Subject: Probability and Statistics

Class XVII_2: K-Nearest Neighbors

UdeA

Bioengineering

Francisco José Campuzano Cardona

Bioengineerer, MSc in Engineering

Los métodos de aprendizaje automático estudiados hasta ahora, son métodos estadísticos clásicos.

Posterior a estos métodos, han aparecido otros más sofisticados que son guiados por los datos en sí, y no buscan imponer un modelo absoluto para la predicción. Estos métodos son conocidos como ***métodos de aprendizaje automático estadístico,*** y son los métodos en los que nos concentraremos en los restante del curso. Aquí entran algunos métodos de regresión y clasificación, métodos de agrupamiento y métodos de reducción de dimensión.

UdeA

K-Nearest Neibhbors (KNN): (K vecinos más cercanos)

Este algoritmo es bastante simple, que puede ser usado como un método de clasificación o de regresión. Las *features* deben ser numéricas.

1. Se ubica el punto que se desea predecir en el espacio, y se define sus k vecinos más cercanos.
2. **En clasificación**: Se determina cual es la clase dominante de entre esos vecinos, y se asigna esa clase.
3. **En regresión**: se determina el promedio de los k vecinos y se asigna esa valor.

UdeA

Distancia

Para determinar los k vecinos más cercanos al punto, se deben tener alguna forma de medir esa distancia:

Distancia euclidiana

$$\rho = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2 + \cdots + (x_{n1} - x_{n2})^2}$$

Distancia Manhattan

$$\rho = |x_{11} - x_{12}| + |x_{21} - x_{22}| + \cdots + |x_{n1} - x_{n2}|$$

UdeA

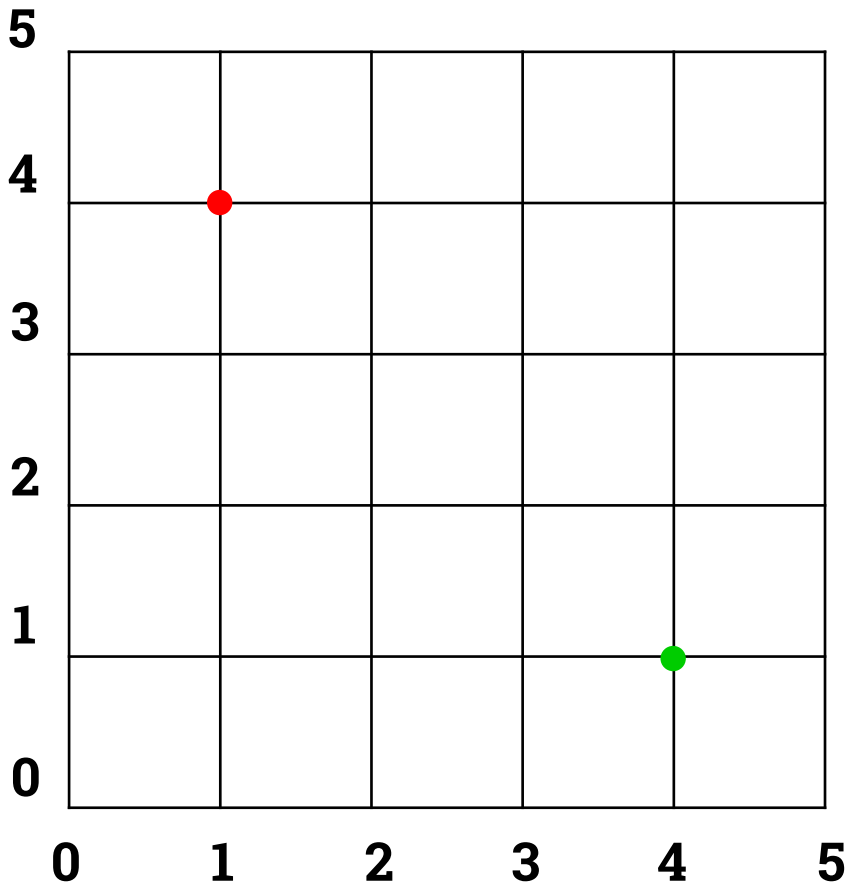
Encontremos la distancia entre el punto rojo y verde:

Distancia euclidiana

$$\rho = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2 + \cdots + (x_{n1} - x_{n2})^2}$$

Distancia Manhattan

$$\rho = |x_{11} - x_{12}| + |x_{21} - x_{22}| + \cdots + |x_{n1} - x_{n2}|$$



UdeA

¿Pero qué pasa con la distancia si los x tienen escalas muy diferentes?

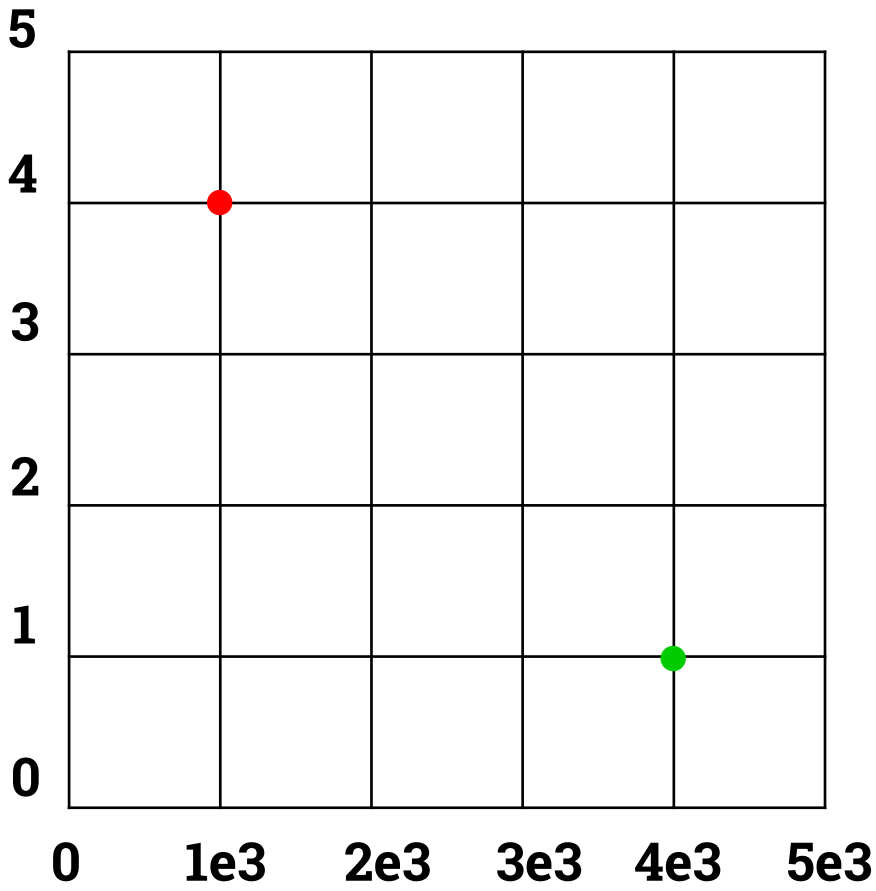
Distancia euclidiana

$$\rho = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2 + \cdots + (x_{n1} - x_{n2})^2}$$

Distancia Manhattan

$$\rho = |x_{11} - x_{12}| + |x_{21} - x_{22}| + \cdots + |x_{n1} - x_{n2}|$$

La distancia estaría guía casi exclusivamente por los x con magnitud grande.



UdeA

Estandarización o Normalización

Por el problema anterior, a las x se les debe normalizar o estandarizar de la siguiente manera:

$$z = \frac{x - \bar{x}}{\sigma}$$

Los valores de los x transformados con este proceso dan lugar a lo que se conoce como los *z-scores*. De este modo, todas las variables estarán en la misma escala, aportaran de forma similar a la distancia.

Este procedimiento es muy importante.

UdeA

Elección del valor K

No existe una regla para esto. En general los valores pueden oscilar entre 1 y 20. Un numero impar es recomendado para evitar empates. Se puede usar validación cruzada, (curva de validación) (scoring)

La elección más simple sería K=1, pero esta elección sería muy sensible al error, entonces en general un K muy bajo no es útil si los datos tienen mucho error. (Arrojaría un modelo sobreajustado, de alta varianza).

Por su parte, un K muy grande, tendería a clasificar todo como la clase más prevalente, aunque este no sea el caso, es decir pierde la capacidad de diferenciación local. Piense en el caso don K = n con n el número de *records.* (Arrojaría un modelo de alto sesgo).

UdeA

KNN como generador de características

KNN es un método muy simple e intuitivo, pero no suele tener un desempeño impecable si se compara con otros métodos más sofisticados.

En la práctica, KNN puede ser utilizado para introducir características al set de datos.

Se ajusta un modelo KNN y se usa para predecir la clase de cada *record* del set, y la clase predicha se usa como una nueva característica, para ser usado en un modelo más sofisticado.

UdeA

Key Ideas for K-Nearest Neighbors

- *K*-Nearest Neighbors (KNN) classifies a record by assigning it to the class that similar records belong to.
- Similarity (distance) is determined by Euclidian distance or other related metrics.
- The number of nearest neighbors to compare a record to, *K*, is determined by how well the algorithm performs on training data, using different values for *K*.
- Typically, the predictor variables are standardized so that variables of large scale do not dominate the distance metric.
- KNN is often used as a first stage in predictive modeling, and the predicted value is added back into the data as a *predictor* for second-stage (non-KNN) modeling.

Subject: Probability and Statistics

UdeA

Bioengineering

¡Thanks!

Francisco José Campuzano Cardona

Bioengineering. MSc in Engineering