Text Mining

Diego Garrido

Departamento de Ingeniería Industrial
Universidad de Chile

1 de Agosto de 2019

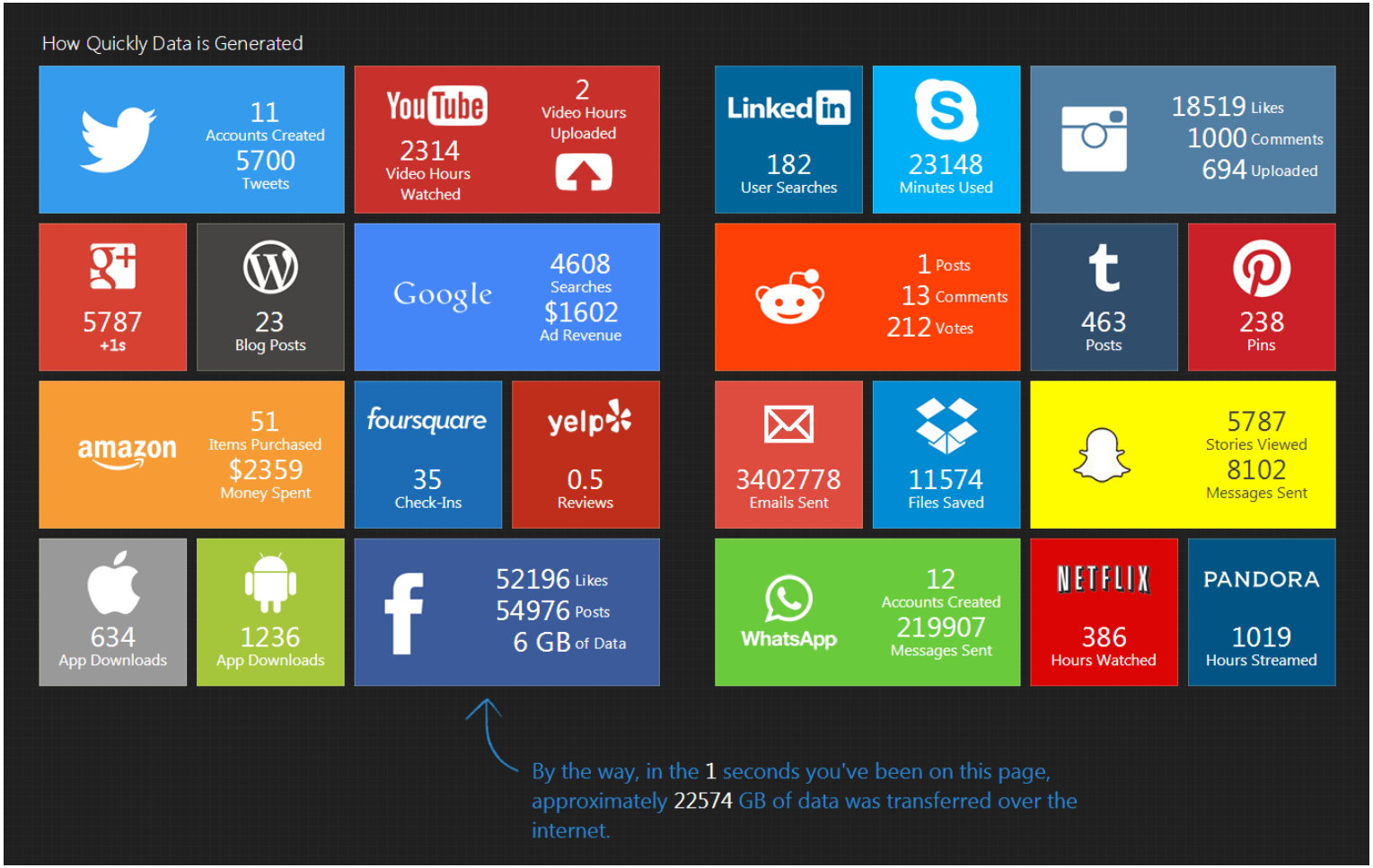
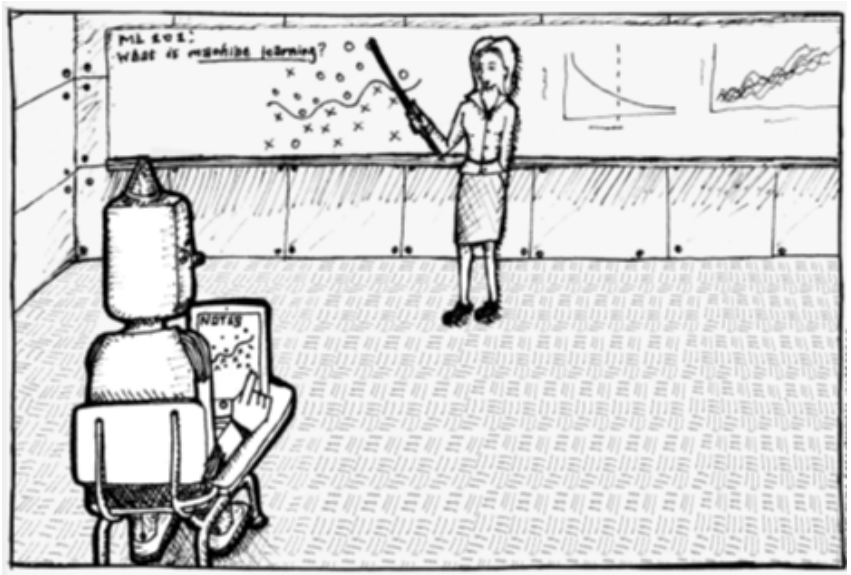


Fig. 1: Cuantos datos son generados por cada segundo en redes sociales. Fuente: Digital Information World, 2014.

"We are drowning in information and starving for knowledge." — John Naisbitt.

Necesidad de automatizar el análisis de datos.

def. Machine Learning: conjunto de métodos que pueden automáticamente extraer conocimiento a partir de experiencia (datos), y luego usar ese conocimiento para predecir el futuro por ejemplo.



Este enfoque evita que operadores humano formalmente especifiquen todo el conocimiento que las computadores necesitan para resolver una tarea.

Machine Learning: Taxonomía

Machine Learning es divido en tres categorías según el enfoque de aprendizaje:

1. **Aprendizaje Supervisado**: El objetivo es aprender un *mapping* $f(x)$ a partir de unos *inputs* x a un *output* y, dado un conjunto etiquetado de pares *input-output* de datos $D = \{(x_i, y_i)\}_{i=1}^{N}$, donde D es el conjunto de entrenamiento y N es el número de ejemplos de entrenamientos. Cada x_i es un vector en $\mathbb{R}^M$, los componentes de este vector son llamados características (*features*), atributos o covariables, en cambio el output y_i es un escalar, se le suele llamar variable respuesta, objetivo, dependiente o simplemente etiqueta.
2. **Aprendizaje No Supervisado**: Aquí solo existen *inputs* $D = \{(x_i\}_{i=1}^{N}$ y el objetivo es encontrar "patrones interesantes en los datos".
3. **Aprendizaje Reforzado**: Aprender a desempeñar una tarea en base a una serie de recompensas o castigos. Ejemplos de aplicaciones: Inteligencia Artificial (IA) que juegan ajedrez, GO o Starcraft.

Aprendizaje Supervisado: Clasificación

- Cuando la variable objetivo y_i es una variable categórica o nominal para un conjunto finito, es decir, $y_i \in \{1, ..., C\}$ el problema es conocido como **clasificación**.
- Ejemplos: predicción de fuga de clientes, riesgo de crédito, detección de fraude, diagnóstico de enfermedades, detección de objetos, detección de SPAM, análisis de sentimientos, etc.

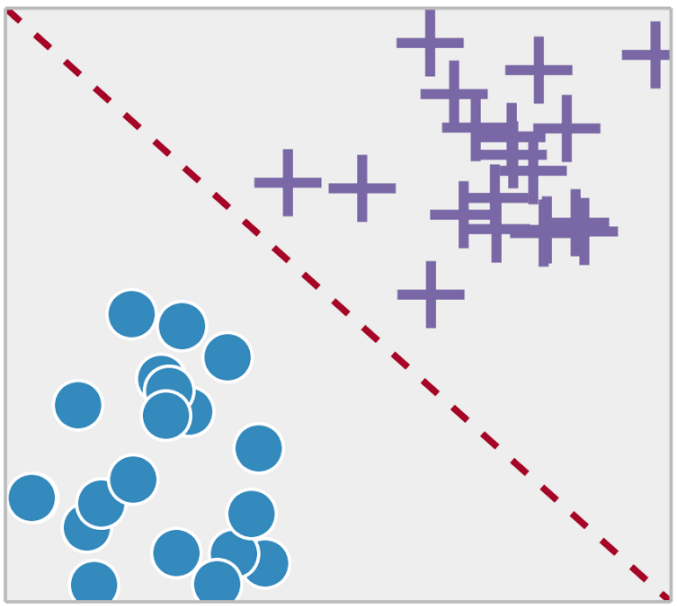


Fig. 2: Problema de clasificación binario con dos atributos.

Aprendizaje Supervisado: Regresión.

- Cuando la variable objetivo y_i es un real o continua en un intervalo el problema es conocido como **regresión**.
- Ejemplos: predicción de demanda, predicción del retorno de un activo financiero, predicción de temperatura, predicción de grado de invalidez, etc.

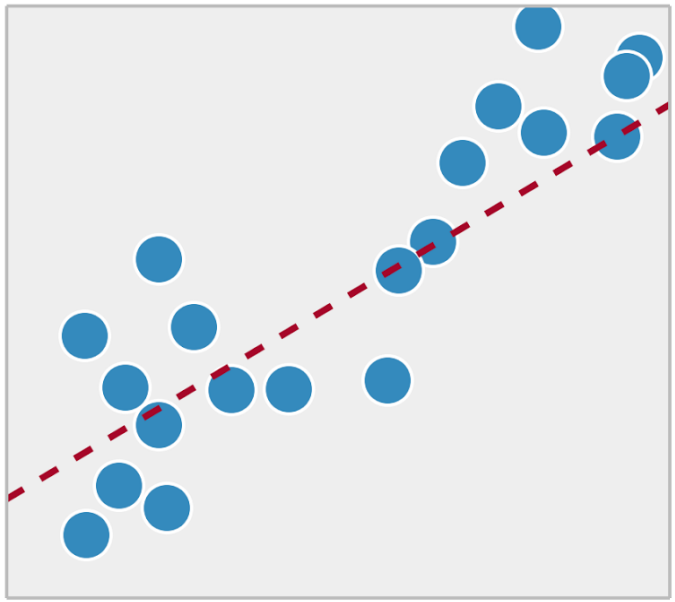


Fig. 3: Problema de regresión con un atributo.

- Consiste en agrupar los datos en grupos (*clusters*).

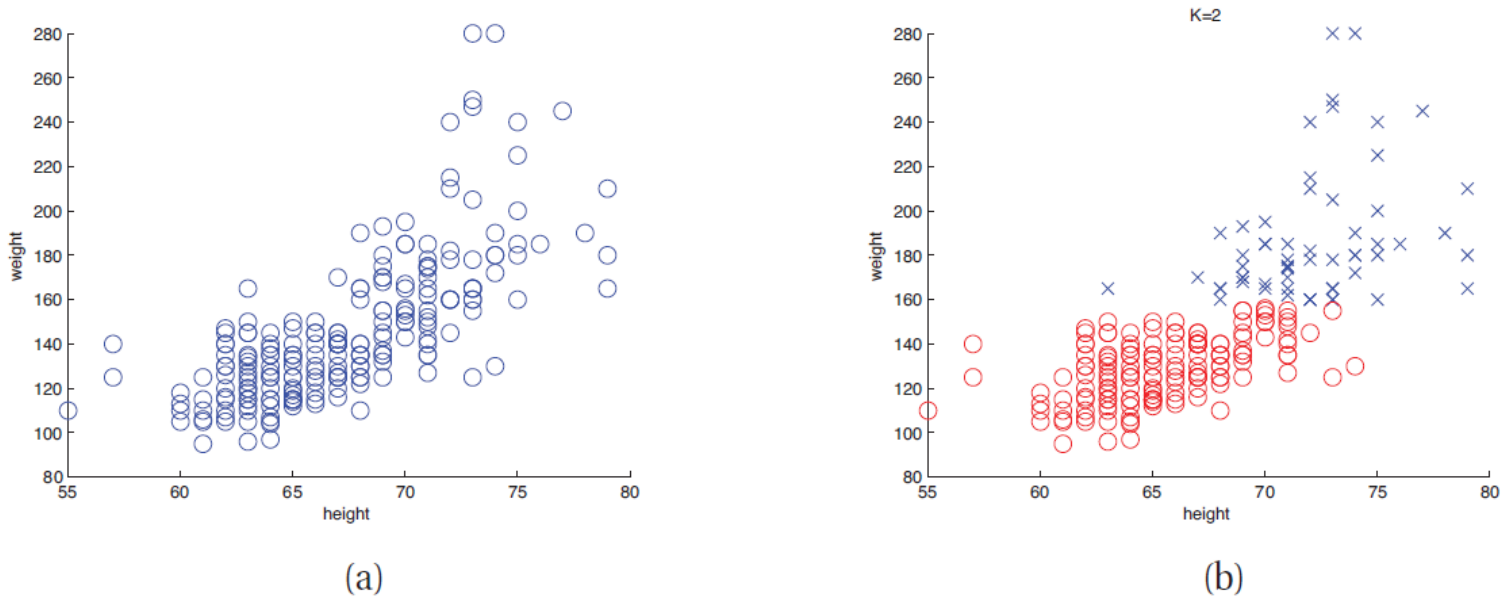


Fig. 4: (a) La altura y peso de algunas personas. (b) Una posible segmentación usando Kmeans con K=2 *clusters*.

- Se trata de codificar un input $x_i \in \mathbb{R}^M$ a un espacio de menor dimensionalidad.

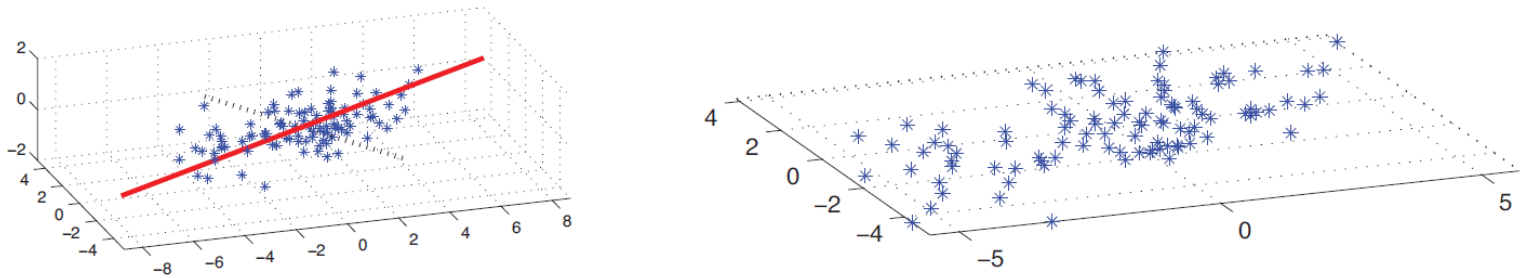


Fig. 5: (a) Un conjunto de puntos en 3 dimensiones. (b) Representación 2 dimensiones usando PCA.

- Consiste en completar datos desconocidos de una matriz, algunos ejemplos de aplicación son: imputación de contenido faltante en imágenes, filtro colaborativo (caso Netflix), análisis de canasta de mercado (qué productos un cliente llevará).

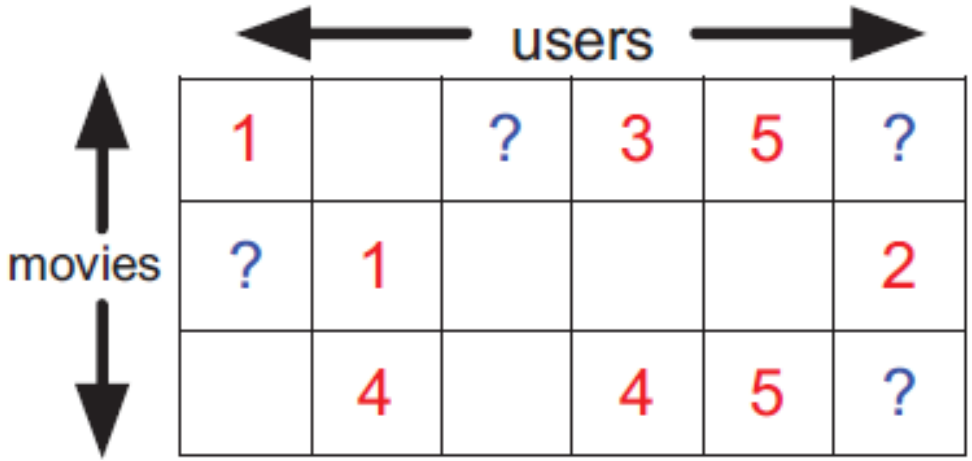


Fig. 6: Ejemplo de *movierating data* caso Netflix. Datos de entrenamiento en rojo, datos de prueba denotados por "?", celdas vacías son desconocidas.

- El objetivo es descubrir patrones en una base de datos que se puede representar como una estructura gráfica, por ejemplo en Social Network Analysis (SNA) una tarea común es encontrar autoridades de un tema en un grafo representado por personas (nodos) y la comunicación entre estas (arcos).

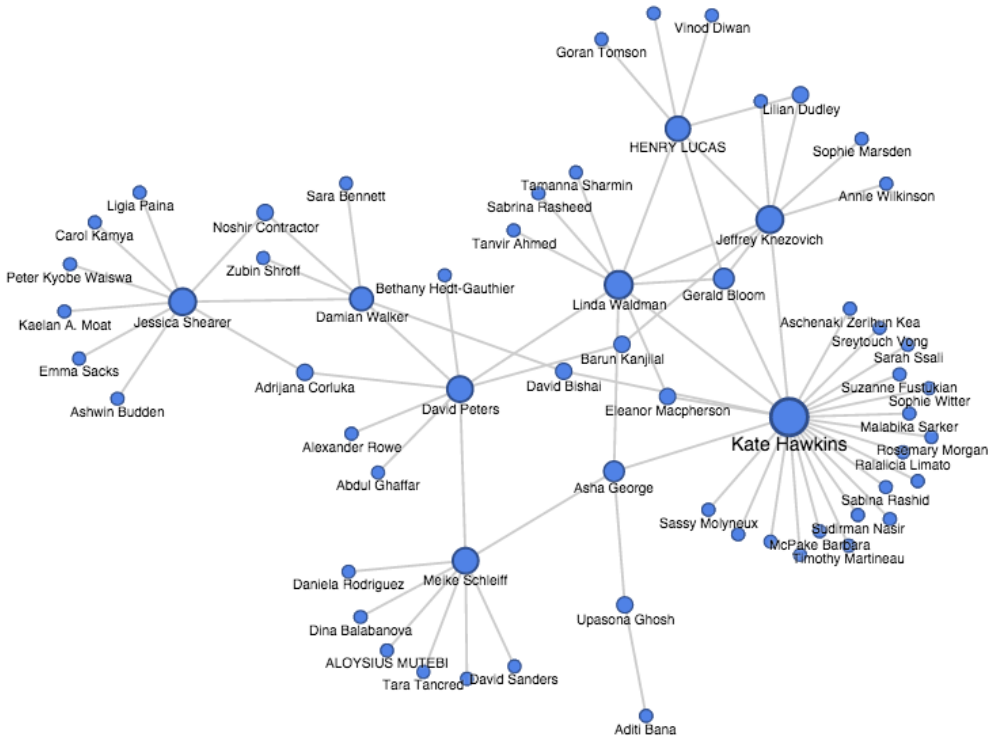


Fig. 7: Ejemplo de red social.

Evaluaciones

1. Laboratorios: El curso tendrá 8 laboratorios que deben ser realizados en grupos de 3 personas. Desde la clase auxiliar correspondiente al laboratorio se cuenta con un plazo máximo de 6 días para entregar este vía u-cursos, por ejemplo si un laboratorio se publicó un día Viernes, entonces, este debe entregarse a más tardar el Jueves de la siguiente semana antes de las 23:59.
2. Proyecto Semestral: El proyecto semestral cuenta con 3 Presentaciones (10%, 20%, 30%) y un Informe final (40%). Los grupos deben ser de 3 personas y debe realizarse con una empresa (leer instructivo publicado en material docente para más información).
3. CTP: El curso tendrá CTPs sobre lecturas y presentaciones, estos últimos serán al final de la clase.
4. Examen: El curso tiene un examen obligatorio que se realizará durante semana de exámenes.

.

Evaluaciones
Condiciones de aprobación y cálculo de nota final

Para aprobar el curso se necesita lo siguiente:

1. Nota control ≥ 4.0
2. Nota promedio laboratorios ≥ 4.0
3. Nota proyecto semestral ≥ 4.0
4. Nota CTP ≥ 5.0

La nota final del curso se determina de la siguiente manera:

1. Nota control 20%
2. Nota promedio laboratorios 40%
3. Nota proyecto semestral 20%.
4. Nota CTP 20%