



Práctica 3 - Clustering y reducción de dimensiones

La práctica 3 de la asignatura *Machine Learning* consiste en la resolución de un problema de **clustering y reducción de dimensiones**. El problema a resolver se describe a continuación.

Requisitos de la práctica

La práctica consta de los siguientes entregables:

1. Una libreta de Python en la cual se realizará el desarrollo de la práctica. La libreta incluirá todo el código de las operaciones realizadas para el preprocesamiento, entrenamiento y validación de los modelos, así como **una sección de conclusiones** en la cual se interpretarán no solo las medidas de calidad obtenidas por el modelo sino también el modelo en sí mismo. Para facilitar el proceso de corrección, **todos los datos deberán ser cargados desde una URL externa**, y no desde el almacenamiento local de la libreta utilizada.
2. Un vídeo corto de **no más de 15 minutos** en el cual se presenten los desarrollos, resultados y conclusiones obtenidas para el problema resuelto. Para la entrega del vídeo, se subirá a la nube y se entregará el enlace.

Será imprescindible realizar los siguientes procesos durante la resolución del problema:

- Preprocesamiento adecuado del conjunto de datos, teniendo en cuenta que hay datos categóricos, numéricos y textuales.
- Visualización del conjunto de datos usando las técnicas de reducción de dimensiones vistas en clase.
- Entrenamiento y validación de los diferentes modelos de clustering vistos en clase. Para aquellos modelos que necesiten a priori el número de clústeres a detectar, tratarlo como un hiperparámetro más.
- Optimización con Grid Search de los hiperparámetros existentes.
- Validación y comparación de los modelos mediante la medida de calidad *silhouette*, ya que no contamos con *ground truth*.
- De cara a evaluar justamente, la semilla para todos aquellos métodos estocásticos será: `random_state=1337`.

Descripción del problema

Disponemos de un conjunto de datos con información sobre los perfiles de la web de contactos OkCupid.

El problema de aprendizaje no supervisado a resolver consiste en determinar qué perfiles de la red social son compatibles entre sí; a fin de cuentas, se trata de una web de contactos.

El conjunto de datos

El conjunto de datos contiene unos 60000 perfiles de usuario, incluyendo información sobre 31 características:

- `age` : edad
- `status` : estado de la relación
- `sex` : sexo
- `orientation` : orientación sexual
- `body_type` : tipo de cuerpo

- `diet` : dieta seguida por el usuario
- `drinks` : ¿bebedor?
- `drugs` : ¿consumidor de drogas?
- `education` : máximo nivel educativo alcanzado
- `ethnicity` : etnia
- `height` : altura
- `income` : ingresos anuales (en dólares americanos \$)
- `job` : empleo/industria
- `last_online` : fecha de la última conexión
- `location` : lugar de residencia
- `offspring` : preferencia con respecto a los hijos
- `pets` : preferencia con respecto a las mascotas
- `religion` : preferencias religiosas
- `sign` : signo del zodiaco
- `smokes` : ¿fumador?
- `speaks` : idiomas
- `essay0 - essay9` : campos de texto libre que el usuario ha llenado en orden arbitrario. Estos campos se corresponden con las siguientes preguntas:
 - Acerca de mí / Auto resumen
 - Objetivos actuales / aspiraciones
 - Mi regla de oro / Mis rasgos
 - Probablemente podría ganarte en / Talento
 - La última serie que he visto / Hobbies
 - Un día perfecto / Momentos
 - Yo valoro / Necesito
 - La cosa más privada que estoy dispuesto a admitir / Secretos
 - Lo que realmente estoy buscando / Citas

Puedes encontrar el conjunto de datos en un fichero comprimido en el siguiente [enlace](#).