*

Tarea 4: aprendizaje no supervisado

Entrega: 26 de noviembre

Profesor: Pablo Estévez V.
Auxiliar: Ignacio Reyes J.
Semestre: Primavera 2021

1. Parte teórica

1. ¿Qué significa que un algoritmo de aprendizaje de máquinas sea no supervisado?

2. El método PCA entrega una nueva base para describir un conjunto de muestras multidimensionales. Al usar PCA como método de reducción de dimensionalidad o método de visualización, los primeros vectores de la base se conservan, mientras que el resto son desechados. ¿Cuál es la justificación para esto? Explique considerando la relación entre los valores propios de la matriz de correlación y la varianza de los datos.

3. Considere el método de Kernel PCA con kernel gaussiano, el cual mapea N muestras a un espacio distinto antes de aplicar PCA. ¿Cuántas dimensiones tiene dicho espacio para el caso del kernel gaussiano?

4. Describa brevemente el algoritmo SOM y explique cómo se interpreta la visualización de la U-Matrix.

2. Parte práctica

Introducción

El objetivo de esta tarea es explorar una base de datos con métodos de aprendizaje no supervisado. La base de datos fue construida a partir de la data del Observatorio de Complejidad Económica del MIT y describe a los distintos países a partir de sus exportaciones. Para cada país se muestran las exportaciones separadas en 97 tipos de productos, indicando para cada producto a qué porcentaje de las exportaciones del país corresponde. Los métodos que se utilizarán para el análisis son PCA, Kernel PCA y SOM. Adjunto a esta tarea encontrará un Jupyter notebook (tarea4.ipynb). Utilícelo para responder los puntos mencionados a continuación. Al final de la sesión usted deberá subir a u-cursos un documento con los resultados y figuras obtenidos.

1. Tal como se menciona en la introducción, cada país está descrito por 97 características que corresponden a tipos de productos exportados. Estos 97 productos pueden ser agrupados en 15 categorías mediante la variable booleana *is_grouped*.

   Visualice los datos tomando las 2 primeras componentes principales de PCA y compare el resultado de utilizar los 97 tipos productos versus agruparlos en 15 categorías. Compare la varianza explicada por las dos componentes principales en ambos casos. A su juicio, ¿cuál de las dos visualizaciones revela más información respecto a los datos?

¿Qué dificultades tiene utilizar la información de los 97 productos para describir a cada país? Para analizar esta situación imagine que tiene una versión más detallada de los mismos datos con 1000 subtipos de productos en vez de 97 tipos. También le puede ser útil pensar a qué distancia están dos países, uno que sólo exporta cobre y otro que sólo exporta hierro, si cobre y hierro están agrupados como "metales".

2. (De aquí en adelante agrupe los productos en 15 categorías) Visualice los datos utilizando PCA. Identifique y describa los grupos de similitud o clusters entre los países. ¿Se observan outliers?

   Para realizar su análisis, formule hipótesis respecto a la característica común que tendrían los países en cada cluster o región. Luego valide su hipótesis mirando las exportaciones de los países en ese cluster usando el Observatorio de Complejidad Económica.

3. Compare el resultado de usar PCA versus Kernel PCA con kernel gaussiano. ¿Hay diferencias cualitativas? ¿Aparecen nuevos clusters? ¿algún cluster se concentra? Haga sus comentarios tanto en términos generales como en referencia a clusters y países específicos.

4. En PCA sin kernel, elija el número de características que capturen al menos el 85 % de la varianza de los datos. Luego construya un mapa auto-organizativo de Kohonen (SOM) a partir de dicha proyección. Visualice los resultados usando la U-Matrix de distancias. Analice los resultados tal como hizo para los modelos anteriores.

5. Usando PCA con kernel gaussiano elija el mismo número de características que usó anteriormente. Luego construya un mapa auto-organizativo de Kohonen (SOM) a partir de dicha proyección. Visualice los resultados usando la U-Matrix de distancias. Analice los resultados tal como hizo para los modelos anteriores. ¿Se observa alguna diferencia al usar la función de kernel?

6. Utilizando todas las visualizaciones anteriores indique a qué países se asemeja Chile. ¿Existe un cluster de países latinoamericanos? ¿existe un cluster de países árabes / medio oriente? ¿por qué en algunos casos los clusters tienen coherencia con la ubicación geográfica y en otros no?

7. ¿A qué país se parece México? De acuerdo a su PIB, ¿es un país desarrollado?. Compare los tipos de exportaciones de México y Chile, comentando respecto a la diversidad de sus matrices productivas.

   ¿Por qué en algunas visualizaciones aparece un cluster con Groenlandia (grl) e Islas Malvinas (flk)? ¿qué tienen en común?

3. Programación

Visualize la base de datos (agrupando los productos en las 15 categorías) utilizando la técnica t-SNE. Utilize la implementación de t-SNE disponible en sklearn. Ajuste la tasa de aprendizaje de tal manera que la proyección se vea razonable. Analize el resultado obtenido y compare con los experimentos anteriores.

Evalúe el impacto de modificar el parámetro *perplexity* del algoritmo probando un valor pequeño, intermedio y grande.

4. Observaciones

- Se recomienda que la extensión del informe no supere las 12 páginas.
- Los modelos y datasets reutilizan los nombres de las variables. Revise que los resultados que está observando corresponden al modelo y datos correctos.