

## Agenda

### “Métodos de Aprendizaje Automático”

APRENDIZAJE AUTOMÁTICO

PROGRAMA DE CIENCIA DE DATOS

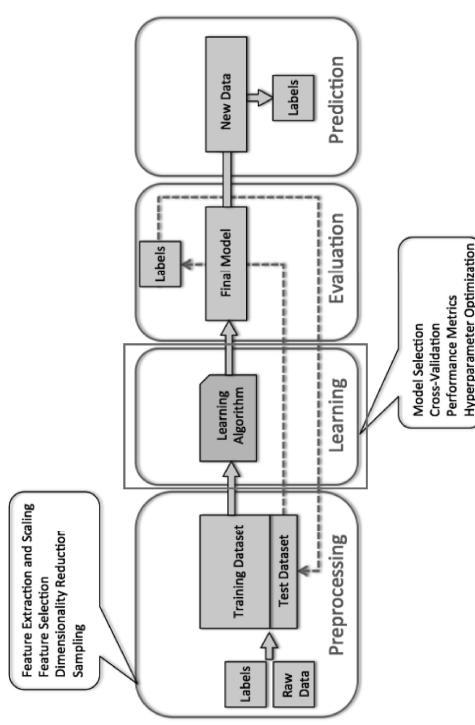
Profesor: MSc. Felipe Meza

TEC | Tecnológico  
de Costa Rica

October 14, 2021

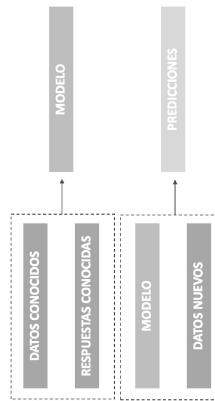


## Metodología de Diseño

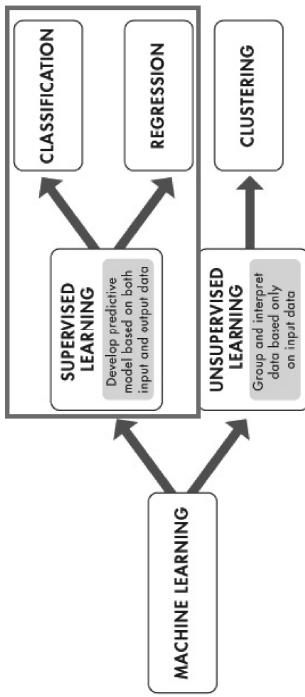


## Aprendizaje Supervisado

- Consiste en predecir una **salida** (clase) a partir de una **entrada** (atributos), y se cuenta con ejemplos “mapeados” de pares entrada/salidas (instancias).  $\{(x_i, y_i)\}_{i=1}^N$
- Se construye un modelo de aprendizaje automático a partir de los pares entrada/salidas (conjunto de **entrenamiento**). El objetivo consiste en utilizar dicho modelo para hacer predicciones precisas con datos nunca antes vistos (conjunto de **pruebas**).



Aprendizaje Supervisado



**"Métodos de Aprendizaje Automático"**

**"Métodos de Aprendizaje Automático"**



## Generalización, Overfitting, Underfitting

- Generalización:
    - Consiste en la capacidad del modelo de aprendizaje de hacer predicciones correctas con el conjunto de pruebas a partir del conjunto de entrenamiento.
    - Si lo hace correctamente se dice que el modelo “generaliza” de conjunto de entrenamiento al conjunto de pruebas.

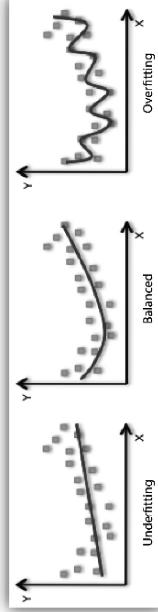
Generalización, Overfitting, Underfitting

- | Age | Number of cars owned | Owns house | Number of children | Marital status | Owns a dog | Bought a boat |
|-----|----------------------|------------|--------------------|----------------|------------|---------------|
| 66  | 1                    | yes        | 2                  | widowed        | no         | yes           |
| 52  | 2                    | yes        | 3                  | married        | no         | yes           |
| 22  | 0                    | no         | 0                  | married        | yes        | no            |
| 25  | 1                    | no         | 1                  | single         | no         | no            |
| 44  | 0                    | no         | 2                  | divorced       | yes        | no            |
| 39  | 1                    | yes        | 2                  | married        | yes        | no            |
| 26  | 1                    | no         | 2                  | single         | no         | no            |
| 40  | 3                    | yes        | 1                  | married        | yes        | no            |
| 53  | 2                    | yes        | 2                  | divorced       | no         | yes           |
| 64  | 2                    | yes        | 3                  | divorced       | no         | no            |
| 58  | 2                    | yes        | 2                  | married        | yes        | yes           |
| 33  | 1                    | no         | 1                  | single         | no         | no            |

  - Podemos experimentar 3 situaciones:
  - Regla 1: Va comprar un bote el que tenga más de 45 años, rde 3 hijos ó no esté divorciado. [muy complejo]
  - Regla 2: Va comprar un bote el que tenga una casa. [muy si
  - Regla 3: Alguna combinación intermedia.

## Generalización, Overfitting, Underfitting

- Overfitting:
    - Escoger un modelo de aprendizaje que sea muy complejo, funciona muy bien con el conjunto de entrenamiento pero no **generaliza** con nuevos datos.
  - Underfitting:
    - Escoger un modelo de aprendizaje que sea muy simple, funciona mal con el conjunto de entrenamiento y con nuevos datos no será capaz de **generalizar**.

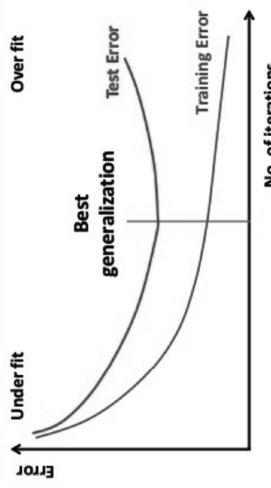


- ```
1 Training Score: 0.99  
2 Testing Score: 0.55  
3  
4  
5 Training Score: 0.57  
6 Testing Score: 0.62  
7  
8  
9 Training Score: 0.82  
10 Testing Score: 0.86  
11
```

## Generalización, Overfitting, Underfitting

```
1 # Overfit  
2 Training Score: 0.99  
3 Testing Score: 0.55  
4  
5 # Underfit  
6 Training Score: 0.57  
7 Testing Score: 0.62  
8  
9 # Fit  
10 Training Score: 0.82  
11 Testing Score: 0.86
```

## Generalización, Overfitting, Underfitting



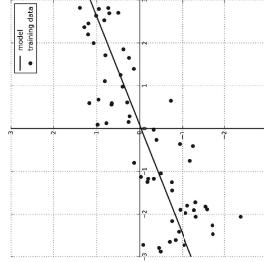


Modelos Lineales

- Son modelos de aprendizaje que hacen uso de la función *lineal* en los atributos de entrada:

$$q + [d]x \times [d]n + \cdots + [1]x \times [1]n + [0]x \times [0]n = \hbar$$

- Para el caso de único atributo:



MODEL 0

$$f_{\mathbf{w}, b}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

COSTO

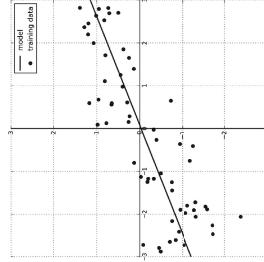
$$\frac{1}{N} \sum_{i=1 \dots N} (f_{\mathbf{w}, b}(\mathbf{x}_i) - y_i)^2$$

Modelos Lineales

- Son modelos de aprendizaje que hacen uso de la *función lineal* en los atributos de entrada.

$$q + [d]x \times [d]n + \cdots + [1]x \times [1]n + [0]x \times [0]n = \hbar$$

- Para el caso de único atributo:



COSTO

$$f_{\mathbf{w}, b}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

ପ୍ରକାଶକ ପତ୍ର ପରିଚୟ

卷之三

ପ୍ରକାଶକ ମୂଲ୍ୟ

-Métodos de Aprendizaje Automático

11

-Métodos de Aprendizaje Automático

18



Regresión Lineal RIDGE y LASSO

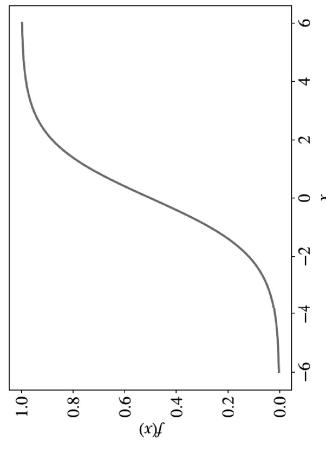
- LASSO (L1)
    - Se pueden obtener valores de  $w$  **iguales** a cero.
    - Reduce la **complejidad** y el **número de variables**.
  - RIDGE (L2)
    - Modelo que persigue obtener valores de  $w$  **cercanos** al cero con el fin de minimizar el efecto de los atributos mientras se mantiene un buen desempeño.
    - El parámetro  $\alpha$  permite balancear la simplicidad y el desempeño. Su valor óptimo depende de las características del conjunto de datos.
    - Reduce la **complejidad** pero no el **número de variables**.

Modelos Lineales



Logistic Regression

- Se busca una función que logre operar como una probabilidad de la forma  $0 \leq p \leq 1$ , esa función es la **logística** o **sigmoid**:



$$f(x) = \frac{1}{1+e^{-x}}$$

Logistic Regression

- El modelo queda de la siguiente forma:

$$f_{w,b}(x) \stackrel{def}{=} \frac{1}{1+e^{-(wx+b)}}$$

- El criterio consiste en maximizar la probabilidad de los datos de entrenamiento de acuerdo al modelo:
$$L_{w,b} \stackrel{def}{=} \prod_{i=1\dots N} f_{w,b}(x_i)^{y_i} (1 - f_{w,b}(x_i))^{(1-y_i)}$$
  - Con el fin de facilitar el uso de la ecuación anterior, se recurre al uso de logaritmos:

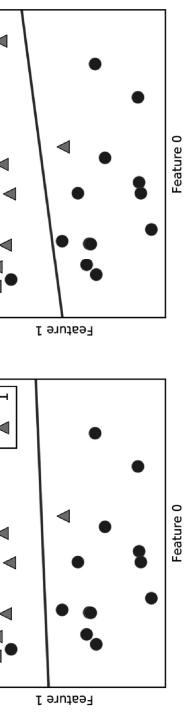
$$\log L_{w,b} \stackrel{def}{=} \ln(L_{w,b}(x)) = \sum_{i=1}^N y_i \ln f_{w,b}(x) + (1 - y_i) \ln(1 - f_{w,b}(x))$$



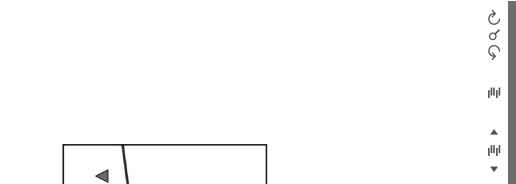


## Logistic Regression y SVM

- Se utiliza el parámetro **C** para controlar la regularización.
- Un valor alto de **C** implica menos regularización, el modelo trata de ajustarse a los datos de training lo mejor posible.
- Un valor bajo de **C** implica encontrar coeficientes  $w$  cercanos a cero.

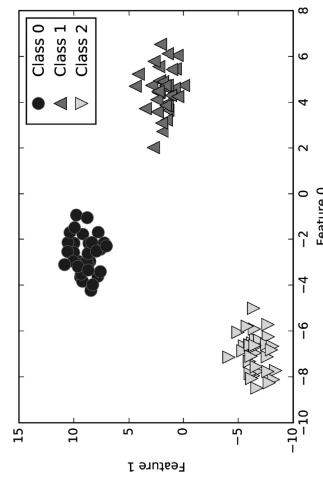


## Logistic Regression y SVM



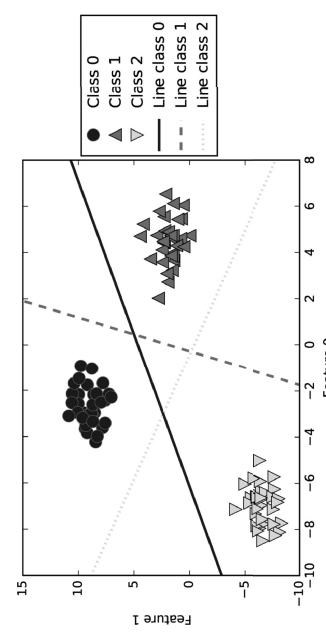
## Modelos Lineales Multi-Clase

- Consiste en usar modelos lineales usados comúnmente en clasificación binaria a clasificación con múltiples clases.



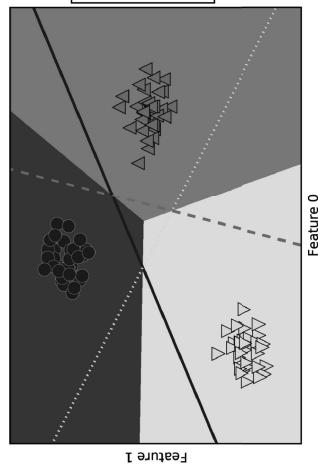
## Modelos Lineales Multi-Clase

- Se recurre al algoritmo *uno vs. el resto*, el cual consiste en separar cada clase del resto como si fuera clasificación binaria.



## Modelos Lineales Multi-Clase

- Los datos de las zonas centrales (en este caso el triángulo central) se asocian con la clase de la línea más cercana, definiendo así una nueva área por clase.



35

"Métodos de Aprendizaje Automático"

## Modelos Lineales - Recordar...

- Los parámetros de regularización corresponden a  $\alpha$  para los modelos de regresión lineal y a  $C$  para la regresión logística y SVM.
- Modelos simples implican un alto  $\alpha$  y un **C** bajo.
- Otro factor importante es la regularización, L1 se escoge cuando se tiene claro que sólo algunos atributos son importantes.
- Los modelos lineales son rápidos de entrenar y de hacer predicciones.
- Escalan muy bien para conjuntos de datos grandes o dispersos.
- Son de fácil comprensión.
- Un detalle importante es que los coeficientes no son fácilmente interpretables.

"Métodos de Aprendizaje Automático"

36