林罡北

Follow 355 Followers About

[Machine Learning] kNN分類演算法

林罡北 Jul 10, 2017 · 5 min read

> 最近在學*Machine Learning*～
> 因為要學的東西太多了
> 文章主要是希望能夠用比較淺白的文字筆記
> 讓自己能夠在忘記時回來快速複習一下

What is kNN algorithm?

KNN演算法全名為「**k nearest neighbor**」
翻成中文意思就是「k個最近的鄰居」
雖然中文看得懂，但光看名字是還不太能了解它的意義啊啊啊

首先，讓我們思考一下這個問題

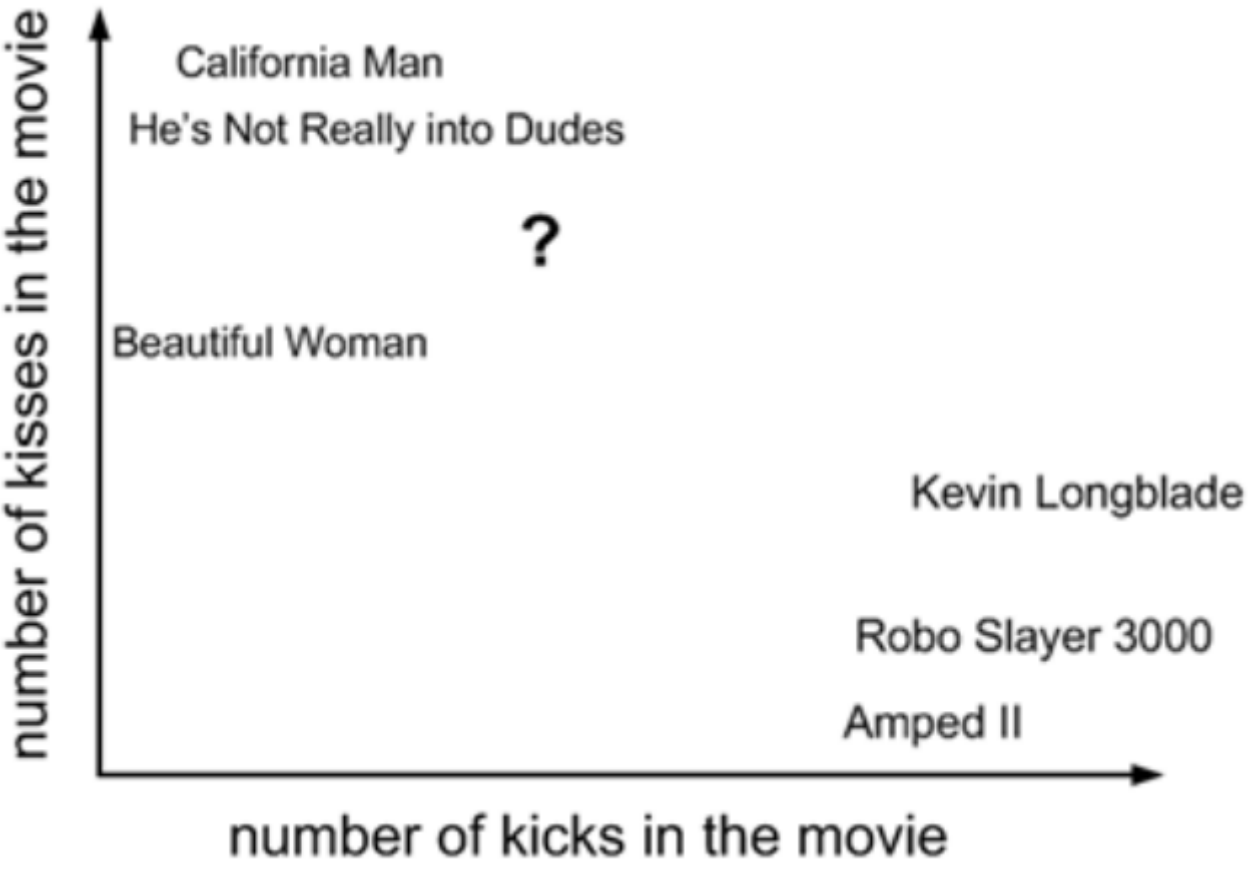
在**Movie Category**這個表格中，紀錄著電影名稱、電影中出現幾次踢腳的動作、電影中出現幾次親吻的畫面、電影的分類這些資料
而所有的電影被歸類成「Romance」與「Action」這兩類
但有一筆資料（？）還沒被分類

Movie title	# of kicks	# of kisses	Type of movie
California Man	3	104	Romance
He's Not Really into Dudes	2	100	Romance
Beautiful Woman	1	81	Romance
Kevin Longblade	101	10	Action
Robo Slayer 3000	99	5	Action

?	18	90	Unknown

Movie Category

接下來，我們用二維平面來呈現「kicks」與「kisses」跟電影的關係



Movie Category On 2-dimensional plane

那麼「？」這部電影的電影類型應該是什麼?

大部分的人在回答這個問題的方法大概是這樣：

1. 看一下「？」附近的電影的類型是什麼
2. 發現Beautiful Woman、California Man的type都是Romace
3. 所以合理推測「？」應該也是Romace類型的電影

而上述找出答案的方法就是kNN演算法的重點
但是對於電腦來說不是「看一下」，而是「算一下」

簡單來說，kNN做的事情就是

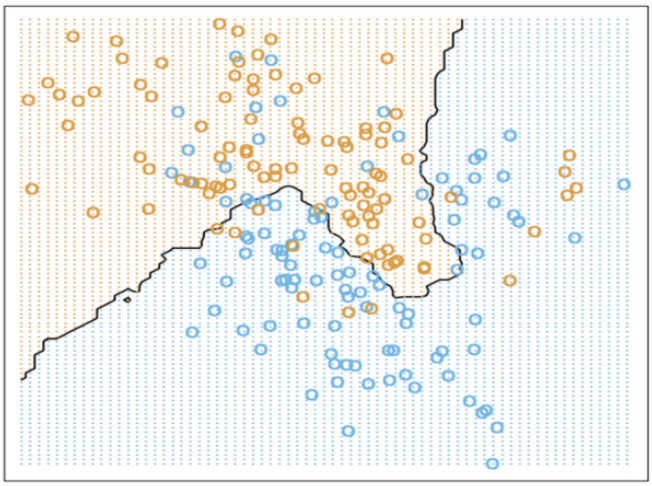
> *Given a test instance* i*, find the* k *closest neighbors and their labels*
> *Predict* i*'s label as the majority of the labels of the* k *nearest neighbors*

簡單的來說就是KNN就是看離你最近的K個點
然後看哪個類別的點最多就把自己也當成那個類別

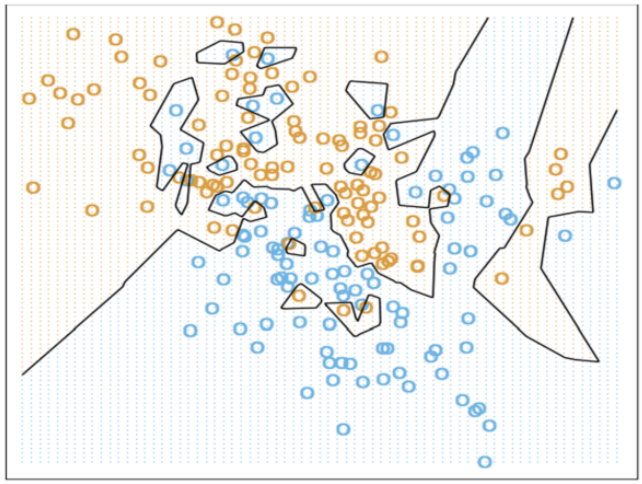
現在我們知道了kNN的意義，接下來要討論一些問題

How to select k?

俗話說，好的k帶你上天堂
選擇一個好的k會讓training出來的model有足夠的彈性
能夠避免掉Overfitting 與 underfitting



15-nearest neighbors



1-nearest neighbors

What if *k* is an even number?

如果k是一個偶數，則有可能碰到無法直接決定類別的時候
需要再去針對該情況做exception handling

例如k=4，結果在i附近的點有2個type A的資料與 2個type B的資料
導致無法判斷結果

What if *k* equals 1?

如上面的右圖，k=1時會造成Overfitting

What if *k* equals the number of the training instances?

如果k=n的話，預測結果一定是資料數量最多的那一類
等於喪失kNN預測的效果

結論：
當k=1 的時候容易Overfitting training data
而k很大的時候容易underfitting training data

優點：精度高、對異常值不敏感、無資料輸入假定。
缺點：時間複雜度高、空間複雜度高，訓練模型依賴訓練集資料且不可丟棄。
適用資料範圍：數值型和標稱型。

kNN need training or not？

> *KNN*屬於機器學習中的監督式學習***(Supervised learning)***，不過一般來說監督式學習是透過資料訓練***(training)***出一個***model***，但是在*KNN*其實並沒有做*training*的動作。*KNN*一般用來做資料的分類，如果你已經有一群分好類別的資料，後來加進去點就可以透過*KNN*的方式指定新增加資料的分類。

引用自：35成群: 輕鬆聊之KNN演算法 — blogger

在網路上有多爬文的朋友可能有看過上面這一段，也有可能看過其他說法，一方認為kNN不需要Training，反之，另一方則認為kNN是需要Training的

kNN的Training主要指的是以某種資料結構儲存各點的關係，達到加速搜尋鄰近k個鄰居點的效果(Ex. ball-tree)

當然如果不做Training當然也可以，只是就變成做real time search這樣

Reference

輕鬆聊之KNN演算法
沒有想像中簡單的簡單分類器 Knn