

PROYECTO

Clase Práctica [APRENDIZAJE SUPERVISADO]

Usar conceptos teóricos de aprendizaje supervisado en analítica.

Aprendizaje Supervisado

Uso corto con enfoque práctico

Pasos necesarios para un problema de analítica.

Caja, Tratamiento, modelado

Tareas de Aprendizaje supervisado

Regresión (2 clases)

Clasificación (2 clase)

Enfoque días esperanza de vida.

Tarea → Proyecto admisión universitaria

METODOLOGIA

ANDRES ARISTIBAL

Aprendizaje automático

Aprendizaje supervisado

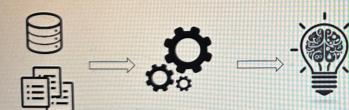
Tareas de regresión

Proyecto de regresión

APRENDIZAJE AUTOMÁTICO

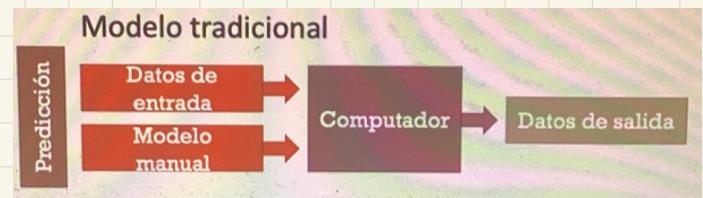
Definición:

El aprendizaje automático es la ciencia que permite a los computadores aprender, sin ser explícitamente programados.¹

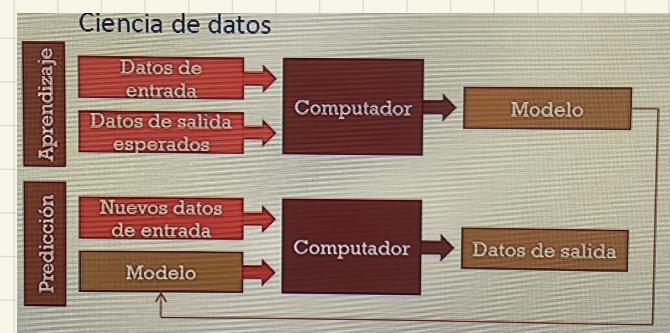


1. Andrew Ng, Stanford University, 2014

Tradicional

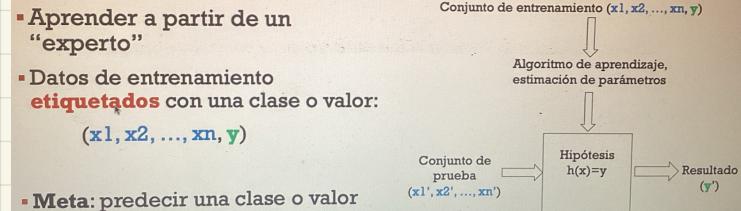


Ciencia de datos



APRENDIZAJE SUPERVISADO

- Aprender a partir de un "experto"
- Datos de entrenamiento **etiquetados** con una clase o valor:
 $(x_1, x_2, \dots, x_n, y)$
- Meta: predecir una clase o valor



Variables independientes

Entrada → Ejemplos (x_1, x_2, \dots, x_n)

salida → Tiene o no cáncer

Clasificación → Valor discreto

Regresión → Valor continuo

Etiqueta, es un valor de verdad dado por un experto

! Importante

(a la variable Y)

Etiqueta categórica → clasificación

Etiqueta numérica → Regresión

Debo buscar con que comparar el resultado
en caso contrario, será aprendizaje no
supervisado.

Predictores, Entrada
Respuesta, salida, etiqueta

TAREAS DE REGRESIÓN

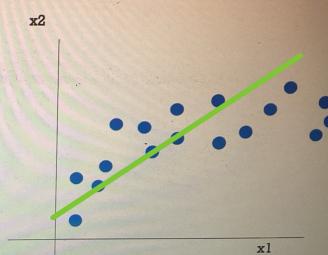
TAREAS DE REGRESIÓN

- Encontrar modelos, f , que permitan predecir **valores continuos**:

- KNN → **k vecinos cercanos**
- Regresión lineal
- Regresión polinómica
- Árboles de regresión
- ...

- Valores **numéricos** de la variable o función objetivo

- Baseline:** medida de evaluación dada por un modelo que predice una medida de tendencia central (e.g. el promedio)
→ "modelo nulo"



Sin que me voy a comparar (si mi modelo lo mejora) en **REGRESIÓN** es con la media, en **CLASIFICACIÓN** es con la moda

Se busca que la línea se ajuste a los datos, los puntos y el valor Predicho (su diferencia) sea mejor

MODELOS DE REGRESIÓN

- Regresión lineal: busca una relación lineal entre los atributos predictores (x_i) y el atributo objetivo (y)

$$y = h_{\theta}(X) = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n + \epsilon$$

Los parámetros θ_j son estimados teniendo como objetivo la minimización de los residuos o diferencias cuadradas entre las predicciones (\hat{y}) y los valores reales (y):

$$\text{minimizar restos de residuales al cuadrado} \quad \sum_{i=1}^m (y - \hat{y})^2 \quad \text{real} \rightarrow \text{predicho}$$

- Regresión Lasso: regularización L1 (suma de los valores absolutos de los coeficientes multiplicados por un lambda)

$$\operatorname{argmin}_{\theta_0} \sum_{i=1}^m (y - \hat{y})^2 + \lambda \sum_{j=1}^m |\theta_j|$$

coeficientes

para evitar sobreajuste

- Regresión Ridge: regularización L2 (suma de los valores cuadrados de los coeficientes multiplicados por un lambda)

$$\operatorname{argmin}_{\theta_0} \sum_{i=1}^m (y - \hat{y})^2 + \lambda \sum_{j=1}^m \theta_j^2$$

Evitar sobreajuste (overfitting)

- Regresión polinomial: busca una relación polinomial entre los atributos predictores y la variable objetivo (Y)

$$Y = h_{\theta}(X) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \dots + \theta_n x^n + \epsilon$$

A partir de unas variables arrojar un resultado lo mas cercano posible

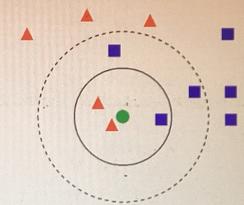
ALGORITMOS

KNN

MODELOS DE REGRESIÓN

KNN:

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Asigna la clase o valor agregado de las instancias conocidas que se encuentran más cerca de la instancia a predecir.



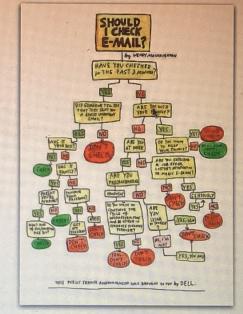
Predicción de un dato nuevo a partir de sus vecinos más cercanos, teniendo en cuenta el **K vecinos** (parametro). Si voy a dar un valor continuo saco promedio de su **K vecinos**.

Árboles de decisión:

MODELOS DE REGRESIÓN

Árboles de decisión:

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Posee una estructura de árbol, donde sus nodos internos representan las características, las ramas las decisiones y cada hoja el resultado.



Basado en nodos representan las características

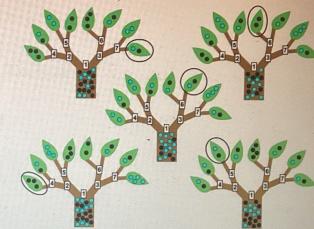
Accuracy (correctitud)

Random Forest:

MODELOS DE REGRESIÓN

Random Forest:

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Basado en el concepto de ensambles, el cual es un proceso de combinación de múltiples regresores o clasificadores para resolver un problema complejo y mejorar el rendimiento del modelo.



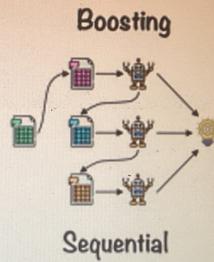
Poder de las mesas

Boosting

MODELOS DE REGRESIÓN

Boosting:

- Algoritmo de aprendizaje supervisado para clasificación y regresión.
- Mejora la exactitud y rendimiento de los modelos de aprendizaje automático convirtiendo múltiples modelos débiles en uno mucho más potente.



Modelo Ruse

resultado métricos a modelo base
resultado métrico del modelo y
compara con base

Uso de pickle para guardar modelo
(Para evitar reentrenamiento)

Selección de características

feature selection

características que generan valor al predicho o clasificado

- Variables (repetitivas) / correlacionadas
↳ se borran (multiplicidad)

Ajustar variables a modelo
Lasso / Ridge

Revisar correlación de las variables independientes y las dependientes

Entre independientes debe estar con baja correlación, puede presentar problemas

matriz de correlación

Ingeniería de características

las variables categóricas no se ajustan a modelos → pasar a numéricas

one hot encoding ó label encoding

el random state es el seed para poder reproducir

busco el error más pequeño

busco el R^2 más alto

Llegar hasta Protocolos

Ingeniería de características

usar características y transformarlos (Dummificación)

Categoricas a numéricas

Teniendo el año, calcular el mensual

Limpieza de datos

Tipos

Faltantes (imputación) ← imputador KNN

Protocolos de evaluación

Hold out ← guarda un pedazo de datos

Métricas de evaluación

Regresión → Error cuadrático medio

Abs

Ran Error cuadrático

R^2