

OPENCOURSEWARE
APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS
GRADO EN ESTADÍSTICA Y EMPRESA
Ricardo Aler



Métodos avanzados para clasificación y regresión : Bagging, Boosting, Support Vector Machines, Redes de Neuronas

Ensembles : Esta parte muestra cómo se pueden mejorar los modelos construyendo conjuntos (ensembles) de modelos básicos.

- Existen básicamente dos tipos de conjuntos: Bagging y Boosting (también Stacking).
- Bagging crea conjuntos de modelos mediante muestreo con reemplazo del conjunto original de datos de entrenamiento. Esto se hace varias veces y se entrena un modelo con cada remuestreo.
- El clasificador final es el conjunto de todos los modelos. La clasificación se realiza a través de voto mayoritario. Para la regresión, se calcula el promedio de todos los modelos.
- Boosting crea modelos secuencialmente. Cada modelo en la secuencia se centra en los errores del modelo anterior.
- Hay variantes de Bagging y Boosting, cuando se usan árboles como modelos base.
- Los Random Forests son una variante de Bagging, donde se usan dos métodos de randomización para introducir árboles diferentes en el ensemble. Por un lado, el remuestreo con reemplazo ya explicado para Bagging. Por otro lado, cada nodo de cada árbol usa un subconjunto de atributos distinto. Más específicamente, se selecciona el mejor atributo para cada nodo en cada árbol, no de entre el conjunto completo de atributos disponibles, sino de un subconjunto aleatorio de ellos. En otras palabras, el proceso de creación del árbol de decisión tiene un componente aleatorio (es decir, el algoritmo determinista original se transforma en estocástico, al elegir aleatoriamente el subconjunto de atributos a evaluar y comparar en cada nodo).
- Gradient Boosted Trees es una variante de Boosting, donde los árboles se usan como modelos base. Se muestra cómo se entrena cada nuevo modelo aprendiendo los pseudo-residuos (la diferencia entre el resultado del ensemble construido hasta el momento y el valor real).

- También se muestra como Boosting podría ser propenso al sobreajuste, y se debe evitar eligiendo correctamente los valores de los hiper-parámetros.

Máquinas de vectores de soporte (SVM) :

- Las máquinas de vectores de soporte son modelos avanzados, válidos tanto para clasificación como regresión.
- Se basan en dos ideas fundamentales, que se pueden entender particularmente bien en problemas de clasificación :
 - Los modelos lineales buscan obtener el hiper-plano de máximo margen (de entre todos los posibles separadores entre dos clases).
 - El truco del kernel (o kernel trick) : para construir modelos no lineales, se proyectan las instancias a un espacio, típicamente de mayor dimensionalidad, en el que se pueda hacer una separación lineal (o aproximadamente lineal). Dicha separación lineal es equivalente a una separación no lineal en el espacio origen. El truco del kernel consiste en que no es necesario hacer dicha proyección de manera explícita, dado que lo único que se necesita es el producto escalar en el espacio destino entre instancias de la proyección.
- Los modelos SVM se ajustan mediante programación cuadrática, optimizando una función de coste que tiene en cuenta dos objetivos : el margen (a maximizar) y errores permitidos en algunas instancias (a minimizar). El hiper-parámetro de coste (C) determina la importancia de ambos objetivos : con valores de C pequeños se obtendrán modelos más simples (con posibilidad de under-fitting) y con valores de C grandes más complejos (con posibilidad de overfitting).
- Los kernels más importantes son el lineal y el gaussiano (el cual dispone de otro hiper-parámetro, la desviación de la gaussiana a considerar)
- El ajuste de ambos hiper-parámetros, coste y desviación, se suele llevar a cabo mediante grid-search.

Redes de Neuronas :

- Las redes de neuronas son modelos no lineales para clasificación y regresión.
- En concreto, las MLP (Multilayer Perceptrons) tienen una arquitectura organizada en varias capas : entrada, capa(s) oculta(s) y salida. Cada capa contiene neuronas. Cada neurona realiza una operación que consiste en la suma ponderada (por pesos) de las entradas, a la cual se le aplica una función no lineal (típicamente la sigmoide o la tangente hiperbólica).
- Las redes se entrenan mediante descenso del gradiente, que consiste en ajustar los pesos de manera iterativa, dirigiéndonos siempre en la dirección en la que el gradiente se reduce. También se llama a este proceso retro-propagación del gradiente, dado que los errores se propagan de atrás (la salida de la red) a adelante (las entradas de la red).

Material asociado

- Transparencias de la clase y algunos ejercicios (consultar la guía del curso).