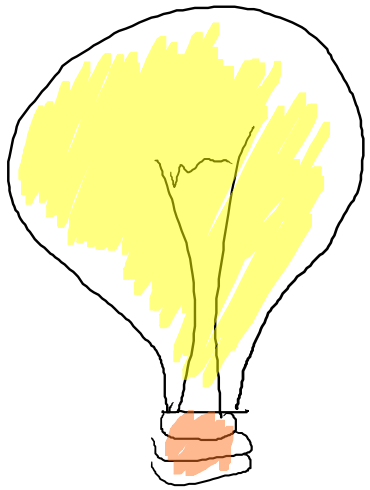
Análisis Multivariado en Ciencias de Datos y Estadística

Instituto Tecnológico Autónomo de México

Primavera 2017

Clase 1: Introducción

¿De qué es esta clase?

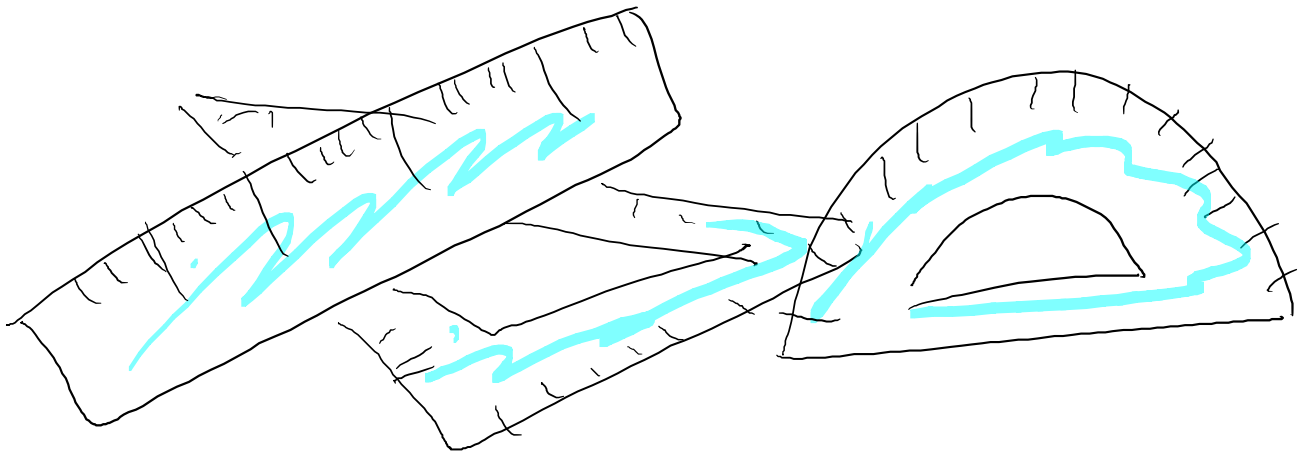


Resumen del curso

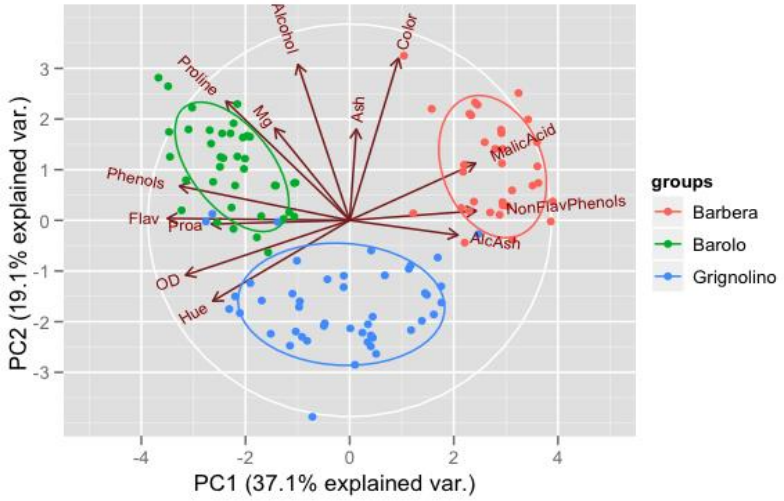
Una imagen dice más que mil palabras.... (aunque una tabla no está tan mal)

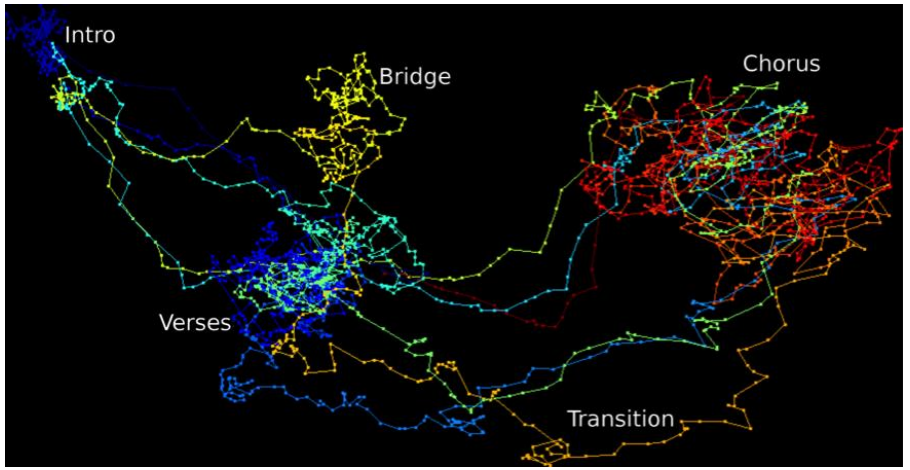
Ojos que no ven corazón que no siente…

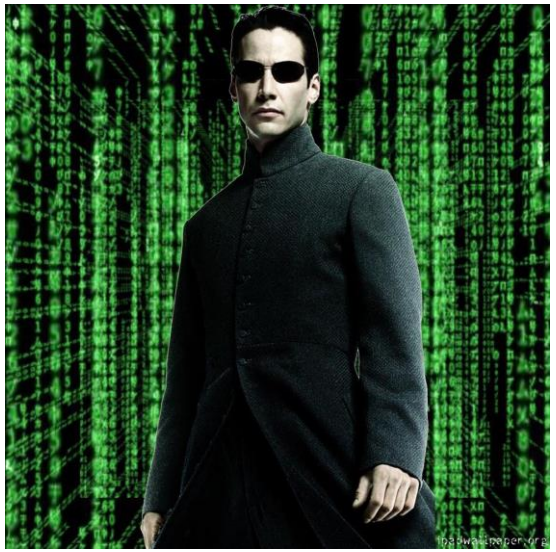
- En este clase veremos métodos para visualizar y explorar datos
- *Advertencia*: No veremos técnicas de limpia y preprocesamiento,
- Es una clase acerca de la *geometría, matemática y estadística* detrás de los datos y de cómo explotarla



Por ejemplo, convertir…

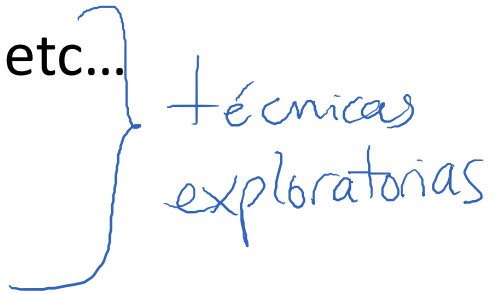






¿Qué herramientas vamos a usar?

- El campo del “análisis multivariado” es un muy amplio y cubre muchos temas de otras áreas (¿quién hace análisis univariado?....)
- Las técnicas que vamos a ver aquí tienen un enfoque geométrico y se centran en herramientas como…
  - Análisis de similitudes, covarianzas entre individuos, atributos, etc…
  - Reducción de Dimensionalidad
  - Análisis de conglomerados (clusters)
  - Resúmenes “topológicos”
- Retos:
  - Entender las matemáticas de los datos
  - Trabajos con datos de muchos tipos: numéricos, categóricos, etc…



¿Qué otras herramientas?...

- Aunque esta es una clase matemática… ustedes son científicos de datos, por lo que también estaremos usando:
  - ggplot
  - The tidyverse (dplyr, tidy, etc…)
  - Creación y documentación de paquetes de R con devtools
  - etc
- Siéntanse libres de reproducir algunos de los métodos del curso en otros lenguajes y compártanlos con el grupo!

Evaluación del curso

- Un componente importante de la evaluación del curso es la participación de los estudiantes, por lo que la asistencia es fundamental y deberán colaborar con tareas asignadas cada clase por el profesor.
- Tendrán una semana para escoger su equipo de trabajo, no podrán cambiar de equipo hasta que todos los equipos hayan tenido al menos una participación.
- La dinámica concreta la discutiremos más adelante
- La evaluación tentativa será de la siguiente forma:
  - Contribución grupal al git del grupo 10%
  - Tareas individuales semanales 20%
  - Tareas grupales semanales 20%
  - Proyecto final grupal 20%
  - Video grupal 30%

I. Tipos de Datos y Espacios Matemáticos

Datos

- Comencemos con lo básico: para nosotros, una base de datos es un arreglo rectangular donde las filas representan individuos y las columnas variables

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & & \\ \vdots & & & \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

n = # individuos

p = # variables

- X^j
- X_i
- X_{ij}

II. Espacios Matemáticos

- Objetos de un mismo tipo viven en un mismo “espacio”
  - E.g., Estrellas -> Galaxias
- Los matemáticos nos gusta hablar de espacios de cualquier cosa
- Los individuos viven en el espacio muestra, espacio de individuos
- Las variables viven en el espacio de variables, espacio de atributos o *feature space*
- El nombre usado depende de la disciplina (Machine Learning, Estadística, etc…)

¿Pero que son estos espacios?

- Pues todo depende del tipo de datos! Pueden ser muy bonitos, numéricos y sencillos, o pueden ser complicados
- Nuestro objetivo es saber la geometría de estos espacios, e.g., ¿qué individuos se parecen entre sí? ¿Qué variable se parecen o son independientes?
- *Teaser*

	Peso	Altura	Sexo
Juan	72kg	1.7m	H
María	55kg	1.6m	M

¿Se parecen?

- Por eso repasaremos los tipos de datos

II. Clasificando Tipos de Datos

- Clasificar los tipos de datos no suena a la actividad más divertida, pero es importante! Pues es uno de los elementos clave para entender qué métodos usar y cuáles sirven y cuáles no!

Tipos de datos estadísticos

- A grandes rasgos tenemos los siguientes tipos
  - Binarias/Booleanas
    - E.g. YES/NO, HOMBRE/MUJER, TRUE/FALSE, 0/1
  - Datos categóricos o nominales: más categorías
    - E.g. País de origen, Etnicidad
  - Datos ordinales: como categóricos pero existe un orden
    - E.g. Malo/Bueno/Buenísimo
    - Es fácil confundirlos con los categóricos… pero tienen más estructura
    - También se confunden con números si se usan escalas como 0-7 (típico encuestas)
  - Datos numéricos
    - Con escala de razón: *los campeones, y=2*x significa que y es dos veces más que x*
      - *Continuas. E.g. Ingreso, tamaño, peso, precio, datos de conteo*
      - *Conteo: E.g. # accidentes*
    - Sin escala de razón/escala de intervalo: *just be careful…*
      - *E.g. temperatura, distancia relativa, latitud, longitud*
- Cualquier técnica depende del tipo de dato!
- Una base de datos que tiene varios tipos de datos se conoce como de base de tipo mixto
- Útil: https://en.wikipedia.org/wiki/Statistical_data_type

Tareas:

1. Investigar y ejemplificar cómo se representa cada tipo de dato en R
2. (Grupal) Investigar cómo se miden distancias entre vectores de un mismo tipo de datos (e.g., distancia entre dos vectores de escala de razón, distancia entre dos vectores categóricos)
    a. Booleanos
    b. Categóricos
    c. Ordinales
    d. Numéricos de escala de intervalo
3. (Grupal) Mostrar ejemplos en R y crear un rmarkdown de los puntos anteriores: usar las base de datos de adults de UCI

III.a La Geometría de los Datos

- Vamos a ver qué forma la geometría de los datos dependiendo de su tipo
- Vamos a comenzar con el caso más sencillo: datos numéricos continuos de escala de razón