Estadística III para Ingenieros de Sistemas

Jose Daniel Ramirez Soto 2023
jdr2162@columbia.edu

Agenda

- Exploración de datos
  - correlación
- modelos de analitica (machine learning-ML)
  - K-means
  - K-nearest neighbors (KNN)
  - Regresión y regresión Logística
  - Árboles
- Caso de uso de modelos supervisado y no supervisado para detectar fraude en el sisben
- Métricas de evaluación un modelo
- Conclusiones

Exploración de datos. Correlación

Permite investigar las relación lineal entre dos variables

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

r=0 variables independientes.
r= +1, > 0.4 (y depende de la variable x), explica cambios positivos
r= +1, <-0.4 (y depende de la variable x), explica cambios negativos

* A Course of Machine Learning http://ciml.info/

Exploración de datos. Correlación

Subject	Age x	Glucose Level y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

= 0.5298

* A Course of Machine Learning http://ciml.info/

modelos de analitica o machine learning

Machine learning es un conjunto de métodos o algoritmos que **entienden o aprenden de los datos** sin ser explícitamente programados. Se dividen en:

- **Supervisado**: Con datos de la variable objetivo. Tipo variable: $f(x) = \tilde{y}, min((y - \tilde{y})^2)$
  - continua: utilizamos modelos de regresión.
  - categórica: utilizamos modelos de clasificación.
- **No supervisado** : No existen etiquetas. Si el objetivo es: $f(x) = \tilde{x}, min((x - \tilde{x})^2)$
  - Crear grupos: Cluster
  - Reducir dimensionalidad o embedding : Representar datos o categorías en números
- **Reinforcement learning**: Aprender del entorno, explorar y explotar

* A Course of Machine Learning http://ciml.info/

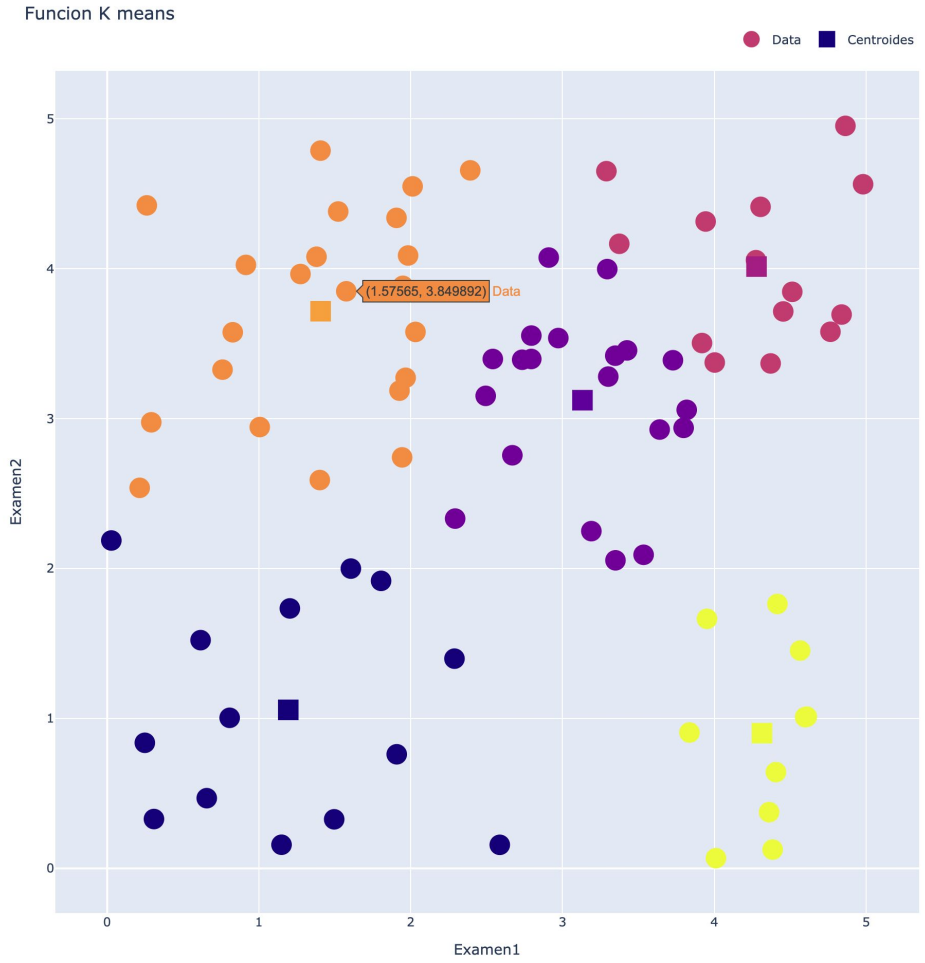
No Supervisado, K-means (Clusters)

Cuando no se tiene etiquetas, el clustering puede ayudarnos a entender cómo crear agrupaciones de los datos basados en su similaridad.

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

Parámetro K es el numero grupos que quiero obtener del cluster

Ejemplo: utilizando datos estudiantes, identificar grupos de estudiantes. ¿Cuándo puedo medir el error?



* A Course of Machine Learning http://ciml.info/dl/v0_99/ciml-v0_99-ch15.pdf

No Supervisado, K-means (Clusters), Algoritmo

1 Asignar una etiqueta random para cada registro

2 Calcular el promedio de los datos con la misma etiqueta, es el centroide y existen K-centroides.

3 Asignar a cada registro la etiqueta del centroide más cercano

4 repetir desde el paso 2 hasta q no exista ningún cambio de etiqueta

* A Course of Machine Learning http://ciml.info/