متوسط التقييمات من الناس اللي سبقوه ، فهاقرب له الافلام اللي اغلب الناس قيموها ايجابا
 - فيه نقطة مهمة ذكية ، لو سرت عندها 50 سنة اشتربت في الموقع ، فهي غالباً مش ه تكون مهتمة بالافلام اللي اغلب الناس رشحوها ، لكن من اللي الناس اللي زيها رشحوها
 - يعني نقطة المتوسط دي كويستة ، بس لو اتعملت segmentation للناس اللي زيبي ، مش كل الناس و خلاص ، وبالتالي هنعمل متوسط للناس المتشابهين مع المستخدم ده ، في جنسه و سنّه و حالته المادية و لغته ، كل ده هيخلني اقدر اجيب افلام اقرب ما تكون للي هو هيرحبها
-