Ciencia de Datos

- **Módulo 3**

Overfitting y *Underfitting*

UTN F r a

GyEE Gestión y Evaluación Educativa

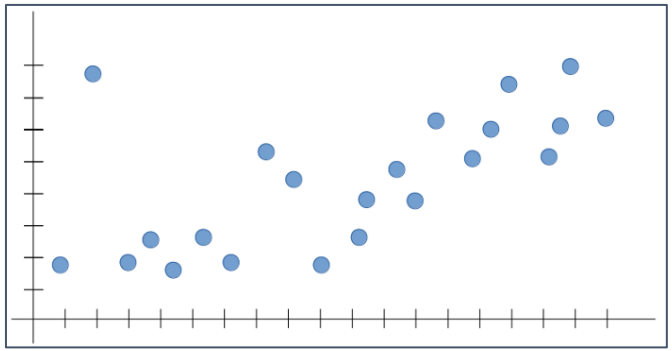
<aprendé programando> Virtual

SECUNDARIA del futuro

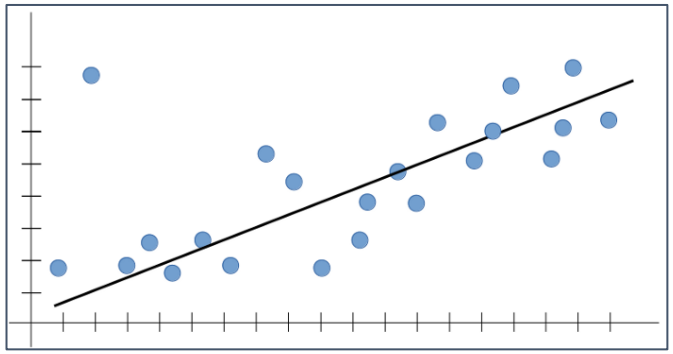
Overfitting y Underfitting

El sobreajuste (overfitting) y el desajuste (underfitting) refieren a deficiencias del modelo que influyen en su rendimiento. Esto significa que saber "qué tan equivocadas" son las predicciones de nuestro modelo es una cuestión de saber qué tan cerca está de sobre ajustarse o desajustarse.

Datos



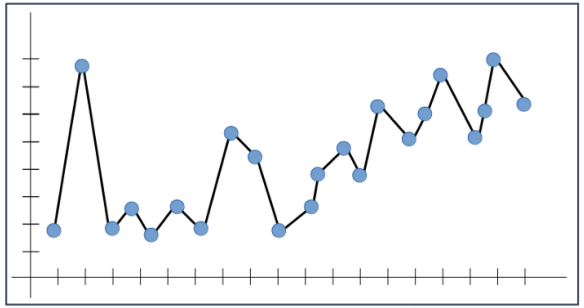
Modelo
(Regresión lineal)



Overfitting

El *overfitting* sucede cuando ejecutamos nuestro algoritmo de entrenamiento en el conjunto de datos disminuyendo la distancia del modelo a los datos con cada iteración. Dejar que este algoritmo de entrenamiento se ejecute durante mucho tiempo conduce a un modelo perfecto para el set de datos con el cual fue entrenado. Sin embargo, esto significa que la línea se ajustará a todos los datos (incluido el ruido), captando patrones secundarios que pueden no ser necesarios para la generalización del modelo.

Modelo
Sobre ajustado

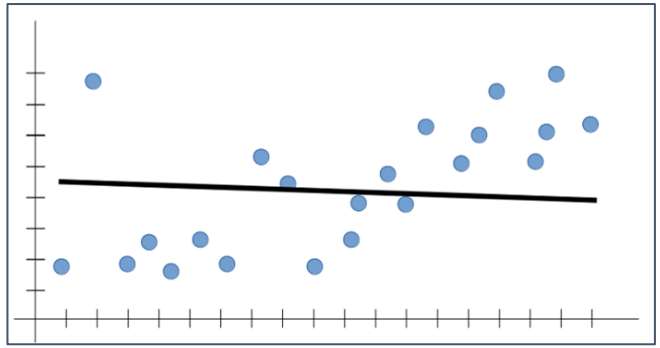


https://www.nopuedocreer.com/wp-content/images/2009/12/Humanbed.jpeg

Underfitting

Queremos que el modelo aprenda de los datos de entrenamiento, pero no queremos que aprenda demasiado (es decir, no queremos sobre ajustarlo). Una solución podría ser detener el entrenamiento antes. Sin embargo, esto podría hacer que el modelo no aprenda suficientes patrones de los datos de entrenamiento y posiblemente ni siquiera capture la tendencia dominante. Este caso se llama Underfitting.

Modelo desajustado



https://www.flickr.com/photos/87969372@N00/163737636

Overfitting y Underfitting

Es decir que *overfitting* sucede cuando el algoritmo está tan bien entrenado con los datos conocidos pero que al darle datos nuevos que no conoce, no puede realizar la predicción de manera correcta.
El *underfitting* es cuando la predicción del modelo es demasiado genérica y no sirve para predecir datos con exactitud.

https://www.nopuedocreer.com/wp-content/images/2009/12/Humanbed.jpeg

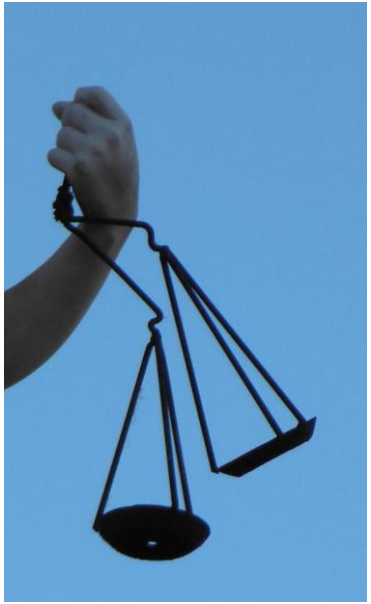
El modelo debe encontrar el equilibrio entre estas dos cosas, que sea lo suficientemente exacto para que sea útil pero que no sea demasiado específico que ya no sirva para realizar predicciones con datos nuevos

https://www.flickr.com/photos/87969372@N00/163737636

Compensación de sesgo-varianza

Entonces, ¿cuál es la medida correcta? Generalmente es deseable un rendimiento que se encuentre entre el sobreajuste y el desajuste. Esta compensación es el aspecto más integral del entrenamiento del modelo de aprendizaje automático.

La predicción está limitada por dos resultados indeseables: alto sesgo y alta varianza. **Detectar si el modelo sufre de alguno de ellos es responsabilidad exclusiva del desarrollador del modelo.**



https://www.flickr.com/photos/mikecogh/7615099958/