



UNIVERSIDAD
NACIONAL
DE COLOMBIA



Universidad Nacional de Colombia

Sede Medellín

Medidas de desempeño del algoritmo de entrenamiento



Profesora: Patricia Jaramillo A. Ph.D

Caso: discriminación salarial, se cuenta con datos de salarios de una empresa y se desea analizar que tanto influye ciertas variables en el salario. Archivo “Discriminación salarial.csv”

Empleado	genero (1 mujer)	Edad	Experiencia previa	experiencia en la empresa	Nivel educativo	Salario anual
1	1	39	5	12	4	57700
2	0	44	12	8	6	76400
3	0	24	0	2	4	44000
4	1	25	2	1	4	41600
5	0	56	5	25	8	163900
6	1	41	9	10	4	72700
7	1	33	6	2	6	60300
8	0	37	11	6	4	63500
9	1	51	12	16	6	131200



Cuáles relaciones podría establecer entre algunas variables?

El proceso de entrenamiento generalmente divide a la muestra de datos en 2 conjuntos:

- De entrenamiento (por ejemplo, el 80% de los datos)
- De prueba de calidad de predicción

La calidad del entrenamiento (aprendizaje) se mide a partir de unas funciones de desempeño diferentes a las de la la calidad de la predicción.

Yo puedo entrenar un modelo que:

Minimice el error entre resultados el modelo y los reales

- Maximice la similaridad entre los resultados el modelo y los reales
- Minimice la distancia entre los elementos de un cluster
- Maximice el numero de casos acertado por el modelo vs la realidad
- Otras

- En aprendizaje Supervisado

- El modelo debe tener la arquitectura y los valores de los parámetros adecuados para que su capacidad de predicción sea máxima.

- En aprendizaje No Supervisado

- El modelo debe tener la arquitectura y los valores de los parámetros adecuados para que la similitud entre datos de un mismo grupo, por ejemplo, sea máxima y la diferencia entre grupos sea máxima también.

Función de desempeño del entrenamiento

- Puede ser **lineal, cuadrática, logístico, exponencial**,..., o más complejo como una **red neuronal o una cascada de modelos**.
- Mide la calidad del modelo.
- Es la función objetivo a maximizar o minimizar mediante técnicas de **optimización**.
- Depende, además, si el modelo es de regresión (salida de valor), clasificación (salida de etiqueta), agrupamiento, etc.
- Puede usarse más de una función, por ejemplo en entrenamiento se quiere maximizar velocidad de aprendizaje y también podría maximizarse precisión predictiva.

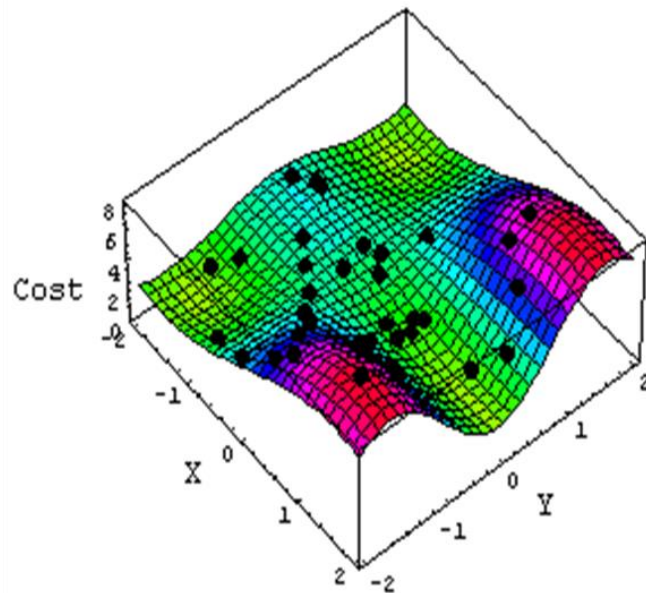
Recuerde que un problema general de optimización corresponde a:

Función objetivo: Optimizar (min o max) $f(x)$

Restricciones: $LB \leq x \leq UB$

$$G(x) \leq b$$

X son variables de decisión

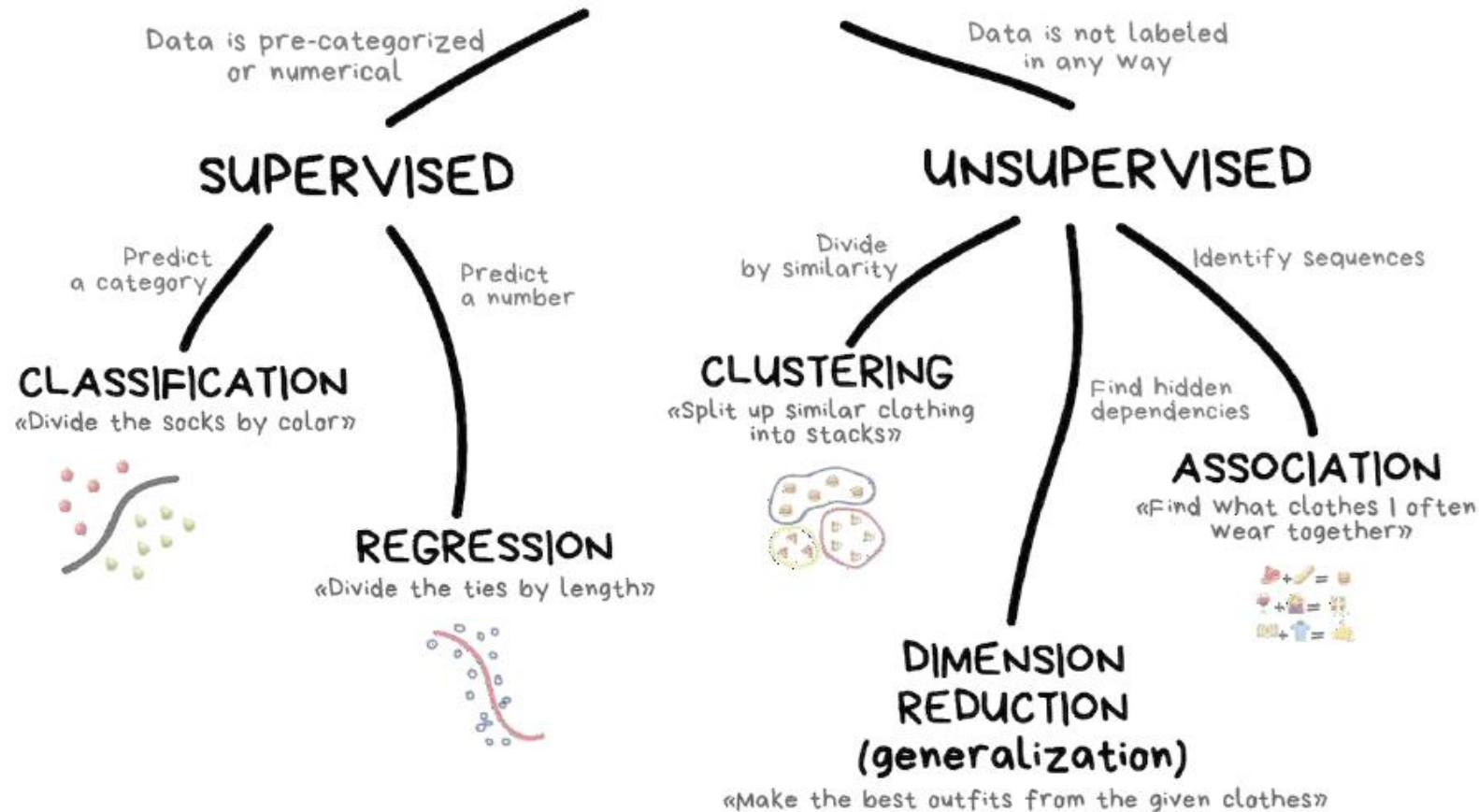


La calidad conseguida depende de:

La calidad de los datos

- Que sean representativos, diversos
 - Que las distribuciones de los casos en cada grupo o etiqueta sean similares: semejante número de casos con similar desviación estándar.
 - Cantidad significativa de casos de entrenamiento.
-
- La función de desempeño usada.
 - El método de optimización aplicado
 - El tiempo computacional usado
 - Etc.

CLASSICAL MACHINE LEARNING



Tomada de: https://vas3k.com/blog/machine_learning/

Cada una de estas herramientas requiere algún método de optimización para su ajuste y entrenamiento

En aprendizaje supervisado

Ejemplo: regresión lineal

Variables de decisión: $y = a + bX$ pesos de las variables de entrada e intercepto b_j ,
 a

Función de desempeño: (Minimizar) error= diferencia entre los valores de salida observados y los proporcionados por el modelo

Hay diferentes maneras de medir el error:

- Error instantáneo
- Error cuadrático medio
- Error cuadrático medio total

Evaluation indicators	Expression
Mean absolute error (MAE)	$\text{MAE} = \frac{1}{n} \sum_{i=1}^n f_i - y_i $
Mean absolute percentage error (MAPE)	$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left \frac{f_i - y_i}{y_i} \right \times 100\%$
Mean square error (MSE)	$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (f_i - y_i)^2$
Determination coefficient (R^2)	$R^2 = 1 - \frac{\sum_1^n (y_i - f_i)^2}{\sum_1^n (y_i - \bar{y})^2}$

n=numero de observaciones, f valor de salida del modelo, y= valor de salida en los datos (observado)

a=?

b=?

Empleado	X2= Edad	Y=Salario anual	Ymodelada =a+bX2	Error Ymodelada-Y	Error2
1	39	57700			
2	44	76400			
3	24	44000			
4	25	41600			
5	56	163900			
6	41	72700			
7	33	60300			
8	37	63500			
9	51	131200			



En aprendizaje supervisado

Ejemplo: entrenamiento de una red neuronal para regresión

Variables de decisión: pesos y sesgos de las conexiones w_{ij} , θ_{ij}

Función de desempeño: (Minimizar) error= diferencia entre los valores de salida observados y los proporcionados por el modelo

Hay diferentes maneras de medir el error:

- Error instantáneo
- Error cuadrático medio
- Error cuadrático medio total

Ejemplo: En un árbol de clasificación

Variables de decisión:

Orden de preguntas, contenido de la pregunta (umbrales de clasificación formuladas en forma de Verdadero/Falso).

Objetivo: (maximizar) número de clasificaciones precisas

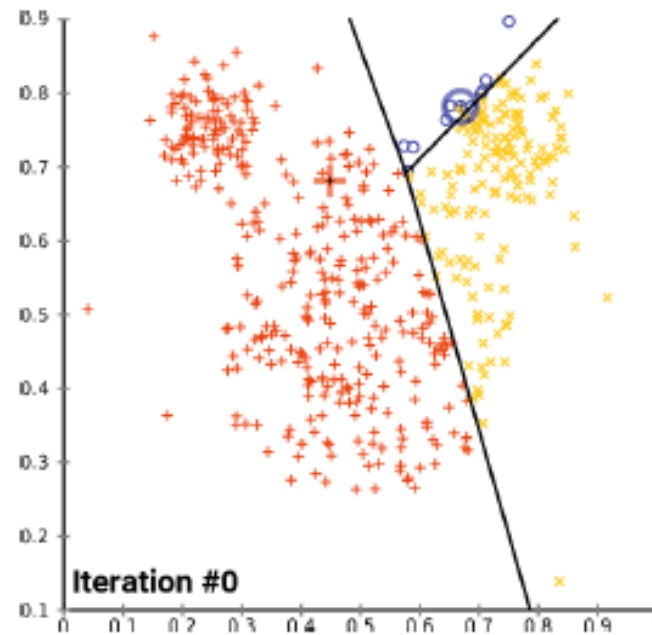
En aprendizaje No supervisado

Ejemplo: En un modelo de clusterización

Los datos que pertenezcan a cada grupo deben tener propiedades similares y muy diferentes a las de los otros grupos y así, para un nuevo caso que se desea clasificar, se hará en la clase más frecuente a la que pertenecen sus K vecinos más cercanos.

- Variables de decisión: grupo al que pertenece cada caso
- Función de desempeño:
 - **(maximizar) Disimilaridad entre grupos:** ejemplo: máxima distancia entre centroides, promedio grupal de distancias entre miembros de grupos
 - (Maximizar) Similaridad entre datos de cada grupo en cuanto a la distribución estadística de los datos conocida como estimación de la densidad.

El algoritmo de agrupamiento (clustering) mas común es el **K-means**: El analista decide cuántos grupos se definen (K). El algoritmo halla, de forma iterativa, k-centros y selecciona los puntos más cercanos al centroide de cada grupo.



Pero la similaridad puede medirse a partir de diferentes funciones

De distancia

Distancia manhattan

Distancia euclidiana

Distancia chevicheff

$$\text{Euclidean} - \sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

$$\text{Manhattan} - \sum_{i=1}^k |x_i - y_i|$$

$$\text{Minkowski} - \left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q}$$

In the case of categorical variables Hamming distance must be used.

Inercia de cluster

$$SSE = \sum_{i=1}^n \sum_{j=1}^k w^{(i,j)} \left\| \mathbf{x}^{(i)} - \boldsymbol{\mu}^{(j)} \right\|_2^2$$

Corresponde a

donde $\mu(j)$ es el centroide del grupo j y $w(i,j)$ es 1 si la muestra $x(i)$ pertenece al cluster j y 0 en caso contrario. Se desea minimizar el factor de inercia.

Cada forma diferente de medir distancias, similitud etc, aporta una calidad diferente del entrenamiento

- Por ejemplo, recuerde que , en clasificación supervisada, se puede usar:
 - Error instantáneo
 - Error cuadrático medio
 - Error cuadrático medio total
- Veamos las diferencias a partir de la simplicidad de una regresión (lineal).