Serie de Machine Learning
PCA aplicado al Genoma Humano

Objetivo: en este mini-proyecto, aplicaremos PCA en un conjunto de datos reales e interpretaremos los resultados.

Descripción: guarde en una carpeta el archivo **p4dataset2020.txt**. En la **misma carpeta** debe guardar en JN "**TP_PCA_ProyectoGenoma**". Los datos representados allí son del proyecto de 1000 genomas. Cada una de las 995 líneas (filas) en el archivo representa un individuo. Las primeras tres columnas representan, respectivamente, el identificador único del individuo, su sexo (1 = hombre, 2 = mujer) y la region del mundo a la que pertenece. Las siguientes 10101 columnas de cada línea son una sub-muestra de nucleobases del genoma del individuo.

Aplicaremos PCA a este conjunto de datos. PCA puede referirse a una serie de cosas relacionadas, asi que para ser explícitos, en este TP, cuando decimos "PCA" queremos decir:

- Los datos deben estar centrados (hay que restar la media de la muestra) pero no normalizados.
- El output debe ser las componentes principales normalizadas a 1 (autovectores normalizados a 1).

Ejercicios: Primero convierta los datos del archivo de las nucleobases[1] a una matriz real (esta parte va a ser provista). Específicamente, convierta los datos genéticos en una matriz binaria X tal que $X_{ij} = 0$ si el i-ésimo individuo tiene su j-ésima nucleobase igual a la **moda nucleobase**[2] de la columna j, y $X_{ij} = 1$ en caso contrario. Tenga en cuenta que todas las mutaciones deben aparecer como un 1, incluso si son mutaciones diferentes, por lo que si la moda de la columna j es "G", entonces si el individuo i tiene una "A", una "C", o una "T", X_{ij} sería 1.

Las primeras 3 columnas del archivo **p4dataset2020.txt** proporcionan "metadatos", y deben ignorarse al crear la matriz binaria X. Examinaremos genotipos para extraer información de fenotipos.

1. (Pre-calentamiento). Si corremos PCA a la matriz binaria X (o a X centrada). ¿Cuál sería la dimensión de los vectores outputs?

[1] Para este trabajo sólo hace falta saber que las nucleobases son las unidades fundamentales del código genético: adenina (A), citosina (C), guanina (G), timina (T) del ADN.

[2] Por "**moda nucleobase**", nos referimos a la nucleobase más frecuente en esa posición (en los 995 puntos de datos o individuos).

2. Examinaremos los primeros 2 componentes principales de X. Estos componentes contienen mucha información sobre nuestro conjunto de datos. Cree un diagrama de dispersión, con cada una de las 995 filas de X proyectadas en los primeros dos componentes principales. En otras palabras, el eje horizontal debe ser el espaneado por $\mathbf{v}_1$, el eje vertical por $\mathbf{v}_2$, y cada individuo debe proyectarse en el subespacio espaneado por $\mathbf{v}_1$ y $\mathbf{v}_2$. El gráfico debe usar un color diferente para cada población e incluir una leyenda.

3. Enumere 1 o 2 hechos básicos sobre el gráfico creado en la parte 2. ¿Puede interpretar los primeros dos componentes principales? ¿Qué aspectos de los datos capturan los dos primeros componentes principales? Sugerencia: piense en historia y geografía.

4. Ahora examinaremos el tercer componente principal de X. Cree otro diagrama de dispersión con cada individuo proyectado en el subespacio correspondiente a los componentes principales primero y tercero. Juegue con diferentes esquemas de etiquetado (con etiquetas derivadas de los metadatos) para explicar los grupos que ve. Su gráfico debe incluir una leyenda.

5. Algo debería llamar la atención en el gráfico del paso anterior. En una oración, ¿qué información captura el tercer componente principal?

6. En esta parte inspeccionará el tercer componente principal. Grafique el índice de nucleobase vs. el valor absoluto del tercer componente principal. ¿Nota algo? ¿Cuál sería una posible explicación? Sugerencia: piense en los cromosomas.

Entregables: diagrama de dispersión para la parte 2. Breve discusión para la parte 3. Gráficos de dispersión para las partes 4 y 6. Una respuesta de una línea cada una para las partes 5 y 6.