IIC2433 — Minería de Datos

Apuntes: tipos de aprendizaje, manejo de datos

1. Problema: Aprender desde los datos

El problema principal que vamos a tratar en este curso es el siguiente.

> Tenemos un espacio $\mathcal{E}$ en el que existen entidades. Una función f mapea a cada entidad de este espacio a otro espacio, de una o pocas dimensiones (por ejemplo, a un número, a un booleano, a una descripción).
>
> Tenemos, además, muchos ejemplos de pares $(x, f(x))$.
>
> ¿Cuál es la función?

Este problema tiene varias componentes.

1.1. Datos (esta clase)

.

¿Qué son exactamente las entidades? ¿Cómo se representan en nuestro espacio? Si son personas, entonces podemos imaginarnos que tenemos todos sus datos demográficos (edad, sexo, etc), y luego la representación de cada persona es una tupla en una tabla de atributos (como las de un censo). Pero quizás tenemos una red social, que además nos dice que entidad es amiga de qué entidad, y aquí la representación se parece más a un grafo. Por otro lado, si las entidades son imágenes, podemos pensar que las entidades se representan como un arreglo con información de cada uno de sus pixeles. Si es texto, son sus palabras, o sus caracteres.

- Vamos a ver que en datos, importa mucho como elegimos representar las entidades. Volviendo a las personas como tuplas en una tabla de atributos. Algunas variables son numéricas (edad), pero otras son categóricas (mujer, hombre, no_binaria). Y otras son texto (dirección), o tal vez son tuplas en si mismas (dia,mes,año).
- Casi siempre nos va a convenir pre-procesar los datos. Las variables categóricas se pueden pasar a booleanos con un one-hot encoding (ver el laboratorio para más información sobre one-hot encodings). Los textos se deben limpiar, pero además quizá nos conviene producir un *embedding* que lleve cada palabra o cada sílaba a un número. Las tuplas con muchos atributos son complejas de trabajar, así que posiblemente nos va a convenir *reducir la dimensionalidad*.

- De forma más general, en este curso también tocaremos el problema de *encontrar nuevos espacios* desde los cuales sea más fácil aprender nuestra función. Por ejemplo, en vez de pensar en secuencias de textos, los transformamos a secuencias de números que cumplen ciertas propiedades. O reducimos la dimensionalidad de una matriz de miles de dimensiones, quedándonos con algunas dimensiones, no necesariamente las mismas de antes pero quizás otras que resulten de transformaciones algebraicas. Estos nuevos espacios son llamados *espacios latente*. O quizás no reducimos la dimensionalidad, pero generamos una dimensión adicional que captura algunas propiedades que nos van a permitir restringirnos a funciones más simples.

1.2. Funciones

.

¿Cómo quiero que se vea mi función? Puede ser que asigne un valor numérico a las entidades (por ejemplo, las entidades son jóvenes de 17/18 años y la función es su puntaje en la PAA/PSU/PTU de matemáticas). Puede que me asocie imágenes a categorías (esta imagen tiene un semáforo, esta un semáforo y un perro, esta nada). O a valores booleanos (un algoritmo que prediga si un banco debe otorgar crédito a una persona).

- El problema de averiguar la función se hace más fácil cuando restringimos el espacio de las mismas. Por ejemplo: encontrar la función lineal que más se parezca a la que tengo. Esta restricción, o "hipótesis", da pie a miles de algoritmos estadísticos o de aprendizaje de máquina distintos, cada uno con sus ventajas y desventajas.

- Es importante saber por qué quiero adivinar la función. Hay muchas posibilidades:

  - Para predecir: Imaginemos que quiero *predecir* el precio del bitcoin mañana. Las entidades los días, y tengo información financiera de cada día (precio del dólar, del euro, de otras criptomonedas). La función toma esa información y me entrega cual es el precio del bitcoin mañana. En este caso, quiero saber la función para poder aplicarla hoy, y saber si me conviene comprar o vender bitcoin.
  - Para clasificar: Imaginemos que quiero *clasificar* todas las imágenes que tienen un semáforo. Las entidades son las imágenes, con información de cada pixel, y la función me entrega un 1 si esa imagen tiene un semáforo, y un cero si no. En este caso, me interesa saber la función para llamarla cada vez que se abre una ventana de captcha/recaptcha.
  - Para explicar: Imaginemos que quiero *entender* qué es lo que más influye en el puntaje de la PTU de los jóvenes. Cada alumno es un vector de varias variables demográficas, y la función asigna a cada variable el puntaje de la PTU. Conozco la función para años anteriores, pero no para el 2021. En este caso, no me importa mucho predecir el puntaje de los alumnos, si no que me interesa saber cuál de todas las variables es la que más incide en el resultado. Por ejemplo: ¿importa el liceo/colegio de donde viene? ¿Importa si tomó o no un preuniversitario? ¿importa

si tuvo clases presenciales?. En este caso, saber la función me va a dar luces de estas respuestas, las que son importantes para economistas o cientistas sociales (por ejemplo, en el ministerio de educación).

- Para producir un ranking: Como en Netflix: las entidades son los usuarios, de los que nuevamente tenemos datos, y un día en particular. La función podemos pensar que le asigna a cada nueva película las ganas de que el usuario quiera verla ese día. En ese caso, me interesa saber la función para recomendarle a cada usuario las películas que sé que tiene hartas ganas de ver.