IMT2200

INTRODUCCIÓN A CIENCIA DE DATOS

2022-2

REGRESIÓN LINEAL

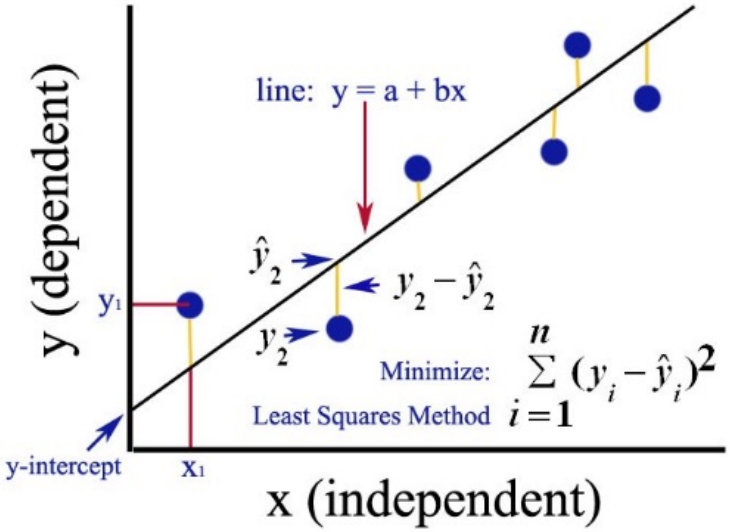
CLASE 17

REGRESIÓN LINEAL

Regresión lineal: modelo estadístico basado en la idea de que la relación entre dos variables puede explicarse mediante la fórmula

$$y = \beta x + \varepsilon, \text{ donde}$$

- y: variable dependiente
- x: variable independiente o predictora
- β: coeficientes de la regresión
- $\varepsilon = \hat{y} - \beta x$: error, o diferencia entre valor predicho y valor real.



- Inferencia ➔ la regresión lineal permite **entender mejor las relaciones entre variables**, y qué variable predictora es capaz de predecir una proporción importante de los cambios en la variable dependiente. Por ahora, *no buscamos predecir y*.

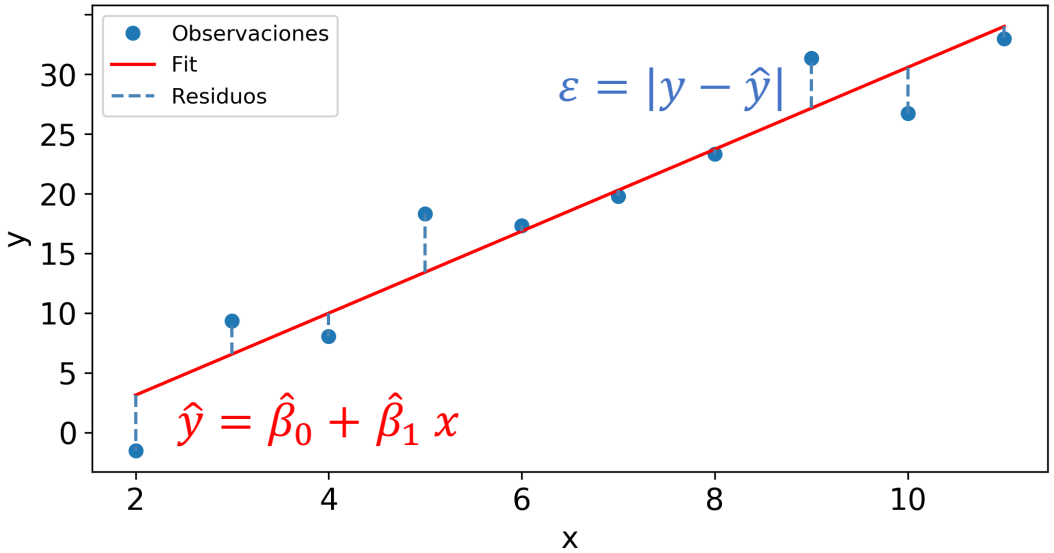
- Correlación ➔➔ causalidad: el que se detecte una correlación entre dos variables, no implica que haya una relación causal.

09:28

REGRESIÓN LINEAL SIMPLE

Si asumimos un modelo lineal para dos variables (x,y), entonces éstas cumplen las ecuaciones:

$$\left.\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}\right\} \quad y = X\beta + \epsilon$$



Error cuadrático medio:

$$\mathrm{MSE} = \mathrm{S}(\beta_0, \beta_1) = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2 = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \hat{\beta}_0 - \hat{\beta}_1 \mathrm{x})^2$$

Estimación de coeficientes de la regresión:

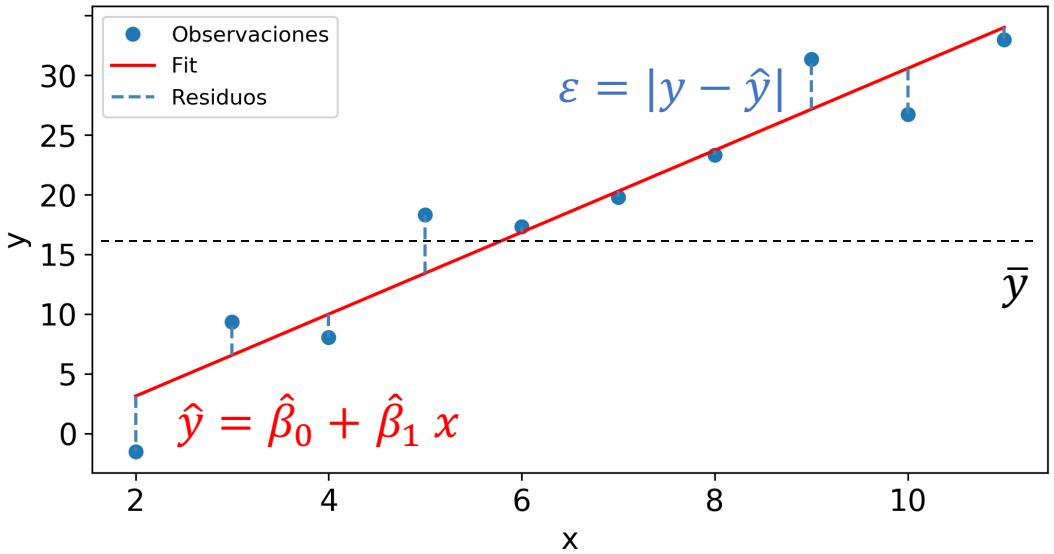
$$\hat{\beta}_0, \hat{\beta}_1 = \operatorname{argmin} \mathrm{S}(\beta_0, \beta_1) \Rightarrow \frac{\partial S}{\partial \beta_0} = 0, \frac{\partial S}{\partial \beta_1} = 0$$

09:28

REGRESIÓN LINEAL SIMPLE

¿Qué representan los coeficientes β?

- x: variable independiente o predictora
- y: variable dependiente o respuesta
- β_0: intercepto
- β_1: pendiente, representa el cambio en y para un cambio unitario en x



¿Qué tan bueno es el ajuste del modelo lineal?

- MSE: mean squared error ➔ se minimiza
- R^2: coeficiente de correlación, indica la fracción de la variación en y que queda explicada por la variación en x.
  - R^2=1 ➔ ajuste perfecto
  - R^2=0 ➔ ajuste ~ promedio de y

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

09:28

REGRESIÓN LINEAL MÚLTIPLE

¿Qué representan los coeficientes β?

- x: variable independiente o predictora
- y: variable dependiente o respuesta
- β: contribución de cada variable x al cambio en y

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_n x_{1p} + \epsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_n x_{2p} + \epsilon_2$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_n x_{2p} + \epsilon_n$$

$$y = X\beta + \epsilon$$

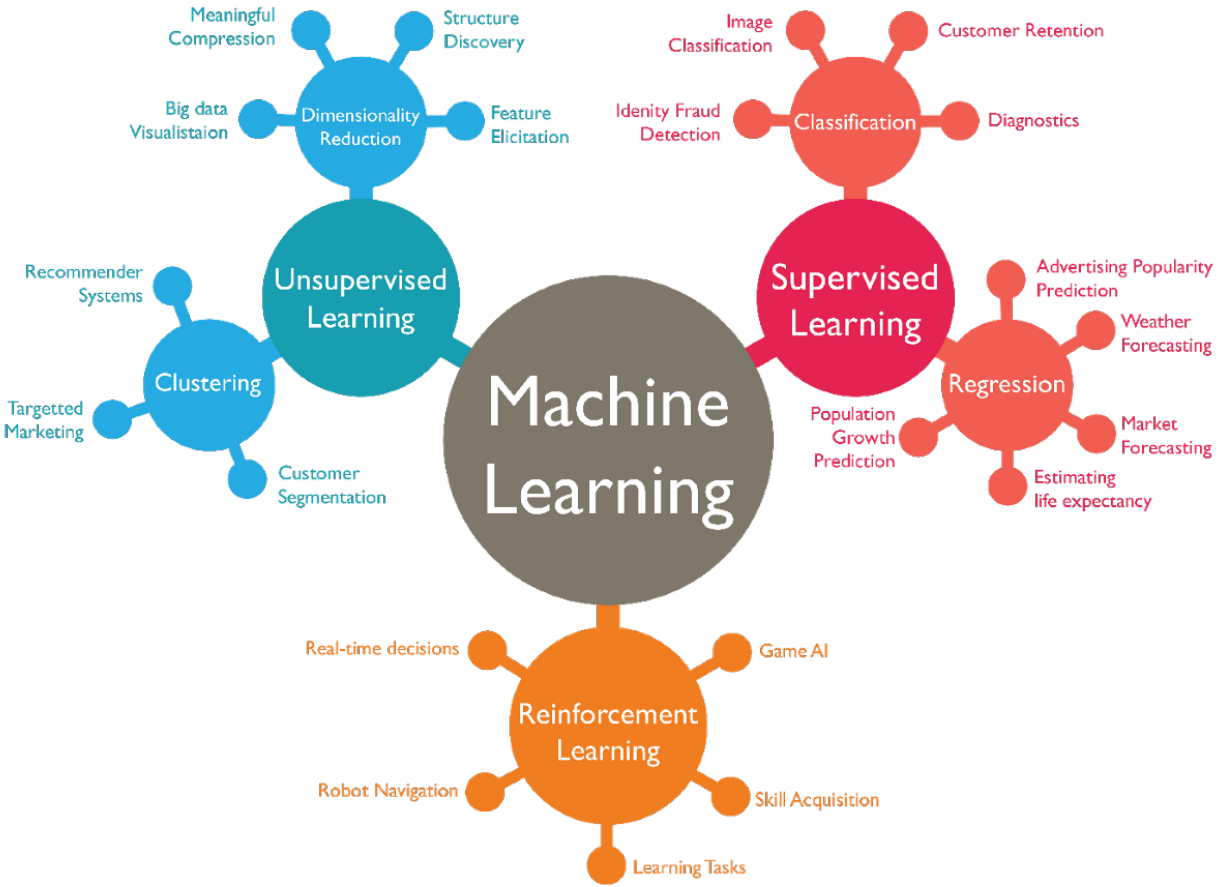
¿Qué tan bueno es el ajuste del modelo lineal? ➔ hay varias métricas

- MSE: mean squared error
- RMSE: root-mean squared error
- R^2: coeficiente de correlación, indica la fracción de la variación en y que queda explicada por la variación en x.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

09:28

ALGORITMOS DE ML



09:28

ALGORITMOS

Procedimiento, o conjunto de pasos o reglas para lograr una tarea (ordenar, buscar, clasificar, etc.)

Tipos de algoritmos en Ciencia de Datos:

- Algoritmos de preparación y procesamiento de datos (ing. de datos)
  - Ej: sorting, MapReduce, Pregel
- Algoritmos de optimización para estimación de parámetros
  - Ej: descenso de gradiente, Newton, mínimos cuadrados
- Algoritmos de aprendizaje de máquina: para predecir, clasificar, o clusterizar.

- Para una tarea dada, pueden proponerse múltiples algoritmos posibles
  - Hay un que puede identificarse como "mejor", en base a métricas de eficiencia y tiempo computacional.

- Desafíos:
  - Comprender qué tipo de algoritmo usar dependiendo del contexto del problema y las suposiciones de base.
  - Implementarlo

09:28

MACHINE LEARNING: TAREAS COMUNES

- Regresión: predicción de un **valor real** para cada ítem
- Clasificación: asignación de una **categoría** a cada ítem de un conjunto
- Clustering: **particionar** un set de ítems en subconjuntos homogéneos
- Ranking: aprender a **ordenar** ítems de acuerdo a algún criterio
  - Ej: buscadores web
- Reducción de dimensionalidad: **transformar** una representación de ítems en una representación con menos dimensiones, pero preservando algunas propiedades de la representación inicial.
  - Ej: preprocesamiento de imágenes digitales para tareas de visión de computador (computer vision)

Aprendizaje Supervisado

Métodos de Regresión

Predictores y Outcomes

Ejemplo:

Predecir el consumo de combustible de un auto a partir de sus características de diseño.

p predictores
j=1,2,...,p

X_j

n observaciones
i=1,2,...,n

	car_name	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2

variable a predecir
outcome
variable dependiente
Respuesta

variables usadas para la predicción
predictores
variable independiente
covariates

$Y = y_1, \dots, y_n$

$X = X_1, \dots, X_p$

$X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj}$

09:28

Predictores y Outcomes

- Para predecir Y, asumimos que se relaciona con X mediante una función desconocida f :

$$Y = f(X) + \varepsilon$$

- Problema de **inferencia** ⇒ encontrar $\hat{f}$, la estimación de f

- Problema de **predicción** ⇒ predecir el valor de Y para distintos sets de observaciones $(x_{i,1}, \dots, x_{i,p})$.

  - No nos interesa la forma de f, sino sólo las predicciones $\hat{y}_i$:

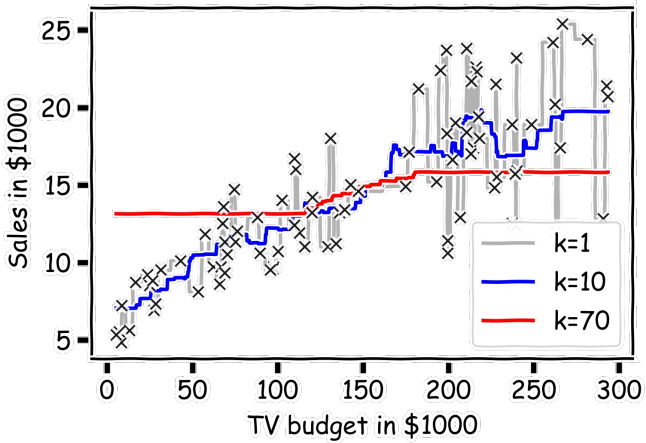
$$\hat{y}_i = \hat{f}(x_{i,1}, \dots, x_{i,p}).$$

Regresión k-Nearest Neighbors (kNN)

- Una forma simple de predecir una respuesta cuantitativa para una observación:
  - usamos el promedio de las respuestas a otras observaciones más cercanas a ella ➔ los k-nearest neighbors

$$\hat{y}_n = \frac{1}{k} \sum_{i=1}^{k} y_{n_i}$$

Donde $\{x_{n1}, \ldots, x_{nk}\}$ son las ***k*** observaciones más similares (cercanas) a x_i



Resumen de Algoritmos de ML

Aprendizaje	Tarea	Algoritmo	Métrica	Parámetros
Supervisado	Regresión	Regresión kNN Regresión lineal	RMSE R^2	k(nº de vecinos)
	Clasificación	Regresión logística Clasificación kNN	Accuracy Recall Precision F-score	k(nº de vecinos)
No supervisado	Clustering	k-means Aglomerativo	SSE (inercia)	k(nº de clusters)
	Reducción de dimensionalidad	PCA	Varianza explicada	Nº de componentes