Machine learning with R

k-Nearest Neighbor KNN

K اقرب جار

خوارزميّة الجار الأقرب k-Nearest Neighbors

•يعتبر المصنف كي -اقرب جا ر (KNN) واحدا من أقدم وأبسط خوارزميات التعلم الخاضع للإشراف. نظراً لسهولة استخدامها واستهلاكها للقليل من الوقت .”

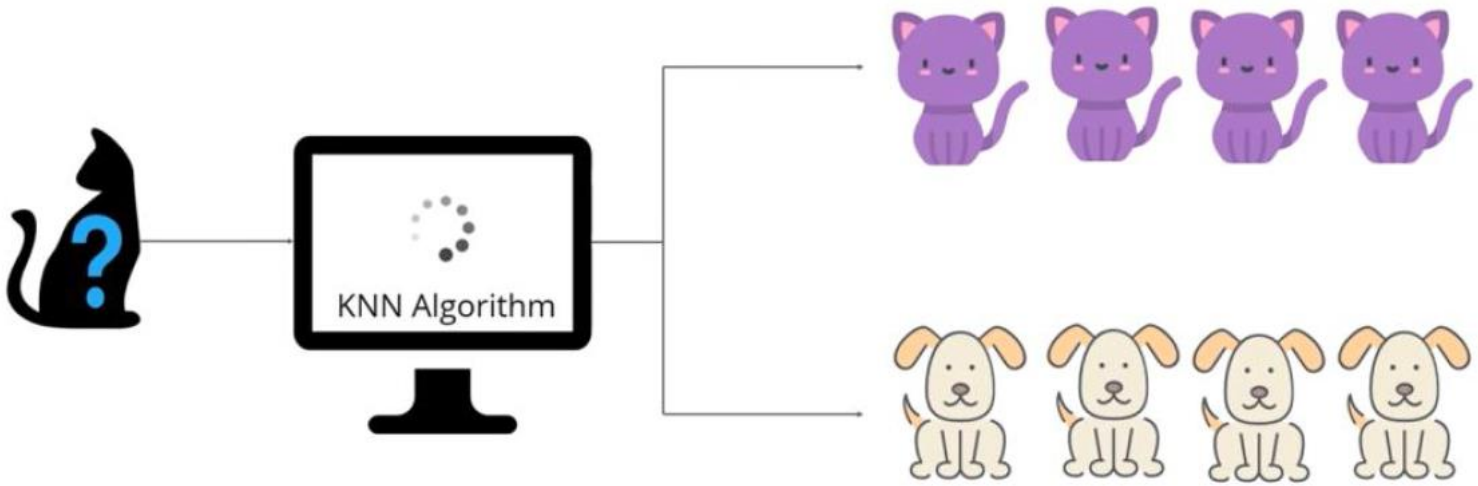
• تعتمد خوارزمية KNN على افتراض أن الأشياء المتشابهة قريبة من بعضها البعض. بالمقارنة مع خوارزميات التصنيف الأخرى ، فإن أقرب جار يستخدم أسلوب التعلم الكسول. بمعنى آخر ، يقوم ببساطة بتخزين العينات في مرحلةالتدريب ولا يفعل شيئا حتى يتم استلام عينات الاختبار. اي ان خوارزميّة الجار الأقرب KNN تستخدم مجموعة البيانات بأكملها كمجموعة تدريب بدلاً من تقسيم مجموعة البيانات إلى مجموعة تدريب واختبار، لأنّ خوارزميّة الجار الأقرب تعمل على فصل بيانات مصنّفة مسبقاً .

•عندما تكون النّتيجة المطلوبة لعنصر بيانات جديد، تنتقل خوارزميّة الجار الأقرب KNN عبر مجموعة البيانات بأكملها للعثور على أقرب مثيلات k إلى العنصر الجديد،

•قيمة k يحدّدها المستخدم. يتمّ حساب التّشابه بين الحالات باستخدام مقاييس مثل مقياس المسافة الإقليديّة ومسافة هامينج.

•

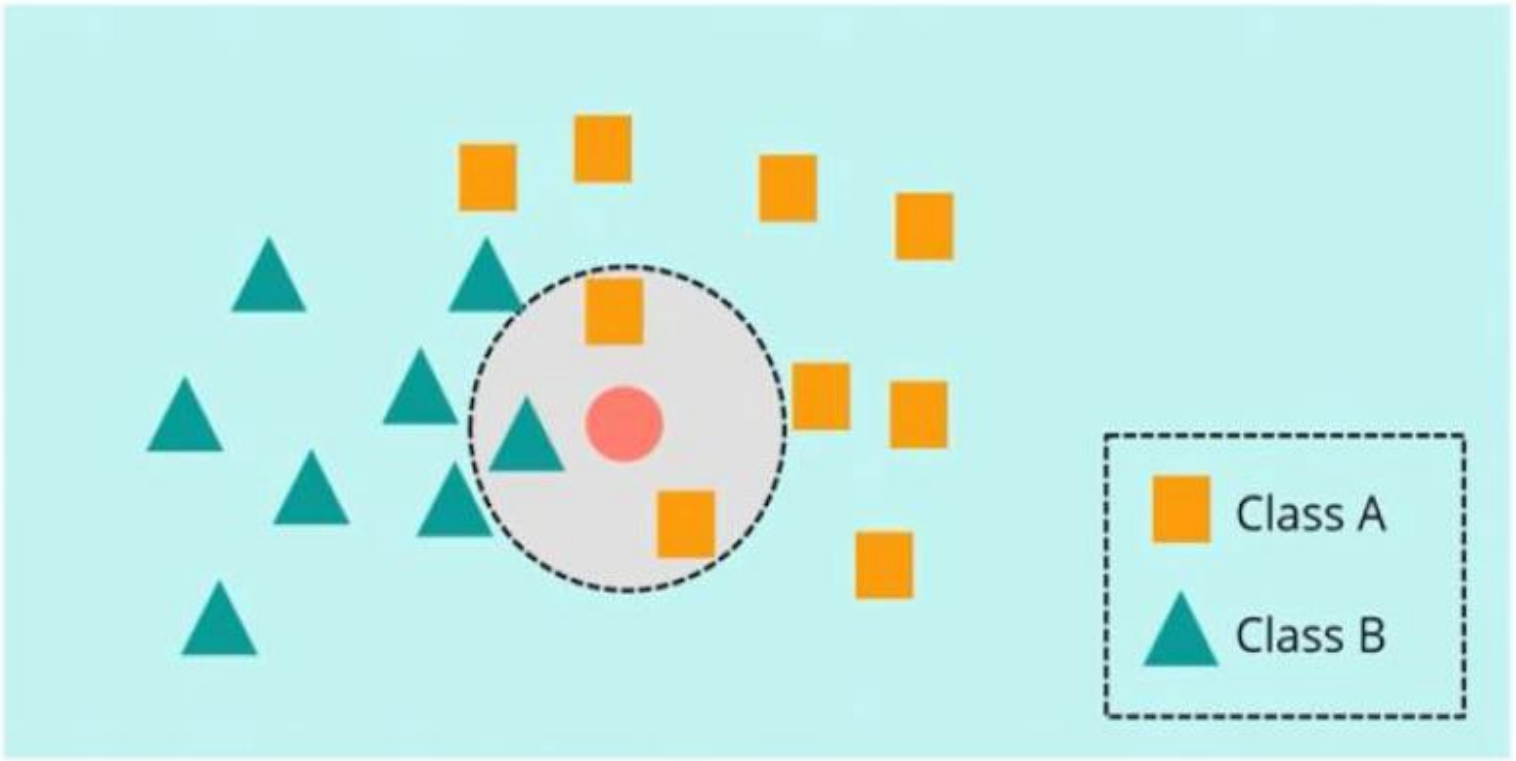
- هل يتم تصنيف الحيوان في مجموعة القطط او الكلاب؟



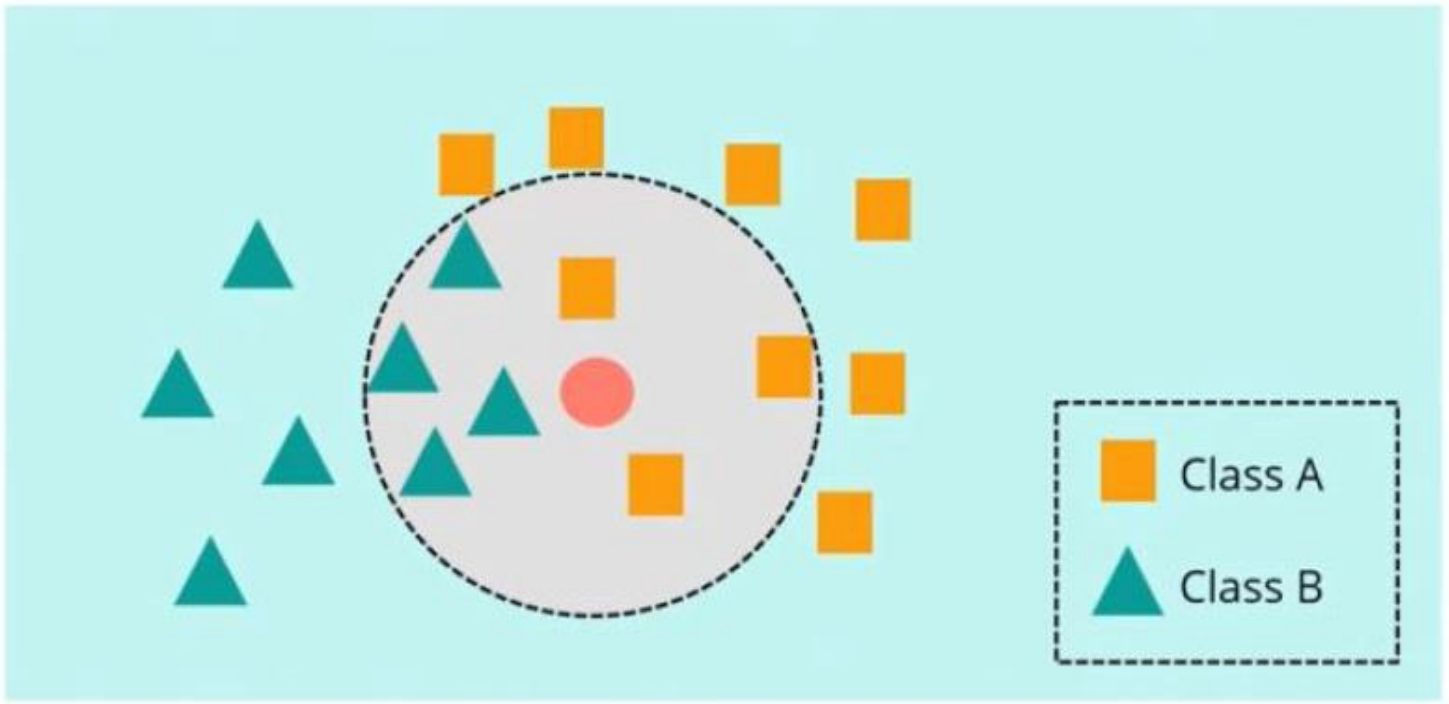
لتصنيف حيوان ما على انه قط او كلب نعتمد على ميزات كل حيوان كما يلي:

CATS	DOGS
	
Sharp Claws, uses to climb	Dull Claws
Smaller length of ears	Bigger length of ears
Meows and purrs	Barks
Doesn't love to play around	Loves to run around

- اختيار اقرب الجيران هنا K=3 هنا نختار الفئة A



- لما تكون K=7 نختار الفئة B



- تلعب المعامل k دورا مهما في تصنيف العينة الجديدة. لأن قيم k المختلفة قد تؤدي إلى نتائج تصنيف مختلفة جدا.

- بالإضافة إلى ذلك ، قد تؤدي حسابات المسافة المختلفة إلى جيران مختلفة ، مما يؤدي إلى نتائج تصنيف مختلفة. ومن ثم ، فإن قيمة K المحددة تحدد دقة التنبؤات وعدد الأخطاء ، لذا فإن اختيار K الصحيح له أهمية أساسية في هذه الخوارزمية. يعتمد اختيار K المثالي على البيانات، لكن الكميات الكبيرة من K تقلل من تأثير الضوضاء على التصنيف ، بينما تقل التمييز بين الحدود والتجمعات.

لاختيار قيمة K

K=3 , K=7 ???

كيفية تحديد معايير التشابه و الاختلاف؟

- يجب أن نستخدم معيار التشابه أو الاختلاف بين نقاط البيانات. هناك العديد من معايير التشابه أوالاختلاف ، بما في ذلك المسافة الإقليدية ، ومسافة مينكوفسكي ، ومسافة هيمينج، وتشابه جيب التمام، و في هذا الدرس نستخدم المسافة الإقليدية .

Distance functions

Euclidean

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

Hamming Distance

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

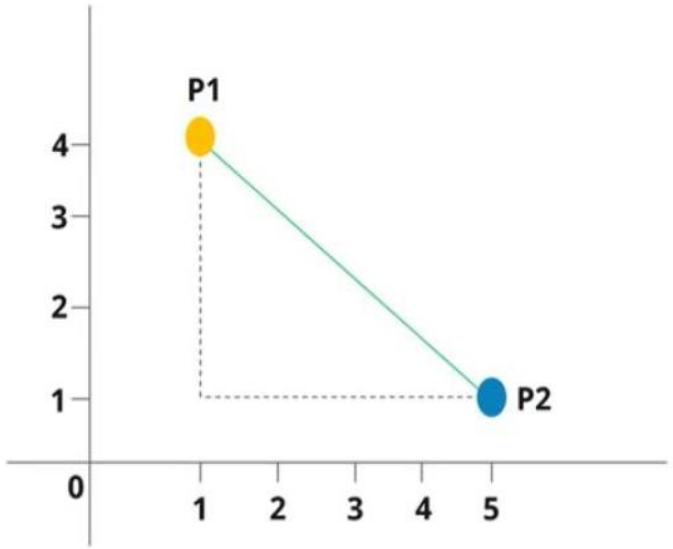
X	Y	Distance
Male	Male	0
Male	Female	1

المسافة الإقليدية هي مقياس للاختلاف بين نقطتي بيانات x_i و x_j. كلما كانت المسافة الإقليدية أكبر ، كلما كانت نقطتا البيانات أكثر اختلافا.

نأخذ الفرق بين القيم المقابلة لتلك الخاصية في العينة x_iو في العينة x_j، نقوم بتربيع هذا الاختلاف ، وفي النهاية يتم أخذ المربع من مجموع عدد المسافات.

عادة ، نقوم بتسوية قيم كل خاصية قبل استخدام المعادلة. يساعد هذا في ضمان أن الميزات ذات المجالات الأولية الكبيرة لا تلغي الميزات ذات المجالات الأولية الأصغر.

طريقة حساب المسافة الاقليدية:



Calculations

Point P1 = (1,4)

Point P2 = (5,1)

Euclidian distance $\sqrt{(5-1)^2+(4-1)^2} = 5$

كيف تعمل خوارزمية كي- أقرب جار

- تحاول KNN التنبؤ بالفئة الصحيحة لبيانات الاختبار عن طريق حساب المسافة بين بيانات الاختبار وجميع نقاط باقي البيانات. عادة بالنسبة لقضايا التصنيف ، يمكن استخدام التصويت للتنبؤ بعينة الاختبار باعتبارها أكثر تصنيفات شيوعا في k عدد الجيران و يتم ذلك من خلال عدة خطوات:

- الخطوة 1: حدد رقم Kللعدد العناصر المجاورة .

- الخطوة 2: المسافة الإقليدية(أو معايير المسافة الأخرى)احسب عدد جيران K .

- الخطوة 3: رتب المسافة وحدد أقرب الجيران K بناء على المسافة الإقليدية الدنياالمحسوبة.

- الخطوة 4: من هذا الجار K، احسب عدد نقاط البيانات في كل فئة.

- الخطوة 5: قم بتعيين نقاط البيانات الجديدة للفئة مع أقصى عدد من الجيران.

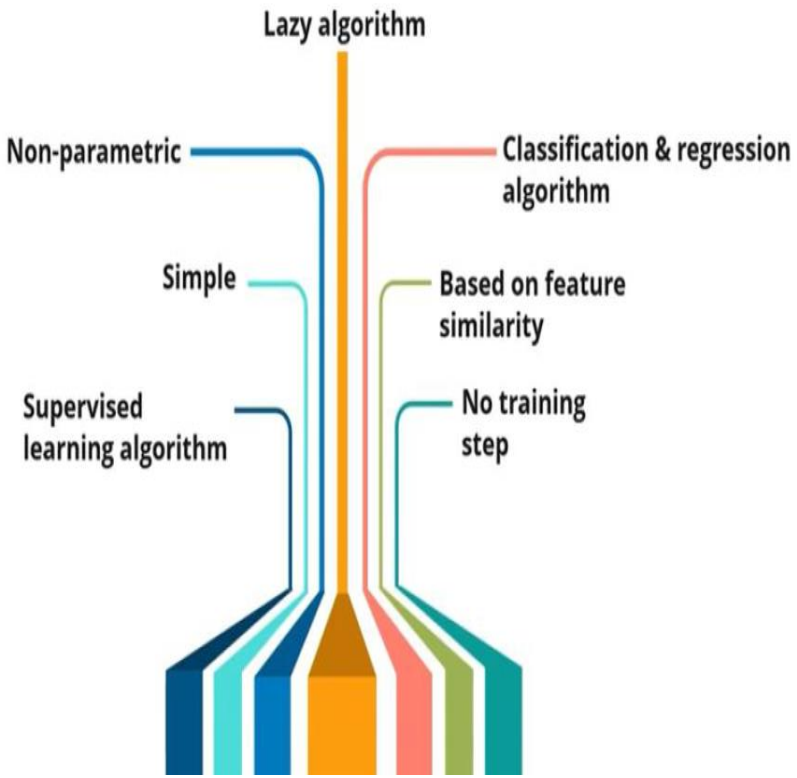
•تعتمد خوارزمية KNN على نوع التعلم على النحو التالي :

•التعلم القائم على العينة: في هذه الطريقة ، لا نتعلم الأوزان من بيانات التدريب للتنبؤ بالمخرجات (مثل الخوارزميات القائمة على النموذج) ولكننا نستخدم عينات تدريب كاملة للتنبؤ بمخرجات البيانات المراد اختبارها .

•التعلم الكسول: لا يتم تعلم النموذج باستخدام بيانات التدريب السابقة

ويتم تأجيل عملية التعلم حتى يتم طلب التنبؤ في العينة الجديدة.

•غير البارامترية: في KNN، لا يوجد شكل محدد مسبقا لدالة التخصيص .



- **مميزات خوارزمية كي- أقرب جار:**

- إنه سهل التنفيذ.

- وقت التدريب صفر (أو القليل جدا)

- ليس لديه افتراضات حول كيفية توزيع البيانات.

- من السهل جدا فهم خوارزمية KNN للمبتدئين في التعلم الآلي.

عيوب خوارزمية كي- أقرب جار:

يجب دائما ضبطه على K، الأمر الذي قد يكون معقدا في بعض الأحيان.

تكلفة الحسابات عالية بسبب حساب المسافة بين نقاط البيا نات لجميع عينات التدريب.

لا يعمل بشكل جيد على البيانات غير المتوازنة. لذلك ، قد يتم تجميع البيانات القليلة بشكل غير صحيح .

خلاصة:

هذه الطريقة موجودة في فئة طرق **التعلم غير البارامترية**

لا يتم تعلم بأي معلمات حول البيانات. و بدون توليد نظريات حول البيانات الأساسية ،

تحد الطرق غير البارامترية من قدرتنا على فهم كيفية استخدام المصنف للبيانات.

من ناحية أخرى ، يسمح هذا للمتعلم بإدخال الأنماط الطبيعية بدلا من محاولة تحويل البيانات إلى شكل وظيفي مسبق وربما متحيز.

مثال حول تحديد الوزن العادي او الناقص

Consider a dataset having two variables: height (cm) & weight (kg) and each point is classified as Normal or Underweight

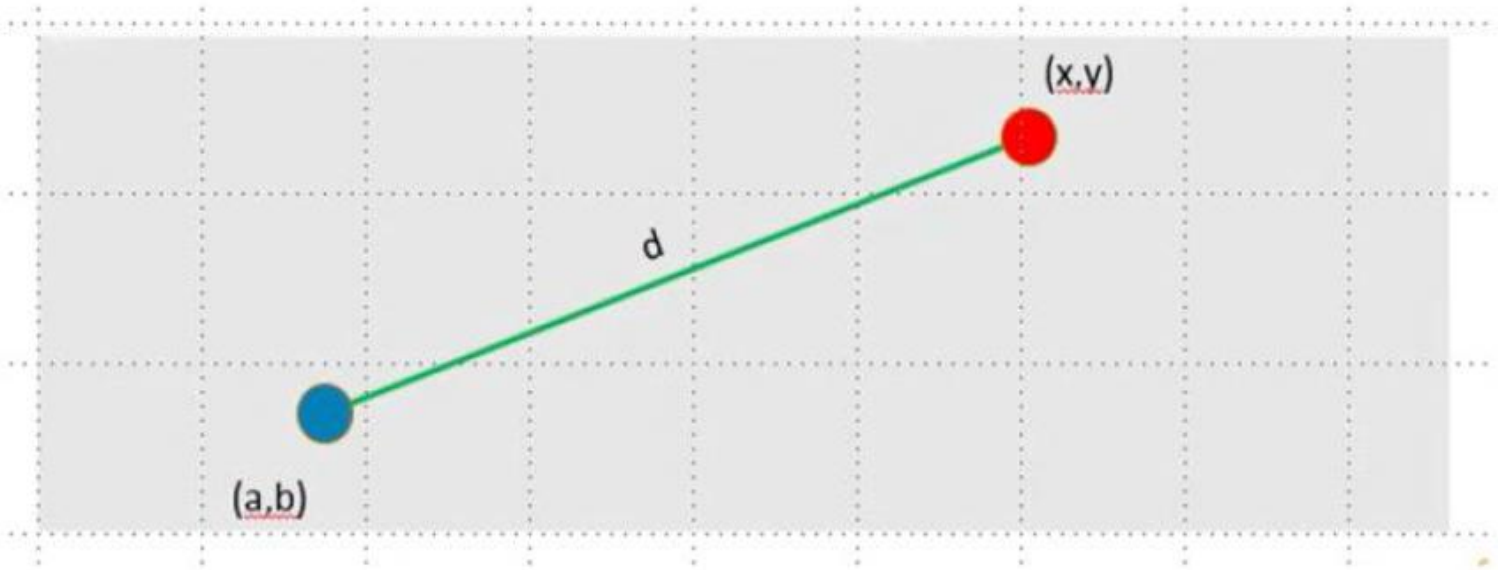
Weight(x2)	Height(y2)	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

- نريد تحديد وزن شخص ما اذا كان عاديا او لا

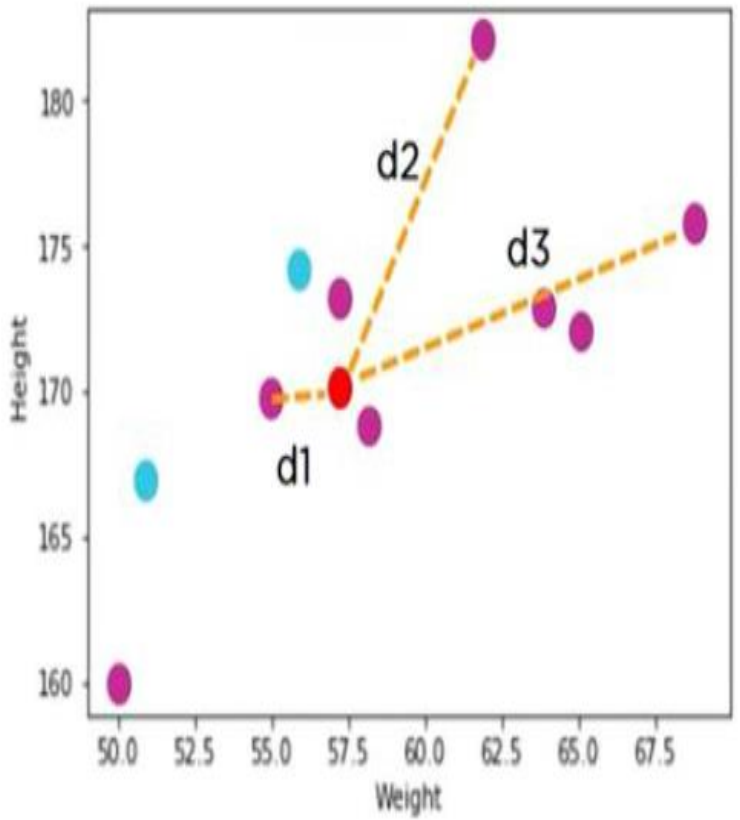
57 kg	170 cm	?

- حساب المسافة الاقليدية

$$dist(d) = \sqrt{(x - a)^2 + (y - b)^2}$$



نقوم بحساب المسافات بالطريقة الاقليدية بنسبة للاوزان



$dist(d1) = \sqrt{(170-167)^2 + (57-51)^2} \sim= 6.7$

$dist(d2) = \sqrt{(170-182)^2 + (57-62)^2} \sim= 13$

$dist(d3) = \sqrt{(170-176)^2 + (57-69)^2} \sim= 13.4$

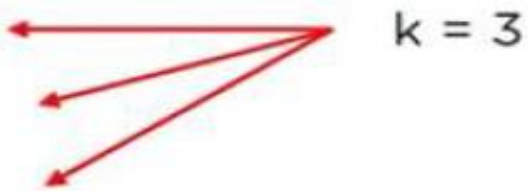
Similarly, we will calculate Euclidean distance of unknown data point from all the points in the dataset

Where (x1, y1) = (57, 170) whose class we have to classify

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

لما K=3 فان هذا الشخص يعتبر ذو وزن عادي
نلاحظ ان الصفة السائدة هي الوزن العادي لهذا اثرت على عملية الاختيار، لو اخترنا K=9

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2



57 kg	170 cm	?

إعداد البيانات للاستخدام مع k-NN

عادة ما يتم تحويل الميزات إلى نطاق قياسي قبل تطبيق خوارزمية k-NN. الأساس المنطقي لهذه الخطوة هو أن صيغة المسافة تعتمد بشكل كبير على كيفية قياس الميزات. على وجه الخصوص ، إذا كانت بعض الميزات لها نطاق قيم أكبر بكثير من غيرها ، فإن قياسات المسافة ستهيمن عليها بشدة الميزات ذات النطاقات الأكبر في مثالنا السابق نلاحظ ان الطول هيمن على الوزن.

خوارزمية KNN وتقليص حجم الخصائص(Feature Scaling)

تعتمد خوارزمية KNN على مقاييس المسافة مثل المسافة الإقليدية للعثور على أقرب الجيران لنقطة بيانات جديدة. عندما تحتوي الخصائص على مقاييس مختلفة تمامًا، يمكن للخصائص ذات المقاييس الأكبر أن تهيمن على حساب المسافة، مما يؤدي إلى اختيار جار غير دقيق وإمكانية ضعف أداء النموذج.

التطبيع مقابل التقييس

التسوية او التطبيع :(Normalization)

يحول الخصائص إلى نطاق معين، عادةً بين 0 و 1 (التقليص بالحد الأدنى والحد الأقصى) أو -1 و 1) التطبيع باستخدام توزيع Z-score).وهذا مناسب لخوارزمية KNN عندما تريد التأكد من أن جميع الخصائص تساهم بشكل متساوٍ في حساب المسافة، بغض النظر عن وحداتها أو توزيعاتها الأصلية.

التقييس :(Standardization)

يحول الخصائص إلى متوسط يساوي 0 وانحراف معياري يساوي 1. يُفضل هذا بشكل عام للخوارزميات مثل KNN التي تستخدم المسافة الإقليدية، حيث يعمل على محاذاة البيانات وتطبيع الانتشار، مما يؤدي إلى حسابات مسافة أكثر دقة. ومع ذلك، يفترض توزيعًاGaussian ا (عاديًا) للخصائص.

•**التطبيع او التسوية:** الطريقة التقليدية لإعادة قياس ميزات k-NN هي **الحد الأدنى للتطبيع**. تقوم هذه العملية بتحويل ميزة بحيث تقع جميع قيمها في نطاق بين 0 و 1. صيغة تطبيع ميزة هي كما يلي:

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

•بشكل أساسي ، تطرح الصيغة الحد الأدنى من الميزة *X* من كل قيمة وتقسم على مدى *X*.

•يمكن تفسير قيم المعالم التي تمت تسويتها على أنها تشير إلى أي مدى ، من 0 بالمائة إلى 100 بالمائة ، انخفضت القيمة الأصلية على طول النطاق بين الحد الأدنى والحد الأقصى الأصليين.

•**التقييس:** يسمى التحول **توحيد درجة z**. تطرح الصيغة التالية القيمة المتوسطة للميزة *X* ، وتقسم الناتج على الانحراف المعياري ل *X*:

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - \text{Mean}(X)}{\text{StdDev}(X)}$$

هذه الصيغة، التي تستند إلى خصائص التوزيع الطبيعي ، تعيد قياس كل قيمة من قيم المعلم من حيث عدد الانحرافات المعيارية التي تقع فوق أو أقل من القيمة المتوسطة. تسمى القيمة الناتجة درجة **z**. تقع درجات z في نطاق غير منضم من الأرقام السلبية والإيجابية. على عكس القيم الطبيعية ، ليس لديهم حد أدنى وحد أقصى محدد مسبقا.

- في خوارزمية KNN (جارة أقرب نقطة او اقرب جار)، متى نستخدم التقييس (standardization) ومتى نستخدم التطبيع (normalization):

نستخدم التطبيع عندما:

لدينا خصائص ذات نطاقات مختلفة على نطاق واسع أو توزيعات غير معروفة. يضمن التطبيع مساهمة جميع الخصائص بشكل متناسب.

إذا كانت البيانات تحتوي على أصفار (خاصة بالنسبة للخصائص ذات القيم الثنائية)، فقد يكون التطبيع بالحد الأدنى والحد الأقصى أكثر ملاءمة لتجنب القيم السلبية بعد التقليص.

نستخدم التقييس عندما:

لدينا افتراض معقول بأن الخصائص لديك تتبع توزيعًا طبيعيًا .(Gaussian) يعمل التقييس على محاذاة البيانات وتطبيع الانتشار، مما يحسن حسابات المسافة في.KNN

خلاصة:

بالنسبة إلىKNN ، يجب أعطاء الأولوية للتقييس إذا كان الشك في وجود توزيع Gaussian في خصائص البيانات.

إذا كنا غير متأكدين من التوزيع أو لدينا خصائص ذات مقاييس مختلفة جدًا، فيجب استخدام التطبيع لضمان المساهمة المتساوية من جميع الخصائص.

تطبيقات خوارزمية KNN

- على الرغم من بساطة هذه الفكرة ، فإن طريقة الجيران الأقرب قوية للغاية. لقد تم استخدامها بنجاح من أجل:

1. تطبيقات الرؤية الحاسوبية، بما في ذلك التعرف البصري على الأحرف(التعرف على الكتابة باليد) والتعرف على الوجه في كل من الصور الثابتة والفيديو
2. توقع ما إذا كان الشخص سيستمتع بفيلم أو وصلة موسيقية
3. تحديد الأنماط في البيانات الجينية ، ربما لاستخدامها في الكشف عن بروتينات أو أمراض معينة

- بشكل عام ، تعد أقرب مصنفات مجاورة مناسبة تماما لمهام التصنيف ، حيث تكون العلاقات بين الميزات والفئات المستهدفة عديدة أو معقدة أو يصعب فهمها للغاية ، ومع ذلك فإن العناصر من نوع الفئة المماثلة تميل إلى أن تكون متجانسة إلى حد ما.

- هناك طريقة أخرى لوضعها وهي القول: إنه إذا كان من الصعب تعريف المفهوم ، لكنك تعرفه عندما تراه ، فقد يكون أقرب الجيران مناسبا. من ناحية أخرى ، إذا كانت البيانات صاخبة وبالتالي لا يوجد تمييز واضح بين المجموعات.

kNN classification syntax

using the `knn()` function in the `class` package

Building the classifier and making predictions:

```
p <- knn(train, test, class, k)
```

- `train` is a data frame containing numeric training data
- `test` is a data frame containing numeric test data
- `class` is a factor vector with the class for each row in the training data
- `k` is an integer indicating the number of nearest neighbors

The function returns a factor vector of predicted classes for each row in the test data frame.

Example:

```
wbcd_pred <- knn(train = wbcd_train, test = wbcd_test,
                 cl = wbcd_train_labels, k = 3)
```

Thank you