# Aprendizaje máquina, generalización y árboles de decisión

martes, 4 de febrero de 2025 08:24 p. m.

## Aprendizaje Máquina

Resulta difícil programar manualmente soluciones para problemas complejos (ej. reconocimiento de objetos o la detección de fraudes), ya que n o entendemos por completo cómo el cerebro realiza estas tareas. El aprendizaje automático ofrece una solución alternativa. En l ugar de programar explícitamente, se alimenta un algoritmo con ejemplos de datos y sus respuestas correctas, permitiendo que el algoritmo apren da el programa por sí mismo.

$\hat{y} = h_{w,w_0}(x) = w_1 x + w_0$ (Hay una infinidad de posibles parámetros)

Dos métodos en el aprendizaje: · Elegir la estructura o funciones ($H$)

· Elegir los parámetros ($\theta \in \mathbb{R}^n$)

$n$ es el número de parámetros.

Función de error (se elige dependiendo del tipo de problema):

$$E_{in}(h) = \frac{1}{m}\sum_{i=1}^{m} loss\left(y^{(i)}, h(x^{(i)})\right)$$

Se eligen parámetros tales que $E_{in} \simeq 0$, el aprendizaje ocurre si $E_{out} \simeq E_{in}$

El objetivo principal es que el modelo aprenda a generalizar a datos no vistos, no solo memorizar los datos de entrenamiento. El aprendizaje es exitoso cuando Eout (Error fuera de muestra) es cercano a cero, lo que es posible si Ein (Error en muestra) también lo es y s i Ein ≈ Eout.

- Error en muestra $E_{in} = \frac{1}{N}\sum_{i=1}^{N} e(y^{(i)}, \hat{y}^{(i)})$

- Error fuera de muestra $E_{out} = E_{x \in X}[e(y, \hat{y})]$.

**Maldición de la Dimensionalidad:** A medida que aumentan las dimensiones de los datos, se requiere exponencialmente más datos para alcanzar resultados similares.

### Método del Vecino más Próximo:

- Método no paramétrico (no requiere ajustes de parámetros). No hay método más simple conceptualmente.
- Guarda todos los datos de entrenamiento y clasifica nuevas instancias basándose en la clase del dato más cercano en el conjun to de entrenamiento.

### Algoritmos de Aprendizaje Máquina:

- Modelos descriptivos
- Modelos lineales generalizados
- Árboles de decisión
- Redes neuronales
- Métodos de ensemble

## Generalización (por qué decimos que el aprendizaje máquina es posible?)

¿Cómo es posible que un modelo entrenado con datos específicos pueda generalizar a nuevos datos? Aprendizaje Probablemente Aproximadamente Correcto (PAC Learning).
**Desigualdad de Hoeffding:** Diferencia entre el error en muestra y el error fuera de muestra, basada en el número de datos. Se discuten las limitaciones de la desigualdad, especialmente cuando el conjunto de entrenamiento es pequeño y el número de hipótesis es alto .
Dicotomías e Hipótesis: Muchas hipótesis pueden ser iguales con respecto al conjunto de entrenamiento, lo que lleva al concep to de dicotomías. Se introduce la función de crecimiento para acotar el número de hipótesis distintas.
Dimensión VC: Se define como el valor más grande para el cual la función de crecimiento es igual a 2N. Representa la compleji dad del modelo y su capacidad para ajustar datos arbitrarios.
Desigualdad de Vapnik--Chervonenkis: Esta desigualdad se presenta como una forma de acotar el error de generalización basándose en la dimensión VC y el tamaño del conjunto de entrenamiento. Se menciona que, aunque se pueden usar los grados de libertad como un a aproximación, no es un cálculo exacto.
La Regla de Oro para la Generalización: Se menciona la necesidad de tener un número de datos de entrenamiento suficiente, que es función de la dimensión VC.
El aprendizaje es posible con el modelo correcto, y un número de datos suficiente.

## Árboles de Decisión

### Estructura

Modelos compuestos por:

- **Nodos** internos que prueban atributos
- **Ramas** que representan los posibles valores de esos atributos
- **Hojas** que asignan una clase

Hipótesis: Los árboles de decisión pueden representar cualquier función de los atributos de entrada, incluso funciones comple jas con muchas variables. Para el caso de funciones Booleanas se indica que cada ruta a una hoja representa una fila de la tabla de verdad.

Aprendizaje de Árboles de Decisión: El aprendizaje del árbol más simple es un problema NP-completo, por lo que se recurre a un algoritmo voraz que selecciona el mejor atributo para dividir recursivamente el conjunto de datos.

Medición de Incertidumbre: Eentropía para medir la incertidumbre en las distribuciones de probabilidad. Una buena división es aquella que reduce la entropía. Entropía condicional y la ganancia de información (information gain) como métrica para escoger el mejor atributo para dividir.

Entropía: Se define como la medida de la incertidumbre o aleatoriedad de una variable aleatoria. La alta entropía indica una distribución uniforme y baja predictibilidad, mientras que la baja entropía indica lo contrario.
Criterios de Parada: Se mencionan dos casos base para detener el crecimiento del árbol: cuando todos los datos en un subconju nto tienen la misma salida o cuando los puntos de datos son idénticos en los atributos restantes.

Problema de Sobreajuste (Overfitting): Los árboles de decisión tienden a sobreajustar los datos de entrenamiento (error de en trenamiento cero) y es necesario introducir sesgos hacia árboles más simples usando varias técnicas.

Métodos de Ensamble (Ensemble Methods): Combinar múltiples modelos (como árboles de decisión) para mejorar el rendimiento, en focándose en el balance sesgo/varianza (bias/variance tradeoff).
Bagging (Bootstrap Aggregation): Método que utiliza muestras bootstrap del conjunto de entrenamiento para entrenar múltiples clasificadores y combina sus predicciones.
Random Forests: Método específico para árboles de decisión que introduce dos fuentes de aleatoriedad: bagging y muestreo alea torio de atributos en cada nodo.