Primavera 2014
Flores y Torres
Ayudantes: Villanueva y Villarroel

ILI 286
Laboratorio 1
06 de Septiembre 2014

Entrega: 12/09/2014, antes de las 17:00 impreso y 23:55 en moodle.

Notación

En el presente laboratorio, A será siempre una matriz simétrica con valores en $\mathbb{R}$, con valores propios $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ y vectores propios asociados v_1, v_2, ..., v_n. Utilice la misma notación al explicar algoritmos o realizar demostraciones.

Motivación

Las técnicas de clustering espectral hacen uso de los valores propios de la matriz de similaridad de los datos para realizar una reducción dimensional. Esto permite realizar el clustering de los datos en un espacio de menor dimensión. La matriz de similaridad es una matriz cuadrada y simétrica que consiste en una cuantificación de la similaridad entre 2 pares de puntos del conjunto de datos.

La idea central puede resumirse en los siguientes pasos:

- Establecer la matriz de similaridad entre los n elementos.
- Calcular los k vectores propios dominantes. En este laboratorio, consideraremos simplemente los 2 mayores vectores propios, v_{n-1} y v_n.
- Los k vectores dominantes forman un nuevo conjunto de datos, esta vez de dimensión $k \times n$, sobre el cual se aplican el clustering.

Problema 1: Modificación de Algoritmos Conocidos

Para desarrollar la técnica de clustering espectral es esencial poder calcular los k vectores propios dominantes asociados a la matriz simétrica A. Para ello podemos utilizar las siguientes técnicas:

0.1. Power iteration

Con Power Iteration obtenemos el mayor valor y vector propio asociado, λ_n y v_n respectivamente.
Responda las siguientes preguntas teórica **(15 puntos)**:

- ¿Cual es el mayor valor propio de $A - \lambda_n I_n$? ¿Cual es el mayor valor propio de $A - \lambda_n v_n v_n^T$?
- Describa un algoritmo (pseudo código) que utilice Power Iteration para calcular los 2 primeros vectores propios.
- Estime su costo computacional y la convergencia.

0.2. Rayleigh Quotient

Con este método también se obtiene el mayor valor y vector propio, λ_n, y v_n. Responda las siguientes preguntas teórica **(15 puntos)**:

- Describa un algoritmo (pseudo código) que utilice Rayleigh Quotient para calcular los 2 primeros vectores propios (Recuerde detener su algoritmo cuando alcance la convergencia).
- Estime su costo computacional y la convergencia.
- Compáre con el costo computacional y convergencia del Power Iteration.

0.3. QR parcial

Unshifted QR permite encontrar todos los valores y vectores propios. Sin embargo, si en vez de realizar una factorización QR completa utilizamos una factorización parcial, es posible encontrar los k vectores propios dominantes.
Responda las siguientes preguntas teóricas **(15 puntos)**:

- Describa un algoritmo (pseudo código) que utilice QR para calcular únicamente los 2 primeros vectores propios.
- Estime su costo computacional y la convergencia.
- Compare los costos anteriores con utilizar el algoritmo unshifted QR para obtener los n valores y vectores propios, y luego seleccionar únicamente los 2 mayores descartando todos los otros valores.

Problema 2: Aplicación a Ejemplo

Considere el dataset entregado junto con la tarea, en el cual se tienen 20 películas, y 40 tags asociados a ellos con valores entre 0 y 1 indicando la aplicabilidad del tag. Por ejemplo, la película "Fight Club" tiene asociado 0,90 al tag **Action** y 0,02 al tag **Family**. Llamaremos T la matriz de asociación de tags, con dimensiones 20×40.
Responda y comente las siguientes puntos **(35 puntos)** :

- Desarrolle una función que retorne la matriz A de similitud entre películas, dada por $A = TT^t$ a partir del archivo `associations.dat`.
- Desarrolle una función para cada una de las 3 técnicas descritas anteriormente. Estas deben recibir una matriz A y retornar λ_{n-1} , λ_n, v_{n-1} y v_n. Verifique que se entreguen los mismos valores propios y vectores propios para dos matrices de test, de 5×5 y 10×10 respectivamente, con coeficientes aleatorios y reales entre 0 y 1.
- Ocupando las funciones de los dos puntos anteriores, aplique las 3 técnicas al conjunto de datos. Muestre sus resultados.
- Compare los resultados y los tiempos de ejecución de cada técnica. ¿Cual preferiría en la práctica?

Observación: Todas las funciones de esta sección, deben ir en un archivo con nombre `p2.py` adjunto en la entrega digital de su tarea.

Problema 3: Aplicación a Datos reales

En esta sección aplicaremos la misma lógica de la sección anterior al conjunto de datos provisto por el proyecto Tag Genome, disponible en `http://grouplens.org/datasets/movielens/`. En este caso se cuenta con 11 millones de asignaciones de tags a películas, formando una matriz de 10000 películas a los cuales se asignan valores para cada uno de los 1100 tags. Para que no tenga problemas (de memoria ram) trabajando con estos datos, se utilizará una muestra reducida (5000 películas). Para poder estudiar estos datos de manera eficiente, utilizaremos las librerías nativas de python para calcular valores y vectores propios.

Responda y comente las siguientes puntos **(30 puntos)**:

- Adaptando el código anterior, desarrolle una función para la lectura de datos y que retorne la matriz de similitud de las películas. (El formato de los archivos puede verse en `http://files.grouplens.org/datasets/tag-genome/README.html`).
- Obtenga todos los valores propios de la matriz de similitud (utilice scipy.linalg.eigvalsh). Grafíquelos. ¿Que observa?
- Obtenga los 2 valores y vectores propios dominantes (utilice scipy.linalg.eigh pidiendo sólo los 2 dominantes). Muestre su resultado.
- Utilice las funciones provistas para graficar los resultados. ¿Que observa?

Observación: Todas las funciones de esta sección, y el código para realizar lo solicitado, deben ir en un archivo con nombre `p3.py` adjunto en la entrega digital de su tarea.

Links

- `http://en.wikipedia.org/wiki/Spectral_clustering`
- `http://www.cis.hut.fi/Opinnot/T-61.6020/2008/spectral_kmeans.pdf`
- `http://cs.stanford.edu/people/ang/papers/nips01-spectral.pdf`

Instrucciones:

(a) El laboratorio puede ser realizado en Python o Matlab.

(b) El laboratorio debe ser entregado en LaTeX o publicado en Matlab.

(c) La estructura del laboratorio es la siguiente:

   a) Título, nombre del estudiante, email y rol.
   b) Una pequeña descripción de los experimentos y suposiciones consideradas.
   c) Desarrollo y análisis de resultados.
   d) Conclusiones.
   e) Referencias.
   f) Anexo con el código utilizado.

(d) Si el código utilizado en los experimentos no es el mismo código entregado se evaluará el laboratorio con un 0.

(e) El archivo de entrega debe denominarse Lab2-apellido1-apellido2.tar.gz, y debe contener un directorio llamado Informe que contenga los archivos .pdf y .tex correspondientes y un directorio llamado Códigos con los archivos correspondientes.

(f) El trabajo es personal o en grupos de a 2, no se permite compartir código, pero se sugiere discutir aspectos generales con sus compañeros.

(g) La entrega digital debe ser un archivo labX-rol1-rol2.zip, Donde rol1 y rol2 son los roles de los integrantes (sin dígito verificador) y X es el número del laboratorio. En este debe estar el informe en formato digital (incluir .tex), los cógidos y todos los archivos solicitados.

(h) Si no se siguen estas instrucciones, el laboratorio será evaluado con un 0.

Consideraciones:

Para todos los laboratorios del semestre se debe tener en cuenta, al momento de realizar el informe, lo siguiente:

- Introducción y conclusión: Que sea pertinente al laboratorio. No escriba cosas como "la historia de la Computación Científica...", ni "aprendí mucho". Sea más objetivo. Una buena idea sería plantear brevemente el problema o situación a analizar, objetivos generales y particulares, la estructura del informe y también, si ya tiene conocimiento de lo que se debe hacer, podría realizar una estimación. MÁXIMO: 5 líneas.
- Desarrollo y análisis: Incluya todos los supuestos, fórmulas, algoritmos, desarrollos matemáticos, etc. No ponga "se ve en el código"porque eso es aparte. Incluya gráficos, resultados, cuadros comparativos, y cualquier cosa que le permita realizar un análisis más exacto. Recuerden que los análisis son distintos de las conclusiones, explique a qué se debe las diferencias entre algoritmos. Cuantifique y fundamente sus respuestas, evite el exceso de adjetivos. Sea creativos, existen muchos criterios para comparar y analizar.
- Ortografía: Se descontarán 5 puntos, por cada 5 faltas ortográficas.
- Precisión: Calidad antes que cantidad, no se de vuelta en la misma idea. No deje tanto espacio en blanco e imprima, en lo posible, ambas caras de una hoja.
- Código: En LaTeX hay distintas formas de adjuntar o presentar un código. Una imagen NO es una de ellas.
- Ponderaciones: El código vale el 30 % y el informe un 70 %. Se evalúa también orden y redacción.