Notas de Clase

Jeison Dario Gutiérrez

5 de febrero de 2008

Profesor: Fabio Gonzales Osorio
Charla: An Introductión to Machine Learning

Que es un patrón?

- Reguralidades en los datos: son las repeticiones de fragmentos de los datos dentro del conjunto total a partir de los cuales se puede predecir su comportamiento.
- Relaciones entre los datos: ejemplos. tendecias de mercado, formas de representar las interacciones, funciones matemáticas.
- Redundancia: Datos que se repiten y que no aportan información importante. Se utiliza en el muestreo de señales y en la compresión de datos.
- Modelo Generativo: Los datos vienen de una fuente y esta tiene un comportamiento conocido. La teoría de colas supone que los datos son generados por un modelo exponencial

Generalización

Se necesitan 2^{2^n} funciones para representar las posibles salidas del modelo. Cada ejemplo de entrada descarta la mitad de las funciones posibles

No es posible generalizar únicamente a partir de ejemplos porque se necesitan todos los posibles ejemplos.

Sesgo Inductivo

El problema de aprendizaje es denominado ill-posed lo que quiere decir que siempre hay una posible solución para el mismo problema particulas.

Para poder generar un modelo de aprendizaje es necesario hacer suposiciones adicionales sobre los ejemplos de entrenamiento.

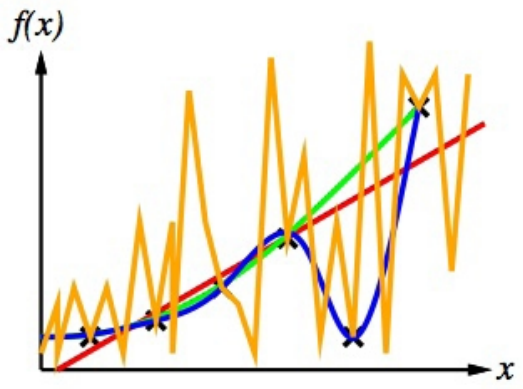

Figura 1: Tipos de modelos

En la figura 1 se observan diferentes tipos de modelos de los cuales el modelo azul se ajusta perfectamente a todos los datos, pero este tiene la desventaja de estar sobre-entrenado y por lo tanto presentara un error de clasificación alto. El modelo color naranja se ajusta perfectamente a los datos pero necesita de una gran cantidad de parámetros. De los modelos verde y rojo se puede decir que necesitan de pocos parametros pero de estos el que mas se ajusta es el verde.

Occam planteo que cuando se tienen distintas soluciones a un problema, de estas la mejor es la mas simple.

Problemas de aprendizaje

Supervizado

El problema principal en este tipo de aprendizaje es encontrar una función que relacione un conjunto de entradas con un conjunto de salidas, por lo tanto es necesario que el conjunto de datos este etiquetado.

Este tipo de aprendizaje es utilizado principalmente en problemas de clasificación y de regresión.

No-supervizado

En este tipo de aprendizaje los conjuntos de datos no vienen etiquetados, por lo tanto el problema fundamental es encontrar la estructura subyacente al conjunto de datos de entrenamiento.

Es utilizado principalmente en problemas de clustering y compresión de datos.

Semi-supervizado

Es similar al aprendizaje no-supervizado solo que en este caso algunos de los datos del conjunto de entrenamiento pueden estar etiquetados.

Activo

Generalmente se usa en el entorno de la programación con agentes. Cuando los agentes no saben si la decisión que tomaron esta bien.

Estos son penalizados o estimulados dependiendo de lo acertado o no de la desición que tomaron. Esto no siempre es posible inmediatamente.

El problema fundamental es definir un conjunto de reglas que permitan maximizar el estímulo positivo dado a los agentes.

On-line

En este tipo de aprendizaje nos encontramos ante grandes volumenes de datos que no pueden ser almacenados en memoria, esto generalmente se da cuando los datos son generados en tiempo real y no es posible almacenarlos.

El problema fundamental es extraer el máximo número de de datos en el menor número de pasos

Algunas Técnicas

- Computacionales
- Estadísticas
- Computacionales-Estadísticas
- Estadísticas