VISUALIZACIÓN DE DATOS

Curso 2020/2021 Borja Arroyo Galende

Propuesta de caso práctico

Elección de los datos

La visualización de datos se puede aplicar a numerosos ámbitos y bien distintos. En mi opinión, estos ámbitos son tales que no sabría hasta dónde llegar para poder optar a una buena calificación. He estado mirando kaggle y las opciones son infinitas y con un nivel de dificultad muy variado. Sin embargo, ante la falta de tiempo y por temor a quedarme corto, he decidido realizar trabajo práctico enfocado en el dataset propuesto por el equipo docente.

Los datos mencionados se componen de dos archivos en formato csv. A mí, personalmente, me resulta más interesante enfocarme en el conjunto de entrenamiento debido a ese atributo extra que en realidad es el más interesante.

Objetivos

En mi humilde opinión, los objetivos de la visualización en tareas de aprendizaje automático suelen ser siempre parecidos. Otro tema es el público objetivo al que se presentan los resultados. Estos objetivos a los que me refiero se enmarcan dentro del típico “pipeline” al que se enfrenta un experto en el área. Por tanto, se podrían definir varios casos de uso, pero siempre ligados a este hecho:

1. Realizar un primer estudio de los datos. Comprender cómo se distribuyen y si existe algún tipo de patrón objetivo en una primera aproximación.
2. Comprobar que el preprocesado y procesado están realizando una tarea adecuada en un enfoque iterativo. Por ejemplo, comprobando los filtros que se van aplicando.
3. Estudiar los resultados obtenidos en el amplio campo conocido como analítica, ya sea una analítica exploratoria: tratar de utilizar la visualización para la toma de decisiones, como el aprendizaje automático y la consiguiente evaluación de modelos.

Probablemente, exista alguna aplicación más, pero creo que estas son sin duda las más importantes en el pipeline del que hablo.

Por todo lo dicho, el objetivo que me propongo es el de realizar estos tres apartados desde un punto de vista orientado a la visualización, es decir, realizar un proyecto de análisis de datos utilizando primordialmente el medio visual. No obstante, en el punto 3, emplearé técnicas de reducción de la dimensionalidad (y probablemente de agrupamiento) para mejorar en gran medida las posibilidades.

Creo que mi elección está muy ligada al concepto de storytelling de los apuntes: utilizar las visualizaciones para contar la historia de los datos a través de un pipeline. Además,

suponiendo el caso de ser una persona que trabaje para la empresa en cuestión, creo que es la herramienta más poderosa en la que apoyar la toma de decisiones de la compañía.

Además de todo lo mencionado hasta ahora, me gustaría realizar el trabajo utilizando un notebook de jupyter (o más de uno si lo veo necesario) donde poder ir representando y explicando las visualizaciones de una forma más “dinámica” que si fuesen scripts de python con imágenes generadas a parte. Otro motivo por el que utilizar jupyter es su capacidad para el refinamiento sucesivo prácticamente interactivo del programador con sus resultados.

Planificación

Definir un cronograma a seguir es una tarea ardua con el poco tiempo del que dispongo. A grandes rasgos mi idea es la siguiente:

1. Segunda quincena de abril: primer análisis exploratorio de los datos. Buscar distribuciones conocidas, relaciones entre variables, etc.
2. Primera quincena de mayo: definir métodos de reducción de la dimensionalidad que ayuden a la visualización.
3. Si dispongo de tiempo, tratar de realizar una segmentación (agrupamiento) de algún atributo con un interés particular.