Tema 9 (II)
Reducción de la dimensión y clustering
Técnicas multivariantes

Dr. Antoni Ferragut

unir
LA UNIVERSIDAD
EN INTERNET

- El aprendizaje no supervisado es de gran interés pues muchos datos vienen sin etiquetar
- Ejemplos de técnicas:
  - **Clustering**: se agrupan observaciones en clusters. Útil para análisis de datos, segmentación de clientes, motores de búsqueda, segmentación de imágenes, reducción de dimensionalidad...
  - **Detección de anomalías**: se aprende cómo son los datos normales para poder detectar datos anormales. Útil en controles da calidad, análisis de registros...
  - **Estimación de la densidad**: se estima la función de densidad. Útil para detección de anomalías (observaciones en regiones con probabilidad muy baja pueden ser anomalías), análisis de datos...

Clustering

- Consiste en realizar grupos de observaciones que comparten características similares
- A cada observación se le asigna un grupo
- La asignación de hace de forma no supervisada
- Se debe determinar cuántos grupos diferentes existen y asignar a cada observación su grupo
- La elección del número de grupos no es trivial

Aplicaciones del Clustering: segmentación de clientes

- Se puede dividir en grupos de clientes a aquellos que tienen un comportamiento similar en cuanto a compras o en cuanto a la navegación dentro de la web
- Muy útil para entender quiénes son los clientes y cuáles son sus preferencias y necesidades, para así poder adaptar las campañas de marketing a cada segmento

Aplicaciones del Clustering: análisis de datos

- Cuando se analiza un nuevo conjunto de datos, puede ser de gran ayuda aplicar algún algoritmo de clustering, y luego analizar cada cluster por separado

Aplicaciones del Clustering: reducción de la dimensionalidad

- Se puede conseguir reducir la dimensión de un conjunto de datos mediante el clustering
- Primero se realiza una división de las observaciones del conjunto de datos en clusters
- Después se obtiene la afinidad, que es una medida que indica cuánto encaja una observación en un determinado cluster, para cada observación con cada uno de los diferentes clusters
- La reducción de la dimensión se consigue transformando las variables originales de las observaciones por el vector de sus valores de afinidad
- Si se han escogido K clusters, el nuevo número de dimensión es K, que suele ser mucho menor que el número de variables originales

Aplicaciones del Clustering: detección de anomalías

- Las observaciones que tengan poca afinidad para todos los clusteres es probable que sean anomalías
- Se pueden detectar comportamientos anómalos, como un número inusual de peticiones (clicks) por segundo en clientes web
- La detección de anomalías es muy útil para detectar defectos en las cadenas de montaje y producción, y para la detección de fraudes

Aplicaciones del Clustering: aprendizaje semi-supervisado

- Si únicamente se tienen unas pocas etiquetas de las observaciones en el conjunto de datos, se puede realizar una clasificación en clusters y después propagar las etiquetas para todas las observaciones que se hayan agrupado en el mismo cluster

Aplicaciones del Clustering: mótores de búsqueda

- Algunos motores de búsqueda permiten buscar imágenes similares a partir de una imagen de referencia
- Para ello es necesario dividir en clusters las imágenes de la base de datos
- Cuando se proporciona la imagen de referencia, el algoritmo debe clasificar esa imagen en uno de los clusters, y devuelve todas las imágenes pertenecientes al mismo

Aplicaciones del Clustering: segmentación de imágenes

- Agrupando los píxeles en clusters según su color y luego reemplazando para cada píxel su color por el de la media del cluster se consigue reducir considerablemente el número de colores distintos de una imagen
- Esta técnica se emplea para la detección de objetos o para sistemas de rastreo

Métodos de creación de clusters

- Dependiendo del tipo de problema de la técnica empleada tenemos diferentes estrategias para crear clusters:
  - A partir de centroides, puntos singulares sobre los cuales se distribuyen las distintas observaciones del cluster
  - Buscando regiones contínuas de observaciones muy cercanas entre sí
  - De manera jerárquica, pudiéndose obtener clusters dentro de clusters
- Una de las técnicas más usadas para la creación de clusters es la de K-medias

- Algoritmo sencillo que crea K clusters a partir de las distancias de las observaciones a los centroides de cada uno de los clusters:
  1. Se generan K centroides, introducidos manualmente o de manera aleatoria
  2. Se le asigna a cada observación del conjunto de datos el cluster del centroide K más cercano
  3. Se computan los nuevos centroides para cada cluster K
  4. Se repiten los pasos 2 y 3 (se reasignan las observaciones a los nuevos clusters formados por los nuevos centroides K y se actulizan dichos centroides) hasta que las observaciones ya no varían más de cluster
- El método siempre llega a una solución
- Es recomendable escoger una buena estrategia para inicializar los centroides

K-medias. Inicialización de los centroides

- Si se tiene alguna idea de cuáles son los centroides, se pueden introducir manualmente; en muchas ocasiones esto no es posible
- Escogemos varios centroides iniciales de manera aleatoria
- Realizamos el algoritmo iterativo para construir los clusters y escoger aquel que proporcione un mejor resultado
- Para valorar qué clustering es mejor existe una métrica denominada inercia que mide las distancias medias al cuadrado entre cada observación y su centroide
- Cuanto menor sea este valor, mejor será el clustering realizado
- Esta es la técnica que emplea por defecto la clase KMeans de la librería scikit-Learn
- En esta librería los valores de los centroides iniciales se escogen de entre las distintas observaciones, pero de tal modo que las observaciones que se escogen como centroides iniciales estén suficientemente alejadas entre sí

K-medias. Número óptimo de clusters

- Para saber cuántos grupos crear, no podemos basarnos únicamente en el valor de la métrica de la inercia, ya que esta suele disminuir a medida que se aumenta el número de clusters
- Pero podemos ver cómo varía ésta con el número de clusters y seleccionar el valor al partir del cual se forma un codo