Tarea 4 teórica: Aprendizaje no supervisado

Fecha de entrega: jueves 17 de octubre, 23:59 (Gradescope)

Profesor: Pablo Estévez V.
Auxiliar: Ignacio Reyes J.
Semestre: Primavera 2019

1. Preguntas

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

   Sol: Se habla de clasificación no supervisada cuando no se conocen a priori las clases, pero las muestras se pueden agrupar en un conjunto finito de categorías mediante relaciones de similaridad.

   El aprendizaje no supervisado es un algoritmo que aprende de los datos de las pruebas, que no ha sido etiquetado, clasificado o categorizado. En lugar de responder al feedback, aquí no hay una medida como la accuracy, el aprendizaje no supervisado identifica elementos comunes en los datos y reacciona en función de la presencia o ausencia de tales elementos comunes en cada nueva pieza de datos.

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.

   Sol:

   La idea general del método PCA es ofrecer una nueva base, llamada componente principal, con el objetivo de maximizar su varianza para poder representar y visualizar mejor los datos. si los datos están normalizados, de media cero, sea x un vector p-dimensional que representa los vectores de entrada y U un vector unidad, $||u|| = 1$:

   $$a = u^T x = x^T u$$
   $$E[a] = u^T E[x] = 0$$
   $$\sigma^2 = E[a^2] = E(u^T x)(x^T u) = u^T E[xx^T]u = u^T Ru$$

   La matriz R_{pxp} es la matriz de correlación de los vectores de entrada.

Desde aquí podemos llegar a:

$$U^T R U = \Lambda$$

Λ es una matriz diagonal definida por los valores propios de R

U es la matriz pxp de los vectores propios asociados.

Por lo tanto, se concluye que encontrar los valores propios de la matriz de correlación R es lo mismo que encontrar las varianzas de la base de datos en las proyecciones respectivas.

El objetivo es encontrar el minimo de vectores ortogonales en espacio de datos que den cuenta de la varianza de la data. Para esto, elegimos los valores propios más grandes de R, descartando así los componentes principales que contribuyen menos.

Uno criterio para determinar cuántas componentes principales se requieren para obtener una representación adecuada de los datos puede ser por la proporción de varianza tomada en cuenta por las primeras r componentes principales, $\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$, hasta 75 % da variancia total.

3. ¿Por qué se dice que PCA es un método lineal de reducción de dimensionalidad? ¿Cómo está construida la matriz asociada a esa transformación lineal?

   Sol: La linealidad en PCA se refiere al hecho de que, para realizar la reducción de dimensionalidad, se proyecta vectores en un espacio de menor dimensión a través de una transformación lineal.

   En PCA, este mapeo viene dado por la multiplicación de x, los vectores de entrada, por la matriz de vectores propios de PCA, V, y, por lo tanto, es manifiestamente lineal. (la multiplicación de matrices es lineal):

   Sea $f : x \to z$ el mapeo de $\Re^n$ para un espacio $\Re^r$, de menor dimensión

$$z = f(x) = V^T x = \sum_{i=1}^{r} v_i * x_i$$

4. 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?. Si los N datos originales son vectores en R 10 , ¿cuántas componentes puede tener PCA si no se utiliza kernel? ¿cuántas componentes puede tener Kernel PCA con kernel gaussiano?

   Sol:

   En el kernel de PCA, los datos se asignan a un espacio de dimensiones muy altas, y la cantidad de PCs solo está limitada por la cantidad de muestras.

   Kernel K, la Gram matriz N x N:

$$K_{ij} = \langle \Phi(x_i)\Phi(x_j) \rangle$$

Así que si tenemos N muestras, asignamos a $\Re^N$, un espacio con N dimensiones. Si los vectores son de $\Re^{10}$, es decir, tienen dimensión 10, el número de valores propios y vectores propios estará limitado a 10, por lo que tendremos solo 10 componentes.

En el caso del kernel PCA para N muestras en $\Re^{10}$, como se explicó anteriormente, el kernel será una matriz NxN. Entonces habrá N valores propios y N vectores propios. El kernel PCA con kernel gaussiano puede tener como máximo N componentes.

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

   Sol: El self-organizing map (SOM) es un tipo de red neuronal que se entrena utilizando el aprendizaje no supervisado para producir una representación de baja dimensión del espacio de entrada, llamado de mapa.

   Básicamente, el algoritmo es el siguiente:

   - Inicializar los prototipos (codebook vectors)
   - Presenta las muestras
     - Busca el prototipo más cercano (el más similar)
     - Actualizar los prototipos más cercanos y sus vecinos para que se vuelvan más similares a la muestra
     - Disminuir un poco el rango de vecindario y el valor de la tasa de aprendizaje
   - Repitir hasta que se cumpla un criterio de detención

   La matriz U contiene en cada celda la distancia euclidiana (en el espacio de entrada) entre celdas vecinas. Los valores pequeños en esta matriz significan que las celdas del SOM están muy juntos en el espacio de entrada, mientras que los valores más grandes significan que las celdas del SOM están muy separados, incluso si están cerca al espacio de salida.