Tema 3
Introducción al aprendizaje automático
Técnicas multivariantes

Dr. Antoni Ferragut

unir
LA UNIVERSIDAD
EN INTERNET

Objetivos

- Programar computadoras para que aprendan a partir de lo datos
- Habilidad de aprender sin ser programadas explícitamente
- No se programan características: el programa aprende mediante ejemplos

Utilidad

- Tratar con problemas cuya resolución requiere un ajuste muy fino o muchas reglas y excepciones
- Resolver problemas complejos no resolubles de manera tradicional
- Se adapta a nuevos datos
- Útil junto a *big data*

Algunas aplicaciones

- Estimación de la demanda de un producto
- Detección de fraude en tarjetas de crédito
- Representación de conjuntos de datos multidimensionales complejos
- Segmentación de clientes según sus patrones de compra

Supervisado

- Se etiquetan los datos y se buscan relaciones
- Ejemplo: tenemos t tipos de clientes; a un nuevo cliente le asignamos un tipo según características

No supervisado

- el sistema trata de aprender sin etiquetar
- Ejemplo: clasificar los clientes sin etiqueta previa, buscando grupos de clientes similares entre ellos

Supervisado:

- Regresión
- Clasificación
- Modelos de ensamblado

No supervisado:

- Reducción de la dimensión
- Clustering

Necesitamos realizar buenas predicciones ante nuevos datos

Basado en casos

- Define el tipo de un nuevo ejemplo asignando el valor de casos más cercaos
- Usado en problemas de clasificación

Basado en un modelo

- Partiendo de ejemplos de entrenamiento se construye un modelo para luego hacer las predicciones con él

Ajuste y predicción

- El modelo debe ser capaz de ajustar los datos con los que se crea el modelo
- Después, debe ser capaz de predecir sobre datos no usados para ajustar el modelo
- Por tanto, al crear el modelo hay que guardar una proporción de datos para usarlos en la evaluación de la capacidad predictiva

Sesgo:

- Distorsión de los resultados
- Por los datos: muestra no representativa
- Por el método: por el modelo sencillo o por cómo se obtienen los parámetros
- Si no se puede eliminar, trataremos de reducirlo

Varianza:

- Variabilidad ante una muestra distinta de la misma población

La solución de compromiso:

- En ocasiones, para disminuir la varianza debemos introducir sesgo
- Modelo demasiado sencillo $\Rightarrow$ infraajuste $\Rightarrow$ más sesgo
- Modelo muy complejo $\Rightarrow$ sobreajuste $\Rightarrow$ más varianza

- Para evaluar la capacidad de ajuste de un modelo, definimos una medida de la **bondad del ajuste**
- Dependerá del tipo de problema y del tipo de algoritmo
- En regresión suele medirse con la raíz del error cuadrático medio

Técnicas de validación: entrenamiento y test

- Dividimos la muestra entre datos de entrenamiento y datos de validación
- Existen muchos métodos para realizar esta partición
- El más sencillo: usar una proporción para obtener los parámetros y con el resto comprobar cómo de bueno es el modelo
- Inconveniente: potencia predictiva restringida

Permite incluir información de todos los datos de la muestra en la evaluación del modelo

- Dividimos la muestra en k partes
- Entrenamos el modelo en $k-1$ partes
- Medimos la bondad de la predicción en la k-ésima
- Igual en las otras $k-1$ combinaciones
- Realizamos la media de las bondades estimadas

Para evitar la influencia de la partición podemos repetir el cálculo r veces y promediar resultados

Caso especial: $k = n$ (LOOCV)

Ventajas:

- No necesitamos repetir la CV
- Identificamos datos influyentes

Inconvenientes:

- Computacionalmente costoso
- Sobreajuste: usamos $n - 1$ datos para el entrenamiento del modelo

Usada cuando queremos estimar el error de generalización y ajustar algún hiperparámetro del algoritmo

Utiliza dos bucles:

- Bucle interior para obtener los valores óptimos de los hiperparámetros usando la CV
- Bucle exterior para estimar el error de generalización

- No se realiza partición de los datos
- Obtenemos artificialmente una muestra similar a la original

Los más usados:

- *Jackknife*: elimina una observación
- *Bootstrap*: del mismo tamaño, muestreo con reemplazamiento

Optimismo

Es otra manera de obtener una estimación de la bondad del ajuste

Pasos:

- Entrenar el modelo o algoritmo con todos los datos disponibles
- Obtener el error del entrenamiento
- Corregir este error mediante el optimismo

Optimismo:

- Es la diferencia entre la medida de la bondad del ajuste en el dataset de entrenamiento y la medida de la bondad del ajuste en la generalización
- Permite corregir la bondad del algoritmo sin perder muestra (utiliza *bootstrap*)