## Problema del sobreentrenamiento

Si tengo un conjunto de datos que se parecen mucho entre sí,mi sistema va a ser capaz de aprender muy bien el conjunto de entrenamiento. Lo ideal es tener separado por una parte datos entrenamiento y datos de prueba para realizar predicciones más realistas.

Como mencionamos cuando definimos al Machine Learning , la idea fundamental es encontrar patrones que podamos generalizar para luego poder aplicar esta generalización sobre los casos que todavía no hemos observado y realizar predicciones. Pero también puede ocurrir que durante el entrenamiento solo descubramos casualidades en los datos que se parecen a patrones interesantes, pero que no generalicen. Esto es lo que se conoce con el nombre de sobreentrenamiento o sobreajuste .

El sobreentrenamiento es la tendencia que tienen la mayoría de los algoritmos deMachine Learning a ajustarse a unas características muy específicas de los datos de entrenamiento que no tienen relación causal con la función objetivo que estamos buscando para generalizar. El ejemplo más extremo de un modelo sobreentrenado es un modelo que solo memoriza las respuestas correctas; este modelo al ser utilizado con datos que nunca antes ha visto va a tener un rendimiento azaroso, ya que nunca logró generalizar un patrón para predecir.

## Cómo evitar el sobreentrenamiento

Como mencionamos anteriormente, todos los modelos de Machine Learning tienen tendencia al sobreentrenamiento ; es por esto que debemos aprender a convivir con el mismo y tratar de tomar medidas preventivas para reducirlo lo más posible. Las dos principales estrategias para lidiar con el sobreentrenamiento son: la retención de datos y la validación cruzada.

En el primer caso, la idea es dividir nuestro conjunto de datos , en uno o varios conjuntos de entrenamiento y otro/s conjuntos de evaluación. Es decir, que no le vamos a pasar todos nuestros datos al algoritmo durante el entrenamiento, sino que vamos a retener una parte de los datos de entrenamiento para realizar una evaluación de la efectividad del modelo. Con esto lo que buscamos es evitar que los mismos datos que usamos para entrenar sean los mismos que utilizamos para evaluar. De esta forma vamos a poder analizar con más precisión como el modelo se va comportando a medida que más lo vamos entrenando y poder detectar el punto crítico en el que el modelo deja de generalizar y comienza a *sobreajustarse* a los datos de entrenamiento.

La validación cruzada es un procedimiento más sofisticado que el anterior. En lugar de solo obtener una simple estimación de la efectividad de la generalización ; la idea es realizar un análisis estadístico para obtener otras medidas del rendimiento estimado, como la media y la varianza, y así poder entender cómo se espera que el rendimiento varíe a través de los distintos conjuntos de datos. Esta variación es fundamental para la evaluación de la confianza en la estimación del rendimiento. La validación cruzada también hace un mejor uso de un conjunto de datos limitado; ya que a diferencia de la simple división de los datos en uno el entrenamiento y otro de evaluación; la validación cruzada calcula sus estimaciones sobre todo el conjunto de datos mediante la realización de múltiples divisiones e intercambios sistemáticos entre datos de entrenamiento y datos de evaluación.