uc3m | Universidad Carlos III de Madrid

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

Preproceso de datos

Ideas principales de la clase :

- Primero, el preprocesamiento se introduce como un paso más en la *pipeline* de Machine Learning
- Se puede hacer pre-proceso tanto de las instancias como de los atributos.
- Un ejemplo de pre-proceso de instancias es el rebalanceo de la muestra. Este se puede utilizar en problemas de muestra desbalanceada (problemas de clasificación en los que una de las clases dispone de muchos menos datos (minoritaria) que otra (mayoritaria). En ese caso, se puede rebalancear la muestra con algoritmos de pre-procesado como SMOTE.
- Uno de los preprocesos más importantes de atributos es la selección de atributos.
- La selección de atributos es importante para eliminar atributos redundantes e irrelevantes, y aliviar la maldición de la dimensionalidad.
- La maldición de la dimensionalidad puede ocurrir incluso en clasificadores lineales.
- Una idea importante en la selección de atributos es que la unidad de selección podría no ser un atributo sino un subconjunto de atributos, porque a veces, dos (o más) atributos no funcionan bien cuando se usan de forma aislada, pero funcionan bien cuando se usan juntos.
- Se introducen las técnicas genéricas de *filter* y *wrapper*.
- Se introducen algunos algoritmos de selección de atributos únicos (clasificación): basados en entropía (que ya se usaba para seleccionar el mejor atributo en los árboles de decisión), basado en información mutua, y basada en chi-cuadrado.
- Se presentan los algoritmos *wrapper* para seleccionar subconjuntos.
- Finalmente, se explica la diferencia entre selección de atributos y la transformación de atributos : la primera selecciona los atributos importantes, mientras que la segunda transforma los atributos originales en nuevos.
- Se explican dos algoritmos principales de transformación de atributos: PCA y proyecciones aleatorias.
- PCA es un algoritmo no supervisado y, si bien puede usarse para generar nuevos atributos y reducir la dimensionalidad, debe usarse con cuidado en clasificación, dado que PCA no considera la clase de los datos. Las proyecciones aleatorias se presentan

como un método más rápido para hacer la generación y selección de atributos tipo PCA.

Material asociado

- Transparencias de la clase y algunos ejercicios (consultar la guía del curso).
- Parte del tutorial de MLR está dedicado al pre-proceso de datos.
- Una de las prácticas está orientada al pre-proceso, para el rebalanceo de la muestra.