Economía y ciencia de los datos

Introducción a Machine Learning : Modelos Supervisados

Carlos Alvarado & Pablo González

¿Qué es Machine Learning?

“Machine Learning is the science of getting computers to learn”

Arthur Samuel (Attributed)

“Machine Learning is the study of computer algorithms that allow computer programs to automatically improve through experience.”

Tom Mitchell

“A.I. is in a ‘golden age’ and solving problems that were once in the realm of sci-fi”

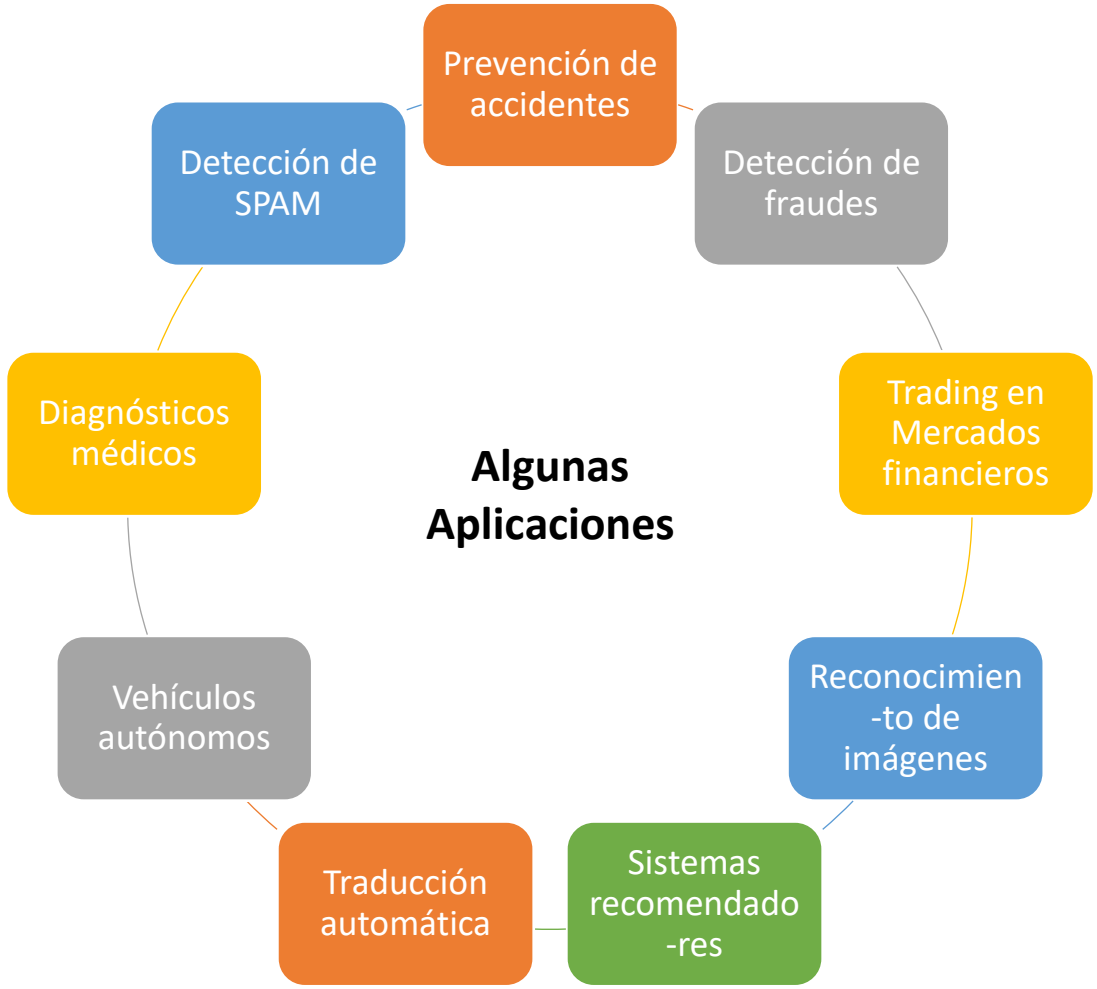
Jeff Bezos

Experiencia previa y aplicaciones

Esta no es la primera vez que ven uno de estos métodos. De hecho, en econometría ya tuvieron experiencia con un modelo base de esta disciplina.

Esta vez, sin embargo, buscaremos expandir el repertorio de algoritmos que conozcan y nos enfocaremos en la aplicación.

** En caso de haberlo olvidado, les recomiendo revisar el siguiente artículo por Jared Wilber: https://mlu-explain.github.io/linear-regression/*



Tipos de aprendizaje

Supervisado

- Requiere de datos etiquetados (respuesta)
- Utilizado en tareas de predicción y clasificación

No Supervisado

- No requiere de datos etiquetados
- Comúnmente utilizado para tareas de clustering

Reforzado

- Inspirado en la psicología conductista
- Fuera del enfoque de este curso.
- Ejemplo: Q-learning

En esta sección nos enfocaremos en el aprendizaje supervisado

Tipos de modelos (Algunos ejemplos)

Paramétricos

- Naive Bayes
- Support Vector Machine (Caso general)
- Random Forest

No-Paramétricos

- KNN
- CART
- Support Vector Machine (RBF kernel)

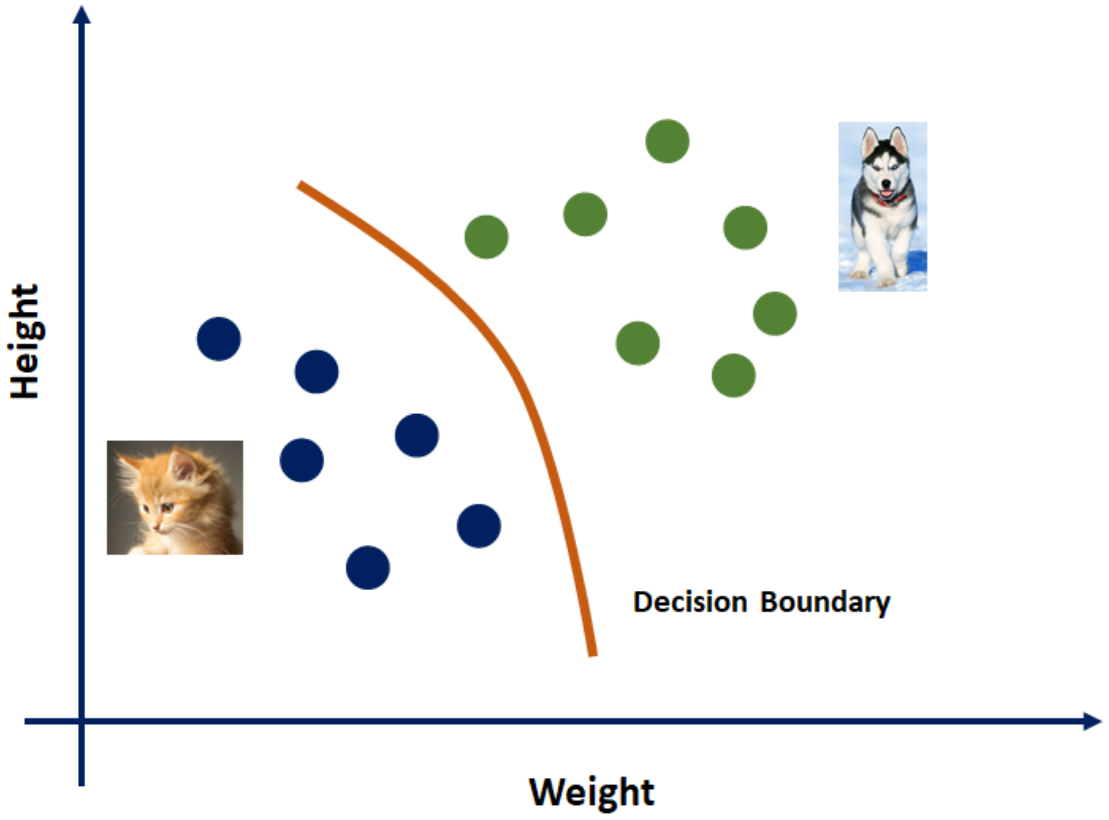
¿Dónde creres que entraría la regression lineal vista en econometría?

Regresión vs Clasificación

Denominamos problemas de regresión a aquellos donde “y” es continua. En cambio, cuando nuestra variable dependiente es categórica, nos enfrentamos a un problema de clasificación.

Algunos algoritmos de clasificación:

- Regresión Logística
- Naive Bayes
- KNN

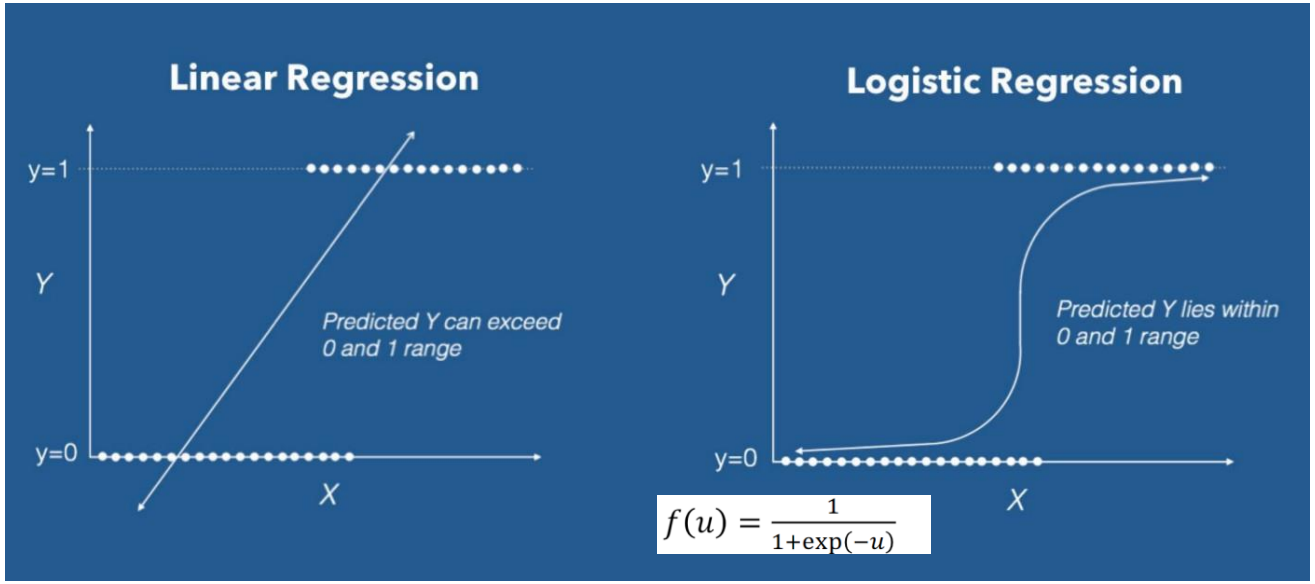


Ej.: Frontera de decisión (modelo discriminante)

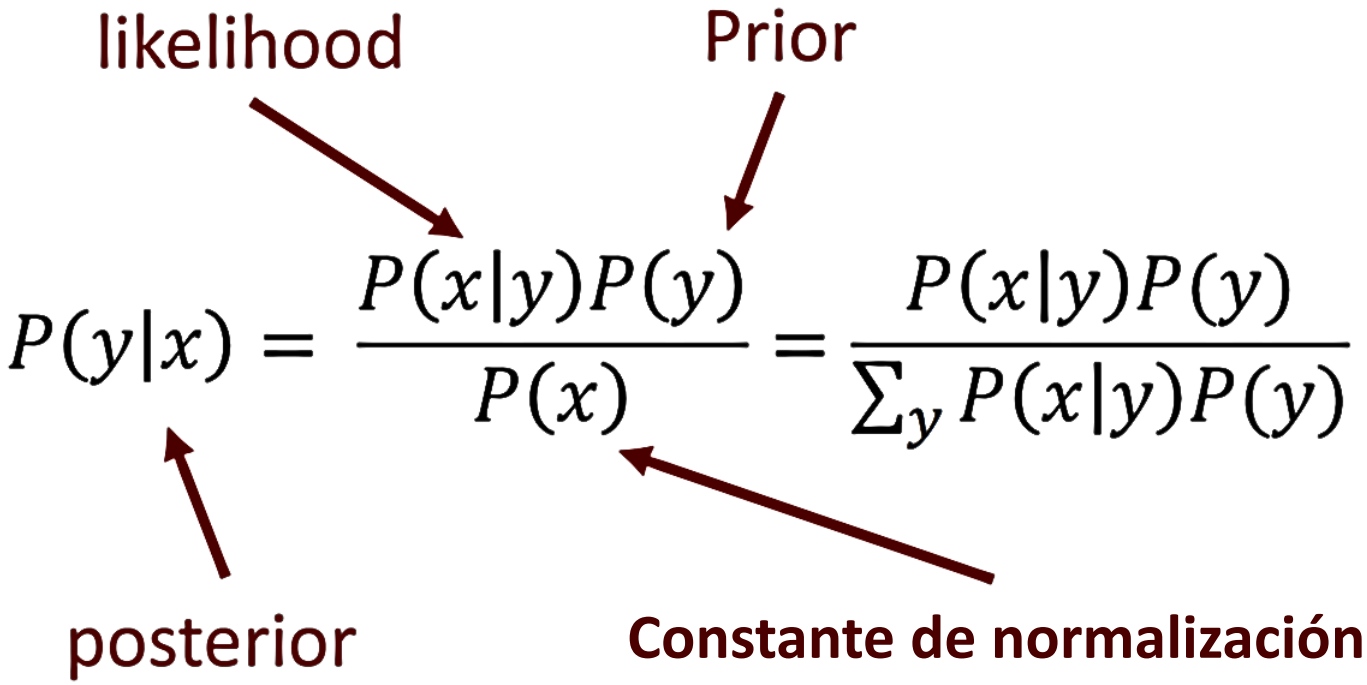
Regresión Logística

Limita el valor y entre 0 y 1, permitiendo clasificar elementos entre dos clases como una probabilidad de que uno pertenezca a la clase respectiva.

$$p(y = 1|x, \theta) = \frac{1}{1 + \exp(-\theta^\top x)} \qquad p(y = 0|x, \theta) = \frac{\exp(-\theta^\top x)}{1 + \exp(-\theta^\top x)}$$



Naïve Bayes



Supuesto:

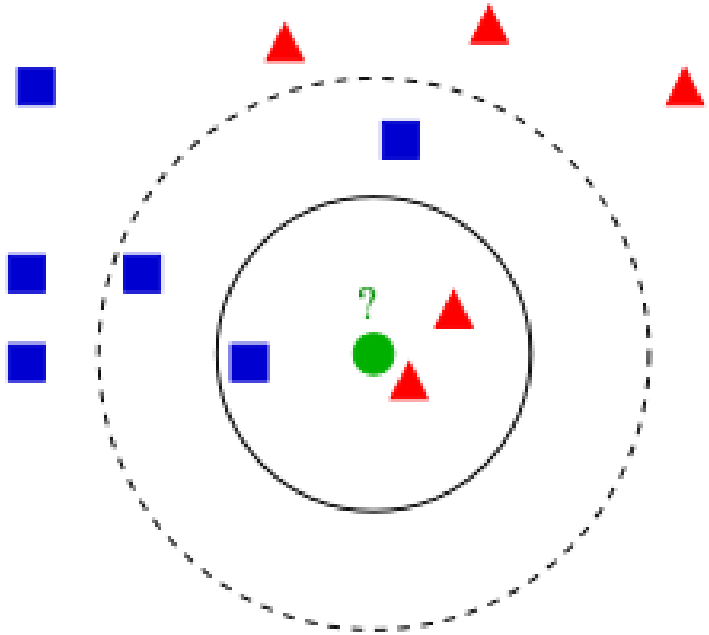
Efecto del valor de una variable en una clase dada, es independiente de los valores de las otras variables.

KNN (K vecinos cercanos)

Clasifica a un punto en función del voto mayoritario de los k vecinos (en la muestra de entrenamiento) más cercanos.

<u>Ventajas:</u>

- No paramétrico
- No linear
- Fácil de kernelizar



Otros algoritmos interesantes

Lasso / Ridge

Support Vector Machine

Decision Trees

Boosting Machines

Random Forest

Y más…

Sobre

scikit learn

scikit-learn

Machine Learning in Python

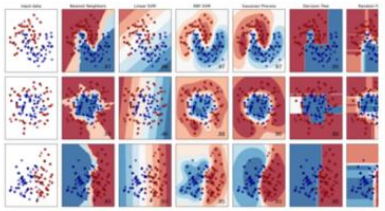
Getting Started | Release Highlights for 1.1 | GitHub

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...

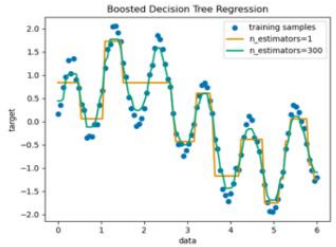


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...

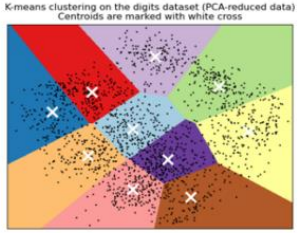


Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...

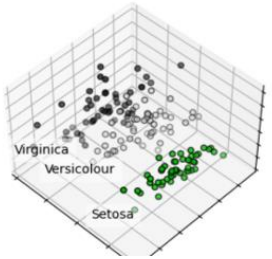


Examples

Dimensionality reduction

Reducing the number of random variables to consider.

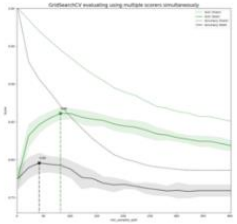
Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization, and more...



Model selection

Comparing, validating and choosing parameters and models.

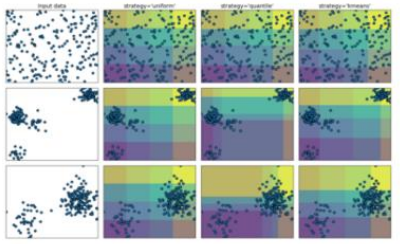
Applications: Improved accuracy via parameter tuning
Algorithms: grid search, cross validation, metrics, and more...



Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.
Algorithms: preprocessing, feature extraction, and more...



Sobre

Documentación

Casos de ejemplo:

- Regresión Lineal
- Regresión Logística
- KNN
- Naïve Bayes
- Random Forest (classifier)
- Random Forest (regressor)

- General