

# DATA SCIENCE



---

Federico Baiocco  
baioccofed@gmail.com  
3512075440



# Clase 20 - Agenda

REGRESIÓN

---

```
from sklearn import tree
```

```
clf = tree.DecisionTreeClassifier()
```

```
from sklearn import tree
```

```
reg = tree.DecisionTreeRegressor()
```

# Regresión



En un problema de regresión, en lugar de predecir variables categóricas, buscamos predecir una variable numérica.

Ejemplos:

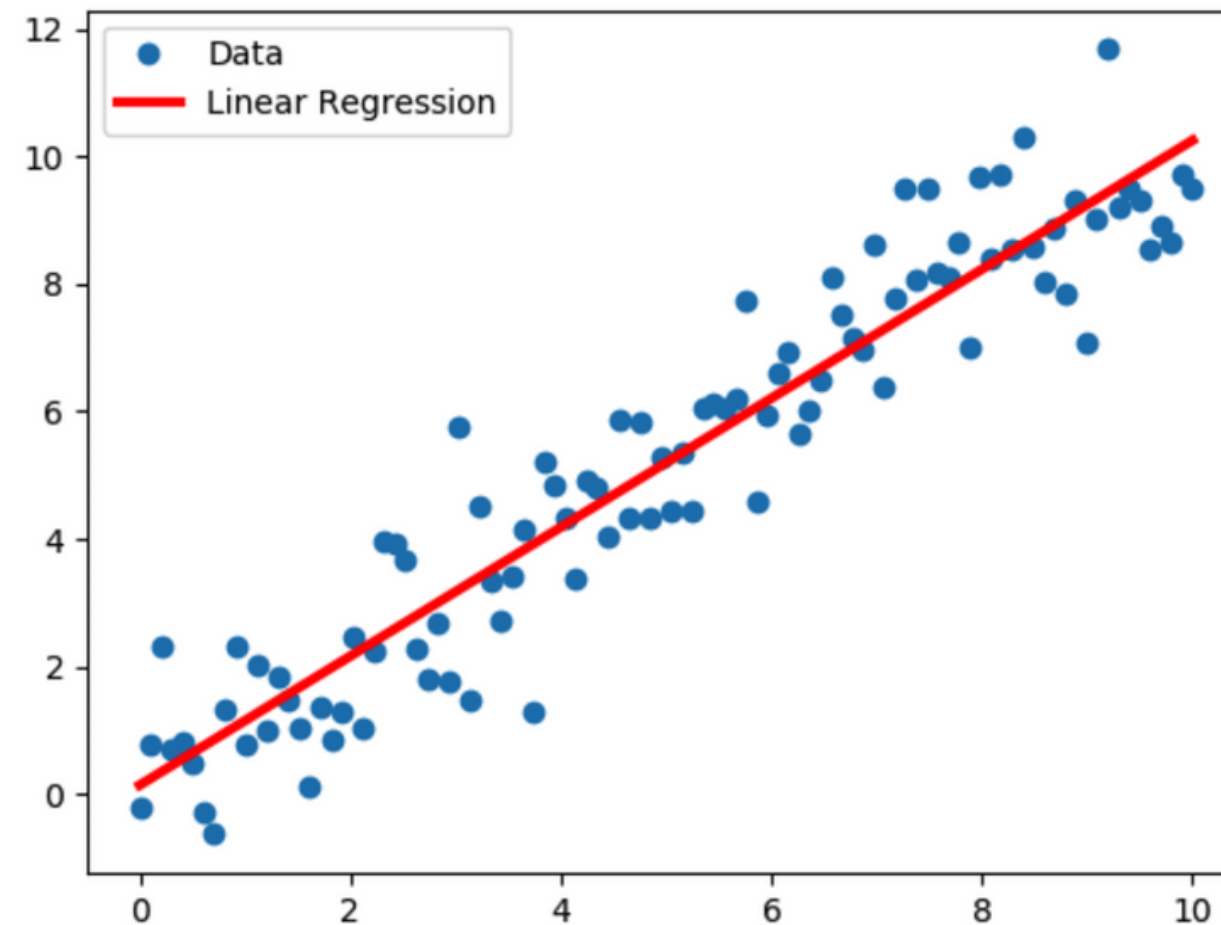
- Precio de una casa
- Precio de acciones
- Nota de un examen

# Regresión



Consiste en predecir un valor numérico "y" en base a los atributos "X" ( $x_1, x_2, x_3, \dots, x_n$ )

El caso más sencillo es una regresión lineal

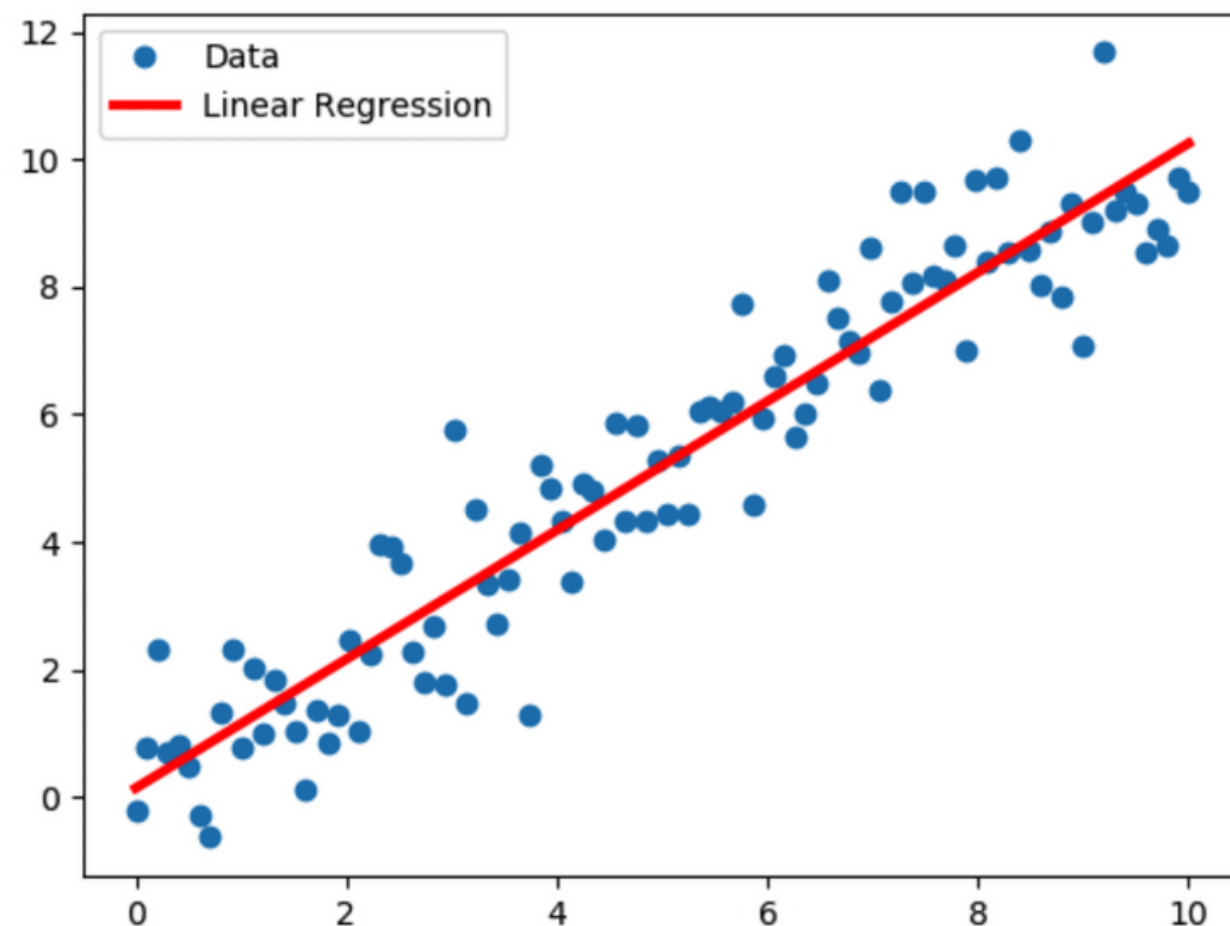


# Regresión



Consiste en predecir un valor numérico "y" en base a los atributos "X" ( $x_1, x_2, x_3, \dots, x_n$ )

El caso más sencillo es una regresión lineal



Buscamos aprender una función

$$Y = aX + b$$

Donde:

- a define la pendiente
- b la ordenada al origen

# Regresión



## `sklearn.linear_model.LinearRegression`

```
class sklearn.linear_model. LinearRegression(*, fit_intercept=True, normalize=False, copy_X=True, n_jobs=None, positive=False)
```

[\[source\]](#)

Ordinary least squares Linear Regression.

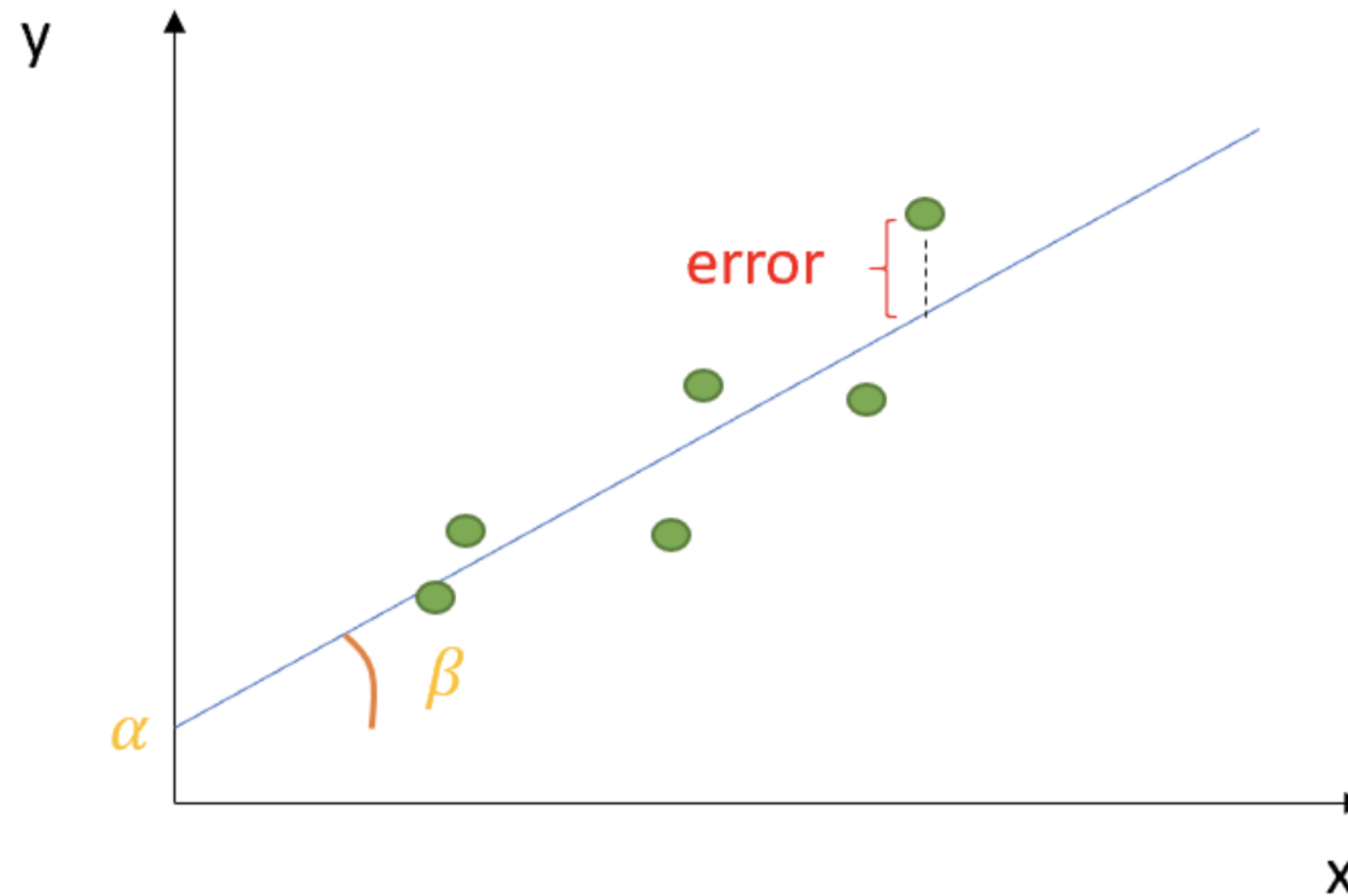
`LinearRegression` fits a linear model with coefficients  $w = (w_1, \dots, w_p)$  to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

¿ Ordinary least squares ?

# Regresión



La idea es encontrar los parámetros  $a$  y  $b$  tal que el error sea mínimo:

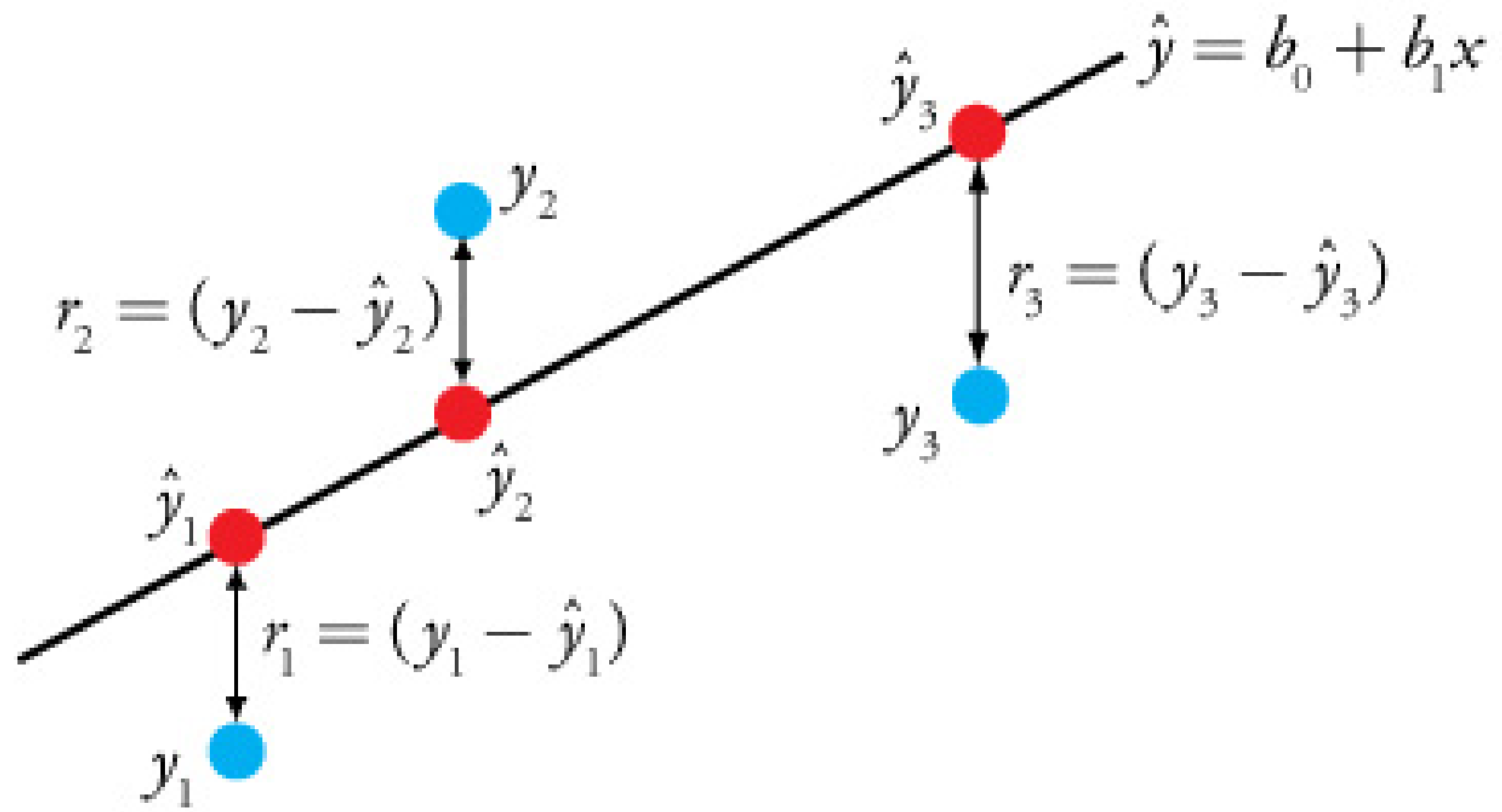




# Regresión - Error

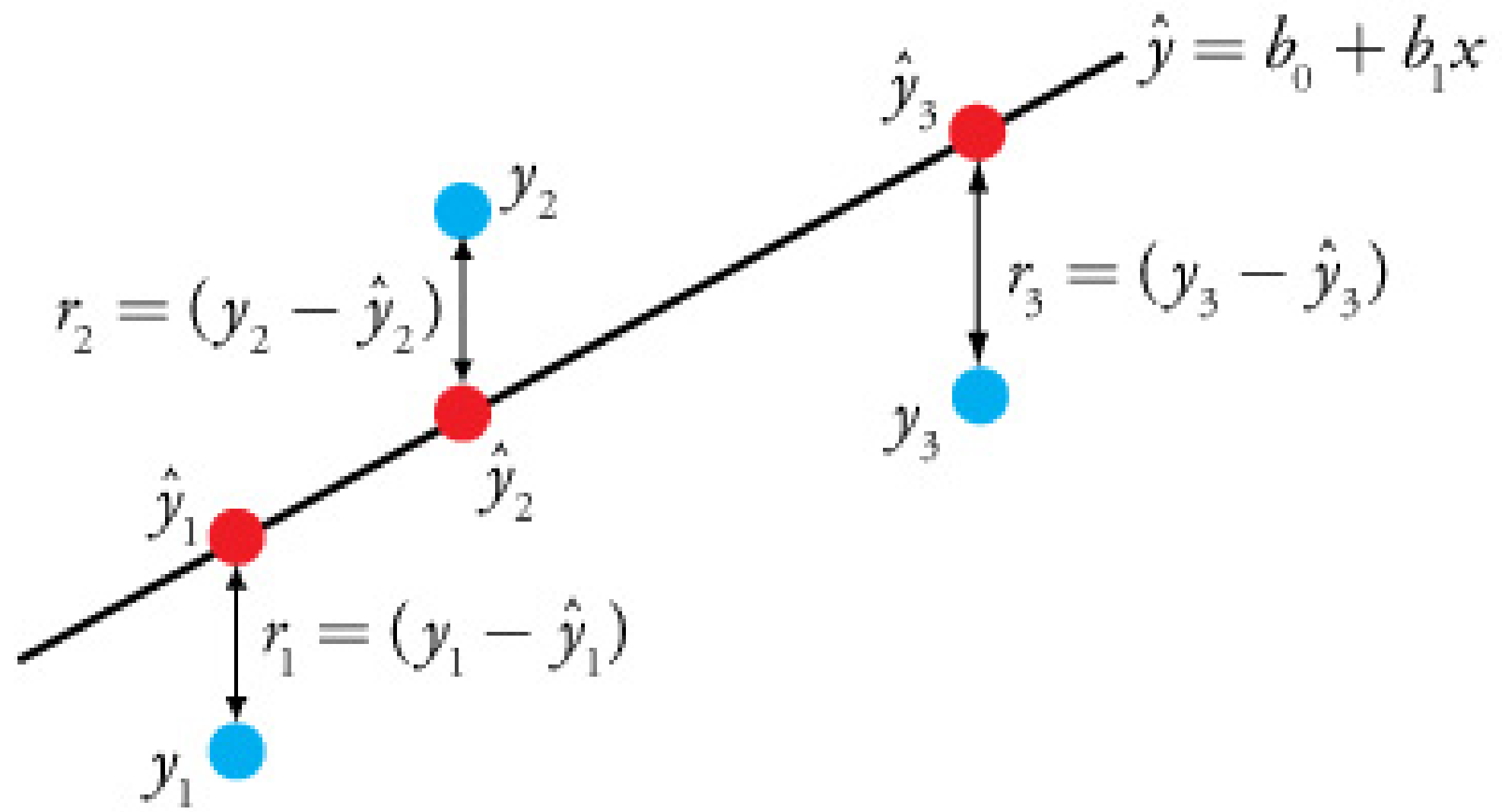


¿ Y cómo medimos el error ?



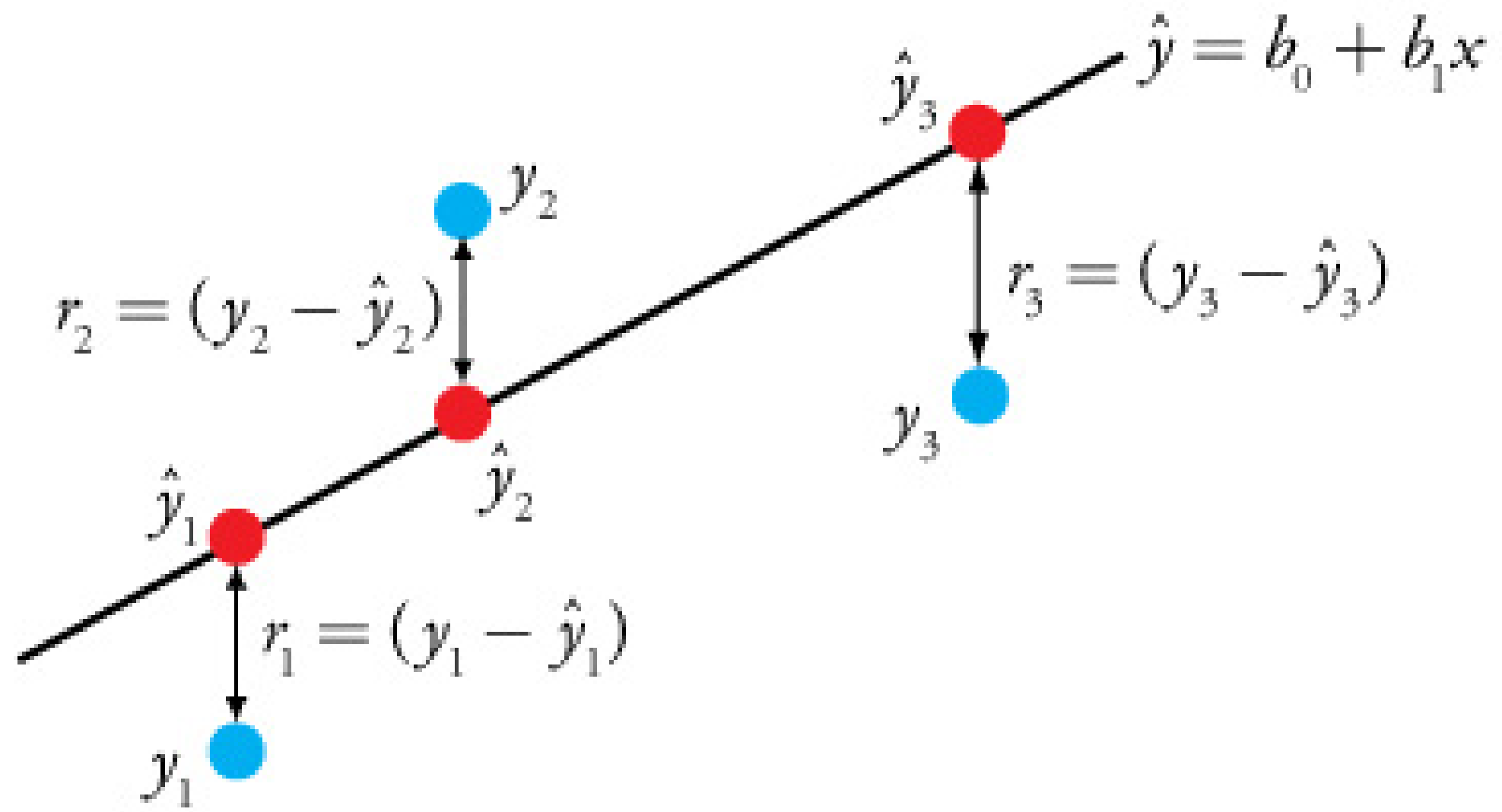
¿ Podemos sumar todos los errores y listo ?

# Regresión - Error



Los errores tienen distintos signos, si los sumamos, se cancelan entre si.

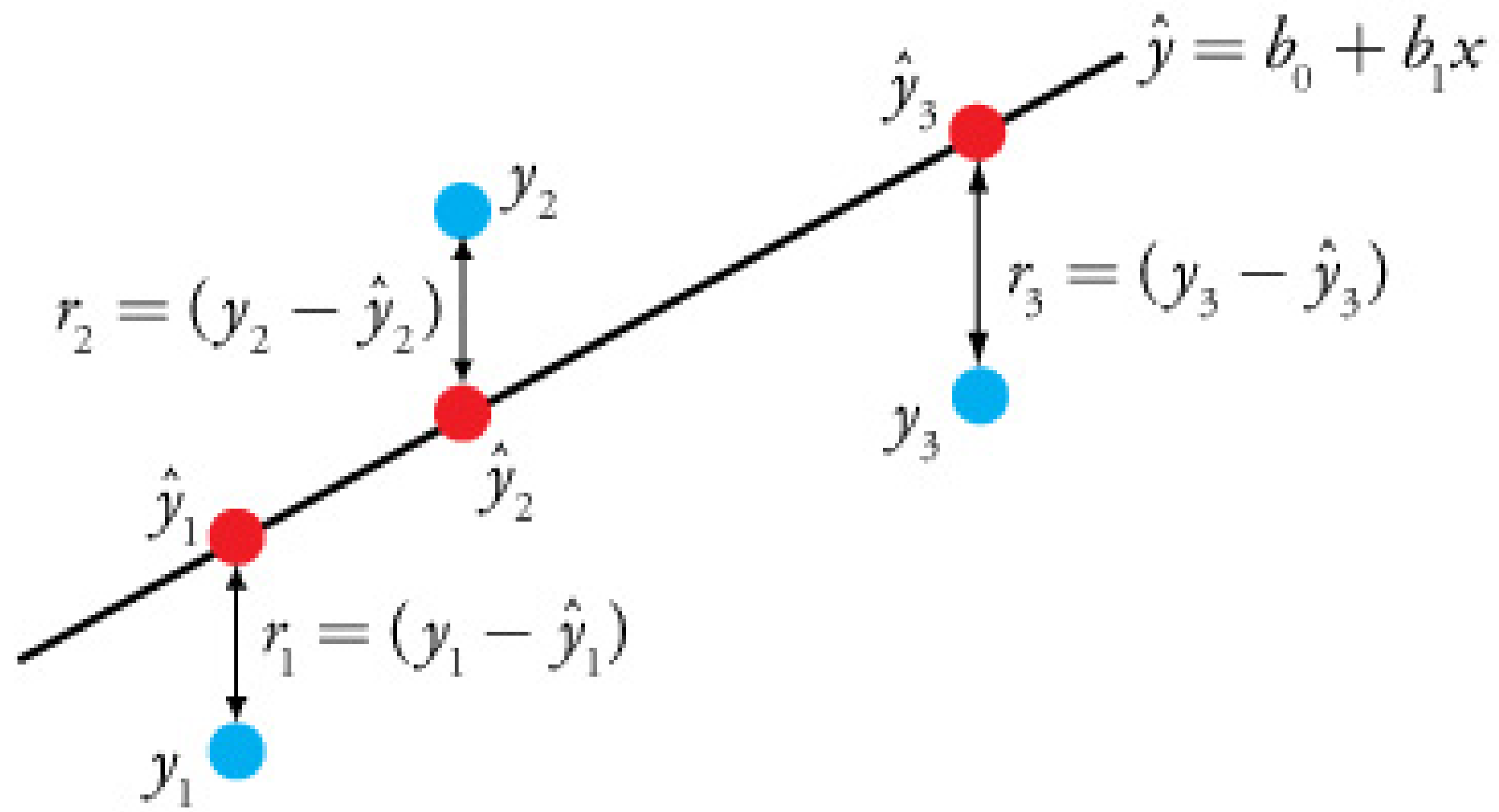
# Regresión - Error




Una solución, puede ser sumar sus valores absolutos:

$$\sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$

# Regresión - Error



$$\sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$

Ahora, sumando valores absolutos no se van a cancelar los errores entre si. Pero mientras más puntos tengamos, más grande va a ser el error 

# Regresión - Error



Para solucionar esto, se utiliza el Mean Absolute Error

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$

# Regresión - Error



Otra métrica que se utiliza para medir el error, es el Mean Squared Error.

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2$$

# Regresión - R squared



El R cuadrado ( $R^2$ , R cuadrado, coeficiente de determinación) se utiliza para medir que porcentaje de la varianza de la variable dependiente (y) es explicada por las variables independientes (X). De forma simple, nos dice qué tan bien se ajusta nuestro modelo a los datos. Toma valores entre 0 y 1.

- Un R cuadrado = 0: indica que la media de la variable dependiente (y) predice los valores de la variable dependiente tan bien como el modelo que entrenamos.
- Mientras más cerca de 1 esté el valor, mejor se ajusta nuestro modelo a la variable que estamos tratando de predecir

$$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Donde

- $y_i$  es el valor observado
- $\hat{y}_i$  es el valor que predecimos
- $\bar{y}$  es la media

# Regresión - Adjusted R squared



Un modelo, mientras más features tenga, mejor R cuadrado va a obtener.

Para tener en cuenta la cantidad de features que estamos utilizando, existe el R cuadrado ajustado.

El R cuadrado ajustado, está ajustado por la cantidad de features que utilizamos. De esta forma, su valor mejora solo si la nueva feature realmente mejora el modelo



# Regresión - Ridge regularization



Los modelos lineales son propensos a overfittear mientras más variables agregamos.

Las técnicas utilizadas para evitar el overfitting se conocen como técnicas de regularización.

En el caso de una regresión lineal, uno de los métodos para regularizar es Ridge o L2 regularization.

En Ridge, lo que se busca es que los coeficientes que se aprenden, sean lo más cercanos a 0 posible. De esta forma, cada feature va a tener el menor efecto posible sobre las predicciones.

Ridge está implementado en sklearn: [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.Ridge.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html)

# Regresión - Ridge regularization



Un hiperparámetro importante en ridge regression es el alpha.

**sklearn.linear\_model.Ridge**

```
class sklearn.linear_model.Ridge(alpha=1.0, *, fit_intercept=True, normalize=False, copy_X=True, max_iter=None,  
tol=0.001, solver='auto', random_state=None)
```

[\[source\]](#)

Con alpha, se controla que tan "restringido" va a estar el modelo.

Mientras más cercano a 1 sea el valor de alpha, mas forzados a estar cerca de 0 serán los coeficientes (menor overfitting).

En general, con Ridge vamos a obtener scores de train menores pero mejores en test.

# Regresión - Lasso regularization



Otra alternativa es Lasso o L1.

En Lasso, también se busca regularizar para evitar el overfitting, pero en este caso, además de empujar los valores para que estén cerca del 0, se hace que algunos coeficientes sean exactamente 0.

Esto hace que algunas features sean ignoradas por el modelo.

Acá, también tenemos el parámetro alpha para indicar que tan fuerte van a ser los coeficientes empujados hacia el 0.

**`sklearn.linear_model.Lasso`**

```
class sklearn.linear_model.Lasso(alpha=1.0, *, fit_intercept=True, normalize=False, precompute=False, copy_X=True,
max_iter=1000, tol=0.0001, warm_start=False, positive=False, random_state=None, selection='cyclic')
```

[\[source\]](#)

# Regresión - Clasificación



Ahora, en la práctica: ¿Qué cambia entre una regresión y una clasificación?

Los modelos que vimos (KNN, Decision tree) también pueden aplicarse para regresión.

Los conceptos que vimos como underfitting, overfitting, entrenamiento de un modelo, hiperparámetros, etc son lo mismo.

Lo que cambia principalmente son las métricas de evaluación.