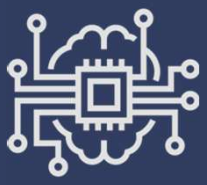


MACHINE LEARNING & DEEP LEARNING

Un enfoque práctico

MSC RENZO CLAURE ARACENA

1

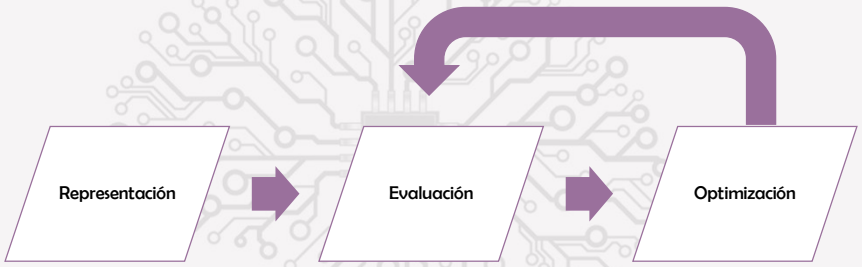
Features Engineering

MSC RENZO CLAURE ARACENA

2

MACHINE LEARNING

Metodología de ML



MSC RENZO CLAURE ARACENA

3

MACHINE LEARNING

COMO REPRESENTAR EL PROBLEMA
PARA QUE PUEDA TRANSFORMARSE EN UN MODELO DE ML

- Representación del problema

ID	Edad	Ocupación	Ingresos	Monto	Recencia	Frecuencia	Intereses
123	23	Empleado	5000	200	20	5	Futbol

MSC RENZO CLAURE ARACENA

4

MACHINE LEARNING

Módulos iniciales requeridos en python

- Scikit-Learn: Open source, cuenta con los algoritmos más utilizados
- SciPy: Análisis estadísticos, algebra lineal, etc.
- Numpy: Estructuras y arreglos matriciales
- Pandas: Manipulación de Datos, diversos orígenes,
- Matplotlib: Librerias para hacer gráficos
- TensorFlow librerías especializadas en Machine y Deep Learning
- Pytorch librerías especializadas en Machine y Deep Learning

MSC RENZO CLAURE ARACENA

5

MACHINE LEARNING

Explorando características de las variables

- Preprocesamiento (manejo de: valores nulos, duplicados, conjuntos de datos desbalanceados, valores atípicos, etc.)
- Codificación de categorías (Codificación de datos, etiquetas y codificación ordinal)
- Dimensionamiento de categorías
- Generación y extracción de categorías
- Selección de categorías relevantes

MSC RENZO CLAURE ARACENA

6

MACHINE LEARNING

Preprocesamiento de variables

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Data Discretization
- Data Normalization

MSC RENZO CLAURE ARACENA

7

MACHINE LEARNING

Proceso de preparación de los datos

MSC RENZO CLAURE ARACENA

8

MACHINE LEARNING

Ejemplo

análisis supervisado con machine learning

- Compra de artículo
- Cliente, ejemplo real
- Dimensiones del cliente
- De que sirve solo tener ejemplos para aprender?
- Es necesario entonces tener ejemplos para comprobar si lo aprendido es correcto

9

MACHINE LEARNING

Empecemos con un ejemplo en python

- Productos

10

MACHINE LEARNING

Repositorios libres

http://archive.ics.uci.edu/ml/index.php

MSC RENZO CLAURE ARACENA

11

MACHINE LEARNING

Glosario de términos

- **Variable independiente:** Ayuda a describir a la variable objetivo, generalmente la denotaremos como X (mayúscula por que es una matriz)
- **Variable dependiente:** También conocida como la variable objetivo, es la variable resultado de la combinación, lineal o no lineal de las variables independientes, generalmente la denotaremos como y (minúscula por que es un vector, la mayor parte de las veces)
- **Muestra:** Grupo representativo del universo
- **Muestra de Entrenamiento:** Es una muestra que se utilizará para que los modelos aprendan, generalmente llamaremos a la matriz de entrenamiento de muestra de variables independientes como: **X_train** y al vector de muestra de entrenamiento para la variable objetivo como: **y_train**
- **Muestra de Comprobación:** Es una muestra que se utilizará para evaluar los modelos, generalmente llamaremos a la matriz de muestra de validación de variables independientes como: **X_test** y al vector de muestra de validación para la variable objetivo como: **y_test**

MSC RENZO CLAURE ARACENA

12

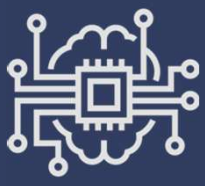
MACHINE LEARNING

Glosario de términos

- **Matriz de datos:** Matriz que contiene las variables independientes y dependientes y los registros o instancias.
- **Registros o instancias:** es la representación de cada individuo en la matriz de datos, es decir es cada fila de la matriz, que contiene los datos de cada instancia o individuo.
- **Modelo:** Es una solución que permitirá resolver el problema planteado
- **Método de evaluación:** es la técnica o indicador con el que se medirán el rendimiento de un modelo
- **Sub ajuste:** underfitting, se refiere a que un modelo tiene bajo rendimiento o precisión, por lo tanto tiene un bajo ajuste
- **Sobre ajuste:** overfitting, se refiere a que un modelo tiene demasiado ajuste con respecto de la muestra de comprobación, es decir que no es generalizable
- **Generalización:** Es la capacidad de un modelo de funcionar bien con datos que no sean del entrenamiento, ya sean de validación o comprobación

MSC RENZO CLAURE ARACENA

13



APRENDIZAJE SUPERVISADO

MSC RENZO CLAURE ARACENA

14

MACHINE LEARNING

Principales tareas
del aprendizaje supervisado

- Encontrar un modelo que permita predecir con adecuada eficacia el resultado de la variable objetivo
- Medir el impacto positivo o negativo, fuerte o débil que tienen las variables independientes sobre la variable objetivo

MSC RENZO CLAURE ARACENA

15

MACHINE LEARNING

Veamos el código
usaremos jupiter

- nb_2

MSC RENZO CLAURE ARACENA

16

MACHINE LEARNING

Tipos de aprendizaje supervisado

clasificación y regresión

- Clasificación:
  - Determinar si la variable objetivo de una instancia tendrá un nivel dentro de una variable cualitativa
  - La mayor parte de los casos el resultado buscado es dicotómico:
    - Si o No
    - + o –
    - 1 o 0, etc.
  - También en el caso de multinivel:
    - Tipo de producto: A, B, C…
    - Timo de tratamiento: 1, 2, 3…

MSC RENZO CLAURE ARACENA

17

MACHINE LEARNING

Tipos de aprendizaje supervisado

clasificación y regresión

- Regresión:
  - La variable de respuesta es contínua, buscamos obtener el valor más cercano
    - El peso, cantidad, espesor, volumen, etc.
  - Los modelos se basan generalmente en modelos estadísticos, pero existen otras opciones
  - Los modelos ahora son tan sofisticados que pueden tratar con variables mixtas e inclusive pueden combinarse

MSC RENZO CLAURE ARACENA

18

MACHINE LEARNING

Entrenamiento y comprobación
dividir el universo en muestras

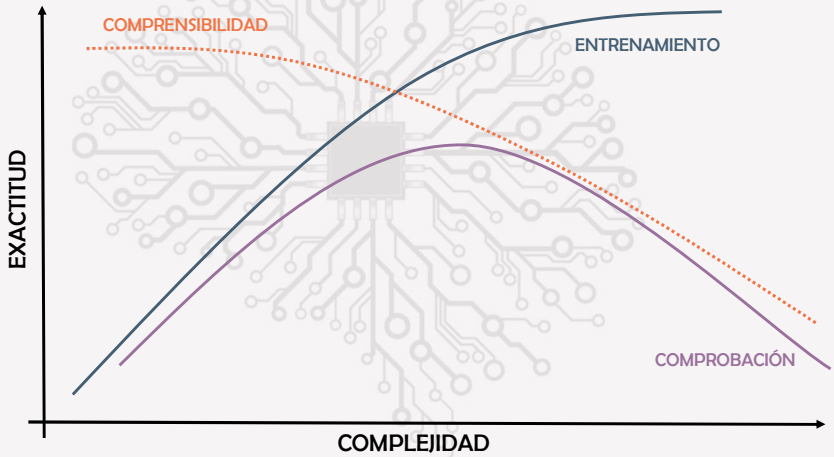
- Los modelos deben entrenarse sobre una base que represente fielmente el comportamiento del universo, es decir que no contenga sesgos, para esto un muestreo aleatorio simple, sin reemplazamiento es suficiente.
- Una vez entrenados los modelos, estos deben comprobarse en una base distinta a la de entrenamiento pero que proviene del mismo universo de la muestra de entrenamiento, es decir, ningún caso utilizado en el entrenamiento debe estar presente en la comprobación.
- El rendimiento del modelo en la muestra de entrenamiento sirve para ajustar y elegir el mejor modelo, el rendimiento en la muestra de comprobación nos da una idea de cómo se comportará el modelo con nuevos casos que no están en el universo.

MSC RENZO CLAURE ARACENA

19

MACHINE LEARNING

Comprensibilidad, precisión y complejidad
de los modelos implementados



MSC RENZO CLAURE ARACENA

20

MACHINE LEARNING

GENERALIZACIÓN, SOBREAJUSTE Y SUBAJUSTE

UNDER FITTING AND OVERFITTING

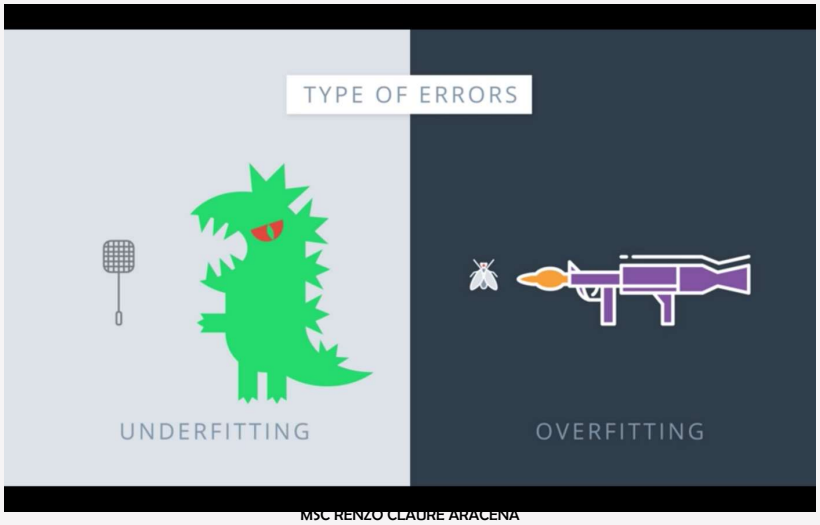
- Generalización:
  - Es la capacidad de un modelo de funcionar con la misma eficiencia en un nuevo entorno, con nuevos datos
  - Aunque se espera que los modelos sean robustos se deben tomar en cuenta ciertos supuestos:
    - La muestra de comprobación fue parte del mismo universo y tomado de forma aleatoria de el
    - La distribución de los datos nuevos sigue el mismo patrón que los datos de entrenamiento
  - Muchas veces los modelos se comportan o tienen un rendimiento muy bajo en la generalización
  - Cuando un modelo es muy complejo, por ejemplo un modelo polinomial, se dice que el el modelo tiene sobre ajuste
  - Por el contrario si el modelo es muy básico, por ejemplo un modelo lineal ante problemas multidimensionales, se dice que tiene un subajuste

21

MACHINE LEARNING

Generalización, sobreajuste y subajuste

under fitting and overfitting

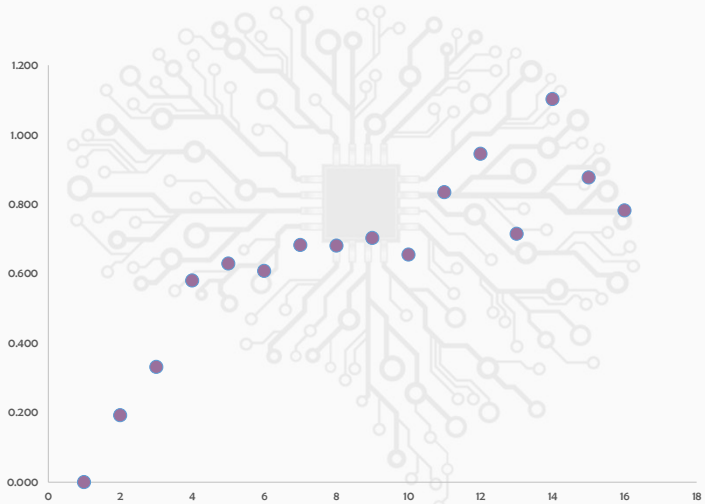


22

MACHINE LEARNING

Generalización, sobreajuste y subajuste

regresión

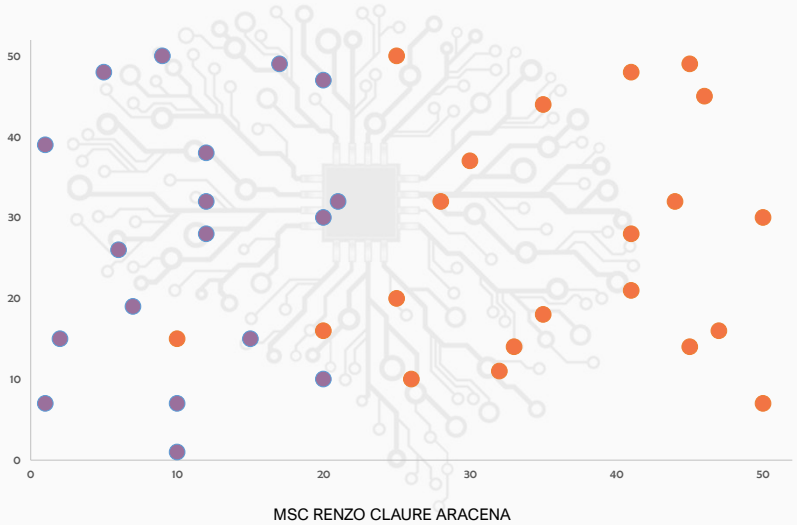


23

MACHINE LEARNING

Generalización, sobreajuste y subajuste

clasificación



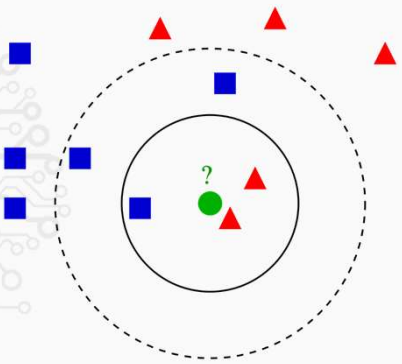
24

K Vecinos Próximos

25

El procedimiento

- Definir la cantidad de vecinos que se tomarán en cuenta (k)
- Determinar las coordenadas del nuevo caso
- Normalizar las coordenadas de las var independientes
- Determinar las distancias a todos los demás puntos, establecer la medida de distancia
- Seleccionar los k más próximos
- Determinar la clase mayoritaria y asignarla al nuevo caso



26

MACHINE LEARNING

Pseudocódigo
para cualquier lenguaje

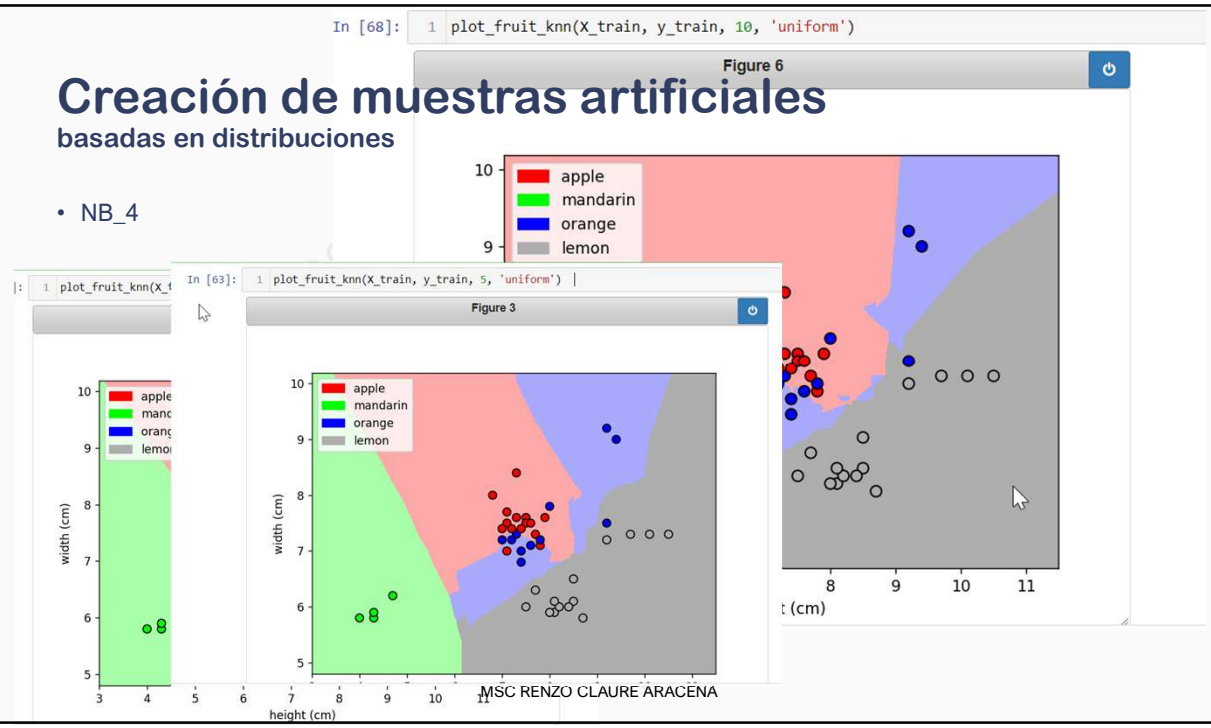
- Leer los datos en un arreglo, X
- Establecer las coordenadas del nuevo caso, P
- Para cada punto en las base de datos
  - Determinar la distancia de cada punto X a P
  - Ordenar las distancias de modo creciente
  - Tomar *k* elementos con la menor distancia
  - Encontrar la clase mayoritaria entre los casos
  - Asignar la clase mayoritaria como predicción del nuevo caso P

27

MACHINE LEARNING

Creación de muestras artificiales
basadas en distribuciones

- NB_4

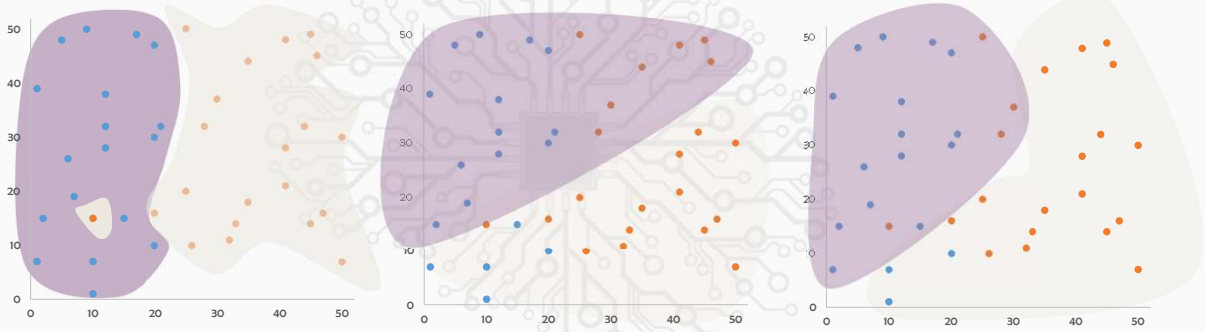


28

MACHINE LEARNING

Efecto en la elección de k

clasificación



MSC RENZO CLAURE ARACENA

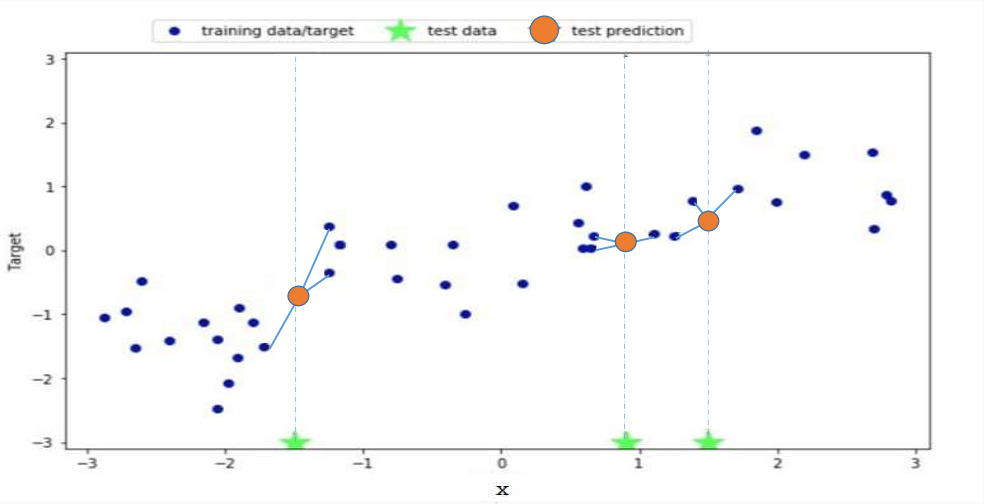
29

MACHINE LEARNING

Regresión con knn

efecto de la elección de k

- NB_5



MSC RENZO CLAURE ARACENA

30

MACHINE LEARNING

Parámetros críticos

vecinos y distancia

- Cantidad de vecinos
  - Mayor cantidad de vecinos reduce la precisión en el entrenamiento, sub-ajuste
  - Menor cantidad de vecinos, reduce la generalización, sobre-ajuste
- Medida de distancia
  - Generalmente se usa la distancia Euclidiana, simple y aplicable a muchos problemas
  - En Casos específicos se requiere otras medidas como Mahalanovis
- En conclusión:
  - KNN es un modelo simple y bastante comprensible
  - Sirve como un Baseline para comparar otros modelos
  - Se conflictua cuando se tiene gran cantidad de variables independientes

31

Regresión lineal

32

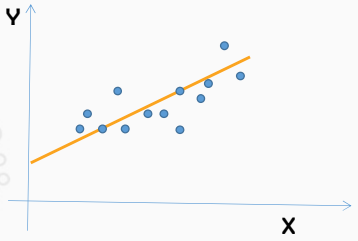
MACHINE LEARNING

Regresión lineal simple

una breve introducción

- Un Modelo Lineal, es una ecuación lineal compuesta de pesos o pendientes, sesgo y factores aleatorios
- Donde β_1 es la pendiente, peso o efecto de la variable *X* sobre la variable *y*
- β_0 es el sesgo, intercepción de la variable y con la ausencia de valor en X
- ε es el efecto de factores aleatorios externos al modelo
- Ejemplo:

$$y = \beta_0 + \beta_1 X + \varepsilon$$



$$Felicidad = \beta_0 + \beta_1 Ingresos + \varepsilon$$

$$Felicidad = 0{,}3 + 5\ Ingresos + \mathbf{0{,}18}$$

MSC RENZO CLAURE ARACENA

33

MACHINE LEARNING

Regresión multilineal

fórmula matricial

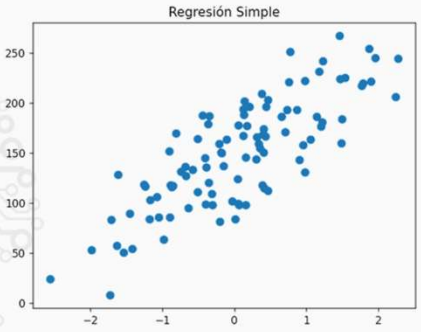
- No hablamos de una sola variable Independiente, si no de un vector:

$$x = x_0 + x_1 + x_2 + ,,, + x_n$$

- La solución es una ecuación matricial del tipo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$



¿Cómo saber qué línea ajusta mejor a todas?
¿Cómo medir esa efectividad?

MSC RENZO CLAURE ARACENA

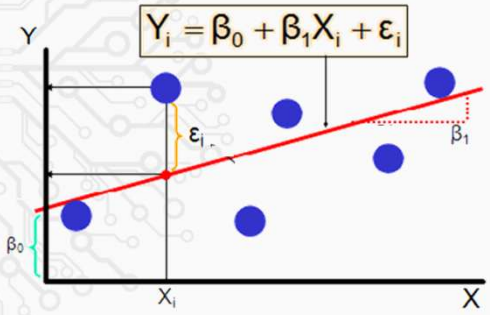
34

MACHINE LEARNING

Mínimos cuadrados

es una alternativa simple y poderosa

- Se basa en la minimización del error de estimación
- El error utilizado es el cuadrado de la diferencia de la estimación y el valor real
- Se dice simple por que no permite modificar la complejidad, siempre será lineal



MSC RENZO CLAURE ARACENA

35

MACHINE LEARNING

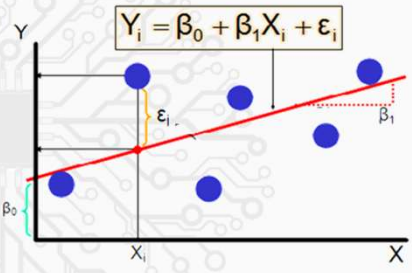
Mínimos cuadrados

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\mathrm{RSS} = \sum_{i=1}^{n} (\widehat{\varepsilon}_i)^2 = \sum_{i=1}^{n} (y_i - (\widehat{\alpha} + \widehat{\beta} x_i))^2$$

$$y = X\beta + e$$

$$\hat{\beta} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y.$$



MSC RENZO CLAURE ARACENA

36

MACHINE LEARNING

Mínimos cuadrados en python

- NB_6

MSC RENZO CLAURE ARACENA

37

MACHINE LEARNING

Regularización ridge

métodos de penalización

- Parte del calculo de MMCC, encontrando los parámetros β y α
- Se añade una penalización por la variación de los parámetros β

$$\text{RSS} = \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2 \; + \alpha \sum_{i=1}^{n} \beta^2$$

- Una vez que se calcularon los parámetros se aplica la penalización con el propósito de reducir los efectos de los valores grandes en los parámetros β
- Esto produce modelos más simples, pero mas generalizables, es decir con mejor rendimiento en la comprobación
- A este proceso de penalización se llama regularización
- La influencia de la penalización depende de un parámetro, α
- A Mayores valores de alpha α, modelos más simples
- Tiene un mayor impacto cuando la muestra de entrenamiento es muy pequeña considerando una mayor cantidad de variables independientes

MSC RENZO CLAURE ARACENA

38

MACHINE LEARNING

Ejemplo de regularizacion

- NB_6

39

MACHINE LEARNING

Normalización

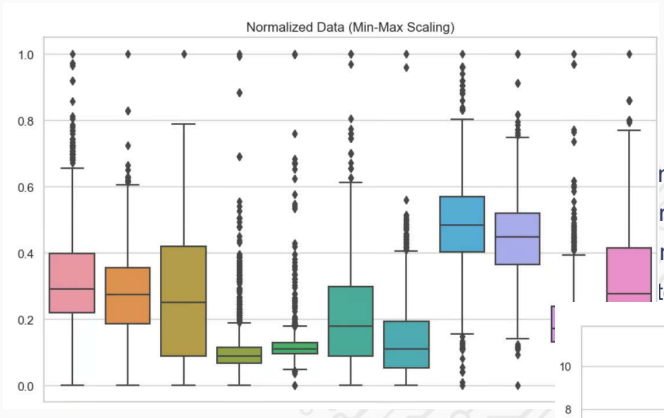
pre procesamiento de los datos

- En casi todos lo problemas reales, la variables tienen dimensiones y escalas distintas
- Las variables con magnitudes desproporcionadamente mayores crean sesgo
- Muchos modelos son influenciados por el sesgo, afectando su efectividad
- La normalización reduce estas influencias colocando a todas las medidas en el mismo rango
- Existen varios tipos de normalización, Ajustada a una curva (Standrd), de Rango, etc.
- Para los ejemplos del curso usaremos la de rango, también la standard en algunos problemas:

$$x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

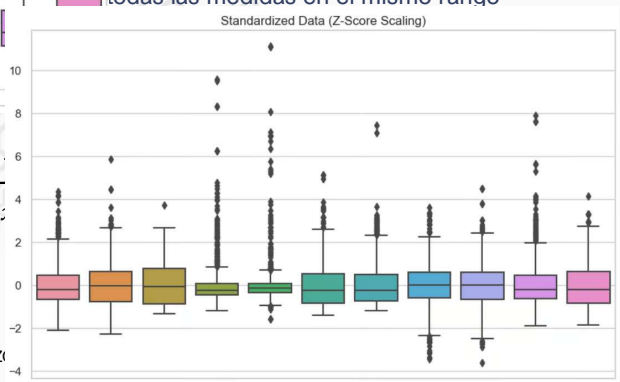
40

MACHINE LEARNING



mensiones y escalas distintas
mayores crean sesgo
ndo su efectividad
todas las medidas en el mismo rango

$$x'_i = \frac{x_i}{x_{ma}}$$



MSC RENZ

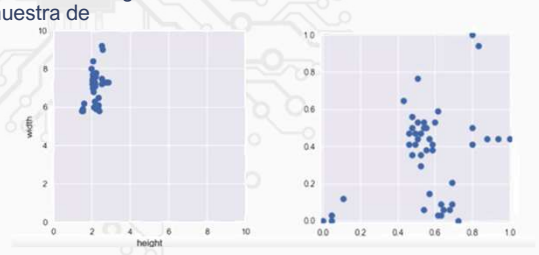
41

MACHINE LEARNING

Normalización

pre procesamiento de los datos

- Procedimiento :
  1. Ajustar las escalas SOLO en la muestra de entrenamiento y aplicar esas escalas en la muestra de comprobación
  2. Si se crean distintas escalas para la muestra de entrenamiento y la de comprobación, se producirá un incremento en la desviación por factores aleatorios
  3. Si se utiliza una solo escala para toda la población esto producirá que se “filtren” las distribuciones o sesgos de la muestra de enteramiento sobre la muestra de comprobación



MSC RENZO CLAURE ARACENA

42

MACHINE LEARNING

Aplicación de la normalización
por rango

- NB_6

MSC RENZO CLAURE ARACENA

43

MACHINE LEARNING

Regularización laso

- La regularización por LASO es muy similar a la Ridge, solo que en lugar de elevar al cuadrado los efectos, obtiene su valor y mantiene el valor de intercepción

$$\text{RSS} = \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2 + \alpha \sum_{i=1}^{n} |\alpha + \beta_i|$$

- LASO puede reducir inclusive a cero el valor de los factores menos influyentes
- La variable alpha controla el efecto de LASO, con un valor por defecto de 1
- Entonces es muy útil cuando se requiere reducir la complejidad de un modelo, reduciendo la cantidad de variables por el descarte de las que son menos influyentes
- Pero RIDGE nos servirá más cuando la mayor parte de las variables son útiles

MSC RENZO CLAURE ARACENA

44

MACHINE LEARNING

Aplicación de laso

- NB_6

45

MACHINE LEARNING

Regresión polinómica

- Muchas veces las variables independientes interactúan entre sí, por lo que incluir este efecto en el modelo puede mejorar significativamente su precisión.
- También es posible mejorar los resultados si los efectos son cuadráticos, o de orden 2, o superiores.
- Poner demasiados grados o factores combinados tenderá a crear modelos más complejos y por ende poco generalizables.
- Será necesario un tipo de regularización, la más frecuente es Ridge

46

MACHINE LEARNING

Regresión polinómica

- NB_6

MSC RENZO CLAURE ARACENA

47

Modelos de regresión lineal para clasificación

MSC RENZO CLAURE ARACENA

48

MACHINE LEARNING

REGRESIÓN LOGÍSTICA

- No buscamos un valor específico, más bien la probabilidad de pertenecía a un nivel de la variable objetivo
- Es necesario cambiar la función
- F deberá estar entre 0 y 1
- Entonces F deberá ser un dist. de probabilidad
- La más popular es la función logística

$$p_i = F\left(\beta_0 + \beta'_1 \mathbf{x}_i\right)$$

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta'_1 \mathbf{x}_i)}}.$$

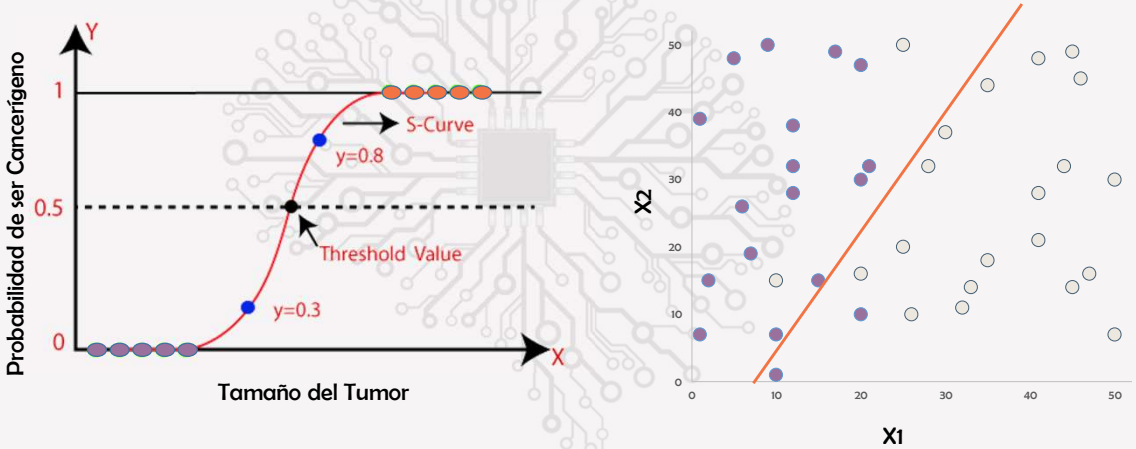
MSC RENZO CLAURE ARACENA

49

MACHINE LEARNING

REGRESION LOGISTICA

BINARIA



MSC RENZO CLAURE ARACENA

50

MACHINE LEARNING

Parámetros de regularización

Regresión Logística

- Al igual que ridge, ya viene configurado por defecto con un parámetro de regularización L2, llamado C
- El parámetro “C”, ajusta la regularización, por defecto está configurado en 1
- Valores más grandes de C, ajustan más o sobreajustan los modelos
- Valores más pequeños de C, suavizan o sub ajustan los modelos
- También juega un papel muy importante la normalización de los datos
- Se comporta mejor con más variables independientes

51

MACHINE LEARNING

Regresión logistica

- NB_7

52

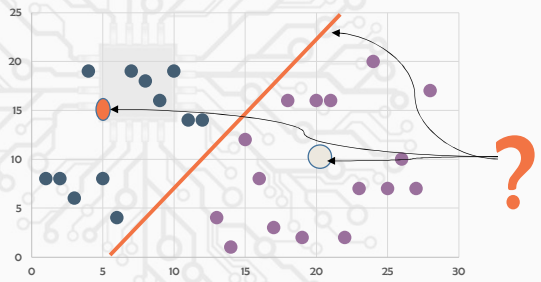
Máquinas de soporte vectorial

MSC RENZO CLAURE ARACENA

53

Máquinas de soporte vectorial
separadores lineales

- Es un algoritmo que divide casos a través de la búsqueda de separadores
- Los primeros separadores utilizados son lo lineales



MACHINE LEARNING

MSC RENZO CLAURE ARACENA

54

MACHINE LEARNING

Máquinas de soporte vectorial

función signo

$1{,}5x_1 - x_2 - 5 = 0$

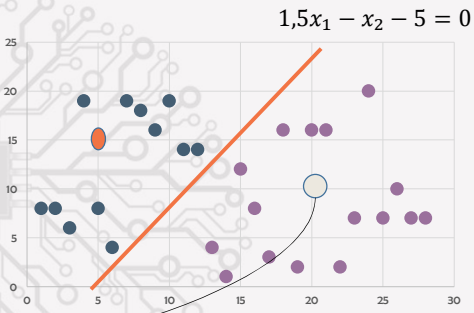
$f(x, b, a) = signo(b * x + a)$

$1{,}5x_1 - x_2 - 5 = 0$

$b = [1{,}5\ , -1]$

$a = (-5)$

$x_1 = 20\ \ x_2 = 10$



$signo\big(1{,}5 * 20 + (-1) * 10 + (-5)\big) = 15 > 0\ entonces:\ +, 1$

MSC RENZO CLAURE ARACENA

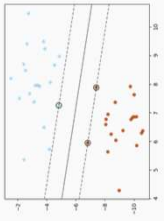
55

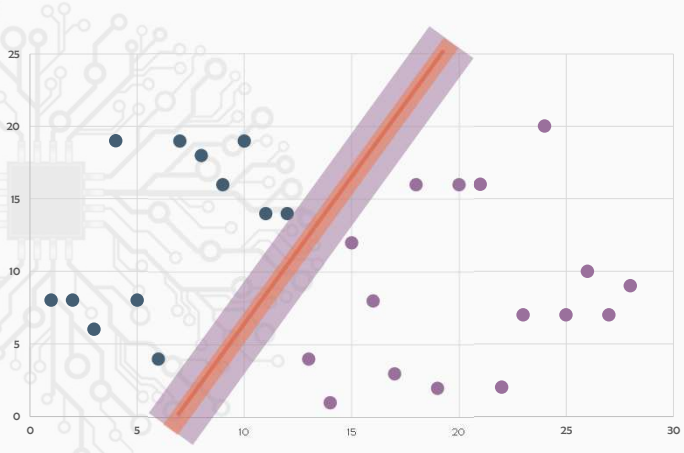
MACHINE LEARNING

Máquinas de soporte vectorial

margen de clasificación

- Es la distancia del punto más cercano a la línea de separación de cada grupo
- Se busca que esa distancia sea la máxima para todos los grupos
- La línea que logre los mayores márgenes es la LSVM





MSC RENZO CLAURE ARACENA

56

MACHINE LEARNING

MÁQUINAS DE SOPORTE VECTORIAL
REGULARIZACIÓN

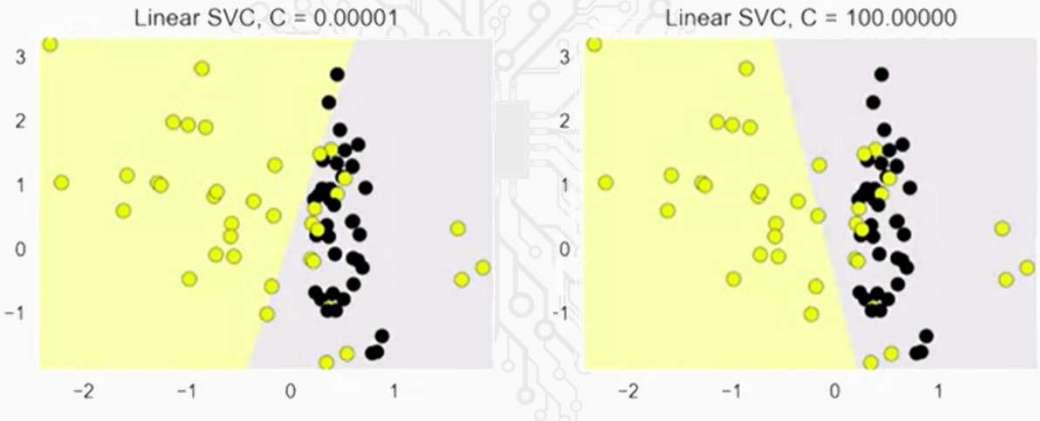
- Se controla por el parámetro C
- Valores de C grandes ajustan más al modelo, tendiendo al sobre ajuste, donde cada punto es importante
- Valores bajos de C, sub ajustan el modelo, mejorando la generalización entonces aumentando la regularización. Es decir es más tolerante a errores

57

MACHINE LEARNING

Máquinas de soporte vectorial

- NB_8



58

MACHINE LEARNING

Ventajas y desventajas de usar svml

Ventajas

- Especialmente bueno en espacios con muchas dimensiones
- Eficiente en memoria
- Funciona bien en muestras con alta variabilidad
- La predicción es fácil de comprender

Desventajas

- Tiene tendencia al sobreajuste
- No brinda una probabilidad, si no una valor determinístico
- No es bueno con muchos datos, por ejemplo más de 1000 casos

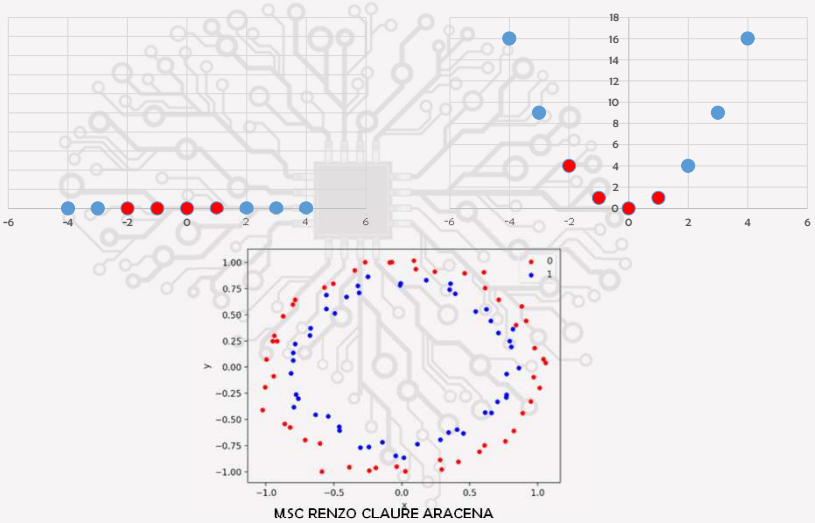
MSC RENZO CLAURE ARACENA

59

MACHINE LEARNING

Máquinas de soporte vectorial

espacios complejos



MSC RENZO CLAURE ARACENA

60

MACHINE LEARNING

Máquinas de soporte vectorial
espacios complejos

- VIDEO

61

MACHINE LEARNING

Máquinas de soporte vectorial
kernels

- Es el método de elevar los datos a un espacio de mayores dimensiones, a través de transformaciones, a este proceso se llama Kernelización
- Con esta transformación los datos son linealmente separables
- El algorimo realiza el mapeo" a través de cálculos lineales
- Dos tipos de funciones son las más aplicadas
  - Polinomicas
  - Gausianas o Funciones de base radial
- La transformación realizada es más simple

62

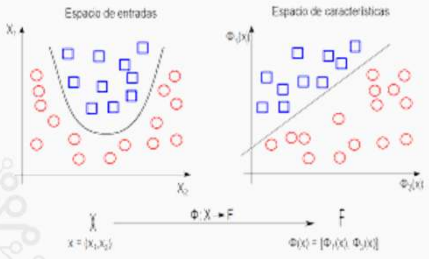
MACHINE LEARNING

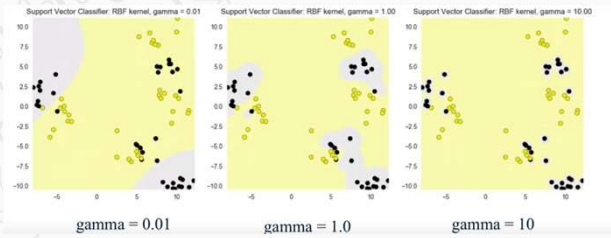
Máquinas de soporte vectorial

radial basis function kernel

- La función kernel nos dice cual es la similaridad de dos puntos en el espacio original, con respecto al nuevo espacio
- RBF:
- A traves de la RBF llevamos del espacio de la izquierda al de la derecha
- Existen otros tipos de Kernel, como el polinómico
- Ahora contamos con un nuevo parámetro llamado gamma y sirve para controlar como influyen los puntos de entrenamiento

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$$





MSC RENZO CLAURE ARACENA

63

MACHINE LEARNING

MÁQUINAS DE SOPORTE VECTORIAL

PARÁMETROS γ Y C

- Un gamma bajo, implica que los márgenes son más amplios, es decir se reduce el sobre ajuste
- Un valor elevado de gamma implica que los el Valor del kernel decae más rápidamente por lo que solo los más cercanos se consideran similares, resultando en límites más complejos
- En cuanto a C, controla la compensación entre la maximización de los limites y la reducción de los no clasificados
- Si gamma es grande, entonces C no tiene casi efecto. Si gamma es pequeño, entonces C se comporta como en un kernel lineal
- Los niveles en los que se mueve gamma son [0,001 a 10] y para C de [0,1, 100]

MSC RENZO CLAURE ARACENA

64

MACHINE LEARNING

Ventajas y desventajas de los svm con kernel

Ventajas

- Funciona muy bien con más Variables que con instancias
- Versátil
- Tiene relativamente buenos resultados en términos de eficacia

Desventajas

- Baja eficiencia por el uso de recursos
- Se vuelve sensible a datos muy dispersos, se debe normalizar a las variables independientes
- Tiende al sobreajuste
- No arroja una probabilidad

MSC RENZO CLAURE ARACENA

65

Validación cruzada

MSC RENZO CLAURE ARACENA

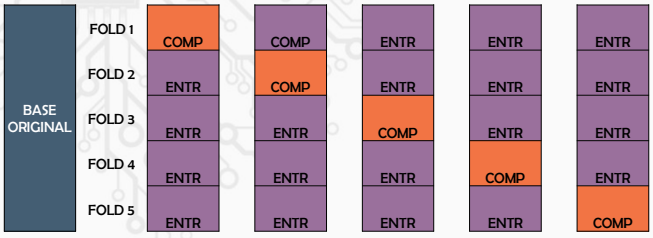
66

MACHINE LEARNING

Validación cruzada

cross validation

- Es una forma más robusta de medir el rendimiento del modelo
- Se logra a través de la generación de varios grupos (folds), extrayendo de cada uno una muestra de entrenamiento y otra de validación, generando los modelos y obteniendo sus scores
- Se debe asegurar que la proporción de la variable objetivo sea la misma que en la población, sklearn ya soluciona esto



MSC RENZO CLAURE ARACENA

67

MACHINE LEARNING

MSC RENZO CLAURE ARACENA

68

Árboles de decisión

MSC RENZO CLAURE ARACENA

69

MACHINE LEARNING

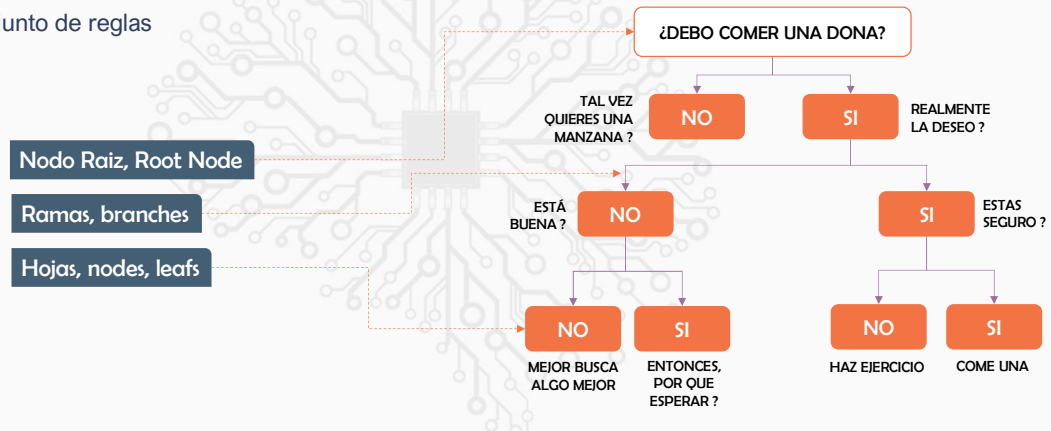
Arboles de decisión
definiciones

- Popular por su simplicidad de comprensión
- Conjunto de reglas

Nodo Raiz, Root Node

Ramas, branches

Hojas, nodes, leafs

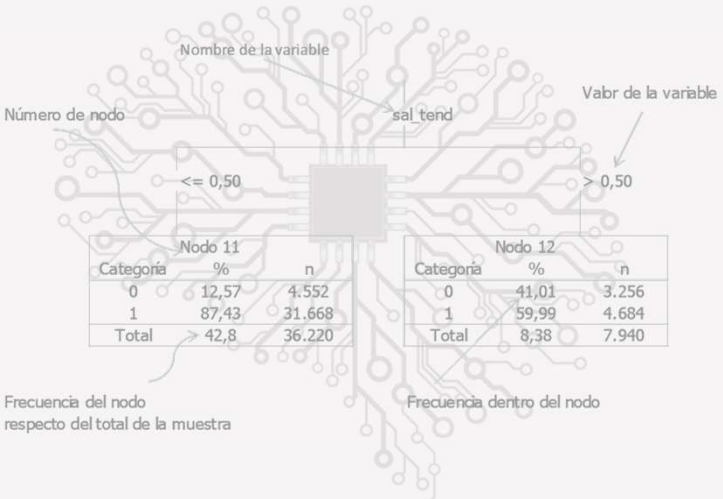


MSC RENZO CLAURE ARACENA

70

MACHINE LEARNING

Componentes de un árbol
una visión general



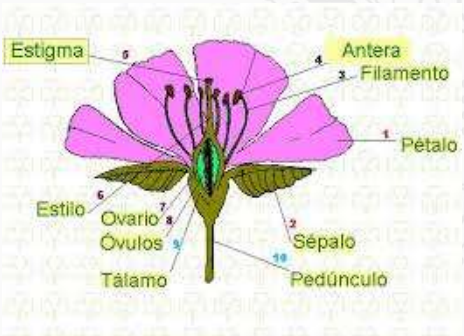
MSC RENZO CLAURE ARACENA

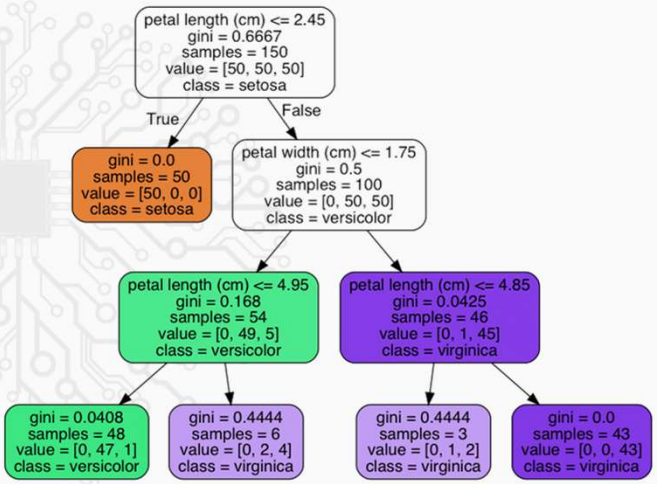
71

MACHINE LEARNING

Arboles de decisión
como leer un árbol completo

- Caso de variables continuas





MSC RENZO CLAURE ARACENA

72

MACHINE LEARNING

Prepoda y post-poda

- Para reducir el sobrer-ajuste de los árboles existen algunas técnicas de poda que pueden aplicarse:
- Prepoda:
  - max_depth: Controla el número máximo de niveles que tendrá un árbol, es la configuración más usada para reducir el sobreajuste
  - min_samples_leaf: estable el número mínimo de casos que una hoja puede tener, por debajo de esta cantidad el árbol ya no realizará más particiones
  - max_leaf_nodes: establece el número total de nodos u hojas en el árbol
- Postpoda: Una vez construido todo el árbol se podrían realizar cortes en ciertas hojas para reducir la complejidad. Scikitlearn no viene con esta configuración.

MSC RENZO CLAURE ARACENA

73

MACHINE LEARNING

Arboles de decisión

código

- NB_10

MSC RENZO CLAURE ARACENA

74

MACHINE LEARNING

Arboles de decisión

calculo de la importancia de la variables

- Es un indicador calculado a partir de la presencia e la variable en el árbol
- Varia entre 1 y 0
  - 0, no tiene efecto eb el resultado de la variable objetivo
  - 1, la variable predice adecuadamente la variable objetivo

75

MACHINE LEARNING

Arboles de decisión

ejercicio en clase

- Utilice la muestra de Vino Rojo
- Primero con las opciones por defecto
- Segundo, con el parámetro max_depth=4 y min_samples_leaf=5
- Compare los resultados y los árboles construidos

76

MACHINE LEARNING

Arboles de decisión

ventajas y desventajas

VENTAJAS

- Fácil interpretación
- Generalmente no se requieren normalizaciones
- Se adaptan bien a datos mixtos de variables categóricas y contínuas

DESVENTAJAS

- Tienden al sobreajuste, inclusive con tratamientos previos
- Los nuevos modelos de árboles ensamblados reducen el sobreajuste, pero a mayor costo de procesamiento

MSC RENZO CLAURE ARACENA

77