Elementos de Machine Learning

Diplomado de Análisis Computacional Estadístico de Datos con Python

Impartido por: Ronaldo Canizales

Bloque A

Aprendizaje supervisado: regresión.

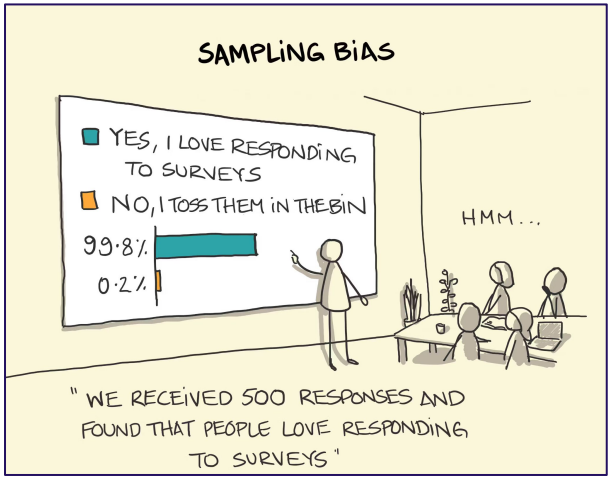
Principales desafíos del ML

- Cantidad insuficiente de datos (entrenamiento + prueba).
  - Principal diferencia con el aprendizaje humano.
  - *"The unreasonable effectiveness of data"* by Peter Norvig (2009).
- El conjunto de aprendizaje debe poseer las siguientes características:
  - **Ser significativo:**
  - Debe haber un número suficiente de ejemplos.
  - No podrá generalizar adecuadamente.

  - **Ser representativo:**
  - El conjunto debe ser diverso.
  - Se debe evitar tener muchos más ejemplos de un tipo que del resto.

Principales desafíos del ML

- Relativo al “toy example” de ayer:
- Sesgo muestral:



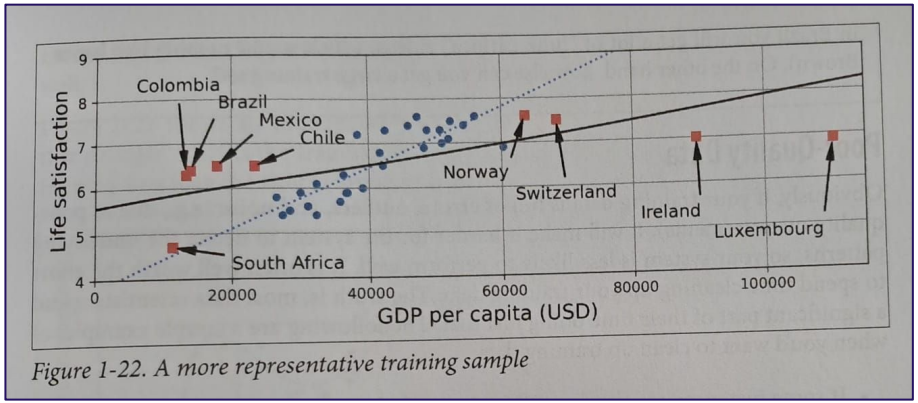
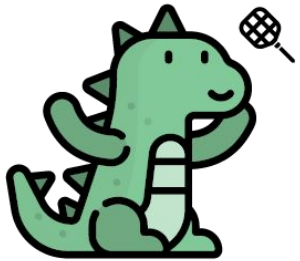


Figure 1-22. A more representative training sample

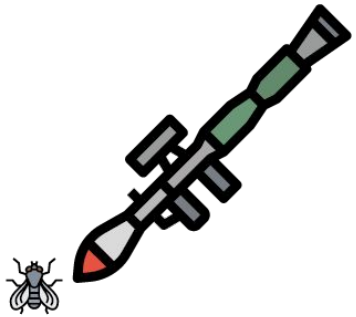
- Elección presidencial EEUU 1936: La revista “Literary Digest” predijo victoria demócrata con 57%, pero en realidad los republicanos ganaron con 62%.
  - A quienes enviaron encuestas?
  - Menos del 25% de los encuestados respondieron.

Principales desafíos del ML

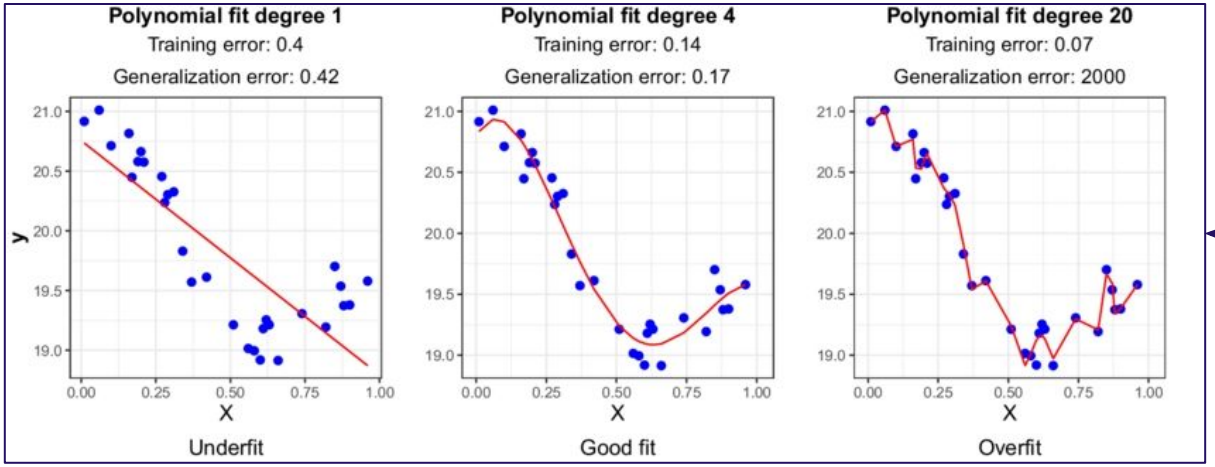
- Datos de entrenamiento de baja calidad:
  - Valores nulos, datos atípicos, y ruido (errores al obtener la data).
- Datos irrelevantes:
  - Es necesario invertir tiempo en “ingeniería de características”.
  - Se puede hacer uso inclusive de metodos de reduccion de la dimensionalidad.
- Overfitting and underfitting.



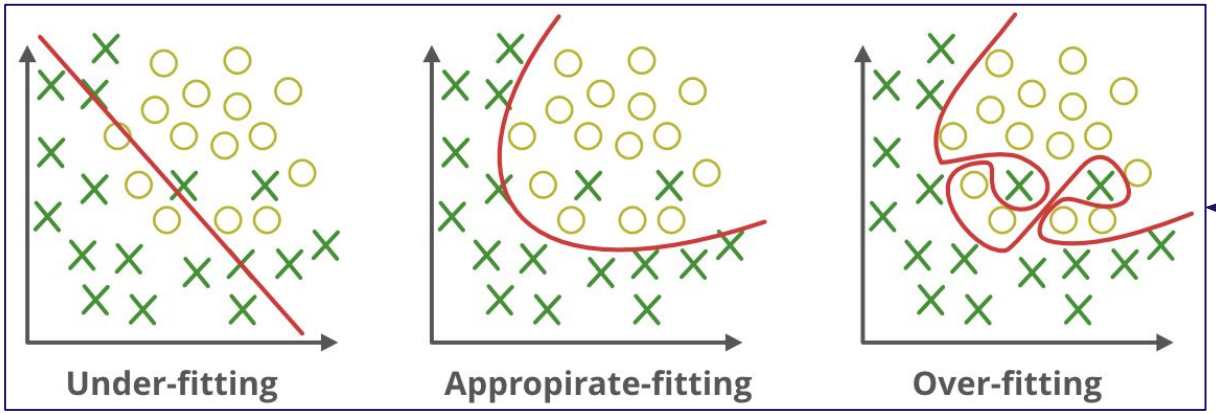
Underfitting



Overfitting



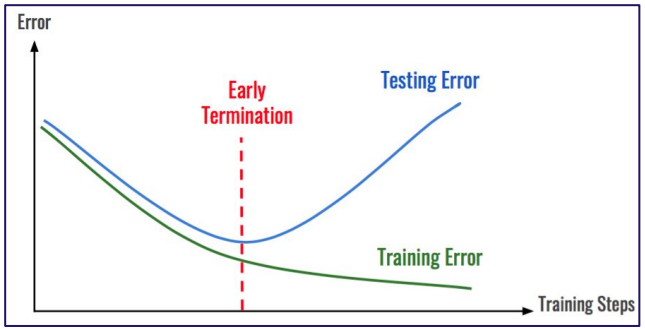
Regresión.



Clasificación

Principales desafíos del ML

- Soluciones al sobreajuste (overfitting):
  - Simplificar el modelo (disminuir la cantidad de parámetros).
  - Obtener más datos de entrenamiento.
  - Reducir el ruido en los datos de entrenamiento (remover datos atipicos o irrelevantes).
  - Early stopping!



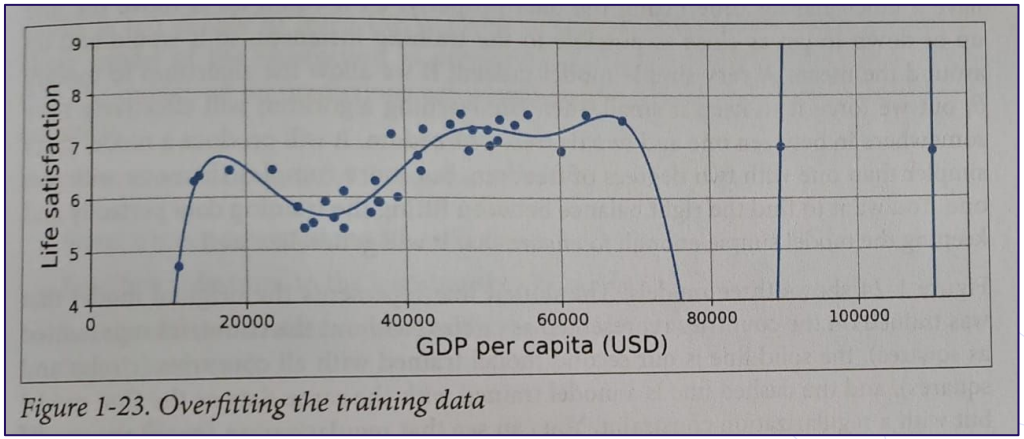
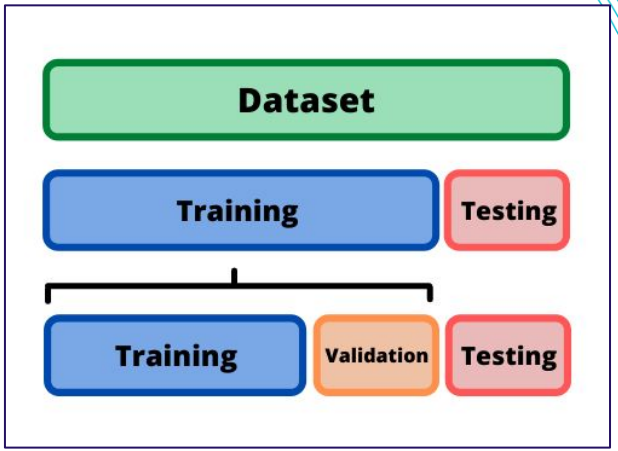
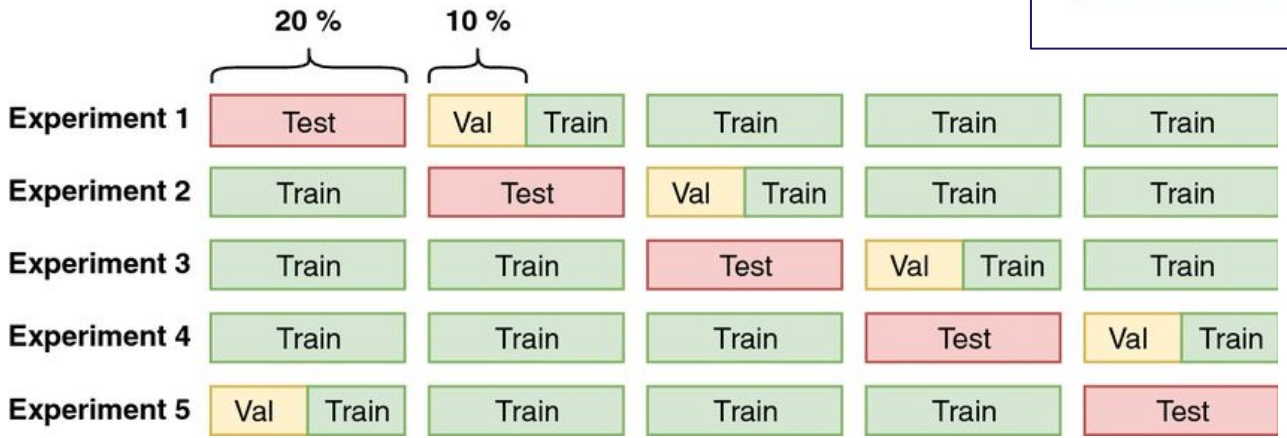


Figure 1-23. Overfitting the training data

Estrategias útiles

- Subconjuntos de datos: entrenamiento, validación y prueba.
- Selección de hyper-parámetros.
- Validación cruzada:





Student Performance

Donated on 11/26/2014

Predict student performance in secondary education (high school).

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Social Science	Classification, Regression
Feature Type	**# Instances**	**# Features**
Integer	649	30

Dataset Information

Additional Information

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

Enlace

Máquinas de soporte vectorial

Su propósito es encontrar un hiperplano que maximice la distancia entre las instancias de entrenamiento.

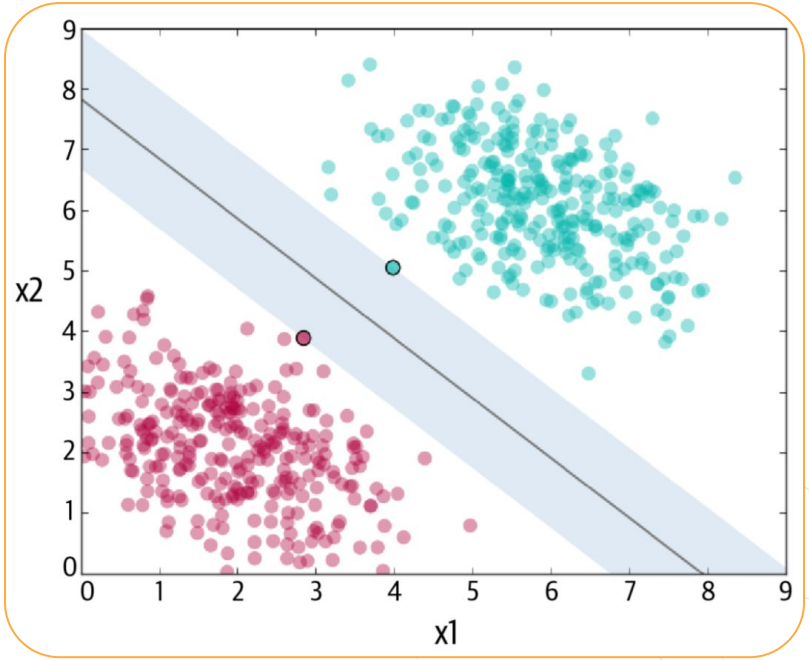
En la práctica, los datos casi nunca pueden ser separados perfectamente por un hiperplano (como en la figura), así que se hace necesario flexibilizar esa regla: permitir que algunos puntos violen la frontera.

Ventajas: es robusto ante el sobreajuste, especialmente en altas dimensiones. Es muy útil para casos no-lineales

Desventajas: no es muy intuitivo. Requiere gran cantidad de memoria cuando se utiliza en conjuntos de datos grandes.

```
from sklearn.svm import SVR
model = SVR()
model.fit(X, Y)
```

```
from sklearn.svm import SVC
model = SVC()
model.fit(X, Y)
```

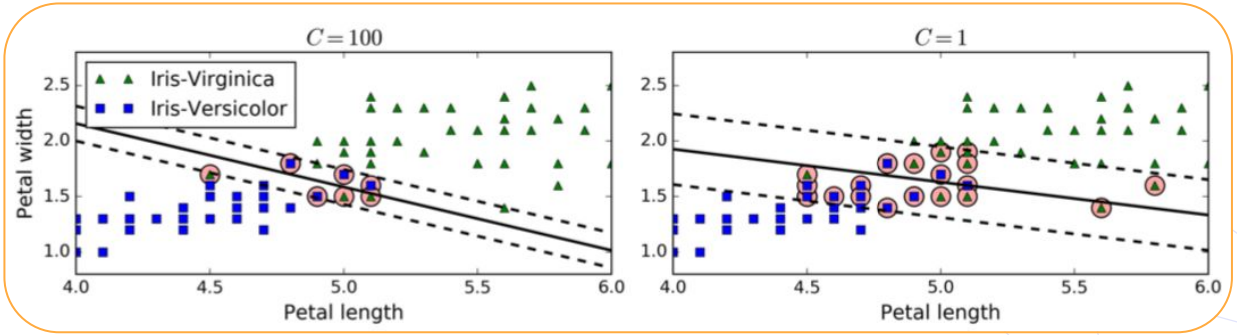


Hiper-parámetros de las MSV

Penalty (C en sklearn)

- Se le especifica al algoritmo qué tan estrictos debemos ser para evitar que hayan violaciones a la regla.
- Valores grandes crearán un hiperplano con un margen más pequeño.
- Valores pequeños conllevarán a márgenes más grandes, y de esa forma se reducirá el sobreajuste.

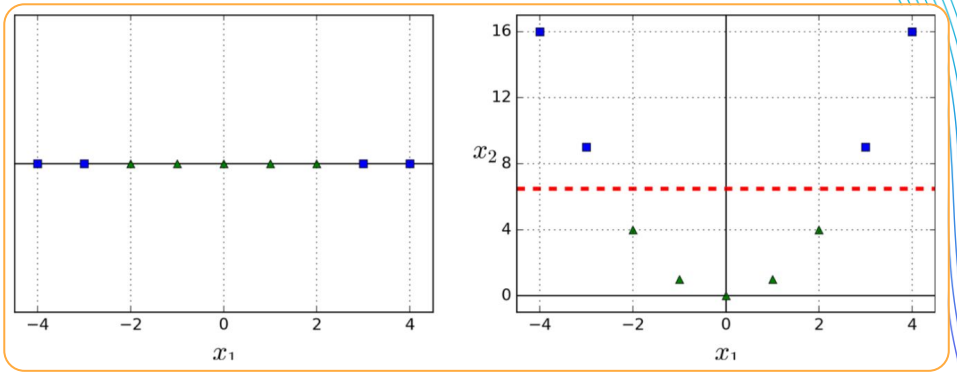
```
SVC(kernel="linear", C=1)
```



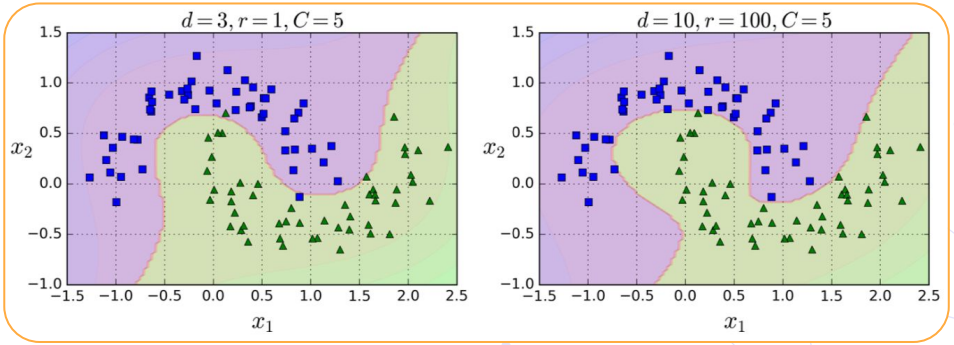
H-params en MSV

Kernels (kernel en sklearn)

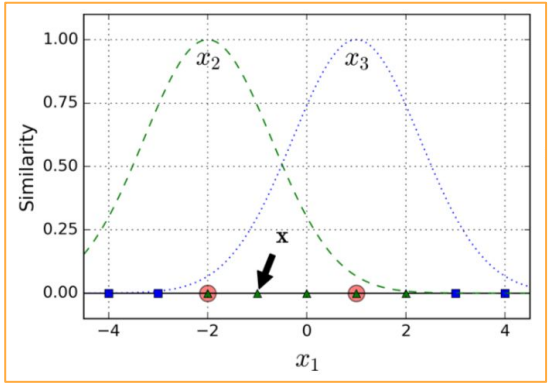
- Existen casos en que además de flexibilizar la regla, es necesario agregar una dimensión extra a los datos.
- Se controla la manera en que las variables de entrada serán proyectadas.
- Existen varios, pero: lineal, polinómico y RBF (Gaussian Radial Basis Function) son los más comunes.

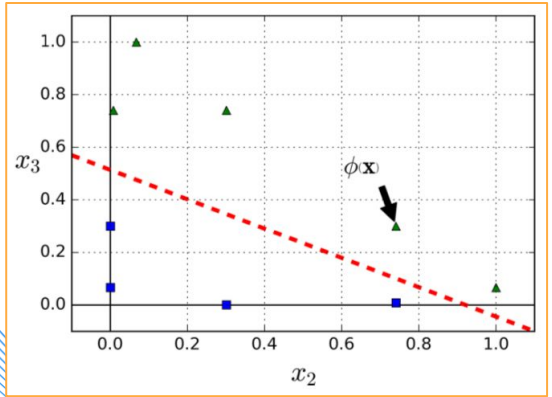



```
SVC(kernel="poly", degree=3, coef0=1, C=5)
```

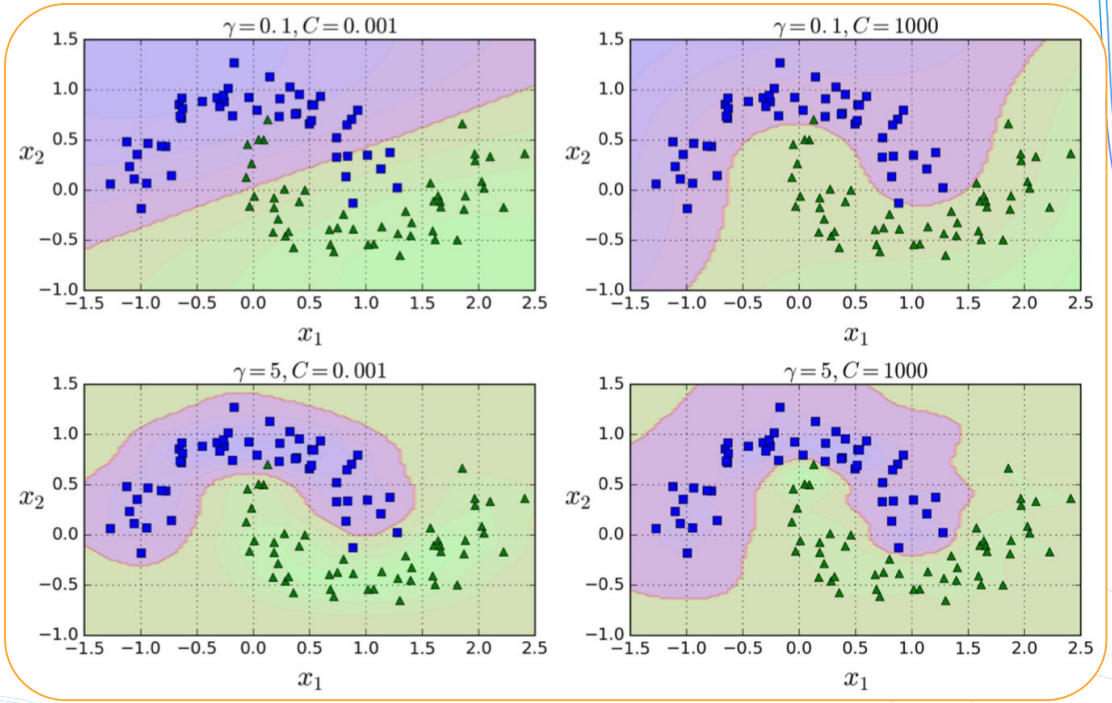


Hiper-parámetros de las MSV





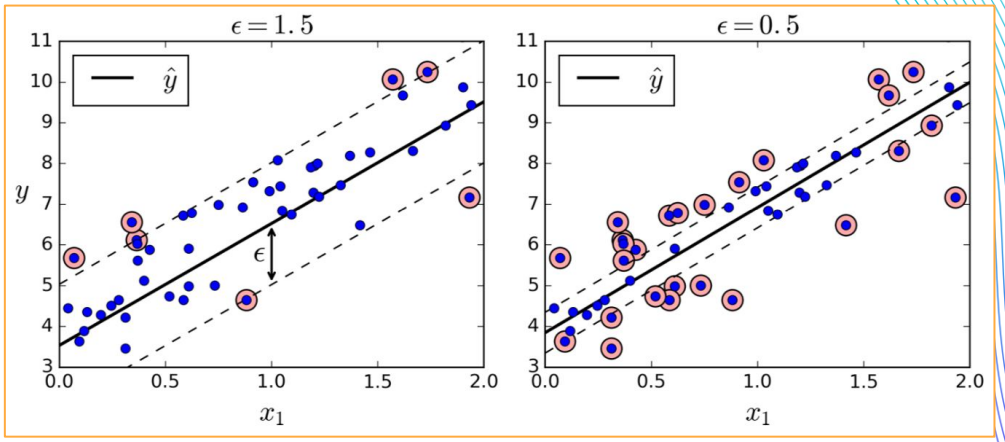

```
SVC(kernel="rbf", gamma=5, C=0.001)
```



H-params en MSV

```python
from sklearn.svm import LinearSVR

svm_reg = LinearSVR(epsilon=1.5)
svm_reg.fit(X, y)
```

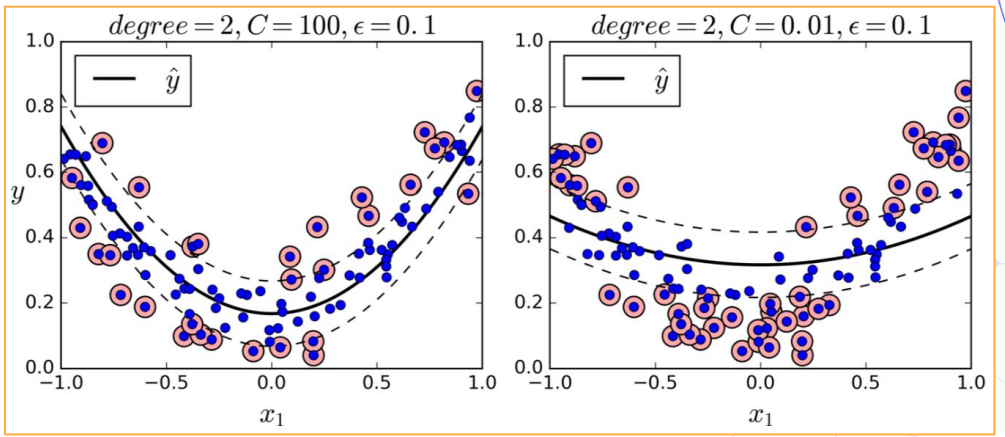


Epsilon (epsilon en sklearn)

- Intenta colocar tantas instancias como sea posible dentro de los márgenes.

```python
from sklearn.svm import SVR

svm_poly_reg = SVR(kernel="poly", degree=2, C=100, epsilon=0.1)
svm_poly_reg.fit(X, y)
```

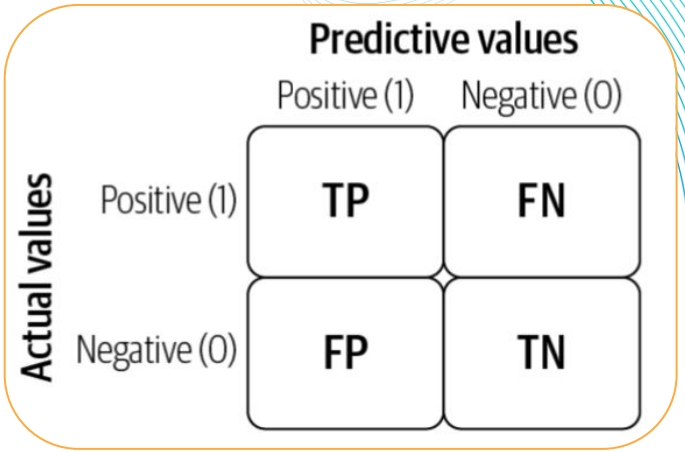


Bloque B

Aprendizaje supervisado: clasificación.

Matriz de confusión

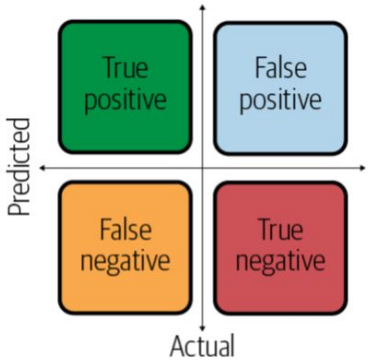
- En algunos escenarios será de mayor importancia un indicador que otro.
- Cuando las clases están balanceadas, entonces el Accuracy es muy útil, caso contrario no es de fiar.



$$\text{Precision} = \frac{\text{True positive}}{\text{Actual results}} \quad \text{or} \quad \frac{\text{True positive}}{\text{True positive + False positive}}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{Predictive results}} \quad \text{or} \quad \frac{\text{True positive}}{\text{True positive + False negative}}$$

$$\text{Accuracy} = \frac{\text{True positive + True negative}}{\text{Total}}$$



Cuando se trate de más de dos clases, las métricas deberán calcularse para cada una de ellas.

Bank Marketing

Donated on 2/13/2012

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Business	Classification
Feature Type	**# Instances**	**# Features**
Categorical, Integer	45211	16

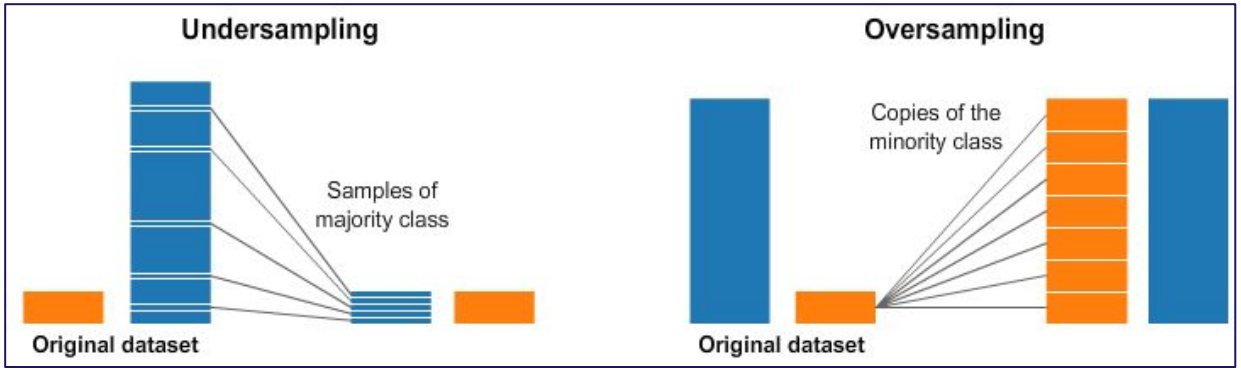
Dataset Information

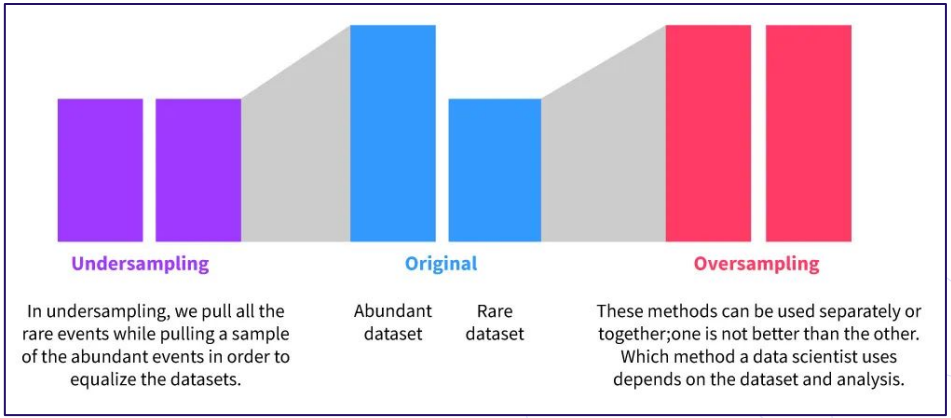
Additional Information

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Enlace

Detección de anomalías





¡Gracias por su atención!

Datos de contacto:
rcanizales@uca.edu.sv
rcanizal@colostate.edu
https://www.cs.colostate.edu/~rcanizal/
www.linkedin.com/in/ronaldo-canizales/
https://x.com/ArmandoCodigos