En esta lección discutiremos el análisis de componentes principales (PCA) y la descomposición de valores singulares (SVD), dos técnicas importantes y relacionadas de reducción de dimensiones. Esto último implica procesos que encuentran subconjuntos de variables en conjuntos de datos que contienen sus esencias. El PCA y el SVD se utilizan tanto en la fase exploratoria como en la etapa de análisis de modelado más formal. Nos centraremos en la fase exploratoria y tocaremos brevemente algunas de las teorías subyacentes.

Ahora hablaremos un poco de teoría. Suponga que tiene miles de multivariantes

variables X\_1, ..., X\_n. Por multivariado nos referimos a que cada X\_i contiene muchos componentes, es decir, X\_i = (X\_ {i1}, ..., X\_ {im}. Sin embargo, estas variables (observaciones) y sus componentes pueden estar correlacionados entre sí.

Como científicos de datos, nos gustaría encontrar un conjunto más pequeño de variables multivariadas que no estén correlacionadas Y que expliquen tanta varianza (o variabilidad) de los datos como sea posible. Este es un enfoque estadístico.

En otras palabras, nos gustaría encontrar la mejor matriz creada con menos variables (es decir, una matriz de rango inferior) que explique los datos originales. Esto está relacionado con la compresión de datos.

Dos soluciones relacionadas a estos problemas son PCA, que significa Análisis de componentes principales y SVD, Descomposición de valores singulares. Esto último simplemente significa que expresamos una matriz X de observaciones (filas) y variables (columnas) como el producto de otras 3 matrices, es decir, X = UDV ^ t. Este último término (V ^ t) representa la transposición de la matriz V.

Aquí, U y V tienen columnas ortogonales (no correlacionadas). Las columnas de U son los vectores singulares de la izquierda de X y las columnas de V son los vectores singulares de la derecha de X. D es una matriz diagonal, por lo que queremos decir que todas sus entradas que no están en la diagonal son 0. Las entradas diagonales de D son el singular valores de X.

> mat

[,1] [,2] [,3]

[1,] 1 2 3

[2,] 2 5 7

svd(mat)

$d

[1] 9.5899624 0.1806108

$u

[,1] [,2]

[1,] -0.3897782 -0.9209087

[2,] -0.9209087 0.3897782

$v

[,1] [,2]

[1,] -0.2327012 -0.7826345

[2,] -0.5614308 0.5928424

[3,] -0.7941320 -0.1897921

Ahora hablaremos un poco sobre PCA, Análisis de componentes principales, "un método simple y no paramétrico para extraer información relevante de conjuntos de datos confusos"

Estamos citando aquí un artículo muy agradable y conciso sobre este tema que se puede encontrar en http://arxiv.org/pdf/1404.1100.pdf. El documento de Jonathon Shlens

of Google Research se llama Tutorial sobre análisis de componentes principales.  
Básicamente, PCA es un método para reducir un conjunto de datos de alta dimensión a sus elementos esenciales (no perder información) y explicar la variabilidad en los datos. No entraremos en los detalles matemáticos aquí (R tiene una función para realizar PCA), pero debe saber que SVD y PCA están estrechamente relacionados.

svd(scale(mat))

$d

[1] 1.732051 0.000000

$u

[,1] [,2]

[1,] -0.7071068 0.7071068

[2,] 0.7071068 0.7071068

$v

[,1] [,2]

[1,] 0.5773503 -0.5773503

[2,] 0.5773503 0.7886751

[3,] 0.5773503 -0.2113249

prcomp(scale(mat))

Standard deviations (1, .., p=2):

[1] 1.732051 0.000000

Rotation (n x k) = (3 x 2):

PC1 PC2

[1,] 0.5773503 -0.5773503

[2,] 0.5773503 0.7886751

[3,] 0.5773503 -0.2113249

observe que los componentes principales de la matriz escalada, que se muestran en el componente Rotación de la salida de prcomp, SON las columnas de V, los valores singulares de la derecha. Por tanto, el PCA de una matriz escalada produce la matriz V (vectores singulares derechos) de la misma matriz escalada

buscar relación entre svd pca y la varianza de los datos

Aquí mostramos nuevamente la matriz de datos agrupados a la izquierda. Junto a él, hemos trazado la primera columna de la matriz U asociada con la matriz de datos escalados. Este es el primer vector singular IZQUIERDO y está asociado con las medias FILA de los datos agrupados. Puede ver la clara separación entre las medias de las 24 filas superiores (alrededor de -0,2) y las 16 inferiores (alrededor de 0,2). No los mostramos, pero tenga en cuenta que las otras columnas de U no muestran este patrón con tanta claridad.

...

| ================================= | 45%

| La pantalla de la derecha muestra la primera columna de la matriz V asociada con la matriz de datos agrupados y escalados. Este es el primer vector singular DERECHO y está asociado con las medias COLUMNA de los datos agrupados. Puede ver la clara separación entre las medias de las 5 columnas de la izquierda (entre -0,1 y 0,1) y las medias de las 5 columnas de la derecha (todas por debajo de -0,4). Al igual que con los vectores singulares de la izquierda, las otras columnas de V no muestran este patrón tan claramente como lo hace este primero.

Ahora probablemente esté convencido de que SVD y PCA son muy interesantes y útiles como herramientas de análisis, pero un problema con ellos que debe conocer es que no pueden hacer frente a los datos FALTANTES. Ninguno de ellos funcionará si falta algún dato en la matriz. (Obtendrá mensajes de error de R en rojo si lo intenta.) Los datos faltantes no son inusuales, por lo que afortunadamente tenemos formas de solucionar este problema. Uno que solo mencionaremos se llama imputación de datos.

...

| Esto usa los k vecinos más cercanos para calcular los valores a usar en lugar de los datos faltantes. Es posible que desee especificar un número entero k que indique cuántos vecinos desea promediar para crear este valor de reemplazo. El paquete de bioconductores (http://bioconductor.org) tiene un paquete de imputación que puede utilizar para completar los datos que faltan. Una función específica en él es impute.knn.