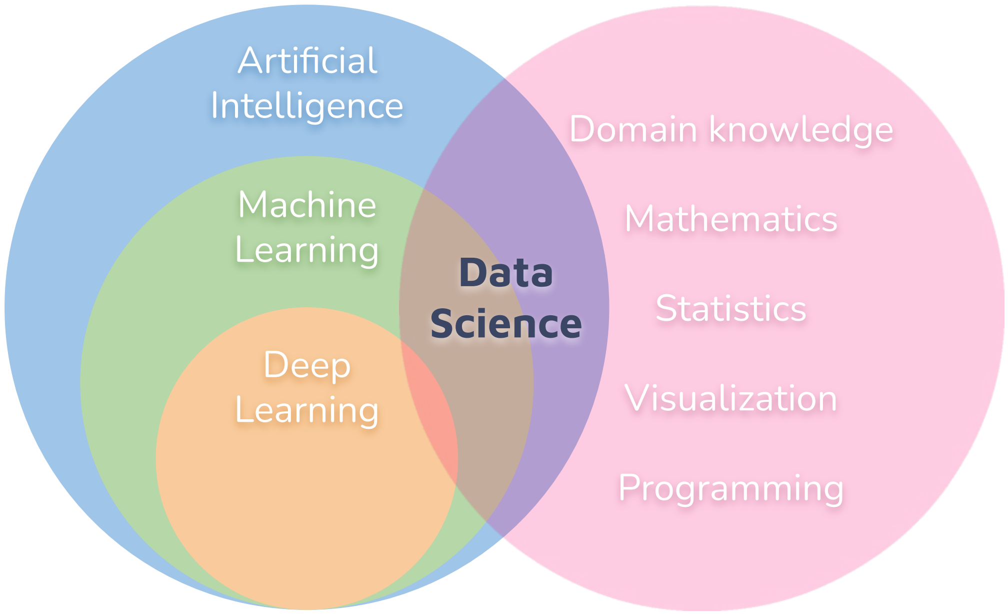
Introducción a la Ciencia de Datos

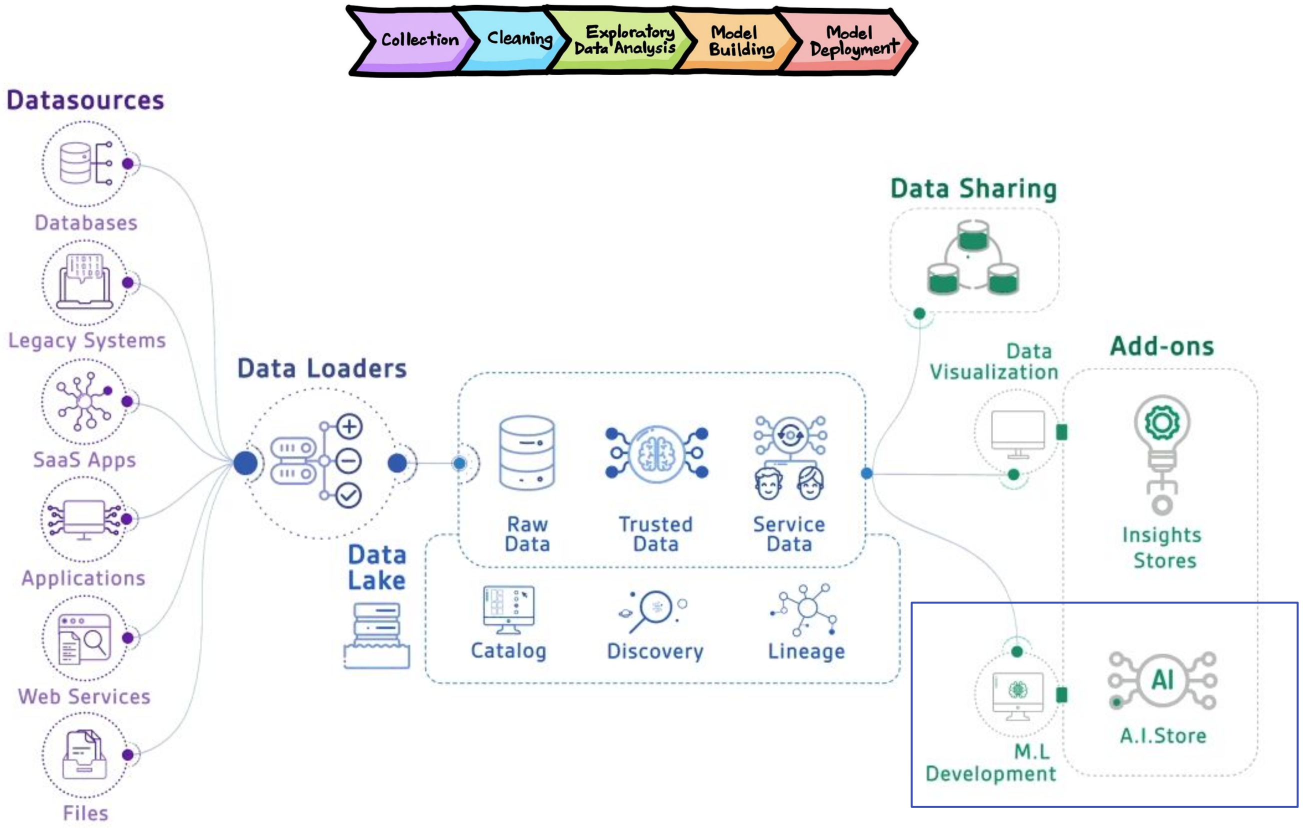
Maestría en Ciencias de la Computación

Dr. Irvin Hussein López Nava

ANIVERSARIO CICESE

CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS


Artificial Intelligence
Machine Learning
Deep Learning
Data Science
Domain knowledge
Mathematics
Statistics
Visualization
Programming



https://www.datasciencecentral.com/data-management/

¿Qué es Machine Learning

aprendizaje de máquina, aprendizaje automático, aprendizaje computacional?

Definiciones

ChatGPT

Es una subdisciplina de la IA que se enfoca en el desarrollo de algoritmos y modelos computacionales que permiten a las máquinas aprender y mejorar su rendimiento en tareas específicas a partir de datos y experiencias previas, sin una programación explícita para cada tarea.

https://chat.openai.com/

Wikipedia

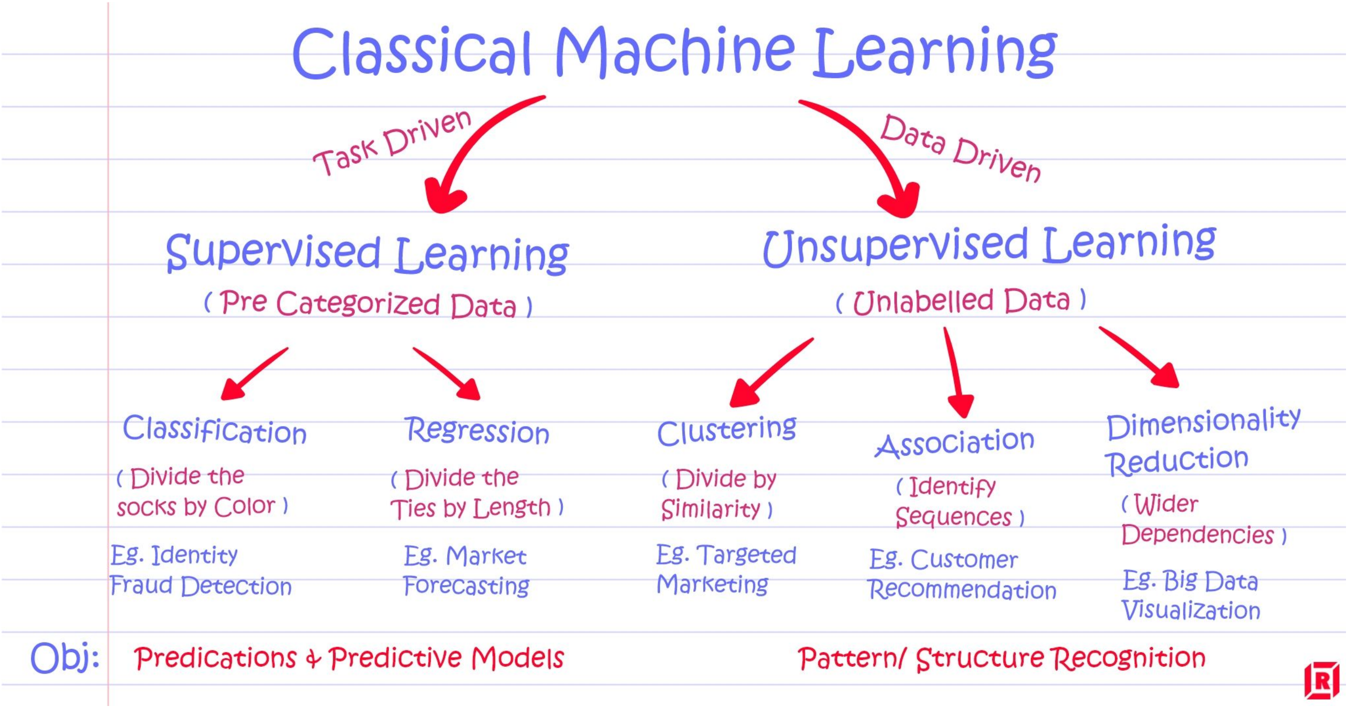
Es un subcampo de las ciencias de la computación y una rama de la IA, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan. Se dice que un agente aprende cuando su desempeño mejora con la experiencia y mediante el uso de datos.

https://**es**.wikipedia.org/wiki/Aprendizaje_automatico

Gemini

Es un subcampo de la IA que se centra en desarrollar sistemas que aprenden de los datos sin ser explícitamente programados para ello. Estos sistemas pueden aprender a realizar tareas como clasificar imágenes, detectar objetos, traducir idiomas y generar texto.

https://gemini.google.com/


Classical Machine Learning
Task Driven
Data Driven
Supervised Learning
(Pre Categorized Data)
Unsupervised Learning
(Unlabelled Data)
Classification
(Divide the socks by Color)
Eg. Identity Fraud Detection
Regression
(Divide the Ties by Length)
Eg. Market Forecasting
Clustering
(Divide by Similarity)
Eg. Targeted Marketing
Association
(Identify Sequences)
Eg. Customer Recommendation
Dimensionality Reduction
(Wider Dependencies)
Eg. Big Data Visualization
Obj:
Predications & Predictive Models
Pattern/ Structure Recognition

03

Regression


PREDICT SINGLE Y VALUE AFTER LINEAR REGRESSION
IGHT IMMA HEAD OUT
imgflip.com

3.1 Regresión lineal

¿Qué es Regresión lineal?

Definiciones

ChatGPT

Es una técnica estadística y un modelo matemático que se utiliza en estadística y aprendizaje automático para analizar y modelar la relación entre una variable dependiente (variable objetivo) y una o más variables independientes (variables predictoras).

https://chat.openai.com/

Wikipedia

Es un enfoque lineal para modelar la relación entre una respuesta escalar y una o más variables explicativas. El caso de una variable explicativa se denomina regresión lineal simple; para más de una, el proceso se llama regresión lineal múltiple.

https://en.wikipedia.org/wiki/Linear_regression

Gemini

Es un modelo estadístico que describe la relación entre una variable dependiente y una o más variables independientes. La variable dependiente es la variable que se desea predecir, y las variables independientes son las variables que se utilizan para predecir la variable dependiente.

https://gemini.google.com/

Regresión lineal en ML

- Recordando que ML es una rama de la IA que se centra en el desarrollo de algoritmos y modelos estadísticos que pueden aprender de los datos y hacer predicciones sobre ellos.
- La **regresión lineal** también es un tipo de algoritmo de aprendizaje automático (**supervisado**) que aprende de los conjuntos de datos etiquetados y asigna puntos de datos a funciones lineales optimizadas, que pueden utilizarse para la predicción en nuevos conjuntos de datos.

Tareas en aprendizaje supervisado

- Es un tipo de aprendizaje automático en el que el algoritmo aprende de datos etiquetados; cuando el valor objetivo (clase) ya se conoce.
- El **aprendizaje supervisado** tiene dos tipos principales de tareas:
  - **Clasificación:** predice la clase del conjunto de datos en función de las variables de entrada independientes. La clase tiene valores categóricos, e.g., gato y perro.
  - **Regresión:** predice las variables de salida **continuas** en función de variables de entrada independientes, como la predicción de los precios de la vivienda, la distancia a la carretera principal, la ubicación, la zona, etc.
- Uno de los tipos más simples de regresión es la **regresión lineal**.

Regresión lineal

- Calcula la relación lineal entre una **variable dependiente** y una o más **variables independientes**.
    - Cuando el número de **variables independientes** es 1, se conoce como regresión lineal univariada y, en el caso de más de una característica, se conoce como regresión lineal multivariada.
- El objetivo es encontrar la mejor ecuación lineal que pueda predecir el valor de la **variable dependiente** en función de las **variables independientes**.
    - La ecuación proporciona una línea recta que representa la relación entre las variables **dependientes** e **independientes**.
    - La pendiente de la línea indica cuánto cambia la variable **dependiente** por un cambio unitario en las variables **independientes**.

¿dónde se puede utilizar?

- La **regresión lineal** se utiliza en muchos campos diferentes, incluidas las finanzas, la economía y la psicología, para comprender y predecir el comportamiento de una variable en particular.
- Por ejemplo, en finanzas, la **regresión lineal** podría usarse para comprender la relación entre el precio de las acciones de una empresa y sus ganancias, o para predecir el valor futuro de una moneda en función de su desempeño pasado.

Regresión, predicción

- En el conjunto de registros de regresión están presentes los valores X e Y y estos valores se usan para aprender una función, por lo que si se desea predecir Y a partir de una X desconocida, se puede usar esta función aprendida.
- En la regresión se requiere encontrar el valor de Y.
  - Por lo tanto, se requiere una función que prediga Y continua en el caso de regresión, dada X como características independientes.

Variables X y Y

- Primero, Y se denomina variable dependiente u objetivo, y X se denomina variable independiente, también conocida como predictor de Y.
- Hay muchos tipos de funciones o módulos que se pueden utilizar para la regresión.
    - Una función lineal es el tipo de función más simple.
- Por su parte, X puede ser una característica única o múltiples características que representan el problema.

Suposiciones

- **Linealidad:** las variables independientes y dependientes tienen una relación lineal entre sí. Esto implica que los cambios en la variable dependiente siguen a los de las variables independientes de forma lineal.
- **Independencia:** las observaciones del conjunto de datos son independientes entre sí, i.e., el valor de la variable dependiente para una observación no depende del valor de la variable dependiente para otra observación.
- **Homocedasticidad:** en todos los niveles de las variables independientes, la varianza de los errores es constante.
- **Normalidad**: Los errores del modelo se distribuyen normalmente.

Función de hipótesis para regresión lineal

- Como se asume anteriormente, la característica independiente es la experiencia, i.e. X, mientras que Y es la variable dependiente.
- Supongamos que existe una relación lineal entre X e Y, entonces esta última se puede predecir usando:

$$\widehat{Y} = \Theta_1 + \Theta_2 X \qquad \widehat{y} = \Theta_1 + \Theta_2 x_i$$

donde $y_i \epsilon Y \ \ (i = 1, 2, \cdots, n)$ son etiquetas de los datos (aprendizaje supervisado)

$x_i \epsilon X \ \ (i = 1, 2, \cdots, n)$ son los datos de entrenamiento independientes de entrada (univariados: una variable de entrada)

$\hat{y}_i \epsilon \hat{Y} \ \ (i = 1, 2, \cdots, n)$ son los valores predichos.

Función de hipótesis para regresión lineal

- El modelo obtiene la mejor línea de ajuste de regresión al encontrar los mejores valores de θ_1 y θ_2.
  - θ_1: interceptor
  - θ_2: coeficiente de x
- Una vez que se encuentran los mejores valores de θ_1 y θ_2, se obtiene la línea de mejor ajuste.
- Entonces, cuando finalmente usemos el modelo para la predicción, predecirá el valor de ***y*** para el valor de entrada de ***x***.

Función de costo

- La función de costo, o función de pérdida, no es más que el error o diferencia entre el valor predicho **Ŷ** y el valor verdadero **Y**.
  - Es el error cuadrático medio (MSE) entre el valor predicho y el valor verdadero.
- La función de costo (J) se puede escribir como:

$$\text{Cost function}(J) = \frac{1}{n}\sum_{n}^{i}(\hat{y}_i - y_i)^2$$

¿Cómo actualizar θ_1 y θ_2 para obtener la línea de mejor ajuste?

- Para lograr la línea de regresión de mejor ajuste, el modelo apunta a predecir el valor objetivo **Ŷ** de modo que la diferencia de error entre el valor predicho **Ŷ** y el valor verdadero **Y** sea mínima.
- Por lo tanto, es muy importante actualizar los valores de θ_1 y θ_2 para alcanzar el mejor valor que minimice el error entre el valor de ***y*** previsto (pred) y el valor de ***y*** verdadero (y).

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

Gradient Descent

- Se puede entrenar un modelo de regresión lineal utilizando el algoritmo de optimización de descenso de gradiente modificando iterativamente los parámetros del modelo para reducir el error cuadrático medio (MSE) del modelo en un conjunto de datos de entrenamiento.
- Para actualizar los valores θ_1 y θ_2 con el fin de reducir la función de costo (minimizando el valor RMSE) y lograr la línea de mejor ajuste, el modelo utiliza Descenso de gradiente.
- La idea es comenzar con valores aleatorios de θ_1 y θ_2 y luego actualizar los valores de forma iterativa, alcanzando el costo mínimo.

Gradient Descent

- Un gradiente no es más que una derivada que define los efectos sobre las salidas de la función con un poco de variación en las entradas.
- Definimos la función de costo (J) con respecto a θ_1

$$
\begin{aligned}
J'_{\theta_1} &= \frac{\partial J(\theta_1, \theta_2)}{\partial \theta_1} \\
&= \frac{\partial}{\partial \theta_1}\left[\frac{1}{n}\left(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i)\left(\frac{\partial}{\partial \theta_1}(\hat{y}_i - y_i)\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i)\left(\frac{\partial}{\partial \theta_1}(\theta_1 + \theta_2 x_i - y_i)\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i)\,(1 + 0 - 0)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(\hat{y}_i - y_i)\,(2)\right] \\
&= \frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)
\end{aligned}
$$

Gradient Descent

- Después, la función de costo(J) con respecto a θ_2

$$
\begin{aligned}
J'_{\theta_2} &= \frac{\partial J(\theta_1,\theta_2)}{\partial \theta_2} \\
&= \frac{\partial}{\partial \theta_2}\left[\frac{1}{n}\left(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i)\left(\frac{\partial}{\partial \theta_2}(\hat{y}_i - y_i)\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i)\left(\frac{\partial}{\partial \theta_2}(\theta_1 + \theta_2 x_i - y_i)\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} 2(\hat{y}_i - y_i)\,(0 + x_i - 0)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n} (\hat{y}_i - y_i)\,(2x_i)\right] \\
&= \frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\cdot x_i
\end{aligned}
$$

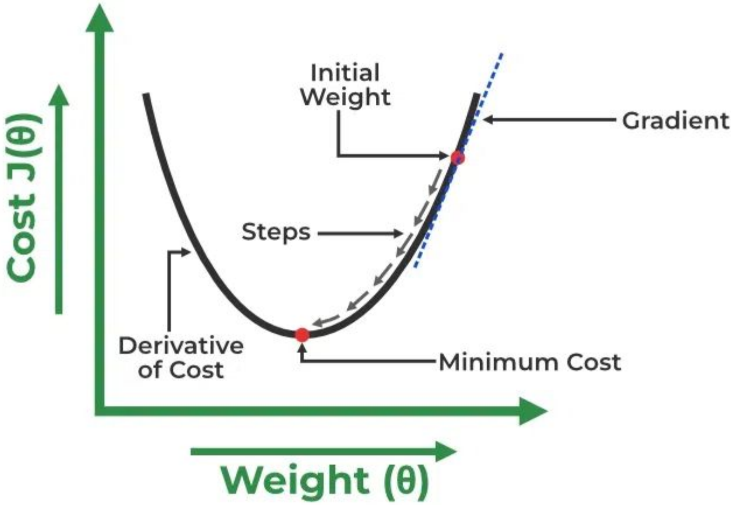
Gradient Descent

- Encontrar los coeficientes de una ecuación lineal que mejor se ajuste a los datos de entrenamiento es el objetivo de la regresión lineal.
- Moviéndose en la dirección del gradiente negativo del error cuadrático medio con respecto a los coeficientes, los coeficientes se pueden cambiar.
- Y la intercepción y el coeficiente respectivo de X dependerán de α como tasa de aprendizaje.

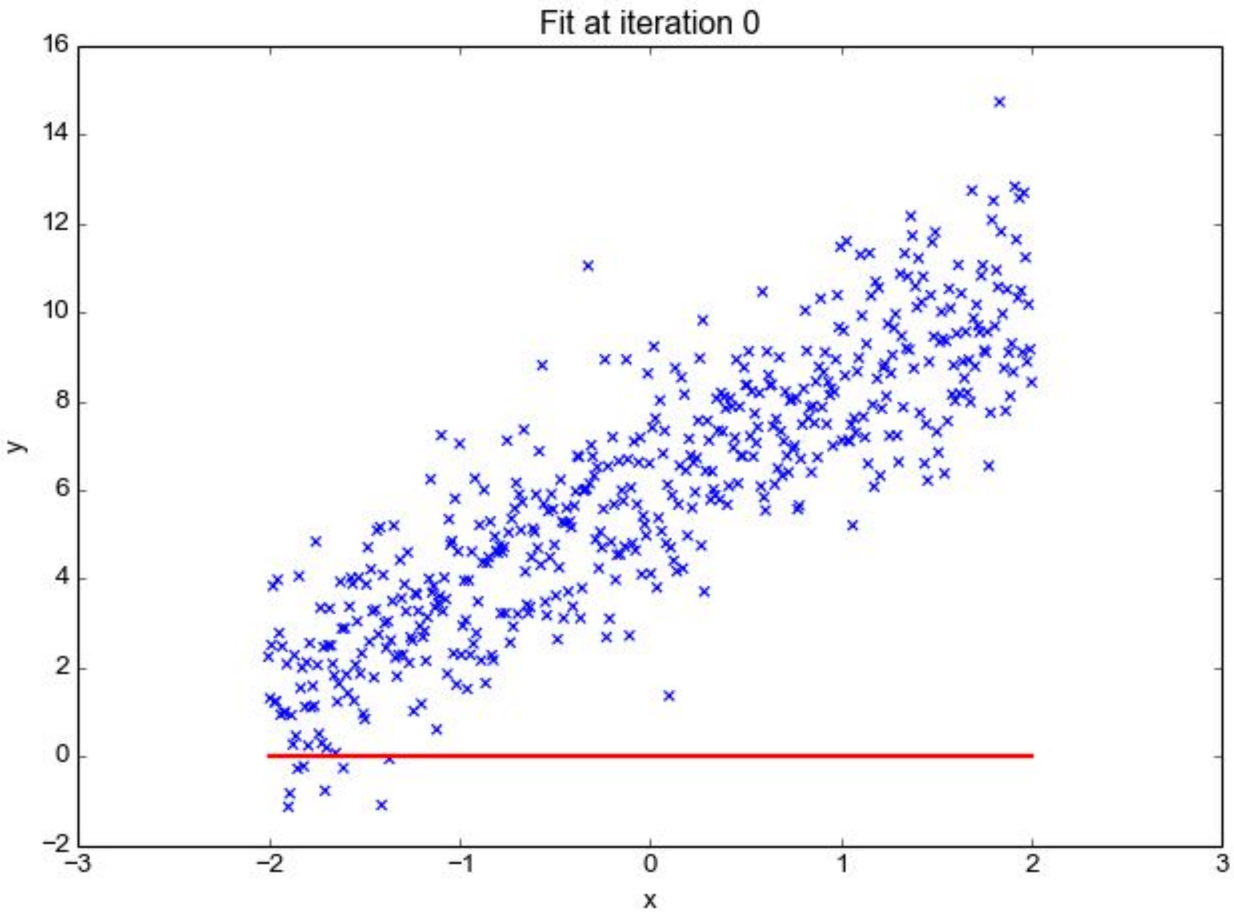
$$\begin{aligned}\theta_1 &= \theta_1 - \alpha\left(J'_{\theta_1}\right)\\ &= \theta_1 - \alpha\left(\frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\right)\end{aligned}$$

$$\begin{aligned}\theta_2 &= \theta_2 - \alpha\left(J'_{\theta_2}\right)\\ &= \theta_2 - \alpha\left(\frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\cdot x_i\right)\end{aligned}$$

En resumen



Cost J(θ)
Initial
Weight
Gradient
Steps
Derivative
of Cost
Minimum Cost
Weight (θ)



Fit at iteration 0
16
14
12
10
8
6
4
2
0
−2
y
−3
−2
−1
0
1
2
3
x

Entonces, ¿regresión lineal es ML?

- Según Tom Mitchell: “Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y medida de desempeño P si su desempeño en las tareas en T, medido por P, mejora con la experiencia E”
- La regresión es ML cuando su tarea es proporcionar un valor estimado a partir de características predictivas en alguna aplicación. Su rendimiento debería mejorar a medida que experimenta más datos.

https://stats.stackexchange.com/questions/268755/when-should-linear-regression-be-called-machine-learning

¿Cómo se utiliza?

- Ajustar una línea
- Predecir un valor
- Clasificar, separar por clases

UC Irvine Machine Learning Repository

Datasets | Contribute Dataset | About Us

Search datasets...

Filters

Search datasets...

- Keywords
- Data Type
- Subject Area
- Task
  - Classification
  - Regression
  - Clustering
  - Other
- # Features
- # Instances
- Feature Type

Browse Datasets

SORT BY # VIEWS, DESC | EXPAND ALL

Iris

A small classic dataset from Fisher, 1936. One of the earliest known datasets used for evaluating classification methods.

Classification | Tabular | 150 Instances | 4 Features

Heart Disease

4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach

Classification | Multivariate | 303 Instances | 13 Features

Adult

Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Classification | Multivariate | 48.84K Instances | 14 Features

Dry Bean Dataset

Images of 13,611 grains of 7 different registered dry beans were taken with a high-resolution camera. A total of 16 features; 12 dimensions an...

Classification | Multivariate | 13.61K Instances | 16 Features

Diabetes

This diabetes dataset is from AIM '94

20 Features

Caso de estudio

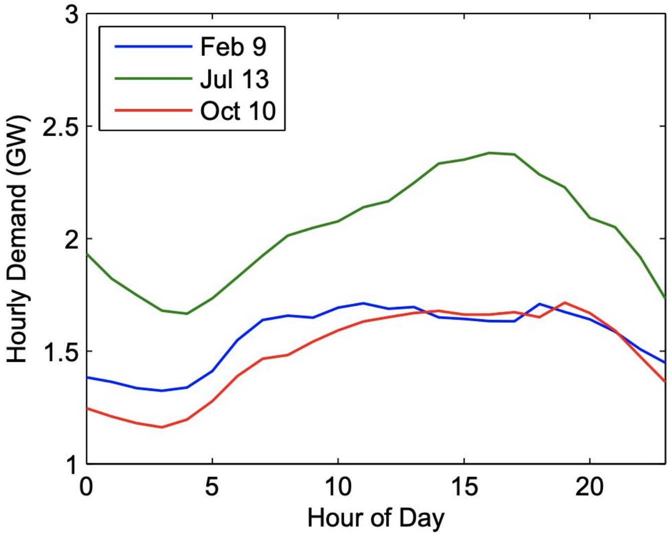
Generación de electricidad renovable en EE.UU.

	Hydropower	Solar[1]	Wind	Geothermal	Biomass	Total Renewables
2004	6.7%	0.0%	0.4%	0.4%	1.3%	8.8%
2005	6.7%	0.0%	0.4%	0.4%	1.3%	8.8%
2006	7.1%	0.0%	0.7%	0.4%	1.3%	9.5%
2007	5.9%	0.0%	0.8%	0.4%	1.3%	8.5%
2008	6.2%	0.1%	1.3%	0.4%	1.3%	9.3%
2009	6.9%	0.1%	1.9%	0.4%	1.4%	10.6%
2010	6.3%	0.1%	2.3%	0.4%	1.4%	10.4%
2011	7.8%	0.2%	2.9%	0.4%	1.4%	12.6%
2012	6.8%	0.3%	3.4%	0.4%	1.4%	12.4%
2013	6.6%	0.5%	4.1%	0.4%	1.5%	13.1%
2014	6.3%	0.8%	4.4%	0.4%	1.6%	13.5%

Planteamiento del problema

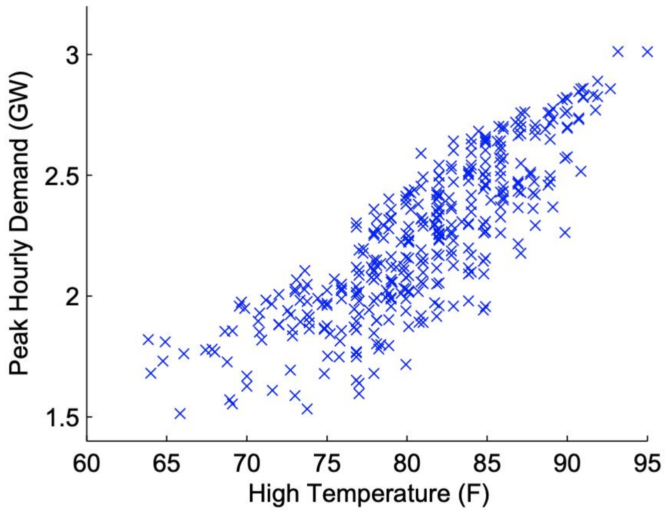
- La energía eólica y solar son intermitentes
- Se necesitan centrales eléctricas tradicionales cuando no hay viento.
  - Muchas plantas de energía (e.g., las nucleares) no se pueden encender/apagar fácilmente ni acelerar/desactivar rápidamente.
- Con pronósticos más precisos, la energía eólica y solar se convierten en alternativas más eficientes.
  - Una previsión precisa ahorró a empresas de servicios públicos entre 6 y 10 millones de dólares al año.
- ¿Se puede pronosticar con precisión la energía que se consumirá mañana?
  - Es difícil de estimar a partir de modelos "a priori".
  - Pero se tienen muchos datos a partir de los cuales construir modelos.

Ejemplo de consumo eléctrico



Predecir la demanda máxima debido a las altas temperaturas

- ¿Cuál será la demanda máxima el día de mañana?
- De saber algo más sobre el mañana (como la temperatura alta), se puede usar para predecir la demanda máxima.



Planteamiento del problema

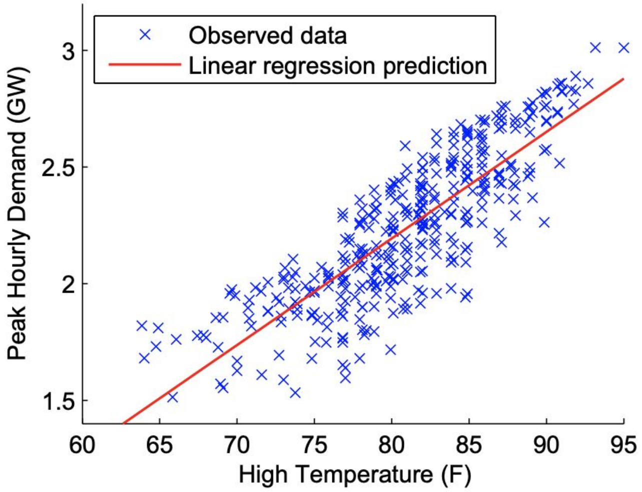
- Un modelo lineal que predice la demanda:

  predicted peak demand =

  $\theta_1 \cdot (\text{high temperature}) + \theta_2$

- Parámetros del modelo:

  $\theta_1, \theta_2 \in \mathbb{R}$ $(\theta_1 = 0.046, \theta_2 = -1.46)$



Modelo lineal simple

- Se puede usar un modelo como el anterior para hacer predicciones.
- ¿Cuál será la demanda máxima mañana?
- Se sabe, por el informe meteorológico, que la temperatura máxima será de 80ºF (ignorar por el momento, que esto también es una predicción).
- Entonces la demanda máxima prevista es:

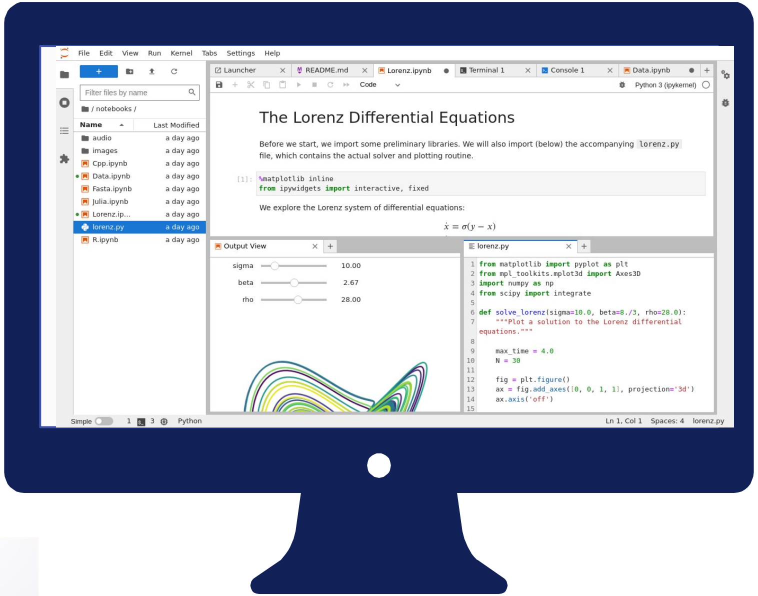
$$\theta_1 \cdot 80 + \theta_2 = 0.046 \cdot 80 - 1.46 = 2{,}19 \text{ GW}$$

Formalización del modelo

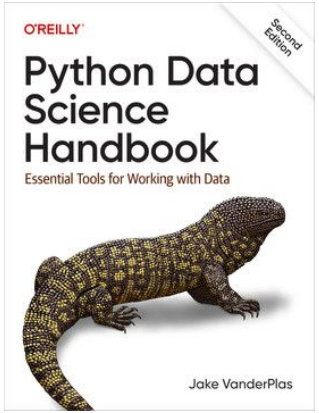
- **Input**: $x_i \in \mathbb{R}^n, \quad i = 1, \ldots, m$
  - E.g.: $x_i \in \mathbb{R}^1 = \{\text{high temperature for day } i\}$
- **Output**: $y_i \in \mathbb{R}$ (*regression* task)
  - E.g.: $y_i \in \mathbb{R} = \{\text{peak demand for day } i\}$
- **Model Parameters**: $\theta \in \mathbb{R}^k$
- **Predicted Output**: $\hat{y}_i \in \mathbb{R}$

E.g.: $\hat{y}_i = \theta_1 \cdot x_i + \theta_2$

(Go to live notebook)



Extra Libro



GitHub

- 05.00-Machine-Learning.ipynb
- 05.01-What-Is-Machine-Learning.ipynb
- 05.02-Introducing-Scikit-Learn.ipynb
- 05.03-Hyperparameters-and-Model-Validation.ipynb
- 05.04-Feature-Engineering.ipynb
- 05.06-Linear-Regression.ipynb

Temas a considerar la sig. edición

https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html#sphx-glr-auto-examples-linear-model-plot-ols-py

Gracias!

¿Alguna pregunta?

hussein@cicese.mx

https://sites.google.com/view/husseinlopeznava

SCAN ME

CREDITS: This presentation was based on a template by **Slidesgo**, and includes icons by **Flaticon**.