Overfitting y Cross-Validation

1. El fenómeno del Overfitting.
2. Cross-Validation (CV)

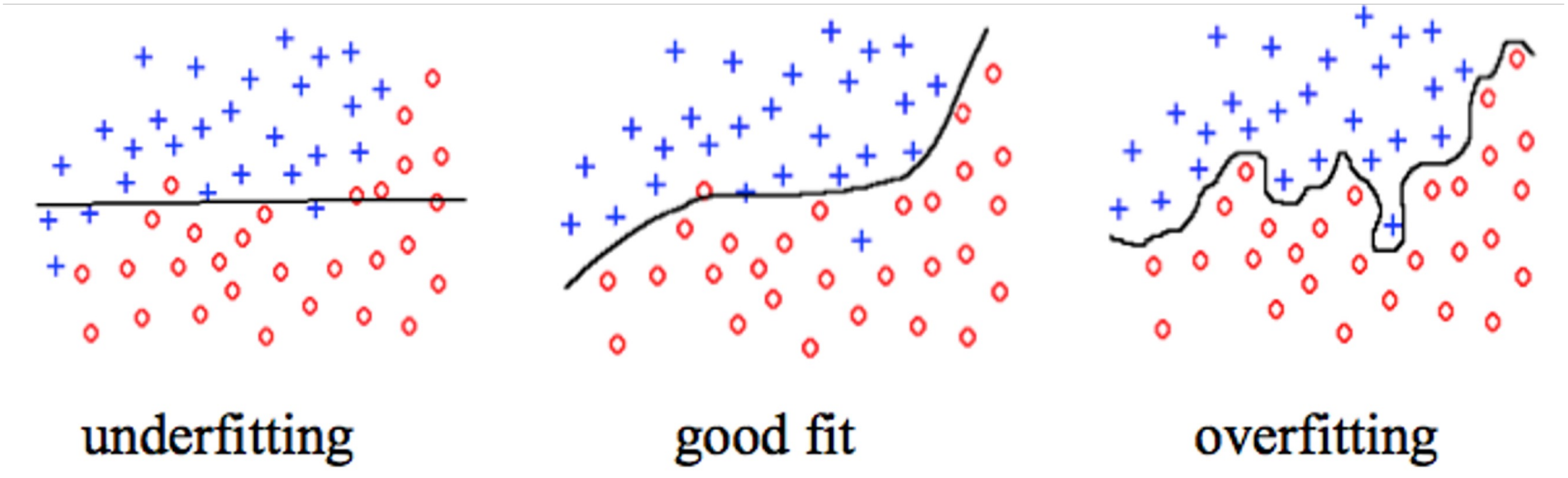
El fenómeno del Overfitting

Overfitting

El término **Overfitting** es traducido como **sobre entrenamiento**. Esto sucede cuando el modelo se adecuada demasiado a los datos de entrenamiento y esto conlleva a una pérdida en la capacidad de generalización.

Lo mismo pasa en sentido contrario: existe el **Underfitting.** Esto sucede cuando el modelo no logra captar los patrones de los datos, sea porque entrenó poco y/o porque le faltan datos.

Puede suceder tanto en regresión como en clasificación



Overfitting

- Una manera de identificar este fenómeno es revisando las métricas de rendimiento del proceso de entrenamiento y testeo.

- Si nos va muy mal en el entrenamiento y en el testeo, entonces posiblemente tengamos **Underfitting**.

  - **Ejemplo?** usamos una SVM de kernel lineal en un dataset que claramente no es separable linealmente. Como en la imagen anterior.

  - **Ejemplo?** Usamos un algoritmo adecuado, pero nos faltaron datos o más iteraciones.

- Si nos va muy bien en el entrenamiento, pero mal en el testeo, entonces posiblemente tengamos **Overfitting.**
  - **Ejemplo?** accuracy en el entrenamiento: 90% y en el testeo: 75%
  - Es más facil caer en el Overfitting, ya que uno tiende a buscar buenos resultados al comienzo. Por eso es importante contrastar ambos resultados (train y test).

Es posible (aunque es raro) que esto sucede por culpa de la aleatoriedad.

¿Por que?

Cross-Validation

CV

La aleatoriedad es una constante en Machine Learning.

Usamos procesos aleatorios constantemente, pero a veces puede jugarnos en contra.

¿Como?

¿Qué podria pasar al dividir el training y testing set?

Piense 1 minuto.

CV

Es posible que al realizar la división del dataset en training y testing set los datos más fáciles de modelar se queden en el train y los más difíciles en el testing set.

Esto implica que seguramente nos irá muy bien en el entrenamiento y cuando nos toque testear el algoritmo nos de muy malos resultados.

También es posible que pase lo contrario.

Para resolver esto, es que existe una técnica llamada Validación Cruzada, o más conocida como:

Cross-Validation (CV)

CV

- Cross-Validation es un método para validar métodos de predicción, como clasificadores o regresores.

- La idea principal es medir la capacidad de generalización del algoritmo. Esta capacidad es obtenida a través del training set y es medida finalmente por el testing set.

- Pero el testing set solo es utilizado 1 vez y al final, por lo que el algoritmo debe saber de alguna manera que tan bien está creando el modelo.

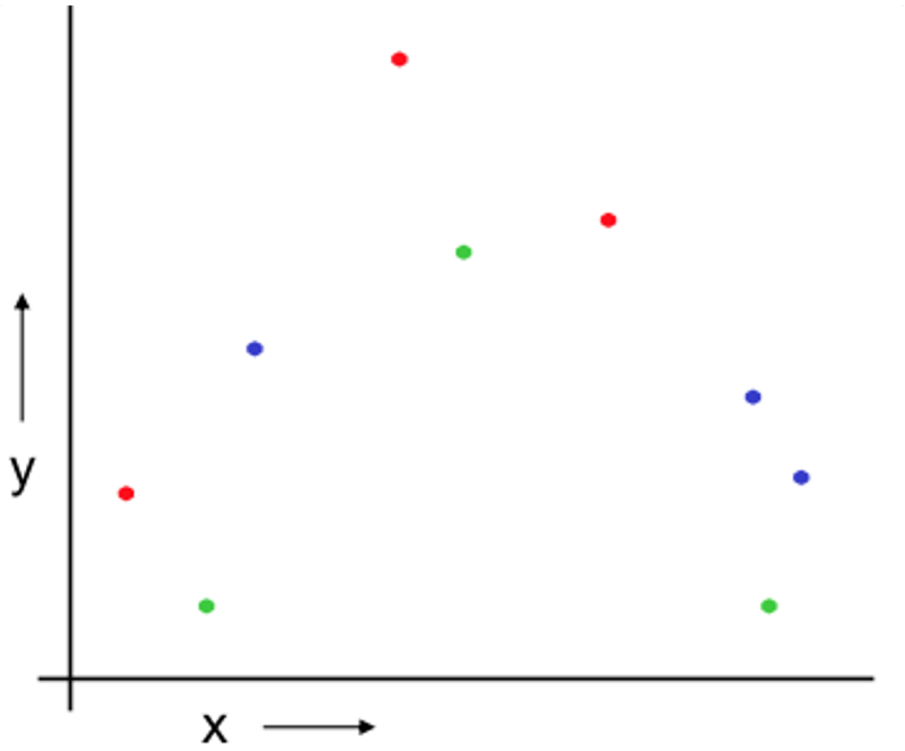
- Si bien, CV es el método elegido por excelencia, existe varias variantes:
  - Leave one out CV (LOOCV).

  - K-fold CV:
    - 5-fold CV
    - 10-fold CV

CV

Haremos un pequeño ejemplo con un problema de regression.

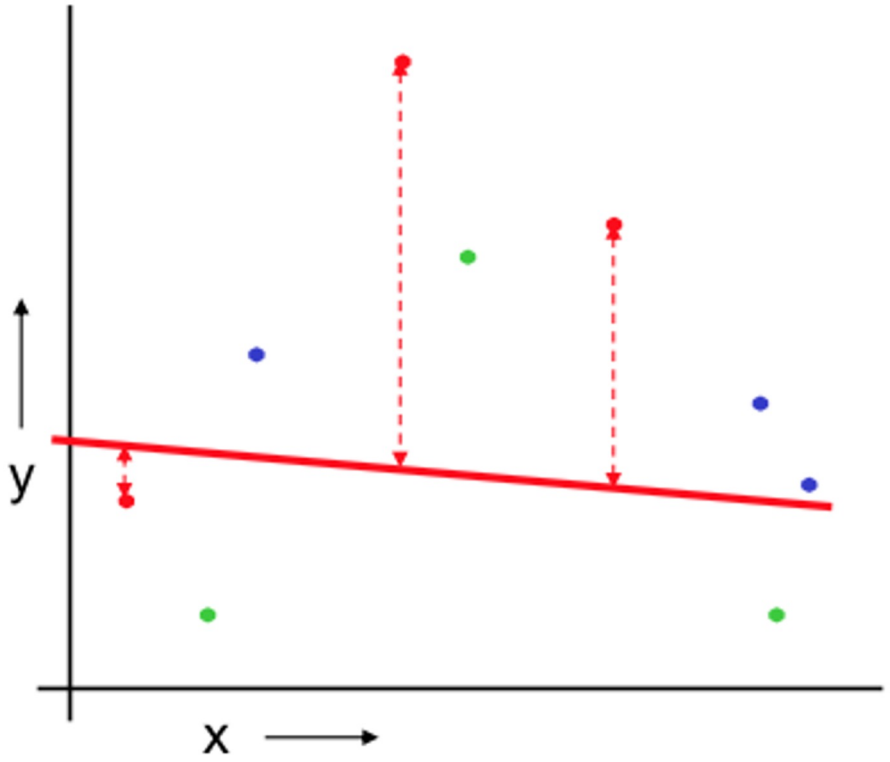
Ejemplo con 3-fold CV (por simplicidad solamente):



- Testing set: puntos rojos
- Training set: todo lo demás

CV

Ejemplo con 3-fold CV :

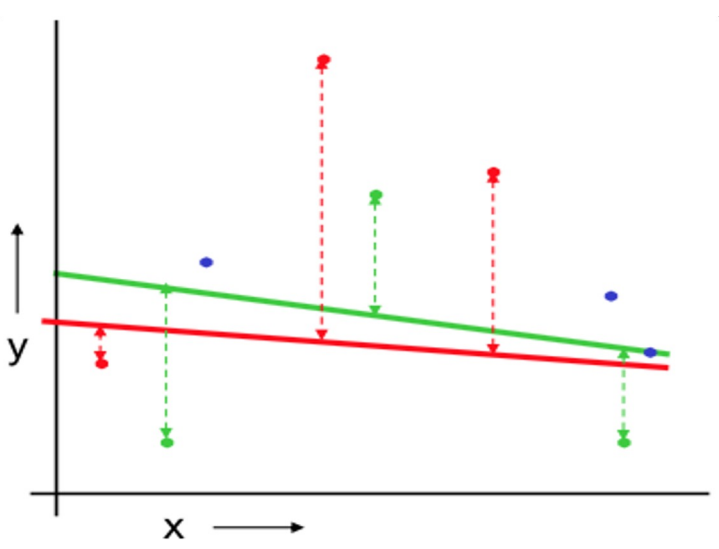


- Testing set: puntos rojos
- Training set: todo lo demás

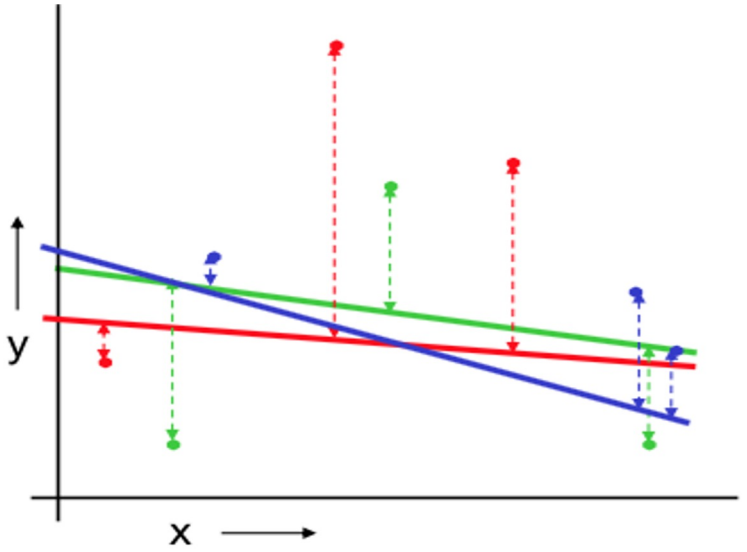
CV

Ejemplo con 3-fold CV :

- Testing set: puntos verdes



- Testing set: puntos azules

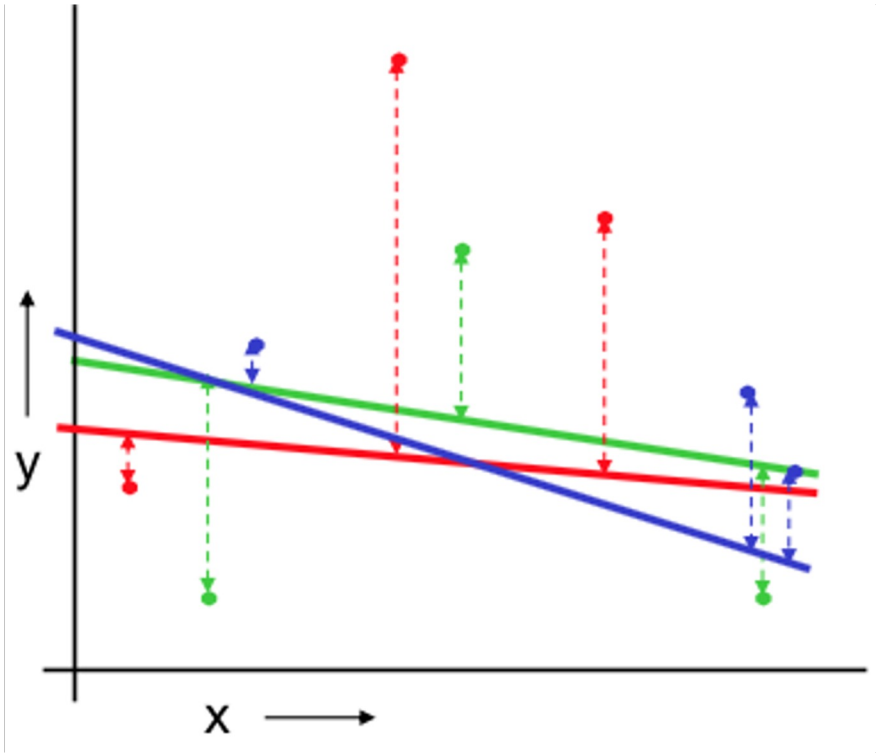


CV

Ejemplo con 3-fold CV:

Entonces se reporta el promedio del error:

$$MSE_{3-fold} = 2.05$$



Se puede hacer lo mismo para cualquier método de predicción.

CV

Lo usual es utilizar 10-fold CV: