



AN END TO END DEEP LEARNING TRAINING V 2.0

by : Muhamed Essam , Ahmed Ismail & Eyad Mohammmd



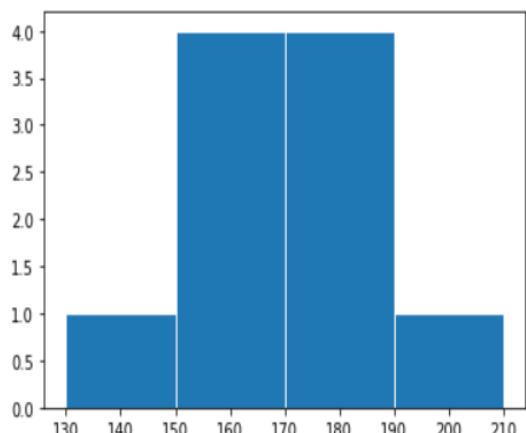
Machine learning

Distribution

تعالوا مع بعض نبدأ رطة ال machine learning ب شويه مفاهيم موم جدا اتنا نفهموا ف تعالوا نشوف مع بعض.

تخيل ان احنا قسمنا طول الناس لفئات مثل اقل من 150 - من 150 ل 170 - من 170 ل 190 اطول من 190 وخدنا مجموعه من الاشخاص وابتدينا نقيس طولهم ف

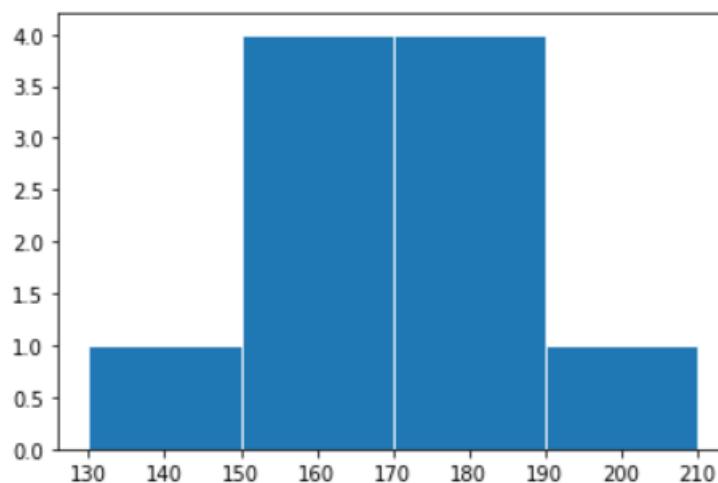
- اول واحد كان طوله 172 ف حطيناه في الفئة من 170 ل 190
- ثاني شخص كان طوله 165 ف حطيناه في الفئة من 150 ل 170
- ثالث شخص كان طوله 169 ف حطيناه في الفئة من 150 ل 170
- رابع شخص كان طوله 175 ف حطيناه في الفئة من 170 ل 190
- الخامس شخص كان طوله 145 ف حطيناه في الفئة اقل من 150
- السادس شخص كان طوله 197 ف حطيناه في الفئة من اطول من 190
- سابع شخص كان طوله 164 ف حطيناه في الفئة من 150 ل 170
- الثامن شخص كان طوله 177 ف حطيناه في الفئة من 170 ل 190
- التاسع شخص كان طوله 159 ف حطيناه في الفئة من 150 ل 170
- العاشر شخص كان طوله 179 ف حطيناه في الفئة من 170 ل 190



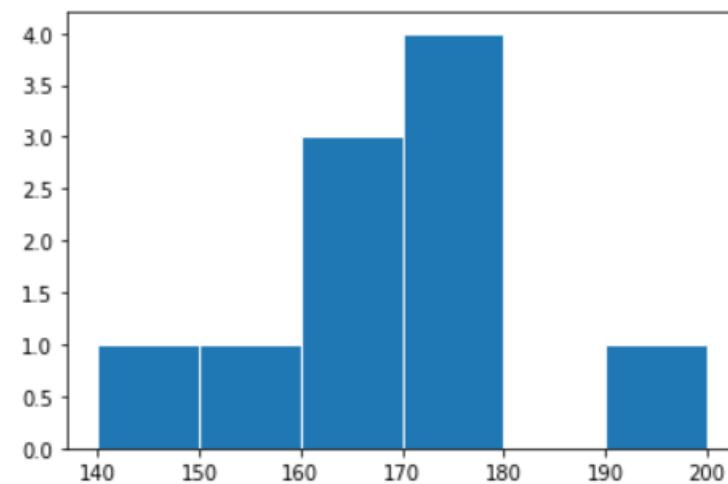
ومن هنا نلاحظ اني بحاول ارسم histogram اللي من خلله بيانلي قد ايه عندي ناس في كل فئه و منها نقدر نقول اني لو جبت اي شخص وقشت طوله غالبا هيطلع في فئه من 2 يا من 150 ل 170 يا من 170 ل 190 اما الفئتين اللي اقل من 150 واللي اطول من 190 ف من الواضح ان احتماليه اني القي دد ف الفئات دي صعب بسبب قلة عددهم.

بس مش ملاحظ حاجه؟؟ اغلب الاشخاص اللي في الفئه من 170 ل 190 هما ف الاساس بين ال 170 و 180 واغلب اللي في الفئه من 150 ل 170 هما ف الاصل بين 160 و 170.

ف تخيل لو صغرت الفئات وخليناها اقل من 150 - من 150 ل 165 - من 165 ل 180 - من 180 ل 195 و اطول من 195 ف هنلاحظ ان التوزيع اختلف شويه بس بقا ادق علشان بقا كل شخص بيتحط ف فئه اصغر بس تفتكري في افضل؟

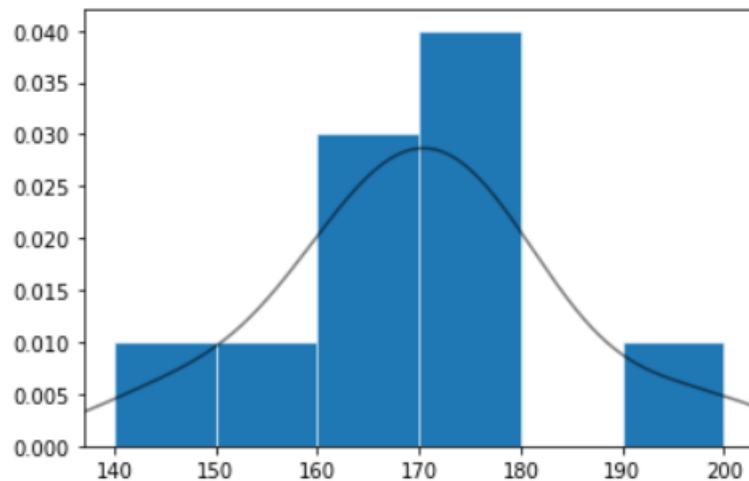


تخيل لو صغرت الفئات و خلتها اقل من 150 - من 150 ل 160 - من 160 ل 170 - من 170 ل 180 و من 180 ل 190 - اطول من 190 ساعتها كل شخص هيتحط ف فئه اصغر وبالتالي هيبيقي التوزيع ادق.



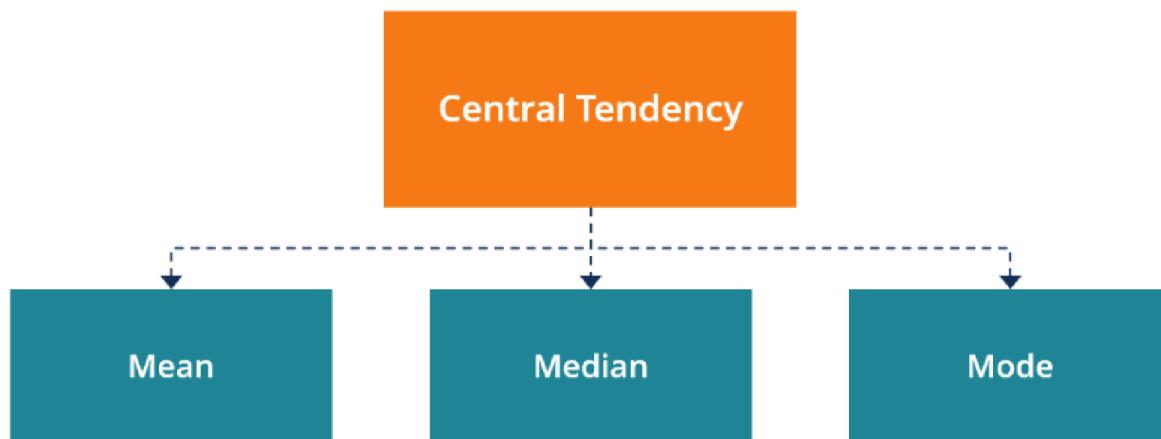
بس هننغير لحد امتى؟؟ لو فضلنا نصغر ف الفئات هنوصل لمصرحله ان كل شخص هيكون في فئه لوحده وده مش مفيد ولو كبرنا الفئات لحد منوصل انهم ييقوا فئتين بس ساعتها بردم مش هستفيد غير بمعلومه وده و هي كام شخص طوله فوق ال average وكام شخص طوله اقل وبالتالي ف انا بجرب اكتر من تقسيمه لحد موصل لكتر تقسيمه توضح شكل الداتا.

ولو بصينا على ال histogram هنلاحظ ان في احتماليه اقل اني الاقي حد اقصى من 150 او اطول من 190 وفي احتماليه كيرم اني الاقي شخص بين 160 لـ 180. ولو لو دبيت اوصل ال histogram بحيث اني ارسم بيها منحنى ف هنلاحظ استفادته مهمه جدا ان حتى وان كنت مقدرتش الاقي حد من 180 لـ 190 ف في ال histogram هنلاحظ انه مبيديش اي احتماليه اني الاقي شخص ف الفئه دي لكن في المنحنى ماذال ليه احتماليه ضعيفه بس موجوده وده واقعي جدا لانك لو بصيت في حوليك هتلاري ان اكيد في ناس طولهم من 180 لـ 190 تكون انك ملتقتش في الشخص اللي قست طولهم دم مش معندهم مش موجودين.



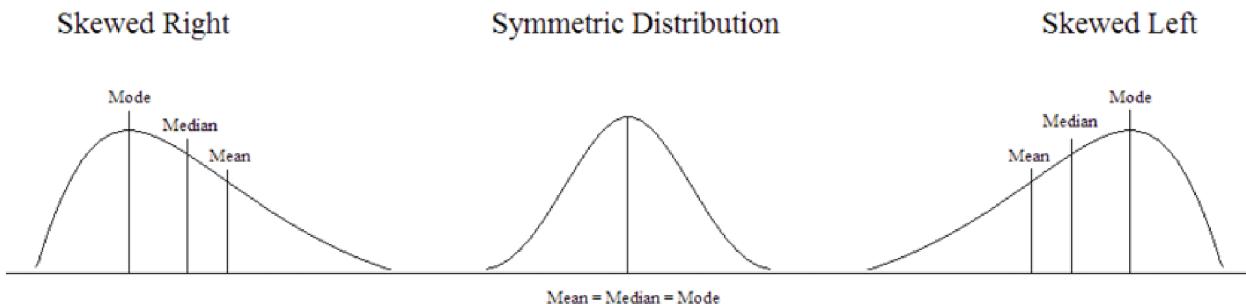
وبخدم نقدر نقول ان ال histogram و ال curve اللي رسمته عليه الاثنين بيقيسوا توزيع طول الشخص بحيث ان المنطقه اللي عاليه في ال histogram او ال curve هي منطقة احتماليه وجود شخص فيها عاليه و المنطقه القليله احتماليه وجود شخص فيها اقل.

Statistical measures



ودلوقتي هنكلمل عن شويه مقاييس احصائيه بتفيدنا كتير ف تخيل معايا لو عندي مجموعه من الداتا هتللا عن طول الاشخاص زي موضخنا في المثال اللي فات ف احنا كان عندنا 179,159,177,164,177,145,169,175,165,172 ف لو قلتلك قولي رقم واحد يعبر عن البيانات اللي ادامك دي هتنقولي ايه؟؟ وهنا هنكلمل عن المتوسط وهو يعتبر اكتر قيمة ممكن تعبّرلي عن البيانات ف لو حسبنا المتوسط اللي بيكون عباره عن مجموع القيم على عددها وهنلاقي انه 170.2 سم طيب نفترض ان كان عندي دد طوله مش طبيعي مثل 220 سم ودم مش من العادي اني الاقيه لانها بتبيقي استثناء لكن حصل ولقيته لو حسبت المتوسط تاني بعد مضفت ال 220 هنلاقيه بقا 174.7 يعني المتوسط اترجع 4.7 سم لمجرد بس ان جاي شخص واحد طوله استثنائي طيب هل دم كده مقاييس كويسي؟؟ الاجابه هي ام هاذال كويسي بس هنا نيجي لمقاييس تاني وهو median ودم يترتيب القيم من الكبير للصغير او العكس وبعددين بيجيب القيمه اللي ف النص ف لو جينا حسبنا ال median قبل منضيف ال 220 هنلاقيه بقا 170.5 وبعد مضفتنا ال 220 هنلاقيه بقا 172 ام اديد هو اثاره برد بالقيمه بس تأثيره اقل كتير زي محسنا شايفين.

طيب لو نفرض كانت الداتا بناعطي categorical زي مثل تقديرات الطلبه A B C D F وعايز اعرف حاجه فيهم تديني معلومه عن المتوسط ف الداتا دي تفتقروا ايه الدل؟؟ نقدر نستخدم هنا ال mode وهو عباره عن ال category اللي اتكررت اكتر ودم عادتا بنستخدمه مع ال .categorical data



طيب دلوقتي نفترض ان عندي مجموعتين من الداتا مجموعه عباره عن 5,10,15 ومجموعه عباره عن 0,10,20 وحسبت لكل وحيث اقارن بينهم ف نبتدى بال المتوسط هلاقى ان رغم اختلاف الداتا الا ان المتوسط واحد بس انا متأكد ان في اختلاف بس اوضحه ازاي؟؟ الاختلاف مش بين ف المتوسط لنه متوزع حولين المتوسط بشكل متتساوي ف المسافه بين ال 5 و ال 10 هي نفس المسافه بين 10 و 15 و المسافه بين 0 و 10 هي نفس المسافه بين 10 و 20 ف بالتالي المتوسط ف الحالتين 10 وبس الاختلاف اللي ملاحظينه احنا ان الداتا مش متوزعه ف نفس ال Range ومن هنا احتجنا مقاييس تحسينا ال dispersion واللي هو مقدار بعد الداتا عن بعضها.

في منهم مقاييس بتاخد ال range زي ال maximum – minimum لكن دم في الغالب مش بنفضل له لأن لو في قيمة extreme زي ال 220 كده هتكبرلي ال range في حين ان من 197 ل 220 مفيش اي قيمة وبكله ف ال Range متخيّل extremes ف ليه مننسبش بعد النقط عن المتوسط ؟؟ ف بكله نقدر نقول ان النقطه اللي بحسب بناءً عليها هي متوسط الداتا و نحسب قد ايه الداتا بعيد عن المتوسط بتاعها ؟؟ ودم مقاييس اسمه absolute deviation أو MAD و بنقول Mean Absolute Deviation لان ببساطه لو ركزنا هنلاحظ ان في قيمة اكبر من المتوسط وقيم اصغر ف انا لو مأخذتش ال absolute ساعتها في قيم هتططلع موجبه وقيم سالبه ف هينقصوا من بعض بس انا مش عايز كده انا عايز اجمع كل المسافات وبنقول Mean لاني مش مهتم اعرف ال total بتاعهم لا انا مهتم اعرف في المتوسط كل نقطه بتبعد عن ال average. ف لو ديينا نشوف المعادله بتاعته هنلاقيها عباره عن اني هاخد كل قيمة اطرحها من المتوسط و هاخد ال absolute value بتاعه الطرح بعددين هجدهم كلهم و اقسم علي عدد الداتا ف بكله اكون حسبت ال .MAD

Formula

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \mu|$$

μ	= average value of the data set
n	= number of data values
x_i	= data values in the set

Fig.x Mean Absolute Deviation

وعندنا ال variance ودم بدل ال absolute بيأخذ المربع ف بيكون عباره عن مربع الفرق بين كل قيمة

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Fig.x Variance Formula

والمتوسط بعددين بنجدهم كلهم وبنقسم علي عدددهم بس الرقم اللي هيطلع دم هو فيمه مربعه

بمعنى اني لو حبيت اوصفها هقول ان غالبا مربع طول الشخص يبعد عن المتوسط ب 200 سم مربع!! حاجه مش حلوم صح؟

ف طلعلنا ال standard deviation و هو عبارة عن الجذر التربيعي لل variance وبكلده ف انا قدرت ارجع تاني للوحدة نفسها بعد مكانه مربعة ف لو ال std طالع مثلا ب 20 سم ف بنقول ان في الغالب طول الشخص يبعد عن المتوسط ب 20 سم بقت منطقية كده صح؟

$$\sigma = \sqrt{\frac{\sum (x - u)^2}{N}}$$

Fig.x Standard Deviation Formula

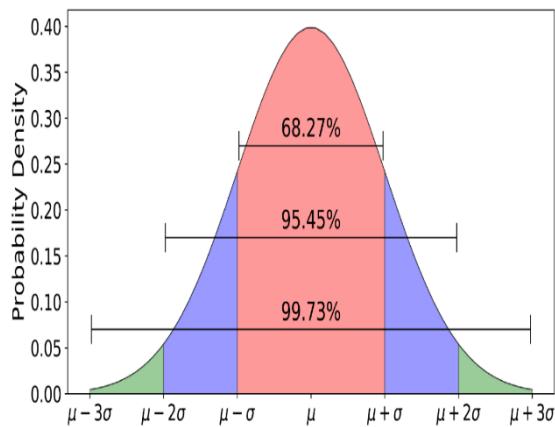
طيب عظيم جدا ب كده نكون عدinya على مقاييس المتوسط و كمان مقاييس ال dispersion او التشتت تعالوا بقا نشوف توزيعه من اكتر التوزيعات اللي بنقابلها.

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element \bar{x} = sample mean n = sample size

لو حابب تعرف اكتر عن $1 - n$ تقدر ترجع ل Bessel Correlation وتفهم اكتر.

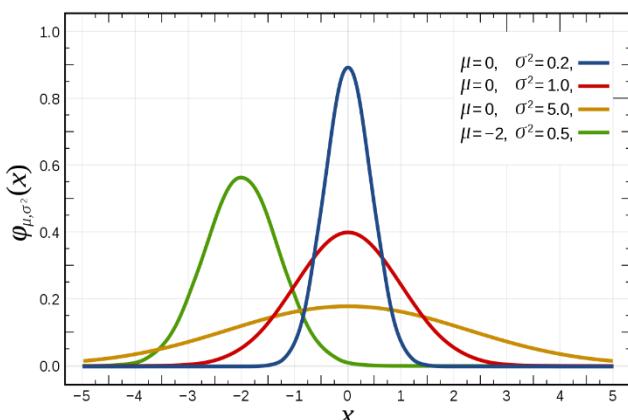
Normal distribution

وهد نوع من انواع bill shaped curves symmetrical لنه bill و شبه ال bill و دم يعبر عن اغلب القياسات الي بنشوفها ف حياتنا لو بقينا هنلقي ان القيم الصغيره جدا او الكبيره جدا احتماليتها ضعيفه و كل منقرب من القيم اللي ف نص المندني الاحتماليه بتزيد لحد منوصل ل نص المندني بظبط اللي واحد اعلي قيمه ولو عكسنا دم علي واقعنا هنلقي ان ف الطول مثلا احتماليه اني الاقي دد قصير جدا او طويل جدا ضعيفه و احتماليه اني الاقي دد طوله ف المتوسط مثلا من 160 ل 180 كبيره جدا. ولو عكسناه علي المرتبات ف احتماليه اني الاقي دد مرتبه كبير جدا او قليل جدا ضعيفه لكن احتماليه اني الاقي دد مرتبه في المتوسط كبيره جدا.



ال distribution normal ديمما بيكون محوره او اعلي الاحتماليه فيه هي ال average وكل مبن بعد عن ال curve بتقل الاحتماليه لحد منوصل لطرف ال average سواء من الاول او الآخر طيب ايه اللي بيفرق من **normal distribution الثاني؟**

اولاً المتوسط و ده ينتمي لـ center



2 ف ده معنام ان:

- 22 من الداتا يبن 18 ل 68%

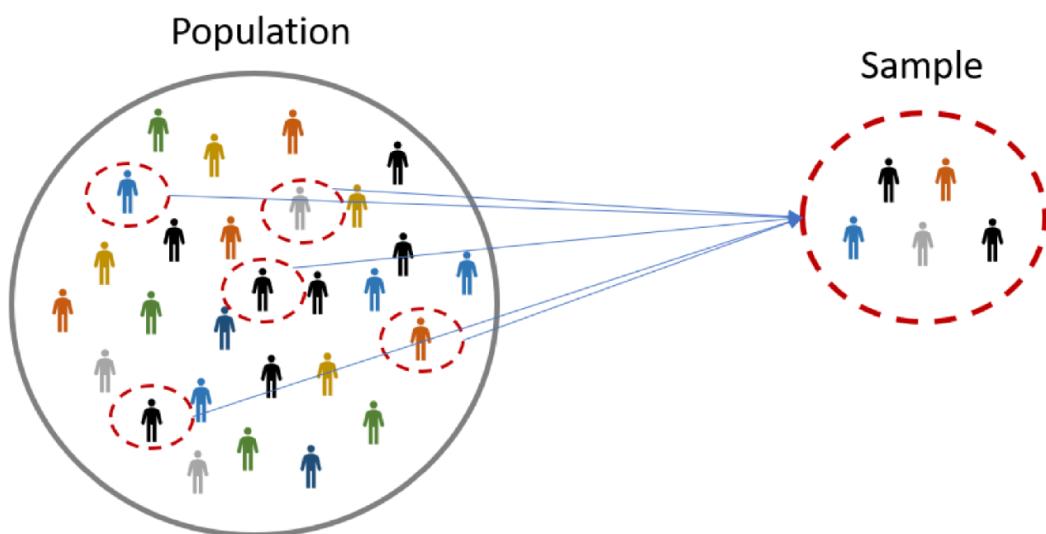
- 95% من الداتا بين 16 و 24

- 26 g من الداتا بين 99.7%

يبقى بكم علشان نرسم normal distribution ف احنا مش محتاجين غير معلومتين وهما ال mean و ال std.

Sampling

لحد دلوقتي كلمنا عظيم جدا لكن في مشكله هنا. نفترك لو عايز ادرس العلاقة بين عدد سنين الخبرة و المرتب في مصر محتاج اجمع داتا من مين؟؟ من كل كائن دي ف مصر شغال وبياخد مرتب متخيلا!! هو ده منطقى اصلا!! ومين هيدفع تمن تجميع الداتا دي كلها!!! ومين معام الوقت الكافي انه يجمع كل ده!! وهنا بتبقى المشكله بتاعي مش منطقى بسبب التكلفه العالية و الوقت اللي محتاجه ف تجميع الداتا ف بنأخذ عينه مثلًا من 2000 موظف و بنسلهم ونجمع منهم الداتا والعينة دي بنسميها ال sample بس لازم نبقي فاهمين حاجه مهمه. النتائج اللي هندسها على ال sample هيكون فيها نسبة error وكل م حجم العينة يقل نسبة ال error تزيد لأن من الطبيعي ان عينه من 2000 شخص مش ه تكون معبره عن اكتر من 20 مليون شخص شغالين ف وبالتالي كل م حجم العينة يزيد كل ما نسبة ال error تقل وكل م يكون في تنوع في العينه بتاعي ما بين شباب و كبار اولاد وبنات كل هيكون ال error اقل وعلشان كده في دراسات بتتكلم عن طريقه سحب العينه بشكل يضم كل اقل error ممكن لكن ده مش موضوعنا حاليا.



هنا بقا نسأل نفسنا سؤال. هو انا امتي اجيب $0 = \text{error}$ ؟؟ لو اخذت كل شخص ف مصر وسألته وبالتالي المقاييس بتاعي بيقي اسمها parameters لأنها محسوبه على ال population كلها. اما لو اخذت عينه ف هيكون الحسابات بتاعي اسمها estimates ويكون فيها نسبة ال error اللي اتكلمنا عنها.

Covariance and correlation

لحد دلوقتي احنا بندرس ف متغير واحد اللي هو الطول مثلاً لكن احنا بنحتاج احنا ندرس علاقه متغير بالثاني علشان نشوف تأثير واحد فيهم علي الثاني زي مثلاً العلاقة بين الطول و الوزن، عدد سنين الخبره والمترتب، والعمر والطول، وهكذا.

هنا بقا يجي دور ال covariance وده بيجاوب علي سؤال مهم جداً وهو تفتكـر هل يوجد علاقه بين عدد سنين الخبره والمترتب؟ ولو في تفتكـر علاقه طردية؟ بمعنى ان لما يزيد عدد سنين الخبره يزيد المترتب ولا عكسـيه؟ بمعنى ان لما يزيد عدد سنين الخبره بيقل المترتب؟ احنا كلنا عارفين انها طردية بس هنا ال covariance هيطلعـنا الاجابـه ف صوره رقم لو كان موجب يعني العلاقة طردية ولو كان سالب يعني عكسـيه ولو كان قريب من الصفر يعني مفيش علاقه عدد سنين الخبره والمترتب.



Covariance Formula

For Population

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N - 1)}$$

الي بتطلعـنا هي

طيب بالنسبة للقيمه

ملهاش اي مدلول عن قوه العلاقه؟؟ في الغالـب لا لاني عارف كويـس ان ال covariance بيتأثر بالقيم نفسها بتاعـه الداتـا ف لو عندي داتـا لمـرتبـات هيطلعـلي ال covariance غالـباً بالـلافـ ولو عندي داتـا بتاعـه العـمر وـالـطـول مثلاً هيطلعـلي covariance غالـباً في حدود المـئـات ف بالتالي ال covariance بيتأثر بـقدر الـدـارـقامـ اللي فـ الدـاتـا كـبـيرـ او صـغـيرـ وـهـنـا يـجيـ لـسـؤـالـ مـفـهـومـ؟ـ هو مـيـنـفـعـشـ اـحـطـ
الـدـاتـاـ فـ؟ـ standard values

علـشـانـ نـحـطـ الدـاتـاـ فـ standard formـ مـحـتـاجـ اـشـوـفـ اـبـهـ الـقـيـمـ الـليـ مـمـكـنـ استـعـمـلـهاـ فـ تـفـتكـرـ
نـسـتـعـمـلـ اـيـهـ؟ـ

احـناـ مـحـتـاجـينـ قـيـمـهـ بـتكـبرـ لهاـ الدـاتـاـ بـتكـبرـ وـبـتصـغـرـ لهاـ الدـاتـاـ بـتصـغـرـ وبـتـاخـدـ كلـ الـقـيـمـ الـليـ فـ الدـاتـاـ فيـ
حسابـاتـهاـ عـلـشـانـ اـعـتمـدـ عـلـيـهاـ وـمـنـ هـنـاـ فـ اـحـناـ نـقـدـرـ نـسـتـعـمـلـ الـ standard deviationـ بـسـ هـسـتـعـمـلهـ
اـزاـيـ؟ـ

نـفـتـرضـ اـنـيـ بـدـرـسـ العـلـاقـهـ بـيـنـ xـ الـيـ هوـ عـدـدـ سنـينـ الخبرـهـ وـyـ وزـيـ مشـفـنـاـ فـ الـ covarianceـ فـ اـنـاـ
بـطـرـحـ كلـ xـ مـنـ الـمـتوـسطـ بـتـاخـدـهاـ وـكـلـ yـ مـنـ الـمـتوـسطـ بـتـاخـدـهاـ وـبـعـدـيـنـ بـنـقـسـمـ عـلـيـ Nـ الـيـ هوـ عـدـدـ الـ

الى معايا. ناقص بقا اني اقسم على std of x و std of y g std of y g std of x علشان ابقي كده عملت correlation بينه وبين القييم بتاعه x و القييم بتاعه y و دم يطلعنا ال standardize

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Fig.x Pearson Correlation Formula

وال strong direct relation correlation ببساطه يطلعنا قيمة بين -1 و 1 بحيث ان ال 1 معنادم relationship و ال 0 العلاقة يتضاعف ف بعدها بقيت بقدر احدد 1- قوام العلاقة بقيمة رقمية تعبر عن قد ايه في ترابط بين x and y.

بس احنا دلوقتي بنقىس الترابط بين y and x بس هل دم يسمحلي اني لو جالي شخص عدد سنين الخبره بتاعته 5 سنين اقدر اتوقع المفروض مرتبه يكون ٥ام؟؟ الإجابة هي ام اقدر اتوقع المرتب بمدلوليه نقطه تانية لكن ساعتها هبقى بتوقع بناءً عليا مدلوليه نقطه وحدة انتها لو انا عايز اتوقع بمدلوليه كل النقط اللي عندي ف تعالوا نتعرف علي اول model هنشتغل عليه.

Simple linear regression

ودلوقتي وقفنا عند سؤال وهو اني علشان اتوقع مرتب شخص عنده 5 سنين خبره بناءً علي كل الداتا اللي عندي ف هنا اناحتاج ارجع بالزمن شويه ونفترض مع بعض معادلة الخط المستقيم

كنتا بنقول $y = mx + c$ فاكرین؟ وكنتا بنقول ان ال m هو ميل المنحنى و c هو الجزء المقطوع من ال axis y وكنتا بنحسبها بطريقه حل المعادلات بمعلومية نقطتين بس حاليا انا معايا مجموعه من النقط اللي ميربطهم خط مستقيم ف اعمل ايه؟

بسط موديل بالنسبة لينا في ال ML هو Simple Linear Regression من ال Model دم انه لو دخله parameters One Label يطلعلي ال Best parameters لميل الخط المستقيم ونقطه التقاطع مع ال axis y علشان تكون معادلة الخط المستقيم اللي عارفينها $y = m*x + c$.

فهو m و b دول بالنسبة parameters اللي عايزين نوصل لافضل قيمة لهم احيانا بنسميهم Beta0 و Beta1 يعني يختلف اسمائهم بس المعادلة واحدة.

بس ف الاخر بعد مبنطاع معادلة الخط المستقيم لو جينا نديله اي input من اللي بنبيت عليهم معادلة الخط المستقيم نفترض

هيطاطعلي نفس ال actual output اكيد لان الخط المستقيم مش هيمبر بكل النقط ف طبيعي هيكون في نسبة Error وها يجي سؤال مهم نحسب ازاي ال error

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

خلونا نتكلم عن مقياس بنستعمله كتير وهو قيمه ال error اللي هي الفرق بين ال y actual و ال y predicted لك كل نقطه و يربع الفرق ويجمع على مستوى كل النقط و في الآخر يقسم علي عدد النقط n وعندنا كمان ال Residual Sum of Squares (RSS) وده بيكون عباره عن مربع قيمه ال error فقط من غير منقسم علي n.

لو عوضنا بالمعادله الاولى في الثانية هنوصل للمعادلات اللي هنشوفها دلوقتي

$$\beta_0 = \text{mean}(y) - \beta_1 \times \text{mean}(x)$$

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \text{mean}(x)) \times (y_i - \text{mean}(y))}{\sum_{i=1}^n (x_i - \text{mean}(x))^2}$$

Beta One

$$\beta_1 = \text{corr}(x, y) \times \frac{\text{stdev}(y)}{\text{stdev}(x)}$$

لو موهتم انك تعرف اكتر عن جت مين انصحك جدا انك تشوف ال Full Derivation.

وبعد ما بنينا ال simple linear regression و قدرنا نسبة ال error دلوقتي لو قلتلك ال error بتاعه 100 هل تقدر تقولي ال model ده حلو ولا وحش؟؟ للأسف لان لو قيم ال y كانت صغيره ساعتها يعني وحش ولو قيم y كانت كبيره ساعتها يعني حلو طيب ايه العمل؟؟

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

من هنا احتاجنا accuracy metric نقدر نقييم فيه ال model اللي هو R2 score وده بيكون عباره عن ال RSS بتاع الموديل بتاعي مقسوم على ال RSS اللي هيطاطل لو انا استغنىت عن الموديل

واستخدمت ال Average prediction ثابت لاي نقطه عندي وبكلد يكون ناتج قسمه ال RSS بقى على ال RSS بقى نسبه ال error اللي جابها ال model مقارنة بال average model طيب عظيم نجيب بقا ال accuracy ازاي؟؟ ببساطه هنطرح ناتج القسمه من 1 .percentage of accuracy الى percentage of error

Formula

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

.from scratch ليك model ومن النقطه دي تقدر تبني اول

الجدير بالذكر ان القواعد اللي احنا بنتكلم عنها فوق دي ساعدتنا ان احنا نوصل لـ Formula بسهوالة one dependent Variable , one independent Variable best fit Line هاين يعني لو كان عندنا أكثر من Variable الموضوع مينفعش يتحل بالطريقة المباشره دي ...

طيب تعالى الأول نلخص الطريقة المباشره دي بتوصلنا ليه !

1. هتجمع داتا لاي مشكلة تكون مفيش غير عامل واحد بس مؤثر في النتيجة يعني One X

and One Y

2. هتجيب الداتا تدخلها عندك في Python Code بقى تعملها Loading يعني.

3. هتشوف لو فيها أي مشاكل أو فراغات في الداتا أو لو احتجت تعمل Feature Scaling .

4. هتببدأ تحسب Betas بقى عن طريق العلاقات اللي كتبناها و اللي هي هتطلب منك انك

تدى للموديل X train , y train عشان يحسب .

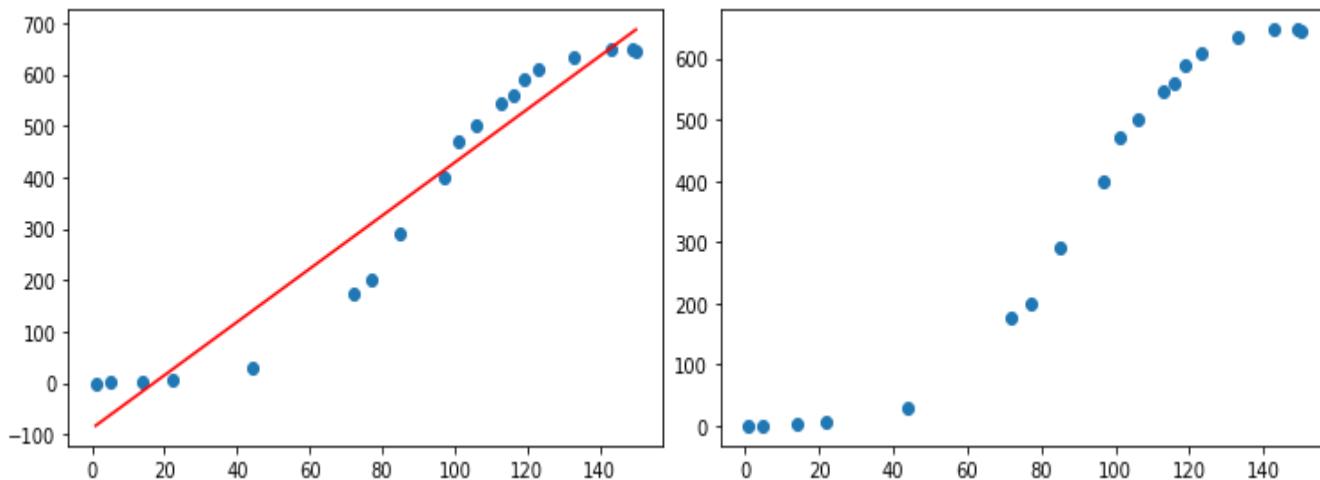
5. تستخدم Betas دي مع X test فالمودل هيطلع Predictions وقارنها بال y اللي معاك وشوف انت كوييس قد ايه عن طريق انك تحسب Error .

6. ممكن تدخل X جديدة خالص وتعوض ب betas هيطلع معاك .

الجدير بالذكر ان الـ 6 خطوات دول فعلا تقدر تطبقها وتطلع علاقة خطية لحل مشكلة مفيهاش غير One X طيب لو في أكثر من X الحل ساعتها هيقي ايه ؟

الحل بالنسبة لنا ان احنا نجأ لأساليب تانية هنتعلمها في Multiple Linear Regression بس قبل مزروح للنقطه دي تفتكرو لو كانت الداتا مش Linear زي م شايفين ف الصوره كده هنعمل ايه ؟

ساعتها معادله الدرجة الاولى مبيقتش كافيه وبقينا محتاجين شكل من اشكال polynomials اللي هي معادلات الدرجة الثانية والثالثة والرابعه وهكذا علشان نقدر ندي ال regression line مروونه اكبر انه يعرف يمشي مع الداتا بشكل كوييس بس تفتكرو دم هيتم ازاي ؟

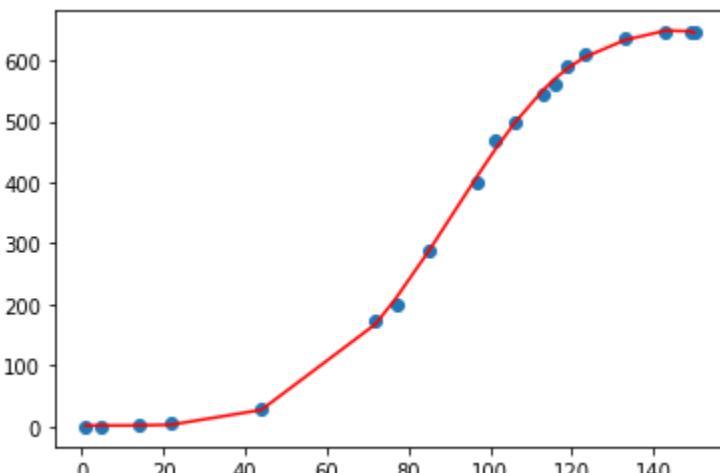


يitem عن طريق حاجه اسمها Polynomial Features

Polynomial Features

فكروم الـ polynomial features هو ان يكون عندي one feature اسمه x مثل وانا ابتدئ اعمل polynomials من الـ x دم يكون عباره عن ... $x^2 x^3 x^4$ وكل متزود generate

كل م الـ complexity بتاعه الداتا بتزيد وكل م الـ linear regression model بقا معاه features من الدرجة الثانية والثالثة والرابعه اللي يقدر من خلالها يخلی ال curve مرن اكتر وقدر يناسب شكل الداتا وهتلقي ال

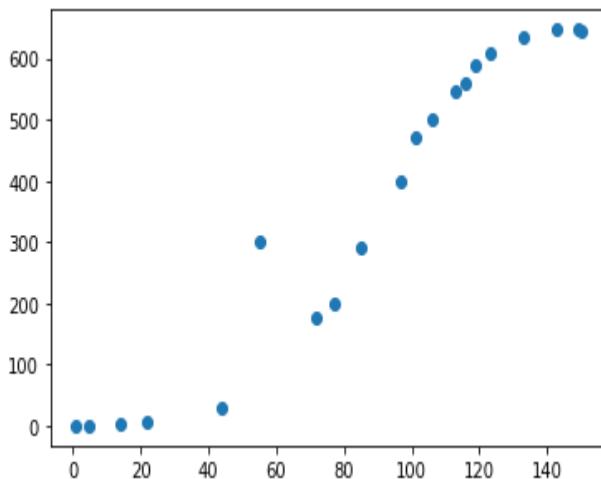
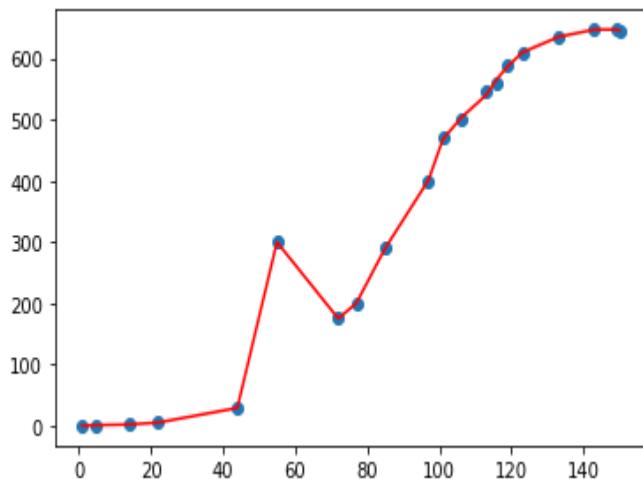


model بتاعك ابتدى يطلع خط فعال متماشى مع شكل الداتا.

بس سؤال هو انا بحط polynomials لحد كام؟ يعني لحد 3 ولحد 4 ولحد 10 ولحد 100 ؟؟ طبعاً انت كل متزود الداتا بتكتّر وبالتالي ال model بيكون complex اكتر وساعتها بيقدر انه يعمل fit على ال data بشكل كوييس.

بس هنا يحضرني مشكله؟؟

لو انا ال curve بتاعي بقا مرن جدا لدرجة انه قادر يمشي علي كل نقطه موجوده في ال data تفتكر
ده ممكن يسبب ايه؟ تعالوا نشوف



ایہ الی حصل ہے؟

انا دلوقتني ضفت نقطه extreme ومش من الطبيعياني الاقيها ودم واضح لان مفيش اي
نقط حوليه ساعتها بقا هتللاقي ان ال polynomial features عملت مشكله وخلت ال graph بتاعي
راح جاب ال extreme في محاوله انه يقلل ال error ولكن تفتكرو انت مكان ال model هتعملني
كم؟؟

الاجابه هي لا لان ببساطه النقطه دي حalle استثنائيه والداتا كلها مش ماشيء معاها ف كان من المنطقى اني اتجاهلها بس ازاى؟

الي خلي ال model يطلع يجيب النقطه دي ويرجع تاني ينزل للداتا هو ان ال model معادل خط مستقيم معامل ال x هو ال slope بتاعها مخلية قادر يدرك برهته واحدنا عارفين ان في معادله الخط المستقيم معامل ال x هو ال slope بتاعها وانا عندي معامل ال x هو ال Beta ف بالتالي لو قدرت اني ادكم قيمة ال Beta انها متزدش بشكل

كبير ف انا حكمت ال Stop انه ميغلاش وبالتالي مش هيكون عند القدرة انه يطلع للنقطه ال extreme points وبخدمه هقدر اني اضيف polynomial features ولكن الحكم على ال Beta بتعده كل

فاحنا دلوقتي يعتبر بقى في ايدينا parameter يخلينا نتحكم في مدى تعقيد model ودم شئ كويس بس الناحية الثانية محتاجين دد ينظم التعقيد دم ويعرفنا انه degree هي الأفضل ودم عن طريق

Regularization Techniques

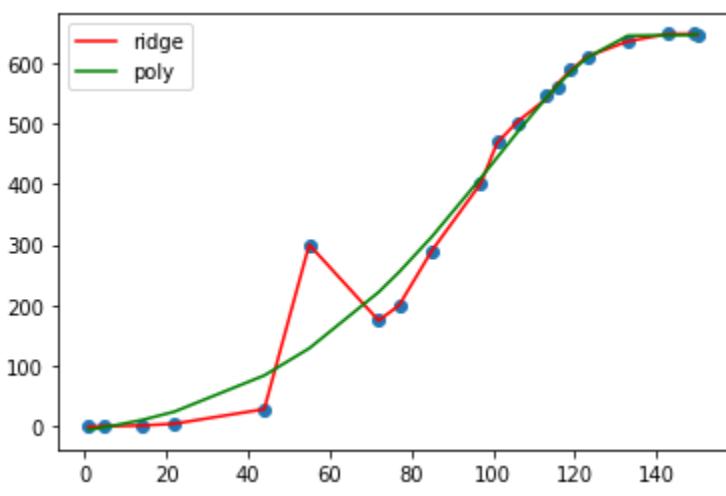
Ridge Regression

فكرة Ridge Regression هو انه يضيف Penalty على ال High Slope ودم عن طريق انه يبعد Ridge Linear g Simple linear Cost function في بالشكل دم يعني لو شوفنا ال

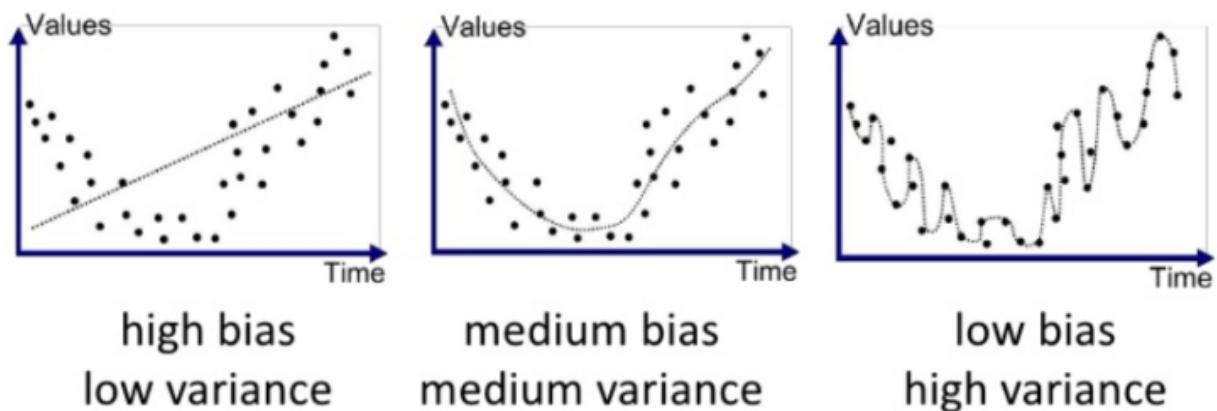
Linear Cost Function = Sum of Squared Error

Ridge Cost function = Sum of Squared Error + lambda * (Slope)^2

الطريقة دي هتلaci ان في ترم زاد وهو خاص بالميل



بسكل عام يعني كل ما يزيد Slope بتاع High Variance دم معناها انه ساعتها بنزود Penalty بتاعته يعني واكتننا بنقول للموديل ان دم ودش أما لو Slope بتاعه قليل يعني ساعتها دم أحسن معناها انه Very low variance لو ولكن بردم طبعا لو دم هيبيقي ودش بردم لتن ساعتها Sum of Squared Error كبير جدا وبالباقي ال model هيفضل ال Medium / Low Variance اللي عامل أفضل



وده طبعاً يندرج تحت Topic عدنا اسمه Ridge and Variance Trade off لو قدرنا شابع Regression هنلقي انه على High Degrees يقل قيمة Coefficients القيمة قريبة من ال 0 ولكن مش صفر

Lasso Regression

واحد من أشهر regularization techniques هو Ridge , Linear لعملاً مقارنة ليه مع الفرق

Linear Cost Function = Sum of Squared Error

Ridge Cost function = Sum of Squared Error + $\lambda \cdot (\text{Slope})^2$

(Lasso Cost function = Sum of Squared Error + $\lambda \cdot |\text{Slope}|$)

الفرق فعلياً هو في ان بدل ما بنربع Coefficients ان احنا بنأخذ لهم ال بس فالفرق الصغير ده بيفرق في ان القيمة كده بقت تقدر تطلع اصفار لل Coefficients وطبعاً كونك بس تقدر تدي ل Feature منهم Coefficient بيساوي 0 ده معناها انك تقدر توقيع ده فعشان كده بنستخدم Lasso Regression عشان يعمل Feature Selection فبيوقع Features اللي ملهاش لازمة ويسيب بس اللي ليه لازمة وطبعاً بيعمل Dimensionality Reduction فبيقلل Complexity يبقى انت كده قدامك طريقين :-

1- عن طريق Poly features تقدر تزود Complexity بقاعد الموديل .

2- عن طريق Lasso , Ridge تقدر تقلل Complexity بقاعد الموديل .

فكلدك بقى عن طريق Hyper parameters tuning تقدر توصل لأفضل حالة تحل معاك المشكلة يعني لو الموديل بقى simple والمشكلة complex تقدر تزود تعقيد الموديل عشان يتاسب مع المشكلة .

وهنا نقدر نقول انا عملنا شغل كبير جدا على feature واحد اللي هو طيب لو عندي أكثر من x مثل
لو بحاول اتوقع سعر البيت بناء على مساحته و عدد الغرف و المنطقة و المحافظة مثل هنا بقا
عندي أكثر من x و فيهم اللي عباره عن ارقام و فيهم اللي عباره عن categorical زي المنطقة مثل
هتكون **cairo** **giza** **alex** وهذا ف تعلموا نشوف مع بعض ايه المشكله و نحلها ازاى.

Data Preprocessing

الفكرة وكل مافيها هي ان ال Machine Learning Models بشكل عام بتقابلها مشاكل انها بتبقى محتاجة الداتا نقية يعني لو فكرنا فيها كده بشكل أو باخر يعني انت مثل عشان Linear Regression يشتغل فهو يحتاج ان يكون كل Features مثل يعني كل الأعمدة القيم اللي فيها أرقام يعني لو مثل لقى في وشه كلمة أو مثل ثانية فاضية كل دم أكيد هيوقفه فبنبدأ لفكرة Pre Processing محل المشكلة

Columns with None Values

لما يكون عندك خانة فاضية أو None يعني مثلاً في عمود .. تفتكر الحل ممكن يكون ايه ؟

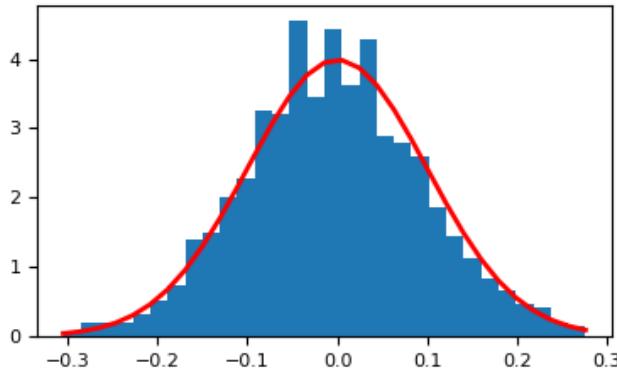
1- انت ممكن تبذل السطر كلها بساع الداتا يعني السطر دم هيتمسح من كل الأعمدة
طيب فرضنا ان كان في 40 عمود وفي السطر دم كلهم موجودين ماعدا القيمة دي في العمود دم ..
ممكن يكون دم مش أفضل حل .

2- انت ممكن تمل الفراغ بقيمة Mean وهي قيمة بتابع العمود نفسه لأن Mean هو الـ Center بتابع الداتا فهو أكثر دلالة حقيقية فيه تحيز لأي جهة

طيب تخيل ان الفراغ دم كان موجود في عمود فيه Categorical Data ساعتها Mean مش هينفع
و Median مش هينفع يبقى هنلأجأ Mode و هو Most Frequent بقى العمود نفسه.

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
10	Louise	Female	8/12/1980	9:01 AM	63241	15.132	True	NaN
11	Julie	Female	10/26/1997	3:19 PM	102508	12.637	True	Legal
12	Brandon	Male	12/1/1980	1:08 AM	112807	17.492	True	Human Resources
13	Gary	Male	1/27/2008	11:40 PM	109831	5.831	False	Sales
14	Kimberly	Female	1/14/1999	7:13 AM	41426	14.543	True	Finance
15	Lillian	Female	6/5/2016	6:09 AM	59414	1.256	False	Product
16	Jeremy	Male	9/21/2010	5:56 AM	90370	7.369	False	Human Resources
17	Shawn	Male	12/7/1986	7:45 PM	111737	6.414	False	Product
18	Diana	Female	10/23/1981	10:27 AM	132940	19.082	False	Client Services

Feature Scaling



زي ما اتكلمنا عنه Feature Scaling يعتبر حل لمشاكل كثيرة جدا وفي نفس الوقت يعتبر أحد أشهر ال Assumptions اللي بنعملها يعني من نظرتك لل histogram بناع الداتا ممكن تفترض ان الداتا دي Normally distributed عليهما بشكل معين وبنعملها وخلاله Standardization

من أشكال ال Transformations بردم اللي ممكن تتم على الداتا هي Power transformation .Box Cox transformation اللي بيتم على الداتا لو حابب تعرف أكثر عنه دور على

Handling Categorical Data

لو الداتا اللي معانا نوعها Categorical Data فالأسف الكود بناعك مش هييفهمها فانت تحتاج تحول ال categories دي لحاجة مفهومة وهي الأرقام ممكن يبقى قدامك أكثر من طريق لو افترضنا ان معاك 2 classes مثلا يعني male , female ساعتها ممكن تفكر انك ترمز لل male ب 0 ولل female ب 1 الفكرة البسيطة دي هتلدك الموقف ودي حاجة احنا بنسميها Label Encoding

طيب لو عندنا مثل 4 ولا 5 حاجات يعني أسمى بلدان مثل أو أسامي أكلات فانا ممكن اخليها بردم 0 و 1 و 2 و 3 و 4 و 5 ولكن لأن الأرقام هتبقى متفاوتة فاحنا على الأغلب مبنجاش ل Label Encoder وبنستخدم بداله One Hot Encoding اللي بكل بساطة بيشيل العمود نفسه ويكسّره لكتير من عمود بنفس عدد different items اللي موجودين في العمود الأساسي .

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

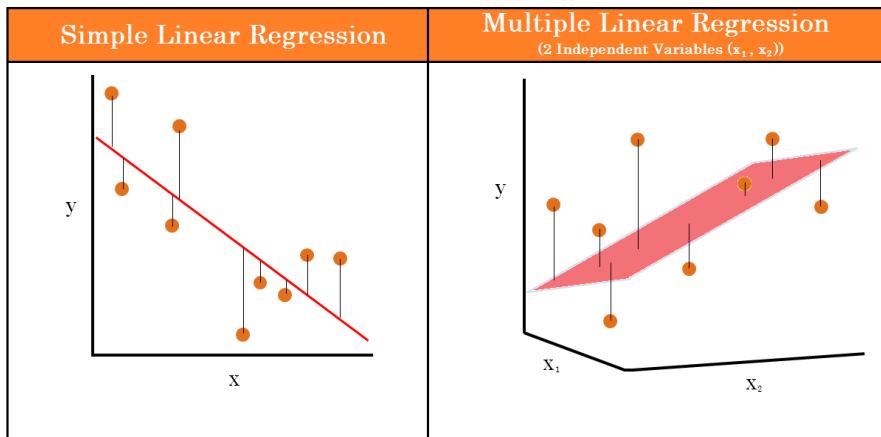


One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

لو بعدنا شوية عن فكرة PreProcessing لأن الموضوع قد يطول شرحه ونرجع تاني للسب اللي
خلاننا نعمل.preprocessing وهو انشا ندي ال data لـ model يتعلم عليها ف تعالوا نشوف الـ Multi
يعمل ايـه.

Multiple Linear Regression



لو حبينا نقارن فكرة Simple معها مفهوم لو معانا بالـ Multiple
فاحنا مثلـ Three Features
بنعتبر ان كل Feature بيمثل Dimension
فاحنا لو تخيلنا كده نفسنا ان في Dimension زيادة
فالنقط بقت متوزعة في Three Dimensional Space

يبقى بالباقي الـ model نفسه اللي بنعمل بيـه Fitting هيـقـى شـكـلـه بدـل ما كان خطـ هـيـقـى
عبارة عن plane بـيـحاـول يـوـسـطـن نفسـه كـدـم بـيـن النـقـط عـشـان يـحـقـق أـقـل Error فيـ Training
betas / weights وـدـم طـبعـا عن طـرـيقـ ان اـحـنا نـعـمل tuning لـكـل coefficients اللي هيـ Testing
باتـاعـتـ المـوـدـيل نفسهـ.

وهنا يحضرني سؤال تفتكـر لو كان حجم الداتـا كـبير جداً و مش معاـك ram تـكفي الداتـا او الـLinear Regression هـيأخذ وقت كـبير جداً وانت مش هـتقدر تستـوي الوقت ده كلـه ساعـتها الحلـ ايـه؟؟
وهـنا اقولـك انـ الـ exact solution مش هـيـقي عمـلي واحـيانـا مـبـيـكونـش مـعاـك الـ resources الـ الكـافـيهـ ولاـ الوقتـ الكـافـيهـ لأنـكـ تـطـلـعـ الـ exact solution فـ بنـجاـ حلـ بـديـلـ يـوصـلـنـا لـ .solution

from exact solution to approximate solution

تخـيلـ انـكـ عـاـيزـ تـعـملـ كـبـاـيـهـ شـايـ لـصـدـيقـ عـزـيزـ عـلـيـكـ بـسـ اـنتـ مـتـعـرـفـشـ درـجـهـ الحرـارـهـ المـنـاسـبـهـ الـليـ صـدـيقـكـ يـقـدـرـ يـشـرـبـ عـنـدـهاـ كـبـاـيـهـ الشـايـ منـ غـيرـ مـتـكـونـ مـضـاـيـقاـهـ فـ تـفـتـكـرـ اـزاـيـ مـمـكـنـ توـصلـ لـ اـلـفـضـلـ درـجـهـ حرـارـهـ؟؟

Exhaustive Search

اـولـ تـفـكـيرـ مـمـكـنـ يـخـطـرـ فـ بـالـكـ هـوـ اـعـمـلـ كـبـاـيـاتـ شـايـ عـنـدـ كـلـ درـجـهـ حرـارـهـ مـمـكـنـهـ وـ اـدـيـعـمـلـهـ كـلـهـمـ يـدـوـقـهـمـ وـبـدـمـ وـبـدـمـ وـيـخـتـارـ الـليـ تـنـاسـبـهـ بـسـ تـفـتـكـرـ دـهـ حلـ عـمـليـ؟؟

يعـنيـ لـوـ مـعـاـكـ مـثـلاـ اختـيـارـاتـ هـتـجـربـهاـ زـيـ مـثـلاـ 50ـ 55ـ 60ـ 65ـ 70ـ 75ـ 80ـ 85ـ 90ـ 95ـ 100ـ سـاعـتهاـ هـيـكـوـنـ حلـ كـوـيـسـ لـانـكـ هـتـجـربـ 11ـ مـرـهـ بـسـ.

لـكـنـ درـجـهـ الحرـارـهـ زـيـ مـحـنـاـ عـارـفـيـنـ هـيـ continuousـ ومـفـيـشـ قـيـمـ مـحـدـدـهـ مـمـكـنـ اـجـرـبـهاـ دـهـ فـيـ عـدـدـ لـاـ نـهـائـيـ مـنـ الـقـيـمـ الـليـ مـمـكـنـ تـجـربـ سـاعـتهاـ فـكـرـهـ اـنـيـ اـجـرـبـ كـلـ الـقـيـمـ الـمـمـكـنـهـ بـقـتـ مـشـ حلـ عـمـليـ نـهـائـيـ.

انـماـ الـحلـ الـاـفـضـلـ هـوـ اـنـ يـكـوـنـ عـنـدـيـ اـسـتـراتـيـجـيـهـ اـدـورـ بـيـهاـ عـلـيـ اـفـضـلـ كـبـاـيـهـ شـايـ وـهـنـاـ بـقـاـ هـنـجاـ لـفـكـرـهـ الـ feedbackـ وـهـيـ بـسـيـطـهـ جـداـ.ـ هـخـتـارـ درـجـهـ حرـارـهـ عـشـواـيـهـ زـيـ مـثـلاـ 20ـ وـ هـعـمـلـ كـبـاـيـهـ شـايـ عـنـدـ درـجـهـ حرـارـهـ 20ـ وـهـدـيـهـالـكـ تـدوـقـهـاـ وـتـقـوـيـ feedbackـ عـنـهـاـ فـ زـيـ مـحـنـاـ عـارـفـيـنـ هـيـرـدـ وـيـقـولـ وـحـشـهـ جـداـ.ـ هـسـأـلـهـ سـؤـالـ بـسـيـطـ وـهـوـ هـلـ عـاـيزـ اـرـفـعـ درـجـهـ الحرـارـهـ وـلـاـ اـقـلـلـهـاـ؟ـ؟ـ فـ هـيـقـوـلـيـ زـوـدـ درـجـهـ الحرـارـهـ

فـ هـعـمـلـ كـبـاـيـهـ شـايـ عـنـدـ درـجـهـ حرـارـهـ مـثـلاـ 30ـ وـادـيـهـالـهـ يـدـوـقـهـاـ فـ هـيـقـوـلـيـ اـفـضـلـ مـنـ الـليـ فـاتـتـ لكنـ درـجـهـ الحرـارـهـ عـاـيزـهـ تـزـيدـ اـكـثـرـ

ف هعمله كبايه شاي تانيه عند درجه حراره 40 واديهاله ف هيقولي زود درجه الحراره اكتر ف هعملها عند 50 ف هيذوقها ويطلب ازود درجه الحراره ف هديله كبايه شاي عند 60 وهنا هيقولي معقوله بس مينفعش تزيد اكتر؟

ف هنعمل كبايه شاي عند 70 وهنا هيقولي لا دي بقت سخنه وبتلسعه ف عايز يقالها شويه هنا هتقلل درجه الحراره ل 65 وتديهاله يذوقها ويقولك زودها حاجه بسيطه ف هتزود درجه الحراره ل 67 وهنا يقولك دي ممتازه شكراء جيلا وهميشرب كبايه الشاي وهو مبسوط.

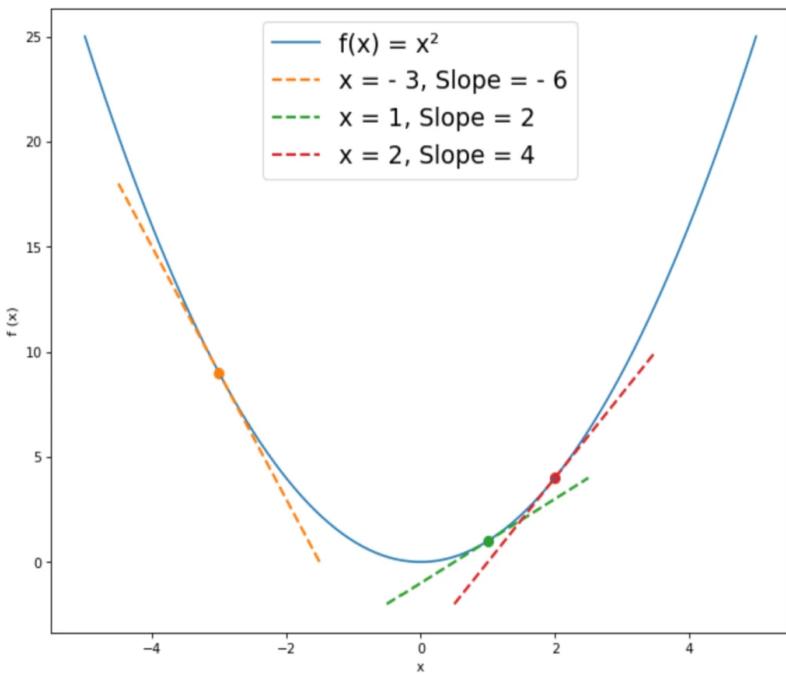
انت عملت ايه دلوقتي؟؟ انت كنت بتجرب بس مجرشash كل درجات الحراره لا احنا كننا بنتحرك بناء على ال feedback بتعاع صديقك ودي كانت استراتيجيه وصلتك لافضل درجه حراره ف وقت اقل.

تعالوا بقا نعمل reflect لدم علي ال regression .
درجة الحراره هي ال feature وقد ايه كبايه الشاي وحشه بالنسبةلك دم ال output .

درجة الحراره اللي ابديت بيه اسمها random initialization
وال error هو النسبة اللي بتعبر عن قد ايه كبايه عند كل درجه حراره كانت بتضيقه
ومحاولتك لانك تلاقي درجه الحراره الافضل اسمها optimization
طيب تعالوا نتعرف علي algorithm قادر انه يطبق نفس الفكره بظبط وهو

Gradient descent

ال gradient descent هو weights بيدور علي افضل قيمه لـ weights تديلك اقل Error عن طريق التجربه و تعديل ال weights بس هنا في سؤال مهم...! هو gradient descent او هيعرف ازي انه يحتاج يقل او يزود ال weights علشان يقلل ال error ولو هيذودها او هيقللها تفتكر يزودها او يقللها ب حام؟؟



الرسمه اللي على الشمال دي عباره عن
شایفین ف احنا محتاجين نشوف ايه هو ال
Weight عند القيمه 0 بس gradient descent
هيعرف ازاي يوصلها؟؟ لازم يكون في علامه
ميزة عند القيمه 0 تخليه لما يشوفها
يعرف انه وصل والعلامه دي هي ال slope
بتاع المنحنى.

هنا نسأل سؤال هو ايه ال slope

ال slope هو ميل الخط المستقيم عند نقطه
مقارنتا بال axis x او بمعني تاني هو ظل الزاويه (tan) اللي بين المنحنى و ال axis x ودم كلام علمي
لي حابب يدور ورامة انما هنا خلونا ناخذ الموضوع ببساطه.

الميل هو معدل التغير اللي بيحصل ف ال y نتيجه للتغير في ال x ف لو قلتاك مثل ان $4x + 5 = y$ ف
ال slope بكم؟؟ ب 5 لأن مع كل تغير ف ال x ال y هينتغير ب 5 اضعاف التغير دم. طيب تفتكرا
slope عند ال $x = 0$ هيكون كام؟؟ هيكون ب 0 لأن النقشه اللي عندها minimum error لو حاولت
ترسم خط مستقيم يلمس المنحنى (مماس للمنحنى) عند ال minimum error هتللاقي ان الخط دم
بيوازي ال axis x وبالتالي ف الميل بتاعه ب 0. يعني احنا دلوقتي علشان نوصل لل
محتاجين نوصل للنقشه اللي ال slope بتاعها ب 0.

طيب السؤال الاول!! هنبدأ منين؟؟ والاجابه هي هنبدأ من عند قيمة عشوائيه ف الاول والسؤال
هنا لو افترضنا اني بدأت من عند -3 - ف انا عارف من شكل المنحنى ان القيمه اللي بتدي
error هي ال 0 ودم معنام ان قيمة ال weight تزيد بس ازاي نحطها ف صوره رياضيه؟؟

بساطه هنجيب ال gradient اللي هنعتبره ال slope ولو حابب تعرف الفرق it.google. لكن خلينا
نعتبرها الاثنين واحد حاليا ونقول ان ال gradient لمعادله هو عباره عن المشتقه الاولى لمعادله
وليها قوانين كتير اللي حابب يدرسها ممكن يدور وهملاقي مصادر كتير.

طيب هو الرسمه دي قلنا هي عباره عن معادله ال error صح؟؟ واتفقنا انها معادله ف ال weight ف انا محتاج اجيبي ميل المعادله عند قيمة ال weight واشوف هطلع ب كام ف لو كانت ب 0 يعني انا وصلت لل minimum error ولو كانت مش ب 0 ساعتها يعني انا محتاج اعدل قيمة ال weight علشان يقرب اكتر من ال 0.

طيب هعدله علي اي اساس؟؟ لو كانت قيمة ال gradient موجبه يعني انا محتاج اقلل قيمة ال weight علشان اوصل لل minimum error ولو كانت سالبة يعني انا عايز ازود قيمة ال weight علشان اوصل لل minimum error ف تعالوا نشوف مع بعض معادله ال error كانت ايه؟

$$RSS = \frac{1}{m} \sum_{i=0}^m (pred - y_i)^2$$

ولو افترضنا انتا هنشتغل علي variable واحد اللي هو x اللي ف المثال باعنى هنعتبر درجه الحراره مثلًا ونفك معادله ال Error بتعاعتنا ف هتبقي:

$$RSS = \frac{1}{m} \sum_{i=0}^m (w_0 + w_1x - y_i)^2$$

طيب احنا قلنا هنأخذ المشتقه بتاعه المعادله بالنسبة لل weight علشان نشوفها تساوي 0 وللا محتاج يتعدل

واحنا محتاجين ال gradient يعني هنأخذ المشتقه الاولى بس ثانية وحدم انا عندي قيمتين معايا w_0 و w_1 ف انا دلوقتي محتاج اعرف هشتاق المعادله بالنسبة لمين؟؟

هنشتقاها بالنسبة ل w_0 علشان نطلع معدل التغير ف w_0 وهطلع كالتالي

$$\frac{\partial Error}{\partial w_0} = \frac{2}{m} \sum_{i=0}^m (w_0 + w_1x - y_i)$$

وهي نشقاً بالنسبة لـ w_1 علشان نجيب معادل التغيير $\frac{\partial J}{\partial w_1}$ وهنطلع كالتالي

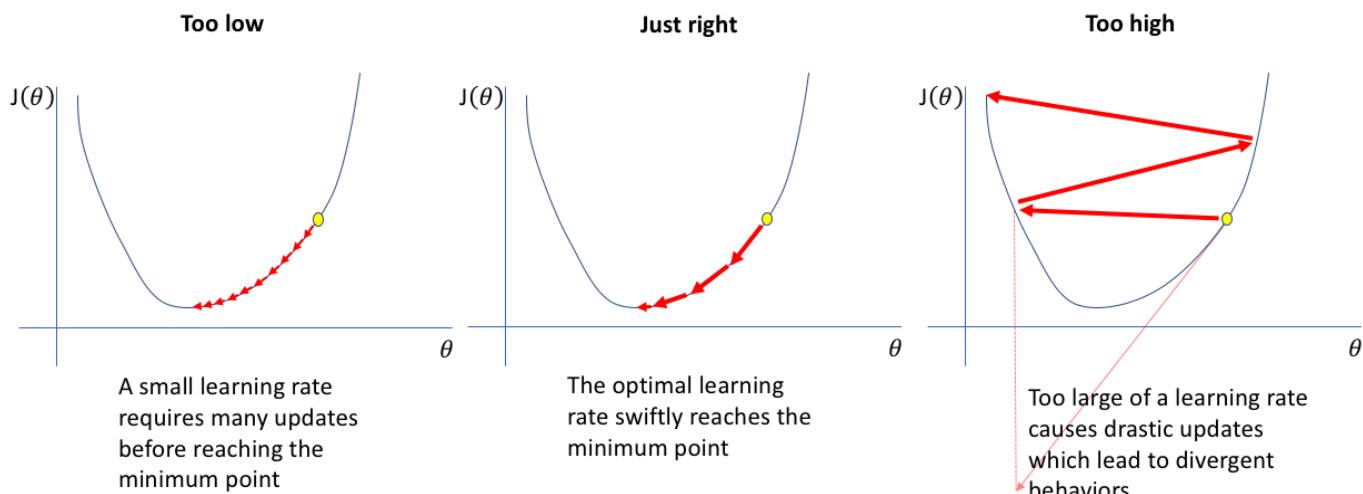
$$\frac{\partial Error}{\partial w_0} = \frac{2}{m} \sum_{i=0}^m (w_0 + w_1 x - y_i) * x$$

طيب كده جبنا معادل التغيير $\frac{\partial J}{\partial w_1}$ لكل weight w_1 ناقص بقى اعرف هعدل ال weight w_1 او w_0 اتساوى

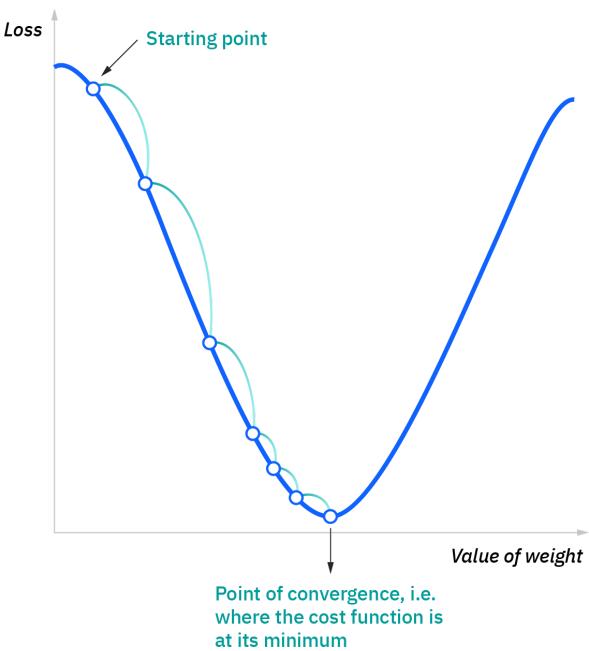
ازاي؟؟ هنمثي على المعادله دي.

$$w_1^{new} = w_1^{old} - \alpha \frac{\partial Error}{\partial w_1} \quad w_0^{new} = w_0^{old} - \alpha \frac{\partial Error}{\partial w_0}$$

لو فكرت فيها هتللاقيها مينطقه لأن كل م بقرب من ال minimum قيمة ال gradient بتقل وبالتالي خطوي هتصغر بس هو ايه ال اللي جنب ال gradient دى؟؟ ده learning rate بنسمهه ال gradient وده بتدكم ايه ف نسبة ال gradient اللي تحتاجها ف كل متغير قيمة ال learning rate وده بتدكم اكابر كل متغير قيمته خطوي بتبقى اصغر ولو خطوي كانت كبيرة ممكن في لحظه اعدل ال minimum واروح النادييه الثانيه ولو خطوي صغير جدا ساعتها هنأخذ وقت كبير علشان نوصل في احنا بنحاول نوصل لقيمة وسط ودي بقى بالتجربه بس في algorithms حل المشكله دي لكن هنعرفها في ال deep learning.



طيب دلوقتي تعالوا نكتب الخطوات اللي ال
gradient descent بيعملها.



1. يفترض قيم عشوائية لل weights.
2. يطلع predictions ويحسب ال error بتاعها.
3. يحسب ال gradient لكل weight.
4. يعمل update لكل weight على حسب قيمة ال gradient بتاعته.
5. يعيد الخطوات من 2 ل 4 لحد موصى لل minimum error.

بس هنا في سؤال مهم وهو هل هنوصل فعلاً لل ??minimum error

الاجابه هي لا احنا مش هنوصل لل minimum لكن هنقرب منه جداً ودم لان كل مونقرب لل خطوتنا بتبقى اصغر لحد منوصل لمدخله ان الخطوه بقت صغيره جداً والتغير في ال error يكون ضعيف جداً يكاد يكون غير مؤثر ساعتها بوقف وبنقول ان انا كده وصلت ل minimum error فربه جداً من ال value

يعني لو افضل درجه حراره هي 67 صاحبك مش هيزعل يعني لو كانت 66.87253 درجه.

وبكله يبقى ال gradient descent قدر يحل المعادله بطريقته بس في سؤال مهم هنا. هو ال gradient descent وهو يحسب ال error على نقطه وبدم ولد على الداتا كلها؟؟ وهذا اقولك ان عندي 3 ممكن استخدامه strategies.

Stochastic gradient descent

وبدم بيأخذ ال data سطر سطر (نقطه نقطه) ويحسب ال error ويحسب ال weights update على كل نقطه يعني عليها وبدم مجزته انه سريع جداً ف ال updates لكن بيأخذ عدد مرات من ال updates جداً وبيتأثر بال outliers.

mini-batch gradient descent

وقد ييأخذ عدد صغير من النقاطه ف كل مرر وييأخذ متوسط ال error على النقطه دي وبناء عليه يعمل update لـ weights وبكله يكون قلل من تأثير ال outliers لأنها بقت نقطه من وسط مجموعه وهي تعتبر ميزته لكن ييأخذ update فيها وقت اكبر لكن ييعمل عدد مرات من ال updates اقل.

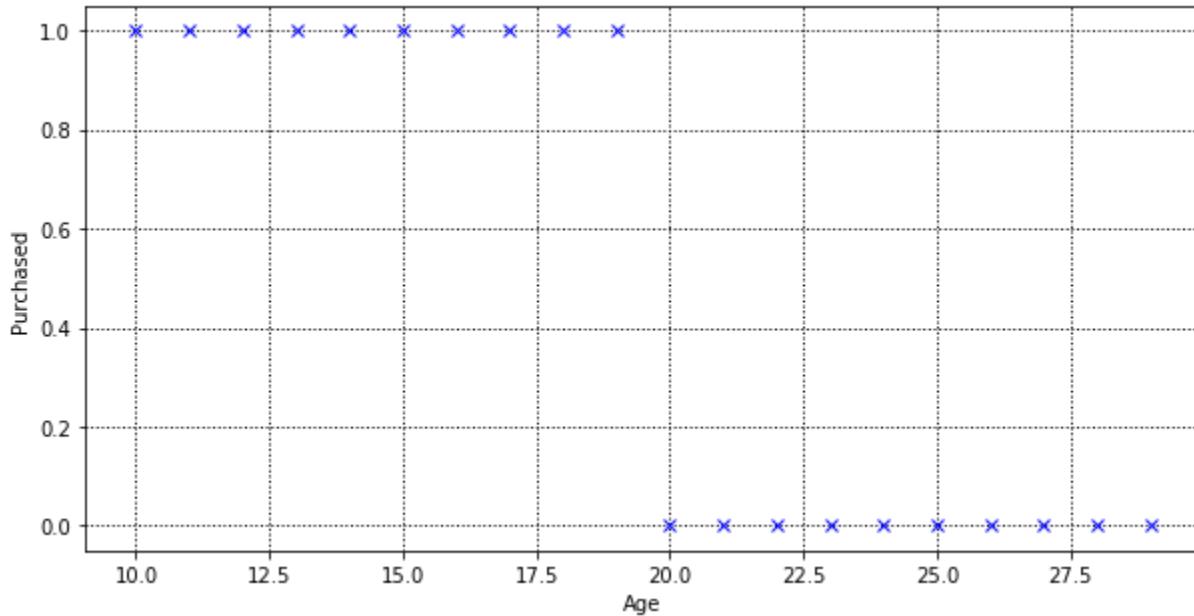
batch gradient descent

وقد ييأخذ ال data كلها ف كل مرر ويحسب متوسط ال error على ال data كلها وبكله ف يكون خطوه ثابتة ورایحه ف اتجاه ال minimum مباشر لكن اكيد ييأخذ update فيها وقت اكبر.

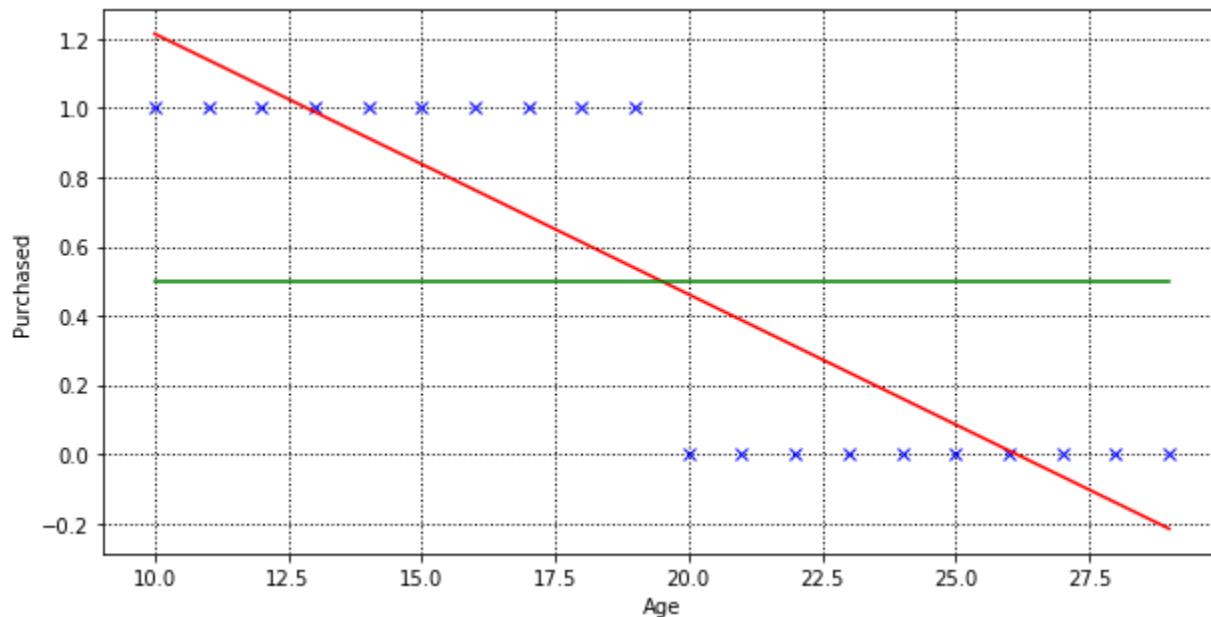
طيب ايه الافضل؟؟ بتعتمد على حجم ال data وعدد ال outliers اللي فيها.

Classification

هل ينفع نستخدم ال linear regression tasks لـ classification tasks؟ فلنفرض ان احنا عندنا داتا perfectly balanced customer data. الداتا دى فيها 20 customer و ال label اشتري او لا لا. عندنا 10 customers بين ال 10 الى 19 سنة ودول اشترووا بالفعل و 10 customers بين ال 20 الى 29 ودول مشتروش. و ال labels بتاعتنا 0 و 1 (0 يعني مشترash و 1 يعني اشتري).

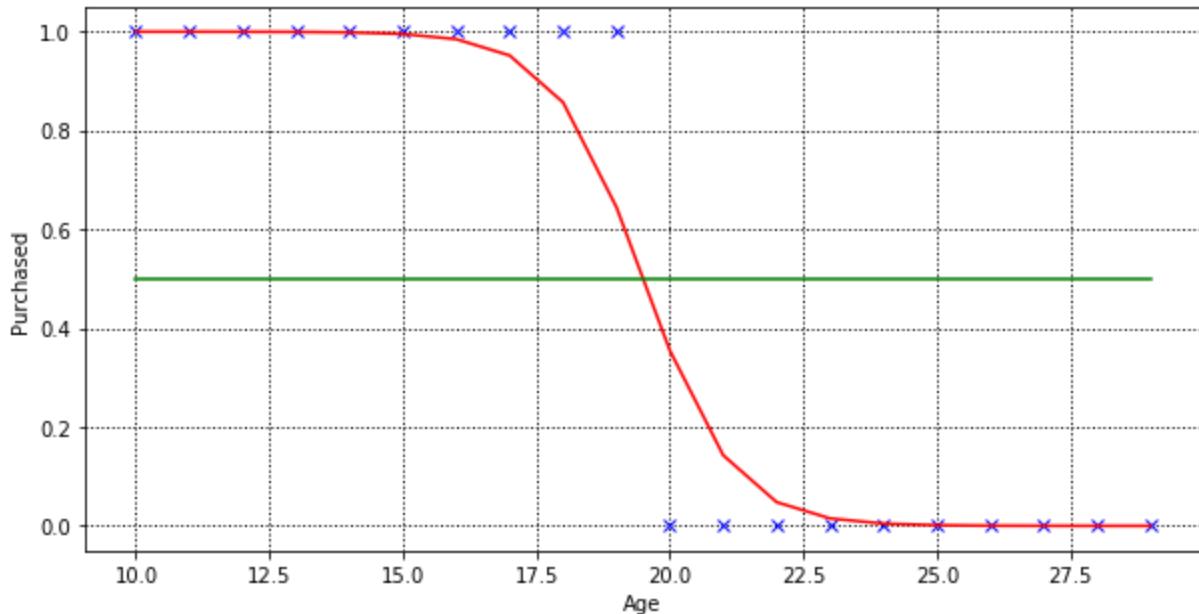


لو جينا عملنا linear regression هنلوقى ان ال best fit line هو اللونه احمر دة.



بنويقلى ان انت دلوقتى اخذت بالك انه مش احسن حاجة. و اول مشكلة عندنا ان ال predicted classification values بتكون continuous واحدنا ال labels بتاعتنا فى ال linear regression output بتكون discrete او بالنسبة للمثال بتاعنا binary. فا احنا منتظرين ان يطلع لنا output فى ال range [0, 1] و نحط threshold معين وليكن 0.5 و نقول لو ال output اكتر من 0.5 يبقى 1 و لو اصغر يبقى 0. لكن ال linear regression ممكن ال output بتاعها يطلع بره ال range [0, 1].

طيب لو كان معانا مثل مش line بقى لا لو معانا مثل curve S زى دة كدة هيكون احسن بكثير.



وهنا نروح لل logistic regression

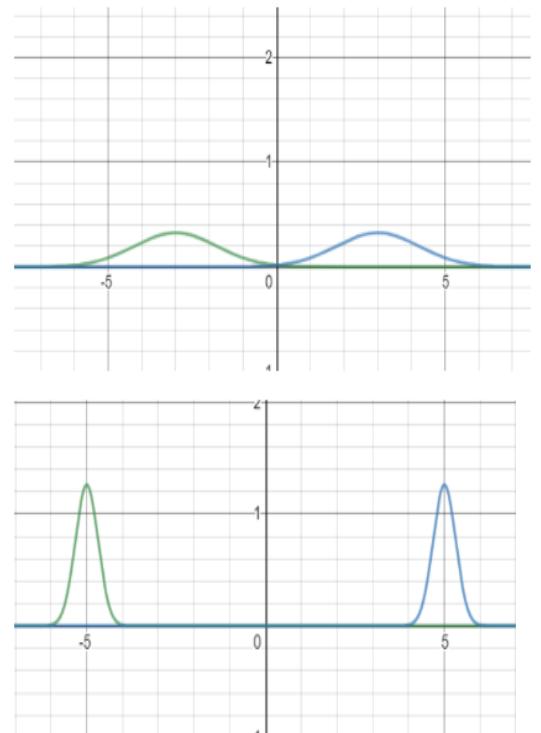
Logistic Regression

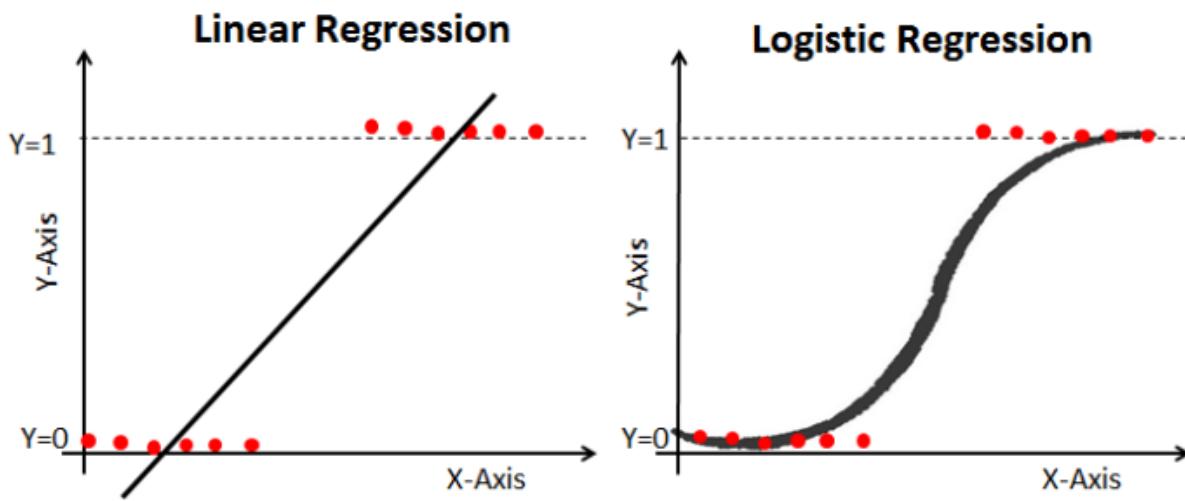
زي مقلنا ان ال label بتاع ال classification يكون binary و في المثال بتاعنا بتكون يعني حاجة من الاثنين يا 0 يا 1. فانت بيبقى عندك احتمال ول يكن

P ان ال output يطلع 1 و احتمال 1 ناقص ال P ان ال output يطلع 0. و على حسب تكرس الداتا نفسها وهل هي بطيء ولا لا يعني على حسب عدد العينات اللي هو يتدرّب عليها فلو افترضنا ان ال Two Classes قد بعض بالظبط بيقى احتمالية ان ال output يبقى حاجة من ال 2 يعتبر مائلة جداً ناحية ال 50 - 50 يعني زي بعض بالظبط. تعالى نبص على الرسمة دي. خلينا نفترض كده ان اليمين اللي هو ال Blue Bell Curve عباره عن احتمالية ان ال output يكون 1 اما ال Green Bell Curve هو احتمالية ان ال output يكون 0 كل ما يكون ال 2 normal distributions بعد عن بعض كل ما تكون معنها انك واثق في اجابتك اكتر وكل ما يكون ال Standard deviation قليل كل ما يكون معنها انك

متأكد اكتر من كلامك وبالباقي هيقي مثل شكل ال Bell Curve زي المنظر اللي قدامنا تحت هنا.

وزي ما شوفنا من شوية ان ال fitting S shaped curve بيعمل better من ال linear.





فأ ال S دة بزمله بال logistic function او بال sigmoid function و هى تعتبر شكل من اشكال الدوال الهندسية. والمعادلة بتاعتها بسيطة جدا بالشكل دة:

$$p = \frac{1}{1 + e^{-y}}$$

يعني لو مثلا معاك خط وعايز تخليه زي حرف ال S فانت يعتبر هتاخدم تعامله شكل من اشكال ال S مثل . يعني Transformation

معادلة Linear
 كانت كده Regression

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

بساطة.

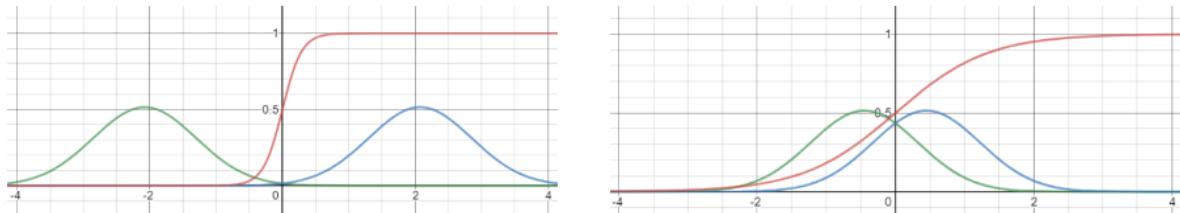
هناخد احنا المعادلة دي كلها نعوض فيها مكان ال y في .Sigmoid Function

$$p = \frac{1}{1 + e^{-y}} \implies p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

طيب ماشي كده عرفنا ازاي هيتحول طب مانا بردم كده لسه عندي weights او لسه عندي normal distributions او اكيد .. طب انا عايز اعرف features اللي فوق يعني weights

دي تفرق في ايه مع رسمنا Sigmoid اللي لما طبقناها على ال linear regression والسمينها Logistic Regression

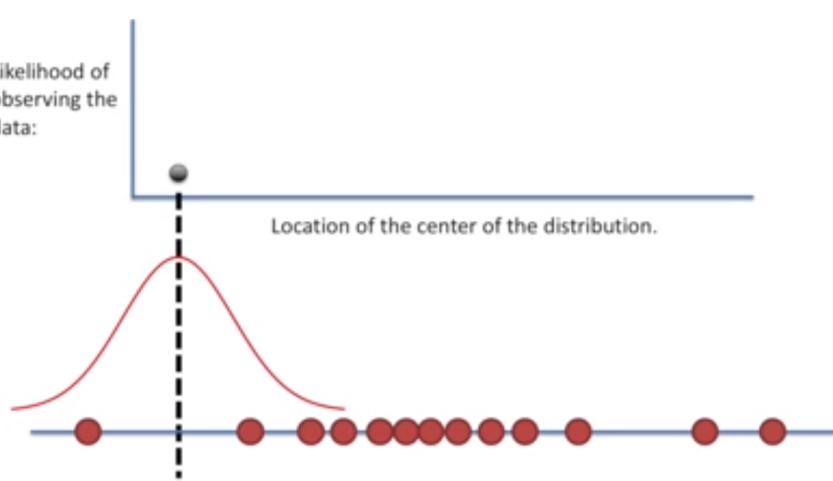
خلينا نبص على الفرق ما بين الحاجتين مع العلم انك ممكن تطبق كل كلامنا ده فقط على ال 2 Binary classification Classes .



في الرسمتين دول بيبينو لنا كل ما هيكون شكل كبير يعني ان ال 0 و 1 متقاربين من بعض كل ما هيكون ال 5 مش بتطابع Sharp classification لا ده بيقي في جزء هايل كبير جدا في النص يطابع قيم مختلفة ما بين ال 0 و ال 1 اما لو ال two normal distributions بعد عن بعض بيقي ال Logistic Regression

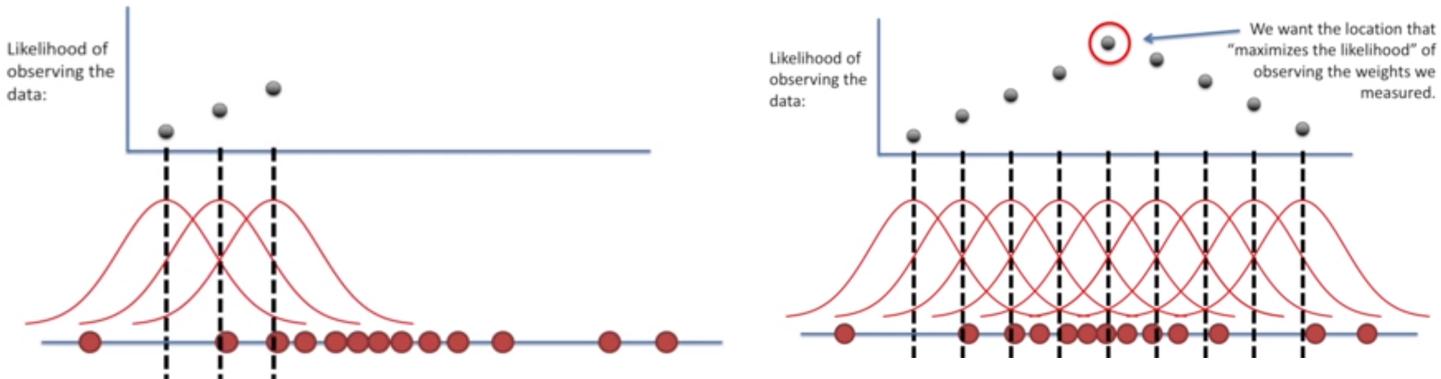
Maximum Likelihood Estimation

من النقط الفارقة جدا ما بين ال Linear Regression وال Logistic Regression ان في ال OLS methods كان هدفنا نقل الفرق ما بين ال best fit line وال نقط عن طريق ان احنا نحاول نصغر ال Sum of Squared Error لاقل قدر ممكن ما بين الخط والنقط أما هنا في ال Maximum Likelihood Estimation فاحنا ماشيين بمبدأ انت نحاول نكبر ال MLE وهو



likelihood Estimation
طيب خلينا نقول ان احنا عرفنا ان ال two normal هتعتمد على ال sigmoid distributions اللي معانا فدلوقتي بقى انت عندك احتمالات كثيرة على حسب شكل الداتا في ان يكون ضيقة أو واسعة Normal distribution طويلة أو قصيرة فعشان ترسم صح بنبدأ نستخدم Likelihood methods زي اللي لو مثلا تخيلت ان ال نقط معاك اهي لونها احمر

ومعاك في ايدك bell curve وبدات تدركه على الداتا كلها واكنك بتدور على best place خد بالك ان انت عندك احتمالات كثيرة لاي مكان ممكن يبقى فيه بس المكان اللي هو فيه مش ييشوف غير اللي تحته من النقط فقط مالوش علاقة بجوانبه فلو شوفنا لما حطيناه في أول موقع likelihood هيللاقى تحت الكيرف نقطتين ولا 3 فالقيمة هتبقى قليلة تعالى نشوف كل ما هندركه بعين هتبقى القيمة فين.



هيللاقى كل ما يكون تحت ال curve عدد اكتر من النقط ويكون شبه equal من الناحيتين ساعتها هتللاقى ال likelihood عاليه وبختار maximum for best position وبناء على ال normal distributions ال 2 اللي معانا هييقى best fit Logistic Regression وهو ده الهدف المطلوب.

Log loss

ID	Actual	Predicted probabilities
ID6	1	0.94
ID1	1	0.90
ID7	1	0.78
ID8	0	0.56
ID2	0	0.51
ID3	1	0.47
ID4	1	0.32
ID5	0	0.10

ال log loss بيقارن بين ال actual g predicted probabilities بعد كدة بتحسب ال score ال هو المفترض 0 او 1. وبينها وبين القيمة المتوقعة يعني بناء على مدى قرب او بعد ال probabilities دى بال actual. خلينا نبص على المثال دة. احنا عندنا 2 columns ال actual value ودمة ال probabilities الحقيقية و ال output predicted probabilities ودمة ال model يقول ان ال object دة احتمالية انه ب 1 قد ايه. بعد كدة هنضيف column جديد و هو ال corrected probabilities. طبعا هتقلى ايه ال corrected probabilities دى هو المودل مطلع

ID	Actual	Predicted probabilities	Corrected Probabilities
ID6	1	0.94	0.94
ID1	1	0.90	0.90
ID7	1	0.78	0.78
ID8	0	0.56	0.44
ID2	0	0.51	0.49
ID3	1	0.47	0.47
ID4	1	0.32	0.32
ID5	0	0.10	0.90

ان دی probabilities او احتماليه ان يكون ال output ب 1
ان دی probabilities او احتماليه ان يكون ال output ب 1
فكل المنهعله ان ادنا هنخلى ال objects الى ال class
بناعها ب 1 عادي زى مهم و الى ال class بناعها ب 0
هت تكون ال probability بناعاتها بتتساوى (1 ناقص ال
corrected probability زى مثلًا ID8)
بناعتها 0.44 = 0.56-1 و هكذا.

وحلوقتى هندسب ال \log لكل .corrected probability
والسبب ان ادنا استخدمنا ال \log هو انه هيفرض نوع من
العقوبة ب value قليلة لو كانت المسافة مبين ال
penalty كـ \log له كانت المسافة كـ \log .

ID	Actual	Predicted probabilities	Corrected Probabilities	Log
ID6	1	0.94	0.94	-0.0268721464
ID1	1	0.90	0.90	-0.0457574906
ID7	1	0.78	0.78	-0.1079053973
ID8	0	0.56	0.44	-0.3565473235
ID2	0	0.51	0.49	-0.30980392
ID3	1	0.47	0.47	-0.3279021421
ID4	1	0.32	0.32	-0.4948500217
ID5	0	0.10	0.90	-0.0457574906

و دلوقتى بعد ما حسبنا ال \log و طبعاً علشان كل ال corrected probabilities بين ال 0 و ال 1. كل ال results بتنعمة ال \log طلعت ب علشان نعادل negative value القيم ال negative دى هنأخذ values لـ negative average.

$$- \frac{1}{N} \sum_{i=1}^N (\log(p_i))$$

فى الآخر الناتج بساع ال \log هيكون 0.214 وده دة ال negative average loss للمثال بتاعنا. و دلوقتى بعد ما شرحنا ال \log loss بالتفصيل نقدر نقول ان ال \log loss نقدر نعرض عنه بالقانون دم.

$$\text{Log loss} = \frac{1}{N} \sum_{i=1}^N - (y_i * \log(p_i) + (1-y_i) * \log(1-p_i))$$

و ال y_i هنا هى ال class 0 probability بتاعة 1 و ال $(1-y_i)$ هى ال class 1 probability و دة ال \log loss يختلف ولو ال actual ب 1 هتلد قى ان نص القانون الاول ال هو دة الميتشغل والنص الثاني هيتصفر.

ولو ال actual ب 0 هتلد قى ان النص الثاني من القانون ال هو دة $(1-y_i) * \log(1-p_i)$ الميتشغل و الاول هيتصفر.

Confusion Matrix

لو عندنا مثلاً 2 classes ال 0 و ال 1. ال columns يعبر عن ال prediction و ال rows يعبر عن ال actual.

		Predicted	
		0	1
Actual	0	30	12
	1	8	56

و g positive ← ال 1 . خلينا نقول على ال 0 ← .negative على ال 0 ← .negative ← 0 . دلوقتى عايزين نشوف امتحان قلت ان ال negative هيبي output negative هنلاقى 30 مرة انا قلت negative و طلعت صحيحة بنسمهه .True Negative طيب امتحان قلت negative و طلعت .False Positive

الكلام على ال positive كام مرة انا قلت positive و طلعت صحيحة هنلاقى انها 56 و دة هو ال .Positive هنلاقى امتحان قلت 12 مرة ان ال output هيطلع Positive و طلعت .False Negative ونفس طيب امتحان قلت 8 مرات negative و طلعت .

طبعاً لو وصلنا ان كل ال False تكون بـ Zero يبقى perfect. طبعاً ده بتقدير ان احنا عندنا العملية over fitting عشان ال Training و ال Testing .

في حاجة كتير بنقدر نحسبها من Confusion matrix وكل parameter منهم بي حاجة مختلفة. زي ال precision و ال Recall وغيرهم. في اللي يركز على ال Positive Predictions بس ويشوف انت جبب كام صحيحة من اصل كام وفي اللي بيتص على ال Actual Positive وفي عوامل كتيرة جداً كل واحد منهم يظهر لك حاجات ممكن تختفي في ال Accuracy metric في ال

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

ملاjkة الماتريكس كنا معروفة يعني الـ

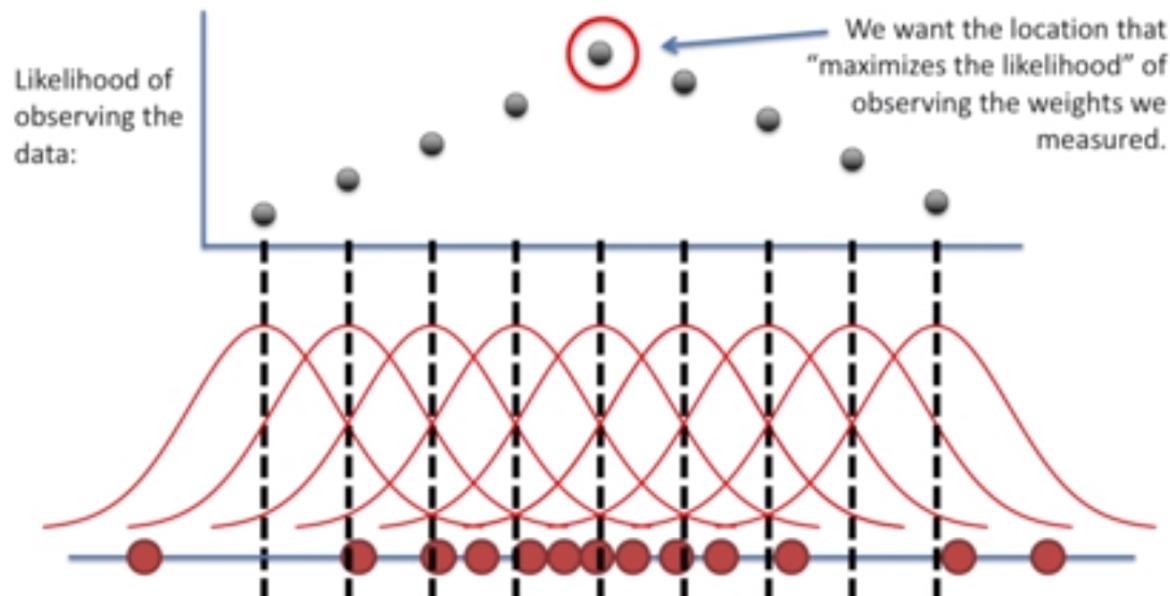
والـ prediction الـ g columns والـ actual

		Actual Result		
		Positive	Negative	
Predicted Result	Positive	1	0	
	Negative	1	998	rows

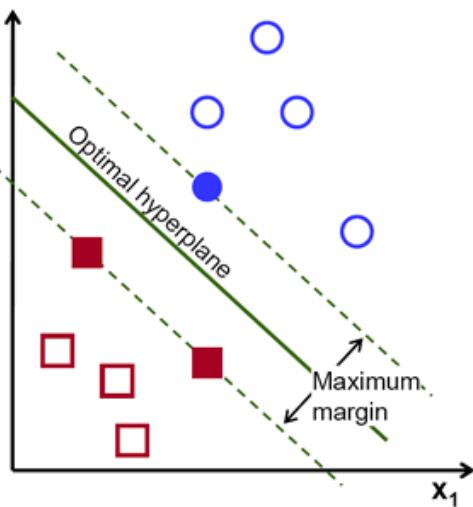
Accuracy = 99.9 %

Precision= 100%

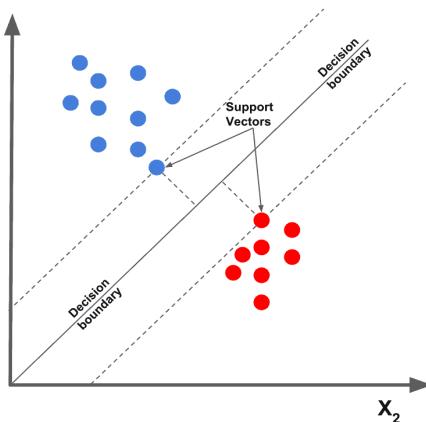
Sensitivity (Recall) = 50%



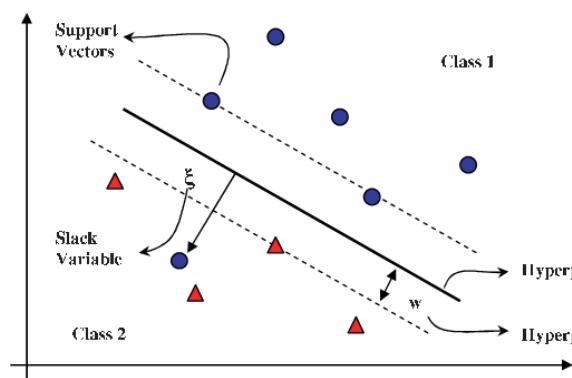
Support Vector Machine



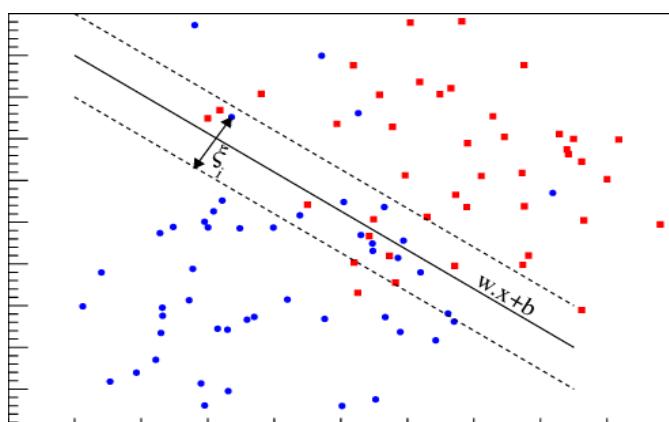
تعالوا دلوقتي نشوف Classifiers تانين ينفع يفيدونا في challenges مختلفة بنقابها زي مثلا لو انا معايا الداتا بالمنظار ده تخيل ان في 2 معانا و هما features .Artist و Computer scientist خلاهم هل الشخص ده فقدمك الداتا بشكل كبير لو قلتلك افضل بعينك مايبي لهم اول حاجة هتعملها هي انك هترسم خط يفصل مايبي لهم و ده ببساطة فكرة .سيحط ال Boundary زي الخط ده وبيبقى Support vector machine



أقل ب dimension من عدد ال features يعني لو عندي 2 features boundary 1D يعني خط ولو بقى 3 Features يبقى ال عبارة عن Plan boundary كثيرة زي انه مثلاً يحيط Boundary ده في مسافة متوسطة ما بين ال 2 Support vectors بقى الي هما اقرب نقطتين بس في سؤال مهم جدا هنا لازم نسأله وهو ان انت تحتاج تعرف ازاي اختار افضل خط او افضل boundary وهل لازم بقى فعلاً ملائم ب انه يكون linear؟ في الحقيقة لا مش لازم بقى linear بس لل Over Fitting فوتقلل فكرة انك تودي نفسك ناحية simplicity.



ماشي طب لو الداتا مش متوزعة بالتساوي يعني لقيت نقط جاية بالشكل ده ؟ يعني لو ركزت هتلaci الخطي اللي بيفصل الازرق عن الاحمر في نقطة زرقاء هناك جنب الاحمر لو انا حاولت اخلي ال موديل over fit يرسم خط معوج كده ويحاول يصلها كده انا بقىت لان ممكن تكون هي اصلاً outlier وبالتالي مكنش اذكي حل اني اروح وراها لا بالعكس ممكن الافضل انه يجعلها غلط . و لما بنسمح ان يكون في misclassification المسافة بين ال decision boundary g و support vector بنسبيتها soft margin.



طيب لو ال data بقى بتاعتنا بالشكل دة. هتللاقى ان مفيش خط linear يقدر يفصل الداتا هنا بيجلنا حاجة اسمها kernel trick هى عبارة عن kernal function و non linear problem الحقيقة اللى بتتحولها ل linear problem طب ازاي؟ خلينى اقلك انها بتحول كل نقطة على ال space two classes على ال space الجديد طيب ازاي برمد. خلينا نشوف ال القدمنا دول وهنعتبر ان اول class اسمه red و الثاني black. خلينا نزود dimension تالت و يكون بساوى $z = x^2 + y^2$

بسخدام الثلاثة dimensions نقدر دلوتى نرسم hyperplane يفصل ال 2 classes.

و من اكتر ال kernels المشهورة هم ال polynomial kernels .Radial Basis Function (RBF) kernel ال formula هو ال default kernel و دى ال RBF بقى بتاعته.

$$K(x, x') = e^{-\gamma ||x - x'||^2}$$

ال gamma انت البتعدددها و لازم تكون اكتر من ال zero. و ال sk-learn لل value

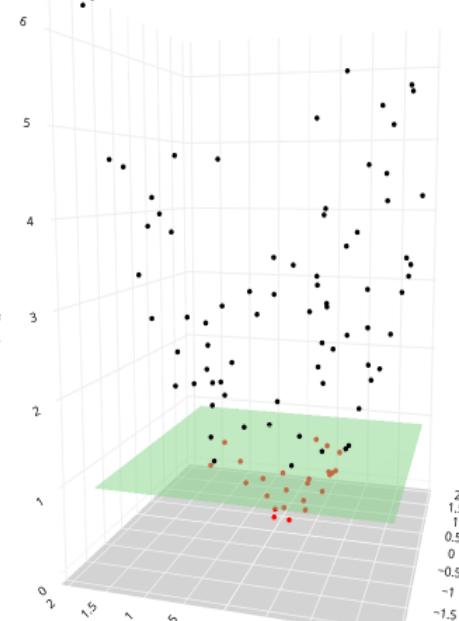
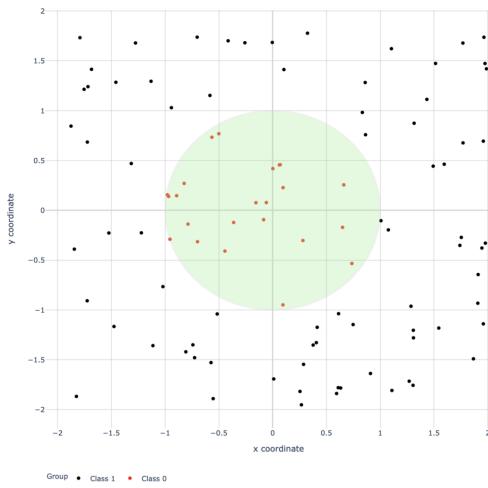
$$\gamma = \frac{1}{n \text{ features} * \sigma^2}$$

و كل ما ال gamma تكبر كل ما يكون النقط القريبة من بعض بشير اكتر على ال model.

اما بقى ال formula فال polynomial kernel بتكون بالشكل دة.

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

و ال d بتعبّر عن ال degree بتاعة ال polynomial.



طبيقات ال Support vector machine كثيرة جدا وتنطبق على عدد ضخم من Features وعدد

ضخم بخدمه من labels مثل بس كده ده كمان

تقدير تستغل مع Regression بشكل كوييس

فده بيخليله واحد من أشهر ال Algorithms في

مجال ال machine learning بس مع ال

Regression هو بيشتغل بطريقة مختلفة

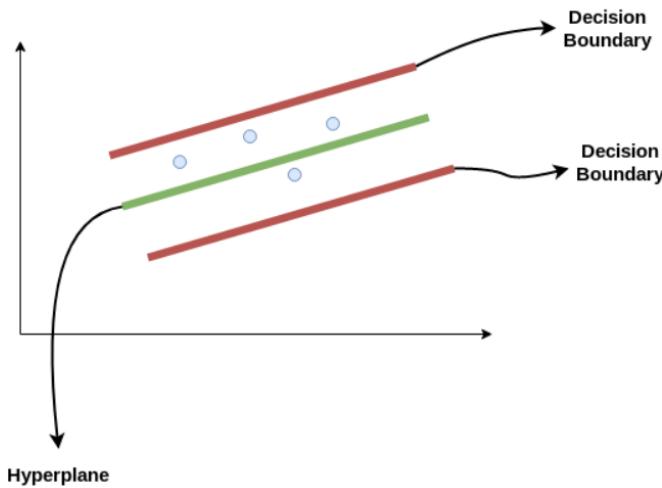
شوية هو بيعمل بخدمه decision g hyperplane

بس يكون الشكل اشبه بال boundary

كدة ال hyperplane دم هو ال

margin هي ال decision boundary

المقبولة لل errors واي data point برهة ال boundary هي البنسبتها ال errors.



K Nearest Neighbours

فكرة KNN بسيطة جدا وهي انه يبقى في مرحلة inference على طول تقدر تقول بيتوافق معك

طول الوقت يعني الطبيعي ان supervised models كانت بتقعد في مرحلة training وبعد كده

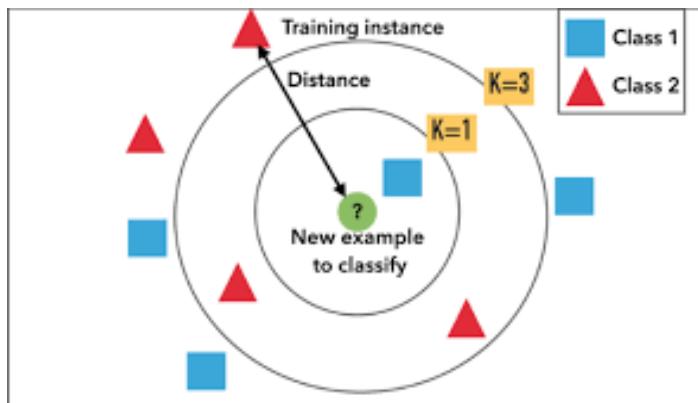
عشان تجوز لمرحلة Prediction دم بقى مختلف شوية انت يعتبر هتعمله Test وبعد كده

هت من غير Training تعالى نشوف ازاي

ال method دي من اسمها بتدور على Nearest Neighbours على حسب عدد K ولازم يكون عدد

فردي يعني لو قلنا أسط حاجة $K=1$ ساعتها انت اصلاً معاك Labeled Data وبتدور على الداتا

بتاعت test بتبدأ تشواف أقرب نقطة ليها ايه وتأخذ نفس ال label بتاعها.



يعني احنا في Two Classes هنا ياما ازرق

ياما احمر دول هما Training Data طيب ايه

Test Data دي ؟ دم ال Question mark

اللي انت عايز تشواف هي المفروض تبقى ايه

هل أحمر ولا أزرق فعل حسب ال K قيمتها كام

لو K فيمتها 1 يبقى هنبع على أقرب نقطة هنلاقيها زرقا يعني الاجابة أزرق يعني Class 1 طيب لو K قيمتها 3 ساعتها بقى في جوا الدايره اللي بتتشيل اقرب 3 نقط دي 2 حمر وواحدة زرقا يعني دلوقتي الاجابة Class 2 طيب هنعرف مين انهي اللي صح ؟ وهل انا ببقى معايا عينة واحدة بس بدور على اجابتها ؟

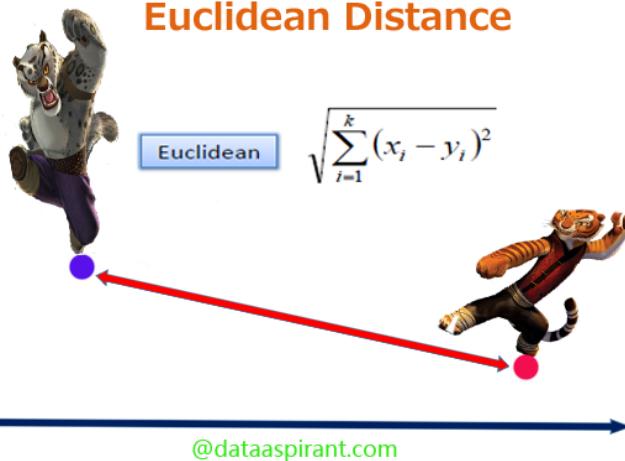
لا انا ببقى معايا عدد كبير وهو 30 % من الداتا فرضا يعني 300 عينة بدور على label بتاعهم وبما انها test set فانا اصلا معايا ال Actual Label يعني معايا Reference وانت محكم ب K واحدة لكل الداتا فبتلشوف انهي K تحلي كل ال Predictions بتاعت ال test set تبقى زي Actual Label يعني مثل ممكن لو حطيت K=1 يعني في 50 واحدة من ال 300 اجابتهم صح ولو K=3 لقيت 150 بس و K=7 لقيت 250 يعني ساعتها هنختار واحدة سليمة ولو جرت K=5 لقيت 150 بس

K = 7

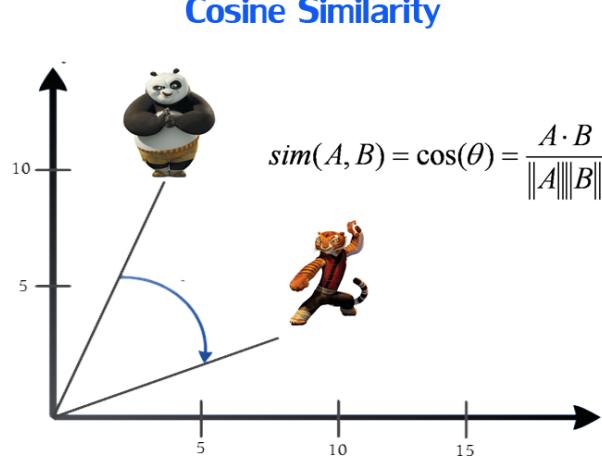
Similarity Distance Measures

في طرق مختلفة انك تقيس المسافة بين النقط وبعضاها منها هتلaci درست منه في الكلية وفي انواع تانية ممكن تكون اول مرة تشوافها بس كلها بسيطة جدا يعني زي

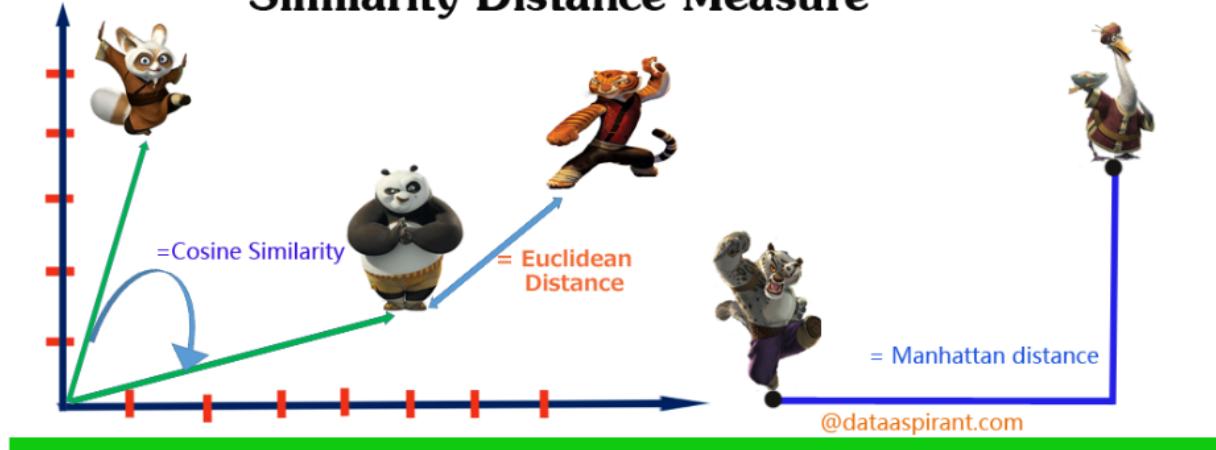
Euclidean Distance



Cosine Similarity



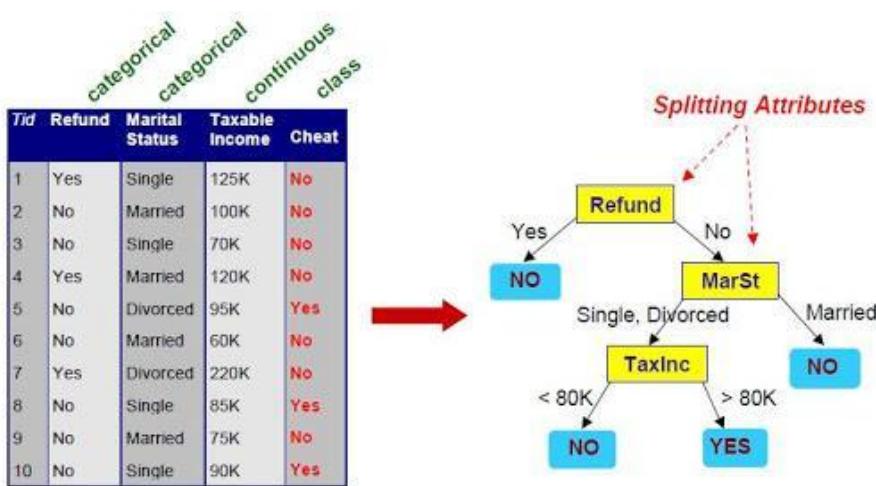
Similarity Distance Measure



لو شوفتهم هتلدقى ان ال Euclidean هو أقصى خط بين نقطتين ال Manhattan ييلزم بالحركات الرأسية والأفقية وال some sort of Cosine هو عبارة عن الزاوية بين ال Two Points لو خليناهم مربوطة بنقطة ال reference vectors .

1- Decision Tree Classifier

ال Model هنا المرة دي عبارة عن "شجرة بسيطة" بس يعتبر مقلوبة يعني الجذر بتاعها فوق والأوراق بتاعتتها تحت هدف ال Tree دي أنها تحاول تسأل أسئلة طبعاً بناءً عن الـ Features اللي معانا بحيث أنها تقسم المشكلة وتحلها في الآخر يعني في نهاية كل الأسئلة في الآخر هيئي في Leaf Node اللي هي تعتبر Classification يعني خلينا نشوف مثال نحاول نفهم منه بشكل أوضح



المثال ده بيوضح بالنسبة لنا ان كل Feature بقى مشابه لسؤال ولو بدأنا من فوق لحت كده هتحس انه يتوصف السطر بتاع الداتا يعتبر.. مثل أول سطر مكتوب فيه

Refund yes , Marital Status : Single , Taxable Income :125 K , Cheat : No

فانت دلوقتي المفترض لو مشيت

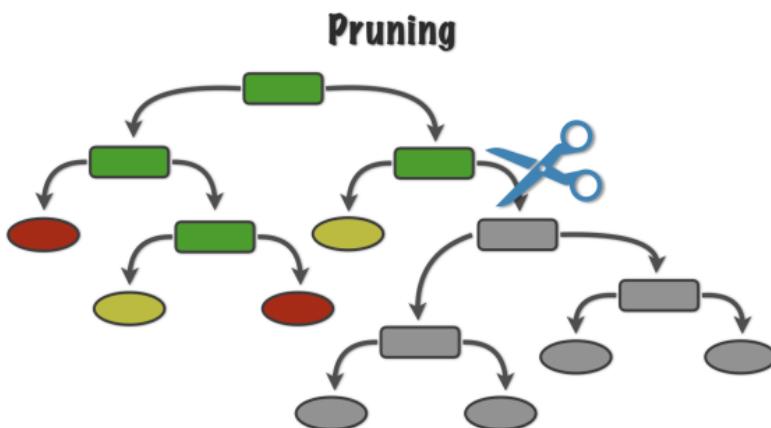
مع " Decision Tree " هتلaci نفسك بتنقل بين ال nodes من فوق لتحت هتلaci ان الاجابة No طيب الكلام حلو أوي وبسيط يعني انت بترسم DT بناءا على الجدول بناءا على Data نفسه تفتكر ايه اول مشكلة ممكن تقابلك هنا ؟

ال Continuous Column بنعرف نتعامل معها عادي جدا بس بنحط Ranges ولكن المشكلة الحقيقة في Decision Tree هي Over fitting انك ببساطة بتبقى ماشي على الداتا اللي موجودة

بالملي يعني لو في outliers او اي حاجة موجودة مش دقيقة انت هتمشي وراها

وبالتالي الموديل بتاعك يعتبر High Variance

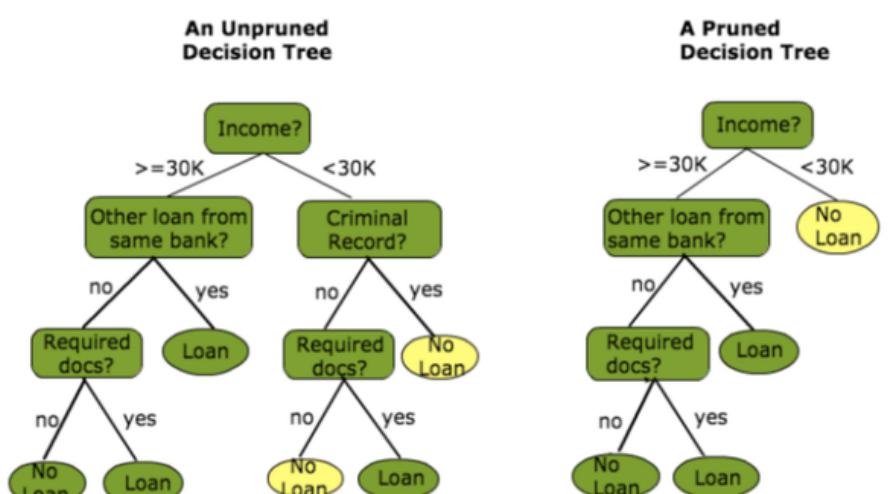
ومن هنا بيظهر بقى اهمية Decision Tree Parameters اللي معنام بسيط جدا انه Pruning شوية فروع من الشجرة .



- طب ليه ؟ هتفرق في ايه ؟

لو كانت الفروع دي موجودة كانت هتسأل يعتبر سؤال تاني او اتنين وا ثلاثة عشان توصل لشاتح معين اما لو قصتهم فانت اعتبرت ان كل اللي في ال Category دي اجابته رقم واحد ومش يحتاج اسأل اسئلة ملهاش لازمة طيب تعالى نشوف مثال بالارقام

ده مثال ل Decision Tree عشان
تعمل Loan Prediction يعني شنباً
هل الشخص ده لو اديته القرض
هيسددهم ولا لا ؟
فلو بصينا على الفرق بين اليدين
والشمال هتلaci ان اليدين حصل



لكل اللي الـ Pruning income بتعده اصغر من 30 ألف وخلينا الناتج انه ميأخذش قرض.
طب هيجي في دماغك سؤال معين وهو ان كان في مثل حالة او اتنين في Training Data set
كانوا اقل من 30 الف ومعندهوش Required Docs Criminal Records وكان معاهem
اختيار كوييس انهم ياخدوا القرض .

هقولك ان كونك تعمل Leaves Branches مخصوص عشان عدد مبدد من samples دم معناها انك over fit يعني هطلع اعلى training accuracy على الاغلب هتشغل في testing لانك عقدت الشجرة جداً ومش بعيد يكون الرجل دم اللي انت عملت الـ branch عشانه كان مجرد outlier وحاجة مش هتكرر.

طيب هل في طرق معين بنعرف من خلاله نقص انهي فرع بالظبط ولا بنقعد نعد العينات ؟
في طرق كتير منها عن طريق Cross Validation Techniques اللي هنتكلم عنها باستفاضة شوية
قدم بس تقدر تقول انتا نحط احتمالات كتيرة ونشوف الاقل Error فيه
وناخده يعني من ادد الاحتمالات مثل انك تقص قد ايه من Tree وعلى اساس Testing Score اختيار .Best Pruning Level

طيب سؤال تاني ! ارتب ازاي الشجرة ؟ يعني انا دلوقتي عايز احط الاسئلة دي اعرف مين مين
الاهم ومين الاقل اهمية ؟ يعني انهي سؤال يبقى فوق ؟
بنجاوب عالسؤال دم بان في طرق كتيرة ندي Score لانهي سؤال احسن ومن أشهرهم Gini Score

عندك 30 طالب و Label انك ت predict مين بيلاعب Cricket فانت عندك اسئلة كتير زي النوع هل
هو ذكر ولا أنثى وهل هو مثل في فصل X او في فصل IX تفكرا انهي احسن ؟

Split on Gender

Students = 30
Play Cricket = 15 (50%)



Female



Students = 10
Play Cricket = 2 (20%)

Male



Students = 20
Play Cricket = 13 (65%)

Split on Class



Class IX



Students = 14
Play Cricket = 6 (43%)

Class X



Students = 16
Play Cricket = 9 (56%)

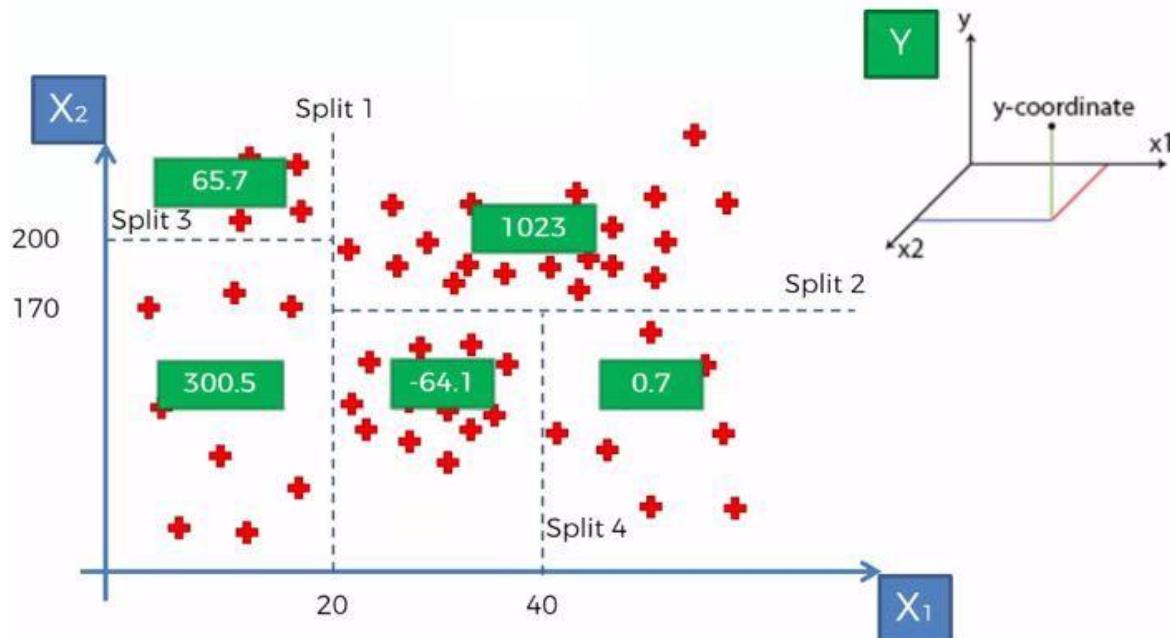
عشان تجاوب صح فانت محتاج تعرف ايه هو الهدف ؟ يعني هل الهدف انك تفصل الناس نصين
كعدد يعني سؤال يخلي في 15 هنا و 15 هنا كطلبة ؟
ولا انك تبص على ال label ? ال Label بيقولك ان 50 % من الناس بتلعب cricket يعني 15 فرد
منهم دم من قبل ما تسأل فانت لو معاك سؤال واحد فانت هتحاول تجيب أكثر سؤال يبقى هو حل
للمشكلة يعني

1- بص على الرسمة تاني فوق عاليمين Split on Class الناتج بتاعها ايه ؟ 9 من الفصل دم و 6 من
الفصل دم تفتكـر لو معنديكش سؤال غير دم هل كده نقدر نقول ان السؤال دم حللنا القضية
وممكن لو عايز اعرف مين بيلاعب cricket بيقى اشوـفه من Class X ولا لا ؟ هل دم صح ؟ ولا
هيقى كلـم بعيد اوـي عن الصح؟ تمام طيب تعالى نـشوف التـاني
2- بـص على الرسمـة اللي على الشـمال Split on Gender النـاتج بتـاعها ؟ 13 من ال males و 2 من
females طـيب كـده معـناها اـيه ؟ معـناها ان ال 15 مـفيـش منهـم غيرـبنـتـين بـس طـيب يعني مـمـكـن
لو بـرسم شـجـرة اـجي اـسـال male or female ولو طـلـع اـقوـم عـاـمل اـيه ؟ أـقول خـلاـص بـيـقـى
اـكـيد مـبـتـاعـبـش و اـعـمـل abstraction حتى لـلـ2 دـوـل واـكـنـهـم outliers فـبـالـتـالـي السـؤـال كانـ مـفـيدـ
وـقـصـرـ معـانـاـ كـثـيرـ.

سؤال مهم هل مـمـكـن اـعـمـل replacement للـشـجـرة كلـها بـسـؤـال male or female ؟ بما ان
كـده يـعـتـبر مـبـيـاعـبـوش ؟ females
الـإـجـابـة لا ! لأنـ مشـ كلـ males بـيـلـعـبـوا دـمـ 13 منـ أـصـلـ 20 فـاـكـيدـ فيـ عـوـاـمـلـ تـانـيـةـ هـتـحـتـاجـ
تسـأـلـهاـ !

Decision Tree For Regression

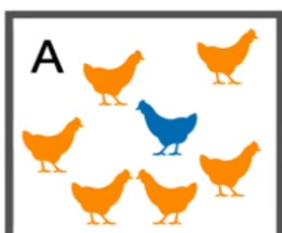
تفـتكـرـ مشـاكـلـ Regression مـمـكـنـ اـزـايـ نـحلـهاـ بـ ؟ Decision Tree هيـ مـبـدـأـهاـ بـسيـطـ هيـ بـتـبـصـ لـلـ Data زـيـ الـيـ قـدـامـناـ مـثـلاـ لوـ معـانـاـ X1 , X2 يعنيـ وـ الـ نقطـ الحـمراـ
ديـ الـ labelـ بـتـبـدـأـ بـدلـ مـاـتـدـورـ مـثـلاـ عـلـيـ الـ best fit lineـ لاـ بـتـدورـ عـلـيـ اـزـايـ تـعـملـ best Splitsـ للـdataـ
وـتـعـبـرـ عـنـ Splitـ كـلـهـ بـسـؤـالـ وـاحـدـ بـاجـابـةـ وـاحـدـةـ يعنيـ خـلـيـنـاـ نـشـوـفـ مـثـالـ كـدهـ سـواـ.



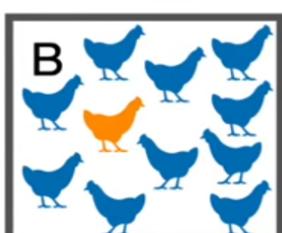
فعندنا هنا على سبيل المثال الجزء اللي تحت على الشمال دم لو عايز تمثله حاجة زي سؤال واحد فهتقول هل X_1 أقل من 20 و X_2 أقل من 200 ؟ لو تهمام يبقى الناتج 300.5 طب الشاتج دم فعلا يعني كل العينات اللي في Split دم كلها اجابتها 300.5 ؟ في الحقيقة لا بس هو دم ال average بتاعها فهو يعتبر يقطع الداتا لجزاء ويفصل الجزاء دي بشوية اسئلة بتحط في

Nodes , Sub nodes

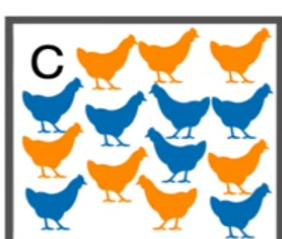
Entropy



تخيل مثلا ان انت معاك 3 مجموعات زي كده. طبعا لو قلتلك اختار من مجموعة A عشوائيا فالاحتمال الاكبر انك تطلع فرخة orange. وبما انه ال probability الاعلى لل orange chicken انت مش هتتفاجئ او اي يعني بس لو طلعلك الفرخة الزرقاء فهيكون نسبة التفاجئ عندك اعلى.



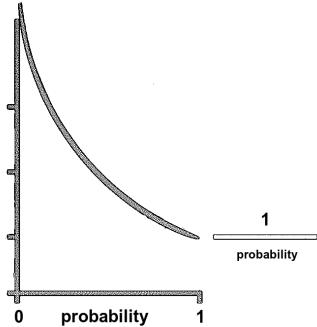
و نفس الكلام في مجموعة B لو طلعلك فرخة زرقاء فانت مش هتتفاجئ ولو طلعلك فرخة orange فه هنا هيكون بردم درجة ال surprising عندك اعلى.



فی المجموعه C بقى فاھتمال انك تطلع فرخة orange هو و انك تطلع فرخة blue. وبما ان ال probabilities متساوية يبقى ال surprises كمان متساوي .

$$\frac{1}{\text{Probability}}$$

هنا لاقى كل ما ال probability تقرب من ال 1 كل ما ال surprises يقل.



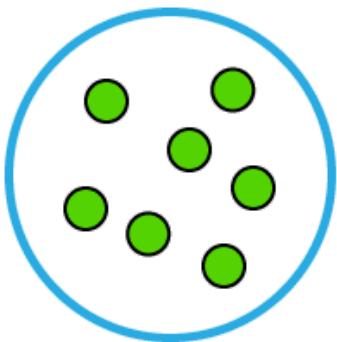
طيب فلنفرض انى معايا مجموعه زى دى هنا في المجموعه دي ال

surprising انى اطلع كوره لونها احمر بصرفا ال surprises هنا هىكون 1/0 ب undefined وده منطقى لأن اصلاً مفيش كور لونها احمر طيب لو جيت حسب ال surprises للكور الخضا هلاقى ان القيمة 1/1 بتساوي 1 بس انا المفترض ال surprises او معدل التفاجئ عندى يكون ب 0 لأن هم كلهم كور خضا ومفيش اى احتمال ان يطلع اي لون تانى ازاي اتفاجئ ان طلع كوره خضا مش منطقى . فهنا جه ال \log و كان الحل و بقت ال formula surprises ال كدة .

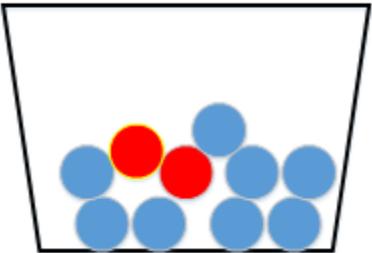
$$\text{Surprise} = \log\left(\frac{1}{\text{Probability}}\right)$$

هي في الحقيقة $\log(1/\text{probability})$ وبكلمة لو ال surprises بتساوي 1 ال surprises هيكون $\log(1) = 0$. ولو ال surprises بتساوي 0 ال surprises هيكون بردم ب undefined واحدنا متفقين ان دة منطقى .

Minimum Impurity



طيب لو معايا box زى ده وقلتلك انك عندك 3 مدعولات انك تجرب تطلع فى كل مرة كورة وقلتلك عايز احسب ال surprising لنه يطلع معاك مرتبين كورة زرقة و مرة كورة لونها احمر و قلتلك ان ال probability بتتعت ال blue ball surprising و ال red ball بتتساوى 0.8 فكل عليك انك تطلع ال بتتساوى 0.8 وكل مرتبة وتجمع النتائج زى كدة



$$\log_2(1/0.8) = 0.32, \log_2(1/0.2) = 2.32$$

بیقی ال پیساوی surprising 2.96 $2.32+0.32+0.32$ پیساوی

طيب لو جيت قلتك انا عايز اكتر العمليه 100 مررر اول حاجة هطلع العدد المتوقع انه يطلع للكور الزرقا (100*0.8) و العدد المتوقع انه يطلع للكور الحمرا (100*0.2) وبعد كدة هضرب كل عدد فى الـ . $72=2.32*(100*0.2)+0.32*(100*0.8)$ surprising

طب لف عان ادص متنبطة || surprising , المرة الفاحدة هتسندة 0.72 = 72/100

Entropy

و يعند الـ i simplify formula وكذلك الـ Entropy $\text{للـ$

$$\text{Entropy} = - \sum p(x) \log(p(x))$$

Ensemble Learning

ايه هو ال Ensemble learning ؟ ال Ensemble learning كذا learning algorithm هو انك تستخدم كذا task معينه بدل متسخدم algorithm واحد. وفي الاغلب هتكون ال predictions بتعاته احسن من انك تستخدم model واحد. خليني اقلك مثال يعني مثلًا عايزين نعمل binary classification هنالك في الصورة دة person ولا لا. وجريت مثلًا



الصوره دى ال accuracy descision tree 65% و طلعت ال logistic accuracy 75% و جريت ال SVM accuracy 60% regression مثلًا. طيب ايه راييك لو جمعنا ال 3 models train لـ 3 input image و بعد كده بتعاتهم models اقدر اعمل vote على ال output. يعني مثلًا دخلت الصورة دى ال SVM و descision tree. person قالوا person. not person logistic regression. ال ساعتها لها اعمل voting بما ان الاغلبية قالوا person يعني الصورة دى person. طيب دة مع ال classification مع ال regression بقى هعمل ايه. فلنفرض. مثلًا انى بحاول اعمل predict لسعر بيت ما بعد ما كل موديل يطلع ال prediction بتاعه هاخد ال mean بتاع ال prediction بتاع ال models كلها و ساعتها هو دة هيكون سعر البيت ال predicted.

طيب ليه اصلًا نستعمل ال Ensemble learning كذا model و ازود على نفسى ؟ computational power

خليني اقلك انتا ممكن نستخدم ال Ensemble Model علشان 3 حاجات:

- اول حاجة low EROR g better accuracy .

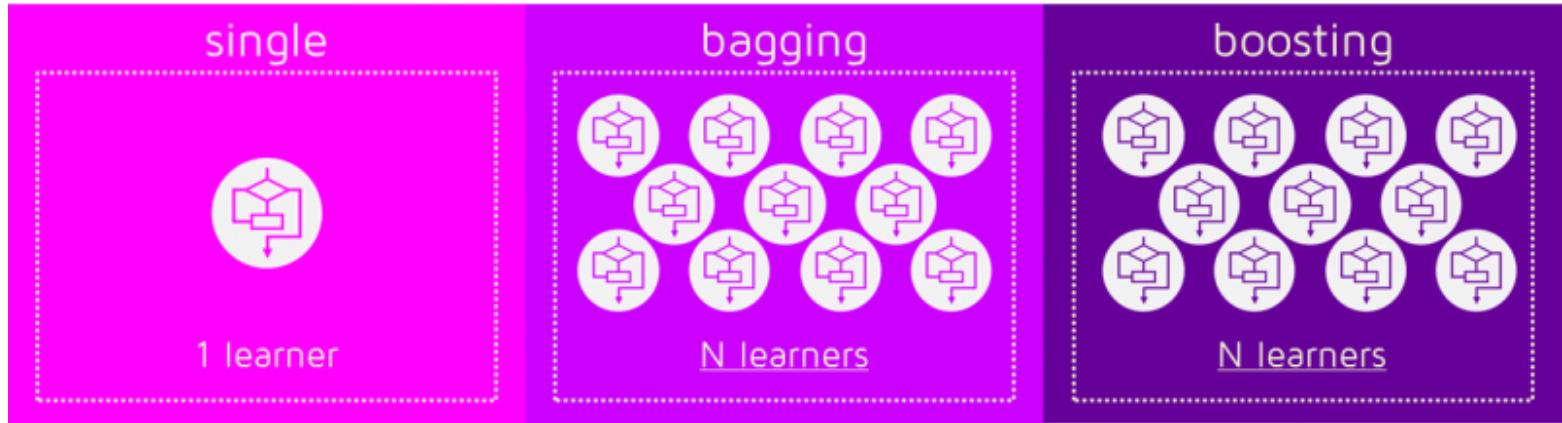
- تاني حاجة Higher consistency يعني consistency overfitting احتمالية ال

- تالت حاجة يقلل ال variance errors و ال bias بتاعه

طيب امتحى ممكن استخدم ال Ensemble learning ؟

خليني اقلك الامثلة البنستخدم فيها ال Ensemble models لما مثلًا تستخدمن single model ي وي results extra training needed. او مثلًا ال overfit حاجة تستخدمنه مع ال classification او ال

خلينا دلوقتى نشوف ال methods المشهورة فى ال regression .جزى ال ensemble learning



Bootstrap Aggregating (Bagging)

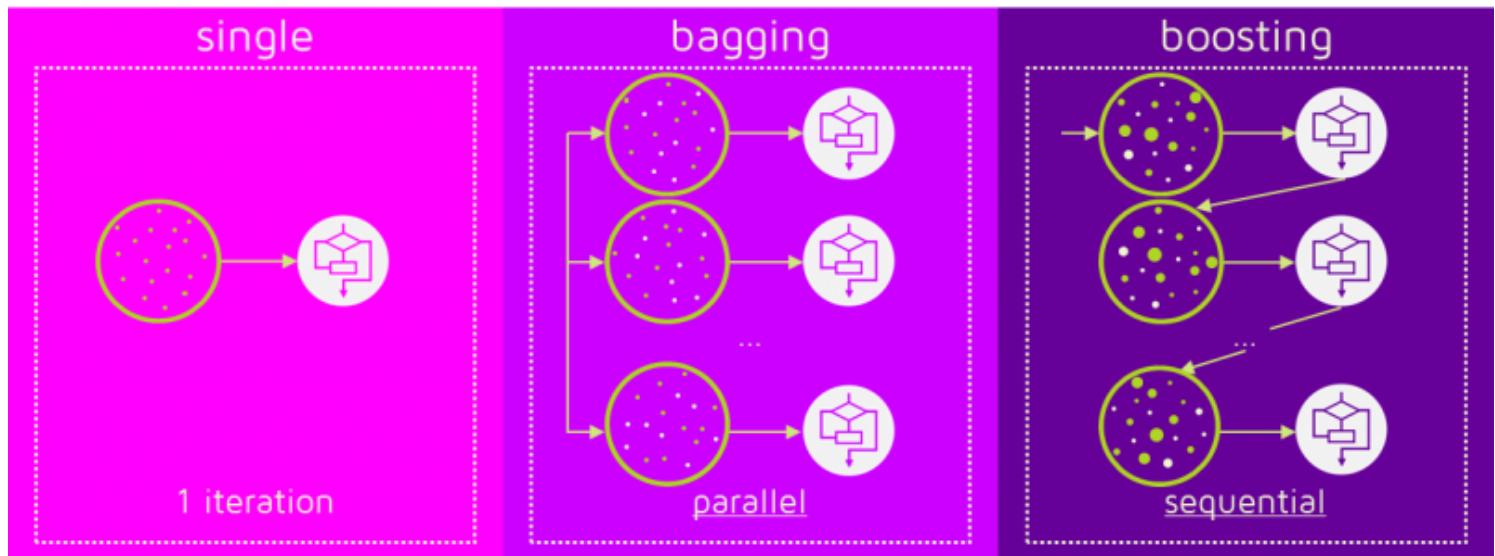
الBagging هو Ensemble method من نفس الـ multiple learners أو multiple models يستخدم ببساطة لـ Ensemble method. وهو ينطوي على بناء several learning algorithms على subsets of dataset، وذلك بعمليات train decision trees على subsets of dataset. training algorithm training dataset. يعني مثلاً لو اخترنا subsets training dataset، بطريقة عشوائية من بين subsets training dataset المعانا، يعني مثلاً لو قلنا ان N يتساوى 10، المعانا sample 100 هنقسمها إلى N من subsets of dataset. يعني مثلاً لو قلنا ان N يتساوى 10، فنأخذ كل subset من الداتا تكون 10 samples. بس هل دمة معنام ان ال learner هياخذ في كلsubset 10 عينات مختلفة؟

الإجابة لا الموضع يختلف شوية إن احنا في Bagging بيقى اختيارنا للعينات عشوائي ومفيش أي مشكلة إن نفس العينة تكرر في أكثر من مرة يعني ممكن نفس ال Row من الداتا يتكرر ول يكن 5 هنات إنشاء عملية ال Training بطريقة عشوائية.

Boosting

ف فى ال boosting بقى فهو يذكر عينات من الداتا بردم ولكن العينات اللي بتذكر دى بتذكر بسبب واضح وهو ان ال Learner غلط فيها فال learner اللي بعدم بنديله النقط اللي غلط فيها فى ال test set ، حذف من training set و هكذا.

فبساطة ال Bagging يتم بشكل Parallel اما ال Boosting فال Sequential يبقى طبعاً أسرع و كونه عشوائي في اختيار الداتا احياناً يبعدك عن over fitting و احياناً يبقى .Boosting Techniques من ال lower accuracy

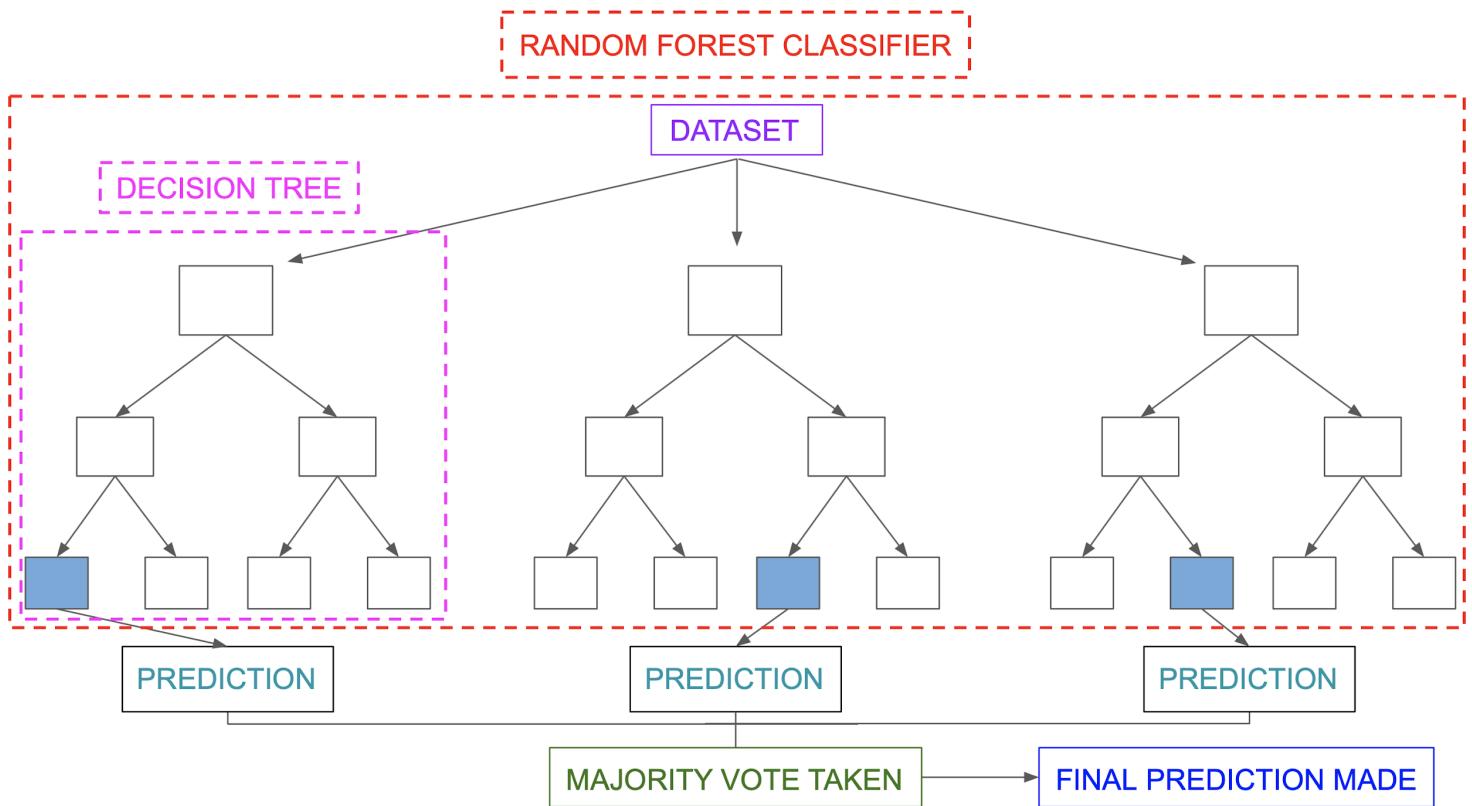


Random Forest Classifier

ال Random Forest واحد من أشهر Machine Learning Techniques في عالم ال Techniques وتقدير تقول أشهر Random Forest وبنستخدمه سواء في Classification أو في Simple g Regression جدا في فوقيه وعندم عدد ضخم من Parameters اللي بتساعده في انه يبقى دقيق جدا في بناءه Predictions

طيب بيشتغل ازاي ؟ تخيل كده انك معاك Decision Tree بس مش واحدة لا مثلاً 200 واحدة وجيست اديت ليهم كلهم نفس الداتا بنفس Parameters ومثبت كمان حتى ال Randomness بتاعتها بالتالي متوقع انه يطلعوا نفس الناتج

طب تخيل بقى لو كل واحدة خدت جزء مختلف من الداتا وفي اجزاء متكرره او وفي حاجات جديدة والموضوع عشوائي و كل واحدة تقدر تبقى طويلة او قصيرة فبالتالي الناتج هتختلف .



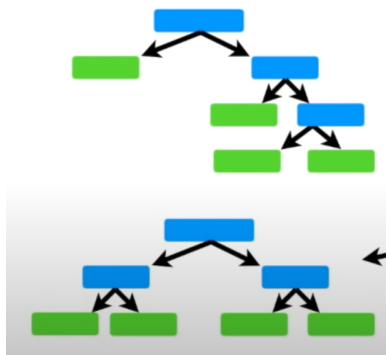
ده بالظبط اللي بيحصل بس السؤال هنا كل واحد اتدرب بشكل مختلف فبالتالي النتائج مختلفة طب فمرحلة Prediction انك تطلع نتيجة من الموديل كل واحد هيطلع نتيجة شكل لنفس العينة هيعمل ايه ؟ ببساطة هي عمل تصويت يعني لو افترضنا ان Classes 0 , 1 و في 80 % قال 0 و 20 % بس اللي قالوا 1 ساعتها هيقول الاجابة 0

ودي من الاسباب اللي بتخليةه يتعامل مع over fitting لنه لو في شجرة كانت over fit فالباقي مش هيقي over fit فهو بيتمشى بمبادئ مختلفه زي Brute force في انه لازم كل Trees تشتعل ويشوف حلولها وفي نفس الوقت بيقى Greedy approach من جوا ال Trees في اختيار اهم ال Nodes .

نضيف على كلامنا انك تقدر تختار العدد اللي تحبه في عدد Trees اللي بنسميهها No.Estimators وال Trees بتختلف عن بعض في الداتا اللي دخللها وفي ال splits بقاعدتها وطولها وعدد عوامل تانية كثير .

فى العادي لما بتبني tree فانت بتبني tree كاملة زي كدة

و ممكن تكون tree اكابر من الثانية عادي جدا.



لكن فى ال forest of trees. فهى فى العادة بتكون عبارة عن two leaves g one node و ال tree المكونة من node

forest of stumps. فا فى الحقيقة دى بتكون two leaves

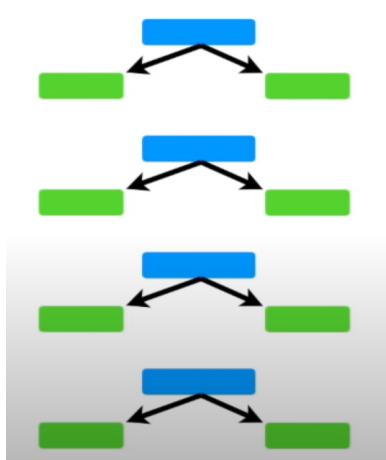
مش forest of trees. ال stumps مش احسن حاجة فى ال classification

يعنى مثلا لو عندنا data من 4 features فا ال descision tree هتاخذ فى

حسابتها ال 4 variables علشان تقدر تطلع prediction. لكن ال

الواحدة تقدر فقط ان هى تستخد 1 variable علشان تعمل decision

وعلشان كدة بتكون ال stumps من ال ناحية ال تعترى "weak learner".



خلينا بردم نبص من ناحية ال random forest فى ال tree كل

ليها نفس نسبة التأثير فى ال vote على ال final classification. لكن فى ال

stumps العملتها ال forest of stumps ليوم

قوة اكتر او تأثير اكتر فى ال final classification result عن باقى ال stumps. كمان فى ال

tree كل decision tree مستقل عن ال trees الثانية. يعني مش فارق امهى Forest

اعملت الاول. لكن فى الناحية الثانية فى ال Forest of Stumps ال بتعميلها AdaBoost فا ترتيب ال

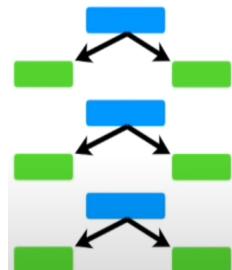
second stumps مهم. علشان ال errors ال بتعميلها ال first stumps بتأثر على ازاى هنعمل ال second

third stumps وهكذا. بنفس الكلام ال errors ال بتعميلها ال second stumps هتأثر على ال third stumps

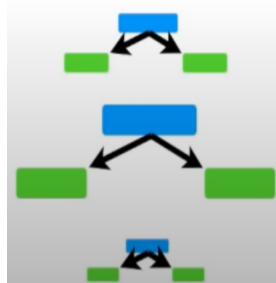
و دلوقتى احنا عرفنا التلت افكار ورا ال AdaBoost

- اول حاجة انها بتجمع عدد كبيير من ال "weak learners" classification

ال stumps دول تقريبا على طول بيكونوا weak learners

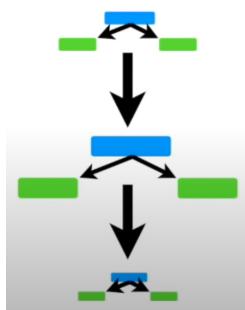


- تانى حاجة وھي ان فی شویة من ال



ليوم ليوم تأثير اقوى في ال stumps

.stumps عن باقى ال classification



- تالت حاجة و هى ان كل stump بتعمل بيافر فيها ال error بساع ال القبها.

Patient Weight	Chest pain	Blocked Arteries	Heart Disease
205	YES	YES	YES
180	NO	YES	YES
210	YES	NO	YES
167	YES	YES	YES
156	NO	YES	NO
125	NO	YES	NO
168	YES	NO	NO
172	YES	YES	NO

خلينا نشوف ازاي ممكن ان احنا نكون

.AdaBoost بال forest of stumps

فالنفرض ان احنا معانا الداتا دي. و احنا

عايزين نعمل بال forest of stumps

عند AdaBoost. علشان نشوف هل المريض دة

عند heart disease ولد لد. و معانا 3

الهم ال chest pain او weight او pain الصدر

و ال patient weight او blocked Arteries او

او وزن المريض. اول حاجة هعملها هي انى

هدى لكل sample من الداتا ال weight

Patient Weight	Chest pain	Blocked Arteries	Heart Disease	weight
205	YES	YES	YES	1/8
180	NO	YES	YES	1/8
210	YES	NO	YES	1/8
167	YES	YES	YES	1/8
156	NO	YES	NO	1/8

المبدئى ليها او اهميتها. و فى الاول هدى لكل ال samples نفس الاصناف و المهمة هو 1 على عدد ال samples العندى كلها.

وبما ان هنا عندى 8 samples فكل واحدة هتكون $\frac{1}{8}$. وبعد كدة لما اعمل اول stump ال weights دى هتتغير علشان تدخلنا نعرف نعمل ال stump بعدها وهكذا. خلينا دلوقتى نبدأ نعمل اول stump. اول حاجة هنشوف انهى feature من الثلاثة العندنا هى احسن واحدة تعمل classification لل features

العندنا. وبما ان كل ال samples

متساوية weights

هنجاهم دلوقتى. بعد كدة

هادحسب ال Gini index لل stumps كلها و هيطلع ان اقل

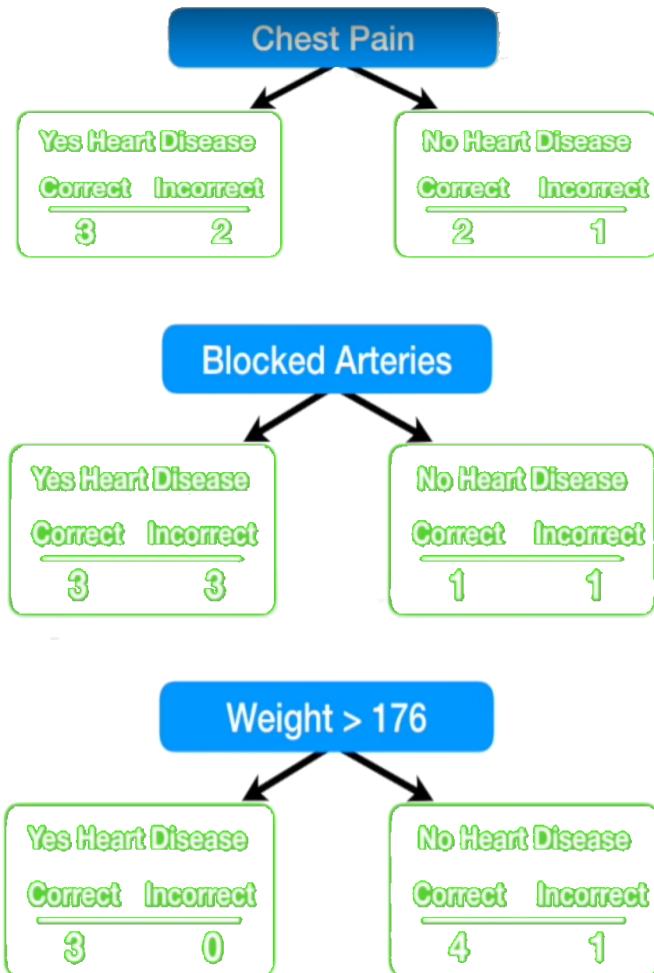
Gini index بقى يتعارض مع وزن المريض ب 0.2

فكرة نقدر نقول ان ال stump دى

هي الـ 8 تكون اول stump فى ال

forest بقى عايزين

نعرف مدى تأثير ال stump دى على



ال اساس جودة ال classification samples عامله لـ final classification . احنا بنعرف هى بتاثير بنسبة قد ايه على ان ال stump دى عملت error واحد بس فامجموع ال error لـ samples دى هيكون مجموع ال weights المرتبطة بال stump classification . فعندها هيكون % و دلوقتى نقدر

$$\text{Amount of Say} = \frac{1}{2} \log\left(\frac{1 - \text{Total Error}}{\text{Total Error}}\right)$$

نستخدم ال formula دى
علشان نحدد ال amount of say

او مقدار تأثير ال stump دى
على ال final classification

كل ما يكون ال stump احسن و ال

total error اقل ال

Amount of say هيكون قيمة كبيرة
و لو ال total error كان 0.5 positive

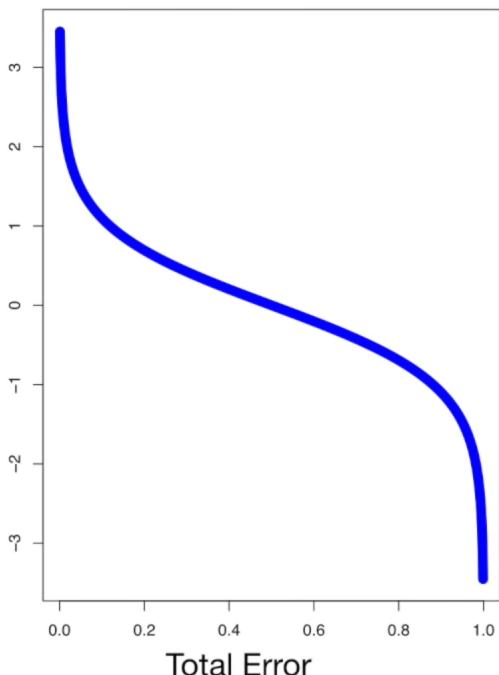
يعنى 50% من ال samples غلط

هيكون ال amount of say بصفر. ولو

ال total error قريب من ال 1 يعني

تقريبا ال stump عمل كل ال

classifications هيكون ال amount of say مربع



كبير. فلو مثلًا الـ stump عملت ان المريض عندما vote heart disease هتخلي الـ vote انه معندهوش negative value فا الـ formula. خلى بالك لو الـ total error كان بـ 0 او 1 الـ disease هتكون بـ undefined علشان كدة في الحقيقة بيضيفوا رقم صغير جدا لـ equation علشان ميصلش كدة.

المهم دلوقتى ان الـ stump بـ amount of say weight ه يكون بـ 0.97 . و دلوقتى عايزين نعدل الـ weights علشان الـ stump الجاية تأخذ فى حسابها الـ errors بـ amount of say weight . فحنا محتاجين ندى الـ classification sample غلط اهمية اكتر او اعلى يعني . و هنسخدم الـ formula دى علشان

New Sample = sample weight $\times e^{\text{amount of say Weight}}$ نجيب الـ weight الجديد. و ه يكون فى

المثال بتاعنا الـ weight الجديد بـ 0.33. و بما ان مجموع الـ weights كلها لازم يكون 1. فاحنا محتاجين ان ادنا نقل الاهمية او

New Sample = sample weight $\times e^{-\text{amount of say Weight}}$ الـ weight بتاع بقىت samples .

الفرق هنا مبين الـ formula دى والقبلها ان هنا ه يكون negative . والنتيج ه يكون 0.05 . amount of say

Patient Weight	Chest pain	Blocked Arteries	Heart Disease	weight	New Weight
205	YES	YES	YES	1/8	0.05
180	NO	YES	YES	1/8	0.05
210	YES	NO	YES	1/8	0.05

بما ان كل ال weights البقية قد بعض لسة مكون

Patient Weight	Chest pain	Blocked Arteries	Heart Disease	weight	New Weight	Norm.
205	YES	YES	YES	1/8	0.05	0.07
180	NO	YES	YES	1/8	0.05	0.07
210	YES	NO	YES	1/8	0.05	0.07
167	YES	YES	YES	1/8	0.33	0.49
156	NO	YES	NO	1/8	0.05	0.07
125	NO	YES	NO	1/8	0.05	0.07
168	YES	NO	NO	1/8	0.05	0.07
172	YES	YES	NO	1/8	0.05	0.07

ه يكونوا بنفس الناتج. ودمة هييقى شكل ال weights الجديدة. بس لو جمعتهم بردم هتلقيهم 0.68 مش 1. فحنا محتاجين نعمل normalization ليوم علشان يبقى مجموع 1. فونقسم كل weight على مجموع ال 0.68 دلوقتى هتلقي مجموع يساوى 1 او ممكن يكون فى بسيط سواء اقل او اكتر من ال 1 حاجة بسيطة.

و دلوقتى اقدر استخدم ال weights الجديدة

.next stump لل

نظرياً احنا ممكن نستخدم ال weights

الجديدة ونحسب ال weighted Gini Indexes

علشان نعرف انهى feature المفروض انه

يتعمل بيها ال stump الجاية.

Patient Weight	Chest pain	Blocked Arteries	Heart Disease	weight
205	YES	YES	YES	0.07
180	NO	YES	YES	0.07
210	YES	NO	YES	0.07
167	YES	YES	YES	0.49
156	NO	YES	NO	0.07
125	NO	YES	NO	0.07
168	YES	NO	NO	0.07
172	YES	YES	NO	0.07

Patient Weight	Chest pain	Blocked Arteries	Heart Disease	weight
205	YES	Yes	NO	0.07
180	NO	YES	YES	0.07
210	YES	NO	YES	0.07
167	YES	YES	YES	0.49
156	NO	YES	NO	0.07
125	NO	YES	NO	0.07
168	YES	NO	NO	0.07
172	YES	YES	NO	0.07

ال weighted Gini index اهمية هيدي
 اكبر انه يعمل classification صح لـ
 دى علشان هى واحدة اعلى sample weight
 بس بدل منستخدم ال Weighted Gini index
 هنعمل collection جديد من الداتا
 ويكون نفس حجم ال samples المعانا و
 يكون يحتوى على duplicate copies من
 ال sample العندھا اعلى weight .

اول حاجة هنعملها علشان نشوف هنختار انى sample من ال original samples المعانا فونختار random number ويكون من range ال 0 لحد 1 . و نبص على ال weights العندنا فوتلاقى ان اول weight بيتساوی 0.07 طبعاً بين ال 0 و ال 0.07 هنختار اول sample و لو بين 0.07 و 0.14 (العنواني) و لو بين 0.14 و 0.21 (العنوانى الثاني) ييقى هختار ال weight الاول + ال weight بناءً على 0.07+ 0.14 = 0.21 و لو بين 0.21 و 0.49 (العنوانى الثالث) يليقى هختار ال weight الثاني + weight الثالث = 0.49 + 0.21 = 0.70

الرابعة وهذا. و هفضل اكتر العملية لحد مييقى عدد ال sample

original الجديدة نفس عدد ال samples

.samples

ودلوقتى بقت دى ال new samples بتعاتى

ولوخدت بالك هتلائق فى اربعة samples

متكررين و هى دى ال sample الكانت

معمولها classification خلط ال

الفات.

وهنرجع تانى ندى كل ال samples اهمية

او weights متساوية زي المرة الفات.

هتقلى طب كدة انا اديتهم كلهم نفس

الاهمية ازاى هياخد باله من انه ميغلاطش

فى ال sample الغلط فيها المرة الفات

هذاك لا هو هيعامل الاربعة

المتكررين على اساس انهم block

وكراهم دة هيذلي فى penalty عاليه انه

ميعملش misclassification ليوم.

ودلوقتى هنرجع للخطوة الاولى ونحاول نوصل لل stump الابسدن

Patient Weight	Chest pain	Blocked Arteries	Heart Disease
156	NO	Yes	NO
167	YES	YES	YES
125	NO	YES	NO
167	YES	YES	YES
<u>167</u>	<u>YES</u>	<u>YES</u>	<u>YES</u>
172	YES	YES	NO
205	YES	YES	YES
167	YES	YES	YES

Patient Weight	Chest pain	Blocked Arteries	Heart Disease	weight
156	NO	Yes	NO	1/8
167	YES	YES	YES	1/8
125	NO	YES	NO	1/8
167	YES	YES	YES	1/8
<u>167</u>	<u>YES</u>	<u>YES</u>	<u>YES</u>	<u>1/8</u>
172	YES	YES	NO	1/8
205	YES	YES	YES	1/8
167	YES	YES	YES	1/8

البتعمل new collection of samples لـ classification . و هتكون

هي ال next stump . وبعد تكرار الكلام دة كذا مرة وخلاص كونا ال final classification بـ بتاعتنا هنأخذ القرار الاخير ازاي او ال forest stumps amount of say بتاع ال total number هنجيب ال القالوا ان المريض عنده total number g heart disease .

القالوا معدوش heart disease . و الرقم الكبير هو صاحب القرار يعني هنا مثل الرقم الكبير كان للنهم قالوا ان المريض عنده heart disease انه عنده final classification يبقى ال heart disease . disease

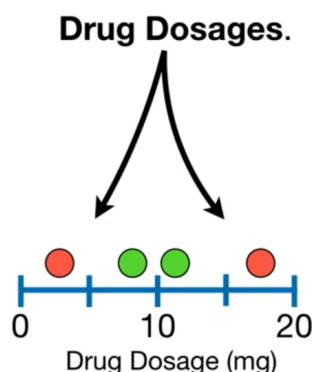


XGBoost

خليبي اقل لك ان ال XGBoost اتعملت علشان نستعملها مع ال large and complicated data sets

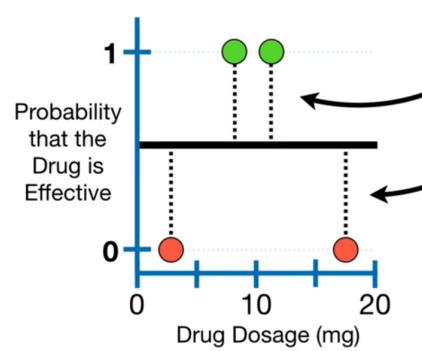
بس scale هن Shruth على for simplicity نقول ان احنا

معانا training data دى بتكون من 4 samples وال samples دى بتعبر عن جرعات مختلفة لدواء معين وعايزين نعرف اذا كانت الجرعة دى



او فعالة ولا لا. فلنفرض مثلاً ان ال red دى هي ال effective green points .not effective هي ال points

اول حاجة بيعملها ال fitting فى ال XGBoost انه يفترض initial prediction تكون اي dosage بس بتكون 0.5 يعني بغض النظر عن ال dosage او الجرعة فال default prediction هتكون 50% ان الدواء دة



او Residuals. ال effective differences بين ال predicted g actual هتقدرا تقلنا ال initial prediction دة كوييس قد ايده. بعد معاملنا ال initial prediction دلوقتى

خلينا نبني اول tree فى ال XGBoost بنبدأ ال tree as a single leaf دى كل ال leaves will be a single leaf. بما ان انا فى اول predictions يتبعها القبلها

فأ ال prediction قبلها ثابت لكل ال samples و هو ال sample residual البالى 0.5. ولما نحسب ال prediction هنلاقي دة شكل ال leaf.

-0.5, 0.5, 0.5, -0.5

دلوقتى هندسب حاجة اسمها ال similarity او ال quality score score

$$\frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

بكل بساطة هو عبارة عن مجموع ال residuals على مجموع لكل sample لـ probability القبليها مضروب في 1 ناقص ال regularization lambda هي probability القبليها بردم. و ال parameter. لو حاسس ان الدنيا مش احسن حاجة في القانون دة متهلاقش و احنا شغالين في المثال هتللاقي الدنيا وضحت.

خلينا دلوقتى نحسب ال similarity score لل leaf العندنا. اول حاجة هنضيف كل ال residual في البسط او ال numerator. لو بطيينا على ال residuals المعانا هنلاقي ان مجموعهم صفر وبالتالي

تربيعهم هيكون بصفه برد و دم

Similarity = 0 **-0.5, 0.5, 0.5, -0.5** .zero ب similarity هيدل ال

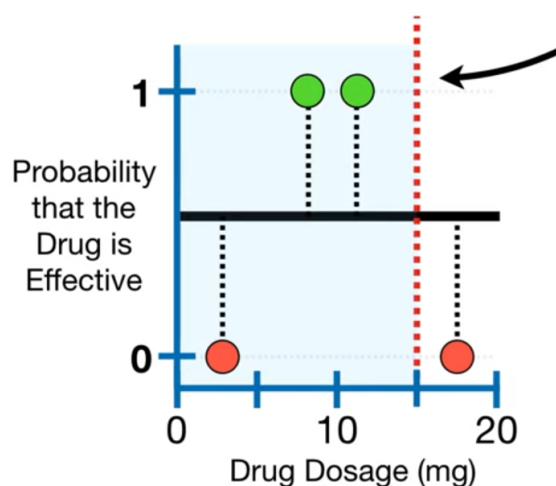


دلوقي عايزين نقرر هل هييقى

احسن لو خلينا ال leaf دى root و قسمناها ل two groups.

خلينا نبدأ بال threshold دة ان الجرعة اصغر من 15 ونقسم بيها.

احنا اختربنا ال threshold انه يكون ب 15 علشان ال 15 هى القيمة



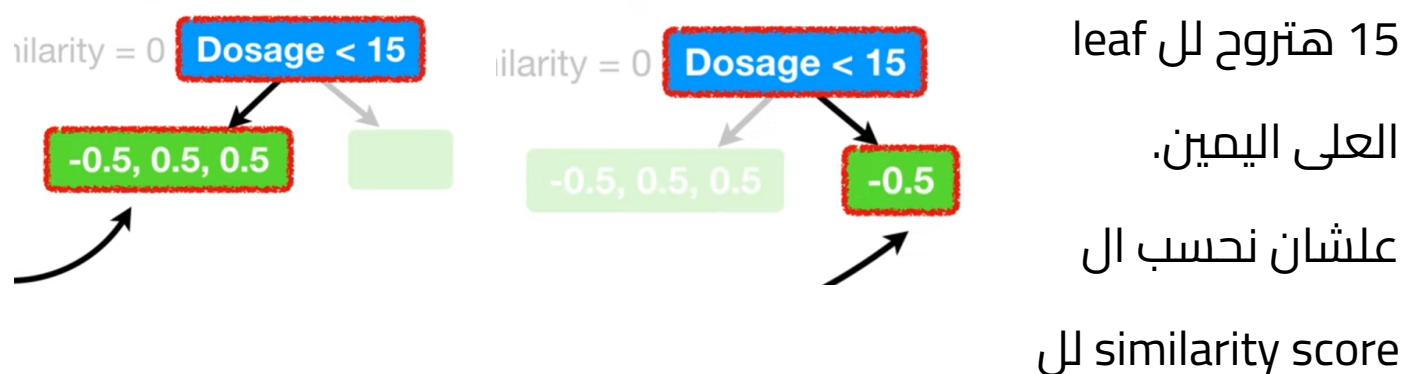
.samples المتوسطة بين اخر 2

دلوقي ال 3 residuals الهم الجرعة

بناعتهم اقل من 15 هيرجعوا في ال

residual leaf على الشمال و ال leaf

للجرعة الاكثر من



اللى في ال leaf على الشمال هونحط ال 3 residuals

البسط و بما انا بنبني اول tree فال probability القبلينا هى ال

. فونجت 0.5 لكل leaf من ال residual على الشمال. وللسولة هنخلى ال lambda ب 0 دلوقتى. وده هيكون شكل القانون.

$$(-0.5 + 0.5 + 0.5)^2$$

$$(0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + 0$$

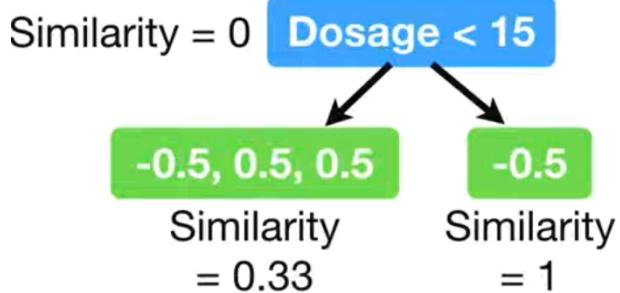
و الناتج بتاع ال leaf دى هيطلع 0.33.

$$\frac{(-0.5)^2}{0.5 \times (1 - 0.5) + \lambda}$$

و ال leaf similarity score لل leaf على اليمين هيكون كدة.

والناتج بتاعها هيكون 1 بعتبار ان

ال lambda ب 0 برمد.



و دلوقتى هنحسب ال gain بتاع ال tree دى.

و ده القانون علشان نحسب ال gain .

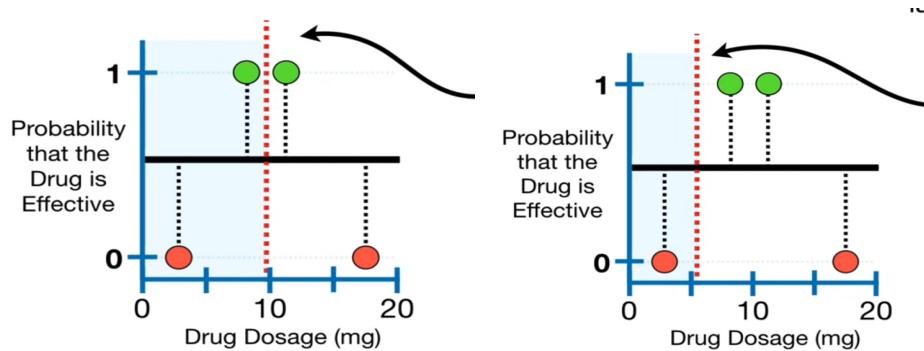


$$\text{Gain} = \text{LeftSimilarity} + \text{RightSimilarity} - \text{RootSimilarity}$$

والناتج هي طاحع

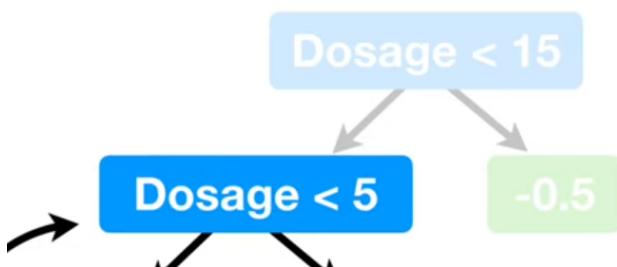
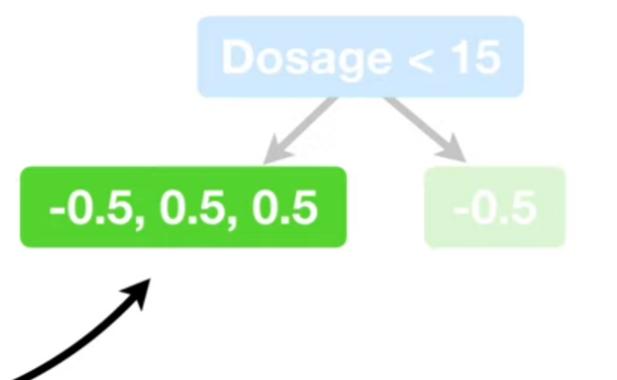
فكل دة بنقول ان ال gain يساوى 1.33 لما قسمنا ب threshold الى dosage اقل من 15.

هنغير ال threshold مررر بقيمة المتوسط بين ال two samples الى في النص و مررر بقيمة المتوسط بين ال two samples الاولين.



و نحسب ال gain عند كل threshold و threshold صاحب اعلى و هو المستخدم. وبعد تجربة هنلائق ان اكتر gain كان لها ال dosage يساوى 15. و دة معنام ان دى هتكون اول branch فى ال tree بتعتبرنا.

دلوقتى ممكن كمان تقسم ال leaf دى بنفس الطريقة.



وهيكون شكلها كدة.

وهكذا بردم لleaf العلی اليمین. دلوقتی بس احنا هنعتبر ان احنا عملین limit لleaf انها تكون 2 levels بس للمسؤوله فوتقف لحد هنا لكن فی الحقيقة ال XGBoost عندم حاجة اسمها minimum cover فی كل leaf. و بنسبيه حاجة اسمها number of residuals cover و ال default cover 1 يعني لو اي leaf ای بتاعه يساوى 1 يعني بتاعها اقل من 1 ال leaf دی بتتشال وال cover دة بيعتمد على ال بتاعها والقانون بتاعه بالشكل دة.

Cover =

$$\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)]$$

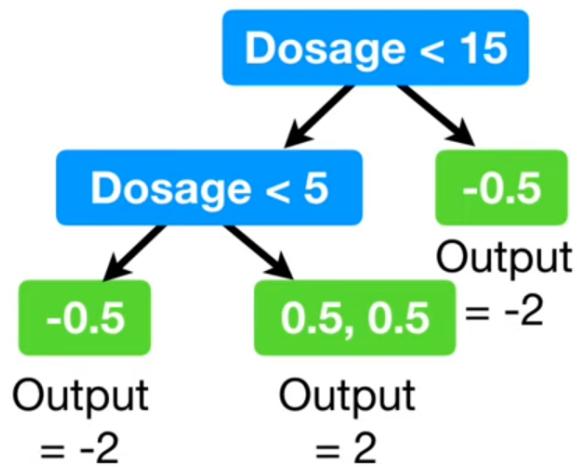
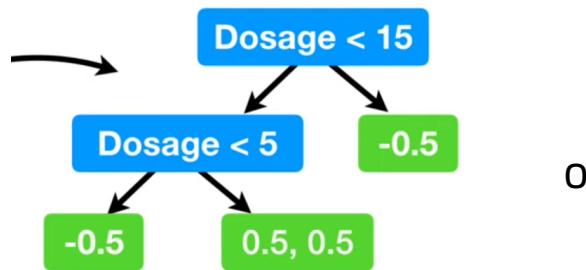
بس احنا دلوقتی هنحط ال minimum value لcover بـ 0. وتقدر تتحكم فيه في الكود من خلال parameter اسمه .min_child_weight

دلوقتی خلينا نتكلم عن ازاي ال xgboost بيعمل pruning لtree. بكل سطوله فی قيمة اسمها gamma احنا بنخترها و بنسبيه ال ناقص ال gain بتاعتك ولو الرقم طاخ .prune هتعمله brunch negative لو

بعد كل الكلم دة خلينا نقول ان ال tree الاولى عندنا بقت بالشكل
دة . خليني افكرك ان الارقام اللي في
ال leaves دى عبارة عن residuals
فاحنا محتاجين نطلع ال output value
لكل leaf من دول. والقانون بتاع ال
output value هيكون بالشكل دة.

$$\frac{\left(\sum \text{Residual}_i \right)}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

هو هو نفس القانون بتاع ال
Output = $\frac{\left(\sum \text{Residual}_i \right)}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$
هذا بس الفرق ان هنا similarity
بس sum of residuals
بتاع power 2
معانا leaf هيبقى معانا
الtree بالشكل دة.
و بكرة نكون بنينا اول tree.



و بما انشا خلصنا اول tree نقدر دلوقتى نعمل predictions جديدة.
لو فاكرا ال prediction الاولاني هو 0.5 فاحنا بردم هنسخدمه
علشان نقدر نعمل new prediction predict . بس محتاجين نعمل
تعديل صغير على ال initial prediction دم و هو ان احنا هندول له
البيحول من probability ل log(odds) value

$$\frac{p}{1-p} = \text{odds}$$

البيحول من probability ل odds هو دم
و منه نقدر نجيب ال log(odds) value

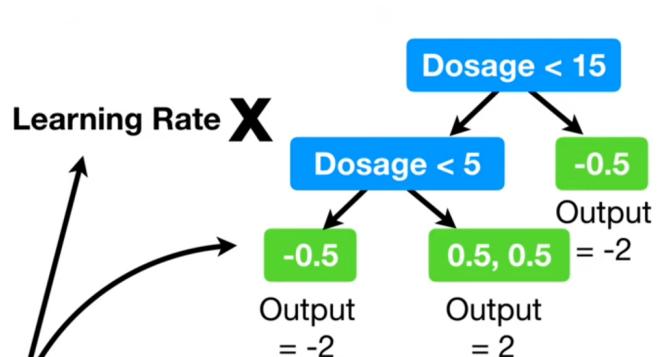
$$\log\left(\frac{p}{1-p}\right) = \text{log(odds)}$$

فخلينا

نتحول اول 0.5 فوتكون النتيجة بتاعتها ب 0 . و علشان اعمل new prediction value
بتبع ال initial prediction output فبنجتمع ال output بتبع ال prediction



Predicted Drug Effectiveness
0.5
Output = $\log(\text{odds}) = 0$



default learning rate و tree مضروب فى ال output بتاع ال learning rate بيكون بـ 0.3 .

فلو عايزين نجيب ال prediction اللى ال dosage value لـ 2 .
log(odds) prediction = 0 + 0.3 * -2 = -0.6

$$\text{Probability} = \frac{e^{\text{log(odds)}}}{1 + e^{\text{log(odds)}}}$$

هندولها ل probability تانى بالقانون دم.

$$\text{Probability} = \frac{e^{-0.6}}{1 + e^{-0.6}}$$

و ال new prediction 0.35 وده هيكون ال probability طبعاً كدة ال residual هيفيل عن الاول لما كان ال prediction بـ 0.35 بعد كدة ممكن تبني tree samples . على ال residuals الجديدة . و تفضل تبني فـ trees لحد ما ال

يكون صغير خالص و يعتبر مبيتغريش او انك توصل لل residual maximum numbers of trees .

CatBoost

ال最新 boosting algorithms هو واحد من ال catboost وبقى open source اظهر اداء احسن من ال CatBoost من سنة 2017. ال computational power او accuracy او gl training الموجودين حالياً من كذا جانب زى فاترة ال categorical data على ال training المحتاجها. كلمة catboost جاءية من Categorical Boosting و دة مش معنام انه متصر بس على ال categorical features و برمجاته يدعم numerical and text features ال يقدر يعمل على handling categorical data بيه لـ . من مميزات ال

Catboost

- طبعاً اول حاجة انه يدعم ال categorical features .
- parameter tuning من غير حتى منعمله high quality .
- GPU version فيه منه .

- overfitting و انه قلل ال data صغيرة

.fast predictions وبطلاع

طيب امتح استخدم ال catboost يشتغل حلو جدا مع

.Heterogeneous data اسمه data نوع من ال

ال Heterogeneous data هى نوع من ال data فيها اختلافات عالية

فى انواع الداتا و تنسيتها. يعني ممكن تكون low quality علشان

مثلها فيها missing values او فيها تكرار عالى للبيانات زى مثل

.Dataset to predict Credit Score

Model Evaluation

يجب دلوقتى لخطوة مهمة جدا وهى Evaluation يعني لو

خلصت ال Model ازاى نقول ان

gl fit ده كويس أو model

under fit gl over fit

عملی ؟

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

الاجابة على حسب نوع ال model لو ال model عبارة عن
Regressor ببساطة لها بسيقى الطريقة
شكليين اما انك تحسب ال Error عن طريق ال Mean Squared
Error أو RootMSE أو MAE وغيرها كثير .
وساعتها بتشوف Prediction - Actual قيمة ال Error وتحسب
بأي طريقة أو انك مثل تلجم ال R Square Score وهو يعتبر شكل
من اشكال ال Accuracy يعني يعرفك الموديل بتاعك بيوصف ال
Problem بنسبة كام في المية فكل ماتقرب من الواحد الصحيح
كل ما يكون ادق وطبعا كل ماتقرب من الصفر معناها ان
الموديل سيء جدا

يعني ملخص اللي فوق ان كل ال Error Metrics اللي اذكرت لو
القيمة قريبة من الصفر دم شئ كوييس ولو ال R Square score
قريب من ال 1 الصحيح يعني شئ كوييس
بس سؤالنا هنا هو انا بحسب Error لمين ل Training set ولا لل
Testing Set في الطبيعي هتقولي ال Testing Set
طب ونعرف ازاي ان الموديل under fit over fit محتاجين

نشوف ال Training طيب هاخد بقى الداتا بتاعت ال
كلها اختبر فيها الموديل ؟ ولا اعمل ايه ؟
الموضوع يختلف على حسب مدى تعقيد
خصوصا لو طويلة جدا زي ما هنلشوف في Deep Learning او ان
تكون حجم الداتا ممكول ساعتها دم مش كلام منطقى فبناجا
للساليب تانية
زي ان احنا بنقسم جزء من Training data نختبر نفسنا فيه وبيبقى
جزء شوفنام قبل كده في مرحلة ال Training دم غير جزء Test
اللي بنختبر فيه ومبيتقاش الموديل شافه قبل كده طيب
اختبارك لجزء Validation دم بسجي ازاي ؟



Validation Set :- The Model got trained on it before

Test Set :- The Model had never seen it before

قبل ما نخشن عالجزء اللي جاين له دم خلينا تتأكد من معلوماتين

1- لو في عندك Valid Set , Test Set , Train Set يبقى اعرف ان
Test g Training Accuracy دي بنقىس بيهما Valid Set
Testing Accuracy بنقىس بيهما Set

2- لو لقيت Train set , Valid Set بس يبقى غالبا هو يقصد
هي Test Set و ساعتها هي بس بمجرد اسم لكن هي set
مش Validation set ولا حاجة لأن من ال Differences بين ال
Validation Evaluation وال Evaluation Validation من نفس الdata
الي هو الموديل شافها ولكن ال Evaluation يتم مع ال
وبالتالي اول مرة يشوفها.

طيب دلوقتي احنا عايزين نختار Validation Set نختارها ازاي من
الdata ؟

هنسخدم أسليب Cross validation عشان نختار أفضل
Validation set من الdata كلها

Cross Validation Techniques

Validation Set Approach

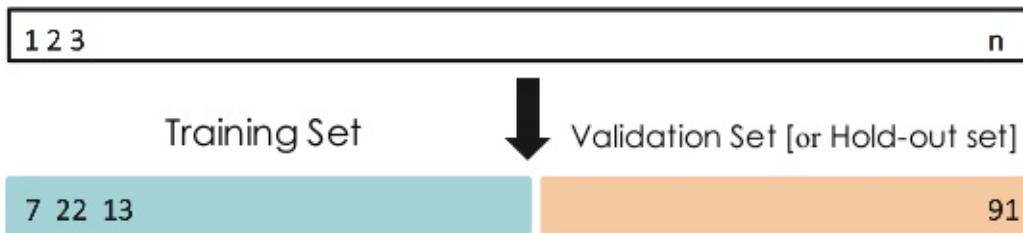
نقدر نقول ان دم most naive انك ببساطة تقسم ال Labeled Data إلى Training Set وTest Set وبنادق . Randomly الموضوع بشكل نصيّن.

Leave one out

Validation set approach

Cross-validation

Randomly
usually almost Half

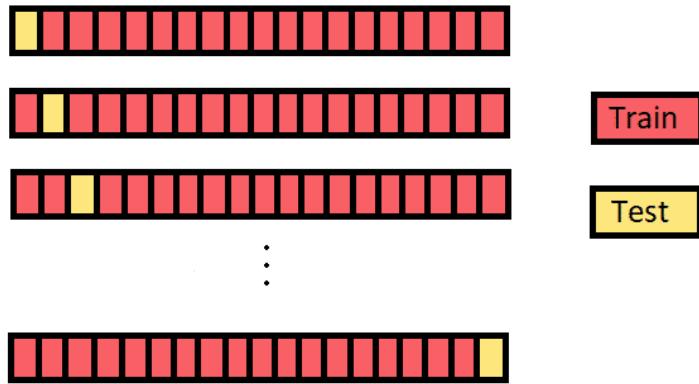


CrossValidation Approach

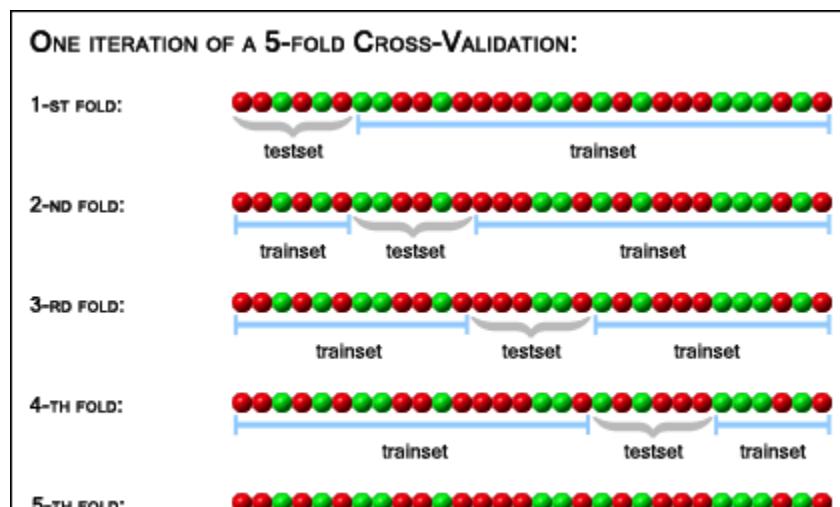
لو شوفنا ان مشكلة اللي فات كانت في ان حجم الداتا كبير جداً مفدر وهو نص الداتا الموديل مبيتدريش عليها فنقدر نقول ان هنا الموضوع معكوس هنا الداتا كلها الموديل يتدرّب عليها وبنخرج

عينة واحدة بس Testing بس بنكرر الموضوع اكتر من مرة وكل مرة نطلع عينة واحدة نختبر الموديل فويا فكل مرة طبعاً يطابع شكل مختلف لأن ممكن تبقى العينة دي مشابهة لحاجة شافها وممكن تكون بعيدة عن اللي شافها فده كان اد البرد . الطرق

K-Fold Cross Validation Approach



الطريقة دي بقى هي اللي بتوزن ما بين ال 2 اللي فوق ان انت عايز تطلع جزء كوييس تختبر الموديل فيها عشان يبقى كوييس وفي نفس الوقت تدرب الموديل على مقدار كبير وفي نفس الوقت بردم مش عايز توقع جزء موم في ال Test ويبقى ضاع منك في ال Training



بنقسم الداتا لعدد معين من Folds اسمه K ول يكن في المثال ده نخليه 5

وبنبدأ نأخذ أول خمس من الداتا نختبر الموديل فيه والباقي تدرب عليه

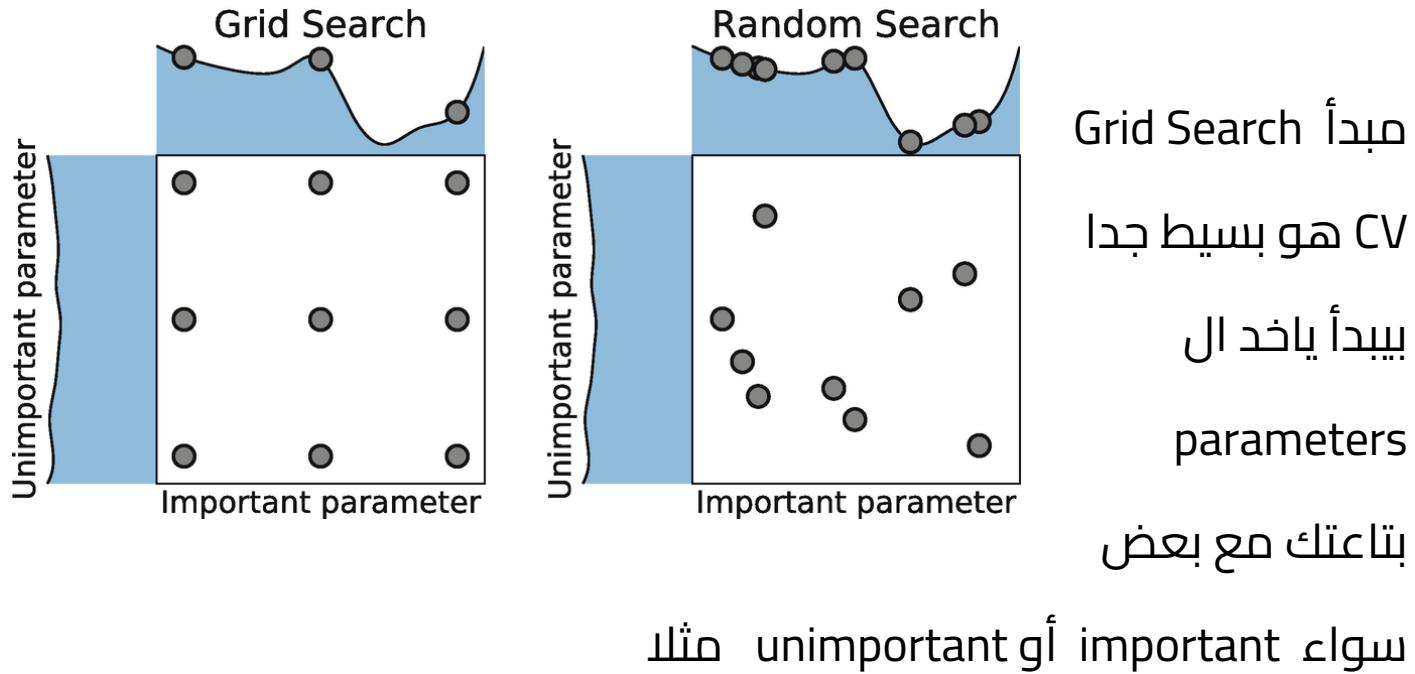
وتاني مرة نأخذ الخمس الثاني ونختبر نفسنا والباقي تدرب عليه مع العلم ان كل مرّة منهـلة لوحدها و كل مرّة بـينحسب لوحدها

الـ Train , Test بـتـاعتـ الـ Accuracy

وفي الآخر بنـشـوف انهـي من دول حقـقـ أعلى Accuracy وبـنـاخـدمـ كـ اختـيـارـ Average Error عـشـانـ نـحـكمـ Train set , Test set عـالمـودـيلـ صـحـ.

الجدير بالذكر ان اـحـناـ بـنـسـتـخـدمـ الـ Cross Validationـ هـمـشـ بـسـ فـيـ اختـيـارـ اـفـضـلـ Test Setـ لـ اـحـناـ كـعـامـ بـنـسـتـخـدمـهـ فـيـ اختـيـارـ Bestـ اختـيـارـ اـفـضـلـ Test Setـ لـ اـحـناـ كـعـامـ بـنـسـتـخـدمـهـ فـيـ اختـيـارـ Bestـ لـأـيـ modelـ يـعـنيـ لوـ اـتـكـلـمـنـاـ عـنـ حاجـةـ زـيـ Parametersـ Randomـ لـأـيـ modelـ يـعـنيـ لوـ اـتـكـلـمـنـاـ عـنـ حاجـةـ زـيـ Parametersـ Forestـ بـيـنـقـىـ لـيـهاـ عـدـدـ ضـخمـ مـنـ Parametersـ فـانتـ مـثـلاـ عـايـزـ تـعـرـفـ انهـيـ عـدـدـ مـنـ الشـجـرـ هـوـ أـفـضـلـ عـدـدـ ؟ـ يـعـنيـ اـدـطـ كـامـ فـيـ Gridـ Techniqueـ اـسـمـهـاـ ؟ـ فـبـنـسـتـخـدمـ Number of Estimatorsـ .Search Cross Validationـ

Grid Search CV vs Random Search Cv



A : [1, 10 ,100]

B : [0.2 , 0.3 ,0.4]

هيدأ يجرب الـ 6 احتمالات الموجودين اللي انت حاطتهم يعني 1 مع 0.2 و 0.3 ... وهكذا لحد مايسوف انهي افضل قيمة Best Error ممكن ساعتها يجعلك لل A ولل B اللي تحقق أقل

Best Parameter g Score

أما Random search فهو ييدي أرقام عشوائية لكل Parameter و بالتالي ممكن توصل ل Accuracy أدق من ال Grid search وممكن لأنها في نوع من انواع العشوائية فيه

