DigitalHouse > Coding School

DATA SCIENCE

UNIDAD 2
MÓDULO 4

Introducción a la clasificación

Septiembre 2017

Introducción a la clasificación

Objetivos

1. Definir el concepto de clasificación

2. Explicar el funcionamiento del algoritmo k-vecinos más cercanos

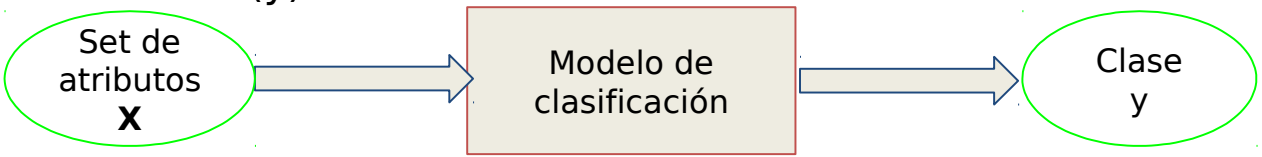
3. Implementar el algoritmo KNN utilizando la biblioteca scikit-learn

Introducción a la clasificación

- Clasificar es la tarea de asignar objetos a categorías que están predefinidas

- Existen numerosos ejemplos de clasificación como clasificar emails en función de su contenido en spam/no spam, identificar tipos de células cancerosas en malignas/benignas , etc.

- El modelo de clasificación es una función que mapea un set de atributos (X) a una clase (y):



Introducción a la clasificación

- Los datos que usamos como input en una clasificación es una colección de **instancias**
- **X** es el set de atributos. Los atributos representan propiedades que pueden tomar valores continuos o discretos
- **y** es es un atributo “especial” denominado **clase** (o atributo “target”).
- En una clasificación habitualmente la clase **debe ser discreta**

Definición formal de Clasificación:

Es la tarea de “entrenar” a una función F, tal que sea capaz de mapear y asignar, en un set de atributos, una clase predefinida.

K- vecinos más cercanos (KNN)

- Entre diferentes algoritmos de clasificación , uno de las más utilizados es K-nearest neighbors (kNN.)

- Este algoritmo se comenzó a usar desde el principio de los 70’s como un método **estadístico no paramétrico** debido a que **evita hacer suposiciones sobre la población** observada.

- Se ha utilizado tanto para estimaciones estadísticas como el reconocimiento de patrones

K- vecinos más cercanos (KNN)

- Idea básica: un nuevo ejemplo se va a clasificar en la clase más frecuente a la que pertenecen sus **K vecinos más cercanos,** por la mayoría de los votos de sus vecinos

- La métrica de la vecindad de vecinos es una **medida de similitud**

  - Input = ejemplo / instancia no conocida
  - Output = (label) membresía a un grupo predeterminado

- Pertenece al grupo de los métodos “non generalizing” o “instance-based” porque simplemente “recuerda” todos los datos de entrenamiento y con eso particiona el espacio para asignar la clasificación.

Introducción a la clasificación

Hiperparámetro K

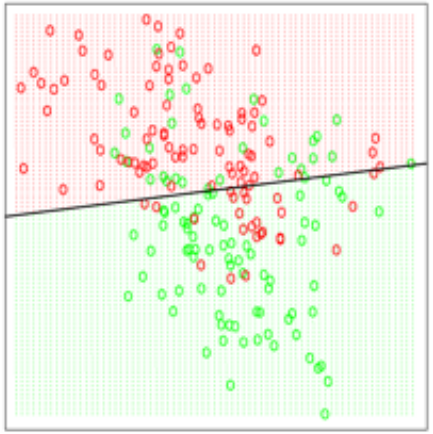
- El valor de K, es uno de los parámetros **a ajustar** en el algoritmo
- Valores elevados de K, disminuye el “ruido” y aumentan la precisión, pero no es garantía de un buen modelo.
- Un método para determinar un valor óptimo de K es mediante cross-validation , corriendo pruebas sobre un conjunto de datos independiente
- Generalmente, un valor óptimo de K se encuentra entre **3 - 10**, sin embargo la cantidad exacta depende fuertemente de los datos y no hay reglas generales.

El **hiperparámetro K** de este algoritmo es el que regula el **trade-off entre sesgo y varianza**

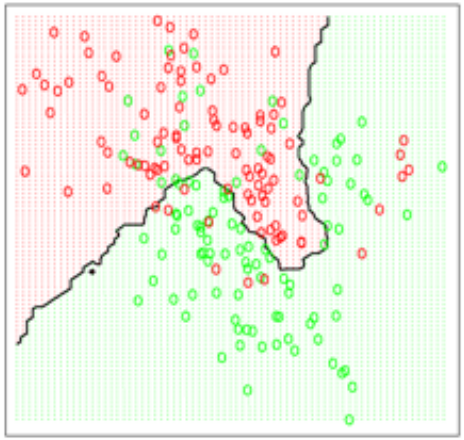
Introducción a la clasificación

Ejemplos de overfitting y underfitting

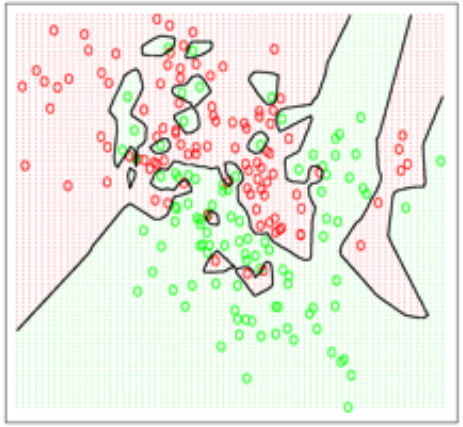
Underfit



Good fit



Overfit

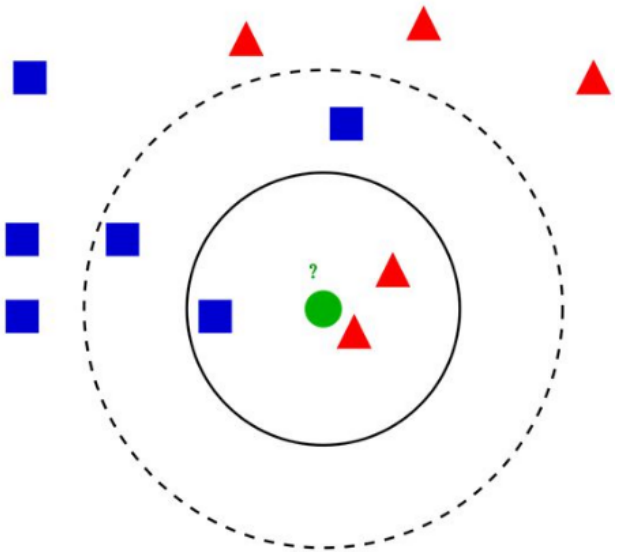


Introducción a la clasificación

Ejemplo gráfico

Intuitivamente clasificar el nuevo ejemplo X (esfera verde)

- hallar el ejemplo más similar a X´
- predecir su clase



DEMO: lógica del algoritmo k-NN

Demo: lógica del algoritmo kNN

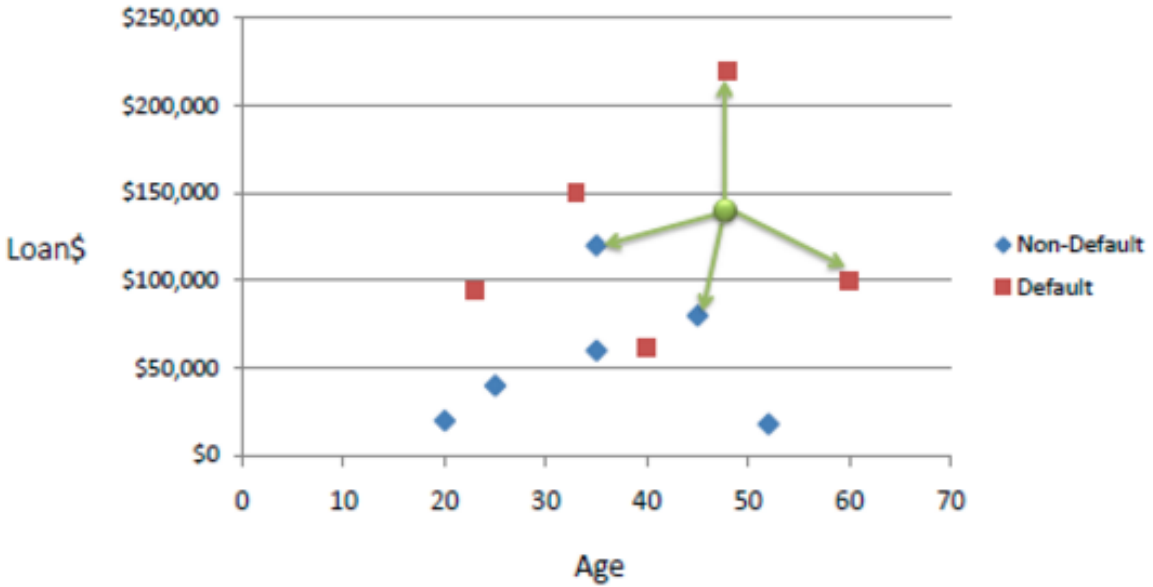
Pseudocódigo:

```
algoritmo knn(x)
   comienzo
       looping a través de todos los puntos de los datos en entrenamiento,
       encontrar los K puntos más cercanos to X
      asignar f(x) = clasificación mayoritaria entre los k puntos más cercanos
   fin
```

Demo: lógica del algoritmo kNN

Observemos los datos siguientes correspondientes a préstamos bancarios:

- tenemos las variables: monto de préstamos, edad de usuario
- queremos conocer si pertenece o no al grupo “default”



Demo: lógica del algoritmo kNN

- El ejemplo tiene como variables Edad = 48 y Monto = $142.000
- La distancia elegida es la Euclídea
- Las categorías de la variable target son Y (yes) y N (no)
- Si elegimos k=1 , el vecino más cercano al ejemplo no conocido (en el set de entrenamiento) tiene asignada la clase Y (grupo)
- Si probamos y cambiamos y asignamos a k el valor 3, los vecinos más cercanos van a ser dos de clase **Y** y uno de clase N.
- Por lo tanto, como resultado al nuevo ejemplo se lo asignará al grupo default.
- En otras palabras se le asignará la clase Y

Demo: lógica del algoritmo kNN

- Si probamos y asignamos a k el valor 3, los vecinos más cercanos van a ser dos de clase Y y uno de N.
- Por lo tanto, como resultado al nuevo ejemplo se lo asignará al grupo default.
- En otras palabras se le asignará la categoría Y

Prácticas

Conclusión

1- ¿Qué es una clasificación?
2- ¿Qué son las variables explicativas y la variable target?
3- ¿Cómo se definen las categorías de una clase?
4- ¿Cómo funciona el algoritmo de KNN?
5- ¿Cómo evaluaríamos un modelo?