Estadística III para Ingenieros de Sistemas

Jose Daniel Ramirez Soto 2023
jdr2162@columbia.edu

Agenda

- anuncios varios
- modelos de analitica (machine learning-ML) Supervisado
    - KNN
    - Regresión
- Práctica de regresión en Python

modelos de analitica o machine learning

Machine learning es un conjunto de métodos o algoritmos que **entienden o aprenden de los datos** sin ser explícitamente programados. Se dividen en:

- **Supervisado**: Con datos de la variable objetivo. Tipo variable: $f(x) = \widetilde{y}, min((y - \widetilde{y})^2)$
  - continua: utilizamos modelos de regresión.
  - categórica: utilizamos modelos de clasificación.
- **No supervisado** : No existen etiquetas. Si el objetivo es: $f(x) = \widetilde{x}, min((x - \widetilde{x})^2)$
  - Crear grupos: Cluster
  - Reducir dimensionalidad o embedding : Representar datos o categorías en números
- **Reinforcement learning**: Aprender del entorno, explorar y explotar

* A Course of Machine Learning http://ciml.info/

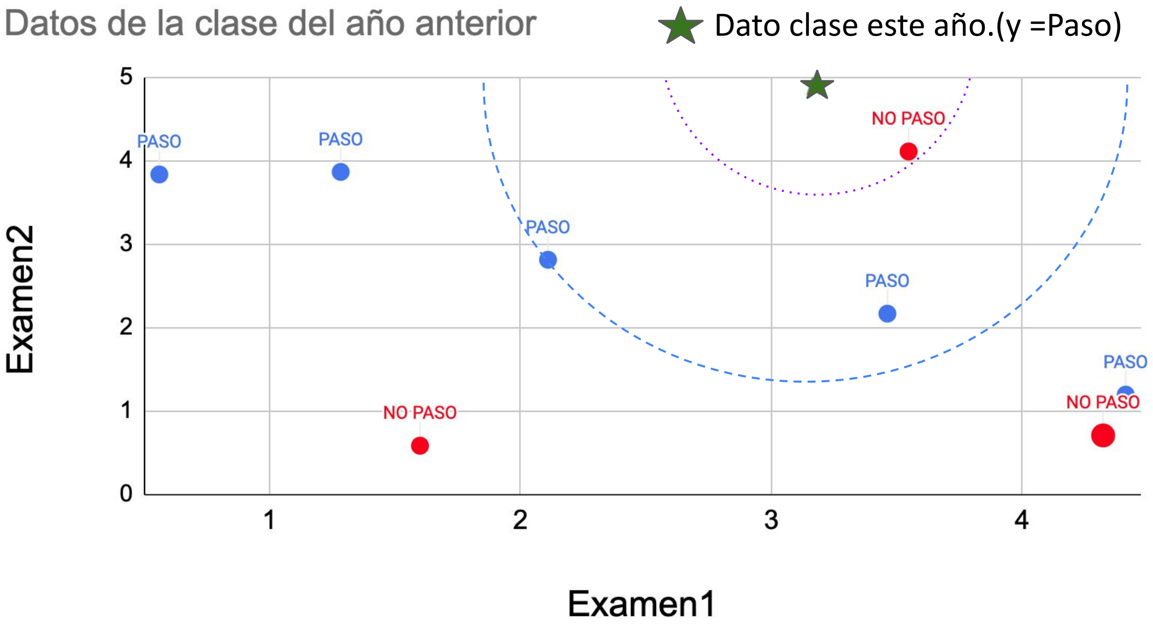
Supervisado, K-nearest neighbors (KNN)

Seleccionar K vecinos más cercanos "Generalizar".

La respuesta depende del tipo de variable objetivo:

- **continúa**: la predicción es el promedio de los vecinos.
- **categórica**: la predicción es la variable objetivo más común. Ej diferentes k:
  - k=1 No pasó (error = 1)
  - k=3 Paso (error = 0)

Ejemplo: utilizando datos históricos del año anterior. Predecir si un alumno pasará el curso este año. ¿Cuándo puedo medir el error?



Parámetro K es el numero de vecinos a considerar en la predicción y se selecciona basado en el error de los datos de validación .
Para datos desbalanceados se puede asignar más pesos a los vecinos con menor representación o hacer un muestreo (sampling) de la clase más común.

* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión

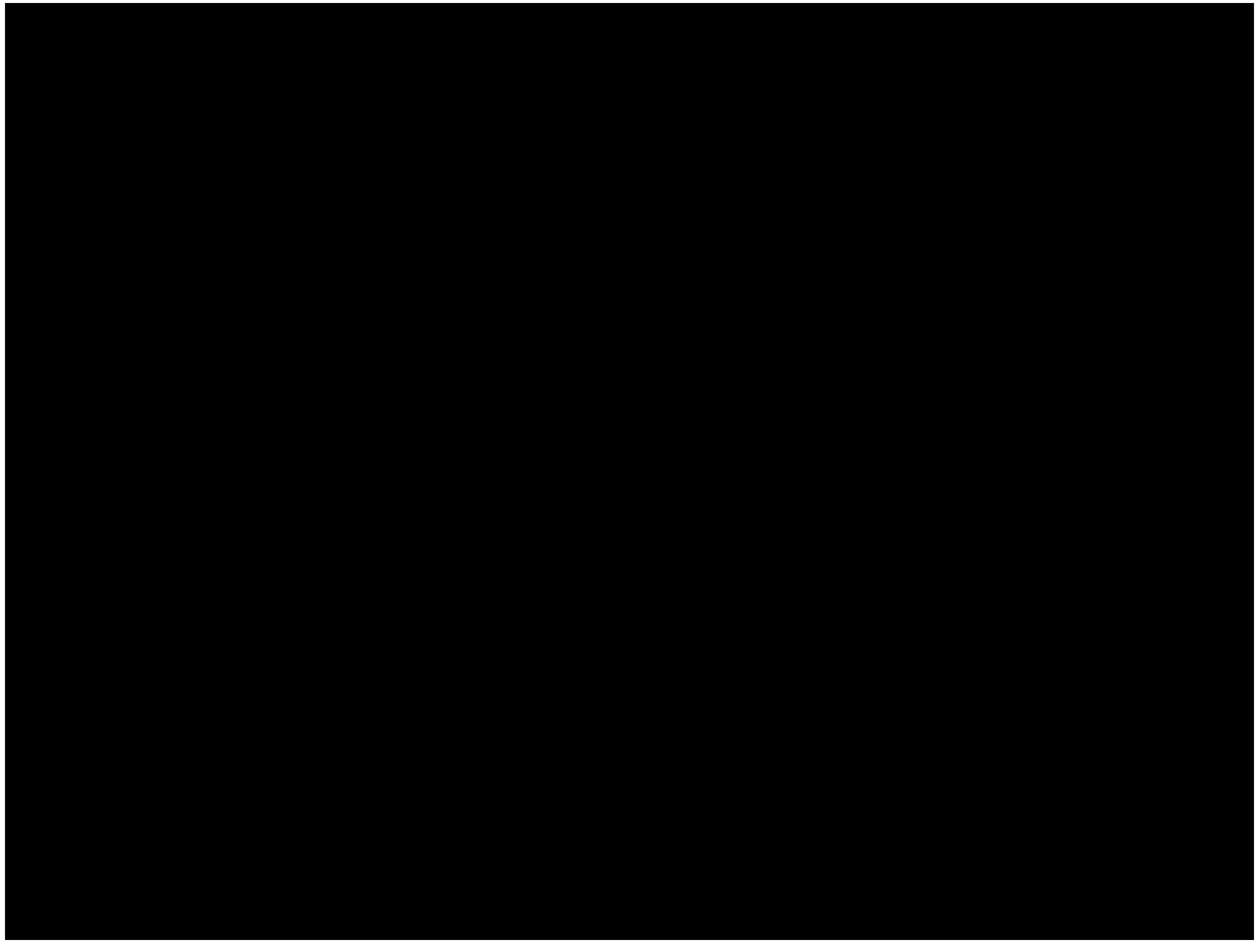
Regresión, encontrar los coeficientes W de una función f(x)=$\hat{y}$, con el objetivo de reducir la distancia de los puntos a la función.

$$\hat{y} = w^T \mathbf{x} + b = \sum_{i=1}^{p} w_i x_i + b \qquad \min_{w \in \mathbb{R}^p, b \in \mathbb{R}} \sum_{i=1}^{n} ||w^T \mathbf{x}_i + b - y_i||^2$$

La predicción es el valor de la función f(x) con los nuevos datos.

$$MSE = \frac{1}{n} \sum_{i=0}^{n} (y_i - \tilde{y}_i)^2$$

Ejemplo: utilizando datos históricos del año anterior. Predecir cuál será la nota final del curso este año.



* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión ejemplo

Tomando los datos de los carros, vamos a crear una regresión utilizando el tamaño del motor.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Medidas de error en los problemas de regresión:

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - \tilde{y}_i)^2 , MAPE = \frac{1}{n}\sum_{i=0}^{n}\frac{|y_i - \tilde{y}_i|}{y_i}$$

Ej y=4.2 , ỹ=3, mse = 1.44 , mape = 0.28

* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión P-value

Para conocer la relevancia de una variable se utilizan hipótesis test :

null hypothesis H0 :No existe relación entre las variables y el coeficiente es 0

$H0 : \beta 1 = 0$

alternative hypothesis Ha : Existe relación entre las variables

$Ha : \beta 1 \neq 0$,

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión

Universidad Católica de Oriente

Resultado de una regresión, R-squared es la proporción de la varianza de la nota final que es explicada por el examen 1 y el examen 2.

Dep. Variable:	nota_final	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	4.086e+05

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2000	0.003	60.705	0.000	0.193	0.207
examen_1	0.5987	0.001	732.319	0.000	0.597	0.600
examen_2	0.3987	0.001	474.276	0.000	0.397	0.400

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2$$

* A Course of Machine Learning http://ciml.info/

Supervisado, Regresión

La predicción es el valor de la función f(x) con los nuevos dato $\hat{y}$

$$\hat{y} = w^T\mathbf{x} + b = \sum_{i=1}^{p} w_i x_i + b \quad = \quad \underbrace{\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,D} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}}_{\mathbf{w}} = \underbrace{\begin{bmatrix} \sum_d x_{1,d} w_d \\ \sum_d x_{2,d} w_d \\ \vdots \\ \sum_d x_{N,d} w_d \end{bmatrix}}_{\hat{\mathbf{Y}}} \approx \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\hat{\mathbf{Y}}}$$

* A Course of Machine Learning http://ciml.info/