UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

PCA Robusto y Extensiones de PCA

Alan Reyes-Figueroa

Introducción a la Ciencia de Datos (Aula 12) 15.febrero.2021

Métodos robustos para PCA

Idea: Evitar que ciertas observaciones tengan mucha influencia en la estimación de las componentes (*e.g.* datos atípicos o datos extremos).

Usualmente hay dos enfoques:

- limitar el efecto de datos típicos
  - ponderar los datos
  - transformar los datos
- eliminar datos atípicos y convertirlos en estimaciones (*e.g.* método masking).

PCA Ponderado

Ponderamos los datos. En lugar de calcular μ y Σ en la forma usual

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \Sigma = \frac{1}{n}\mathbb{X}_c^T\mathbb{X}_c = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T,$$

calculamos la media ponderada de los datos

$$\mu = \frac{1}{\sum_i w_i}\sum_{i=1}^{n} w_i\mathbf{x}_i,$$

y la matriz de covarianca ponderada

$$\Sigma = \frac{1}{\sum_i w_i}\sum_{i=1}^{n} w_i(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T.$$

Si conociéramos Σ, podemos medir cuán lejos está una observación $\mathbf{x}_i$ del centro de la distribución.

PCA Ponderado

Definición

*La **distancia de Mahalanobis** de una distribución $\mathcal{N}(\mu, \Sigma)$ se calcula como*

$$d_{Mah}(\mathbf{x}_i, \mu) = (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu).$$

Entonces, esto genera un algoritmo iterativo en dos pasos:

1. Si conocemos Σ, calculamos las $d_i = d_{Mah}(\mathbf{x}_i, \mu)$, y podemos definir pesos $w_i = f(d_i)$, donde f es una función decreciente.

2. Con los w_i, podemos re-estimar Σ como

$$\Sigma = \frac{1}{\sum_i w_i} \sum_{i=1}^{n} w_i (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T.$$

Se puede mostrar que bajo ciertos supuestos, este algoritmo converge.

Spherical PCA

Debida a S. Marron.

Transformamos los datos. Para ello, se toma una hiperesfera en $\mathbb{R}^d$, centrada en la mediana robusta μ de los datos, y proyectamos cada datos $\mathbf{x}_i$ sobre dicha esfera.

Equivalente a normalizar la distancia a μ no influye.
Tomamos todos los datos, pero limitamos el efecto sobre la estimación.

Para ello en necesario definir la mediana para datos multidimensionales.
Una posible solución que la mediana μ satisface

$$\sum_{i=1}^{n} \frac{\mathbf{x}_i - \mu}{||\mathbf{x}_i - \mu||} = \mathbf{0}.$$

Minimum Volume Ellipsoid

Debida a P. Rousseau.

Localizamos y eliminamos outliers. Para ello, buscamos el elipsoide de volumen mínimo que contenga cierto porcentaje $h\%$ de los datos. Luego estimamos Σ con sólo esta muestra.

Recordemos que en los elipsoides $\{\mathbf{x} : \ \mathbf{x}^T A\mathbf{x} = c\}$, con A simétrica y positiva definida, el volumen es proporcional a $\det A$.

Pregunta: ¿Cómo hallar el subconjunto $H \subset \mathbb{R}^d$ de las $h\%$ observaciones cuya matriz de covarianza tiene determinante mínimo?

Propiedad

Sean μ y Σ estimadas con un subconjunto H de h% observaciones. Definamos H_1 subconjunto las h% observaciones más cercanas a μ en la distancia de Mahalanobis de Σ. Entonces, para $\Sigma_1 estimadaconH_1$

$$\det \Sigma_1 \leq \det \Sigma.$$

Algoritmo:

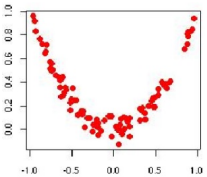
1.) Inicio: Dados μ^o y Σ^o

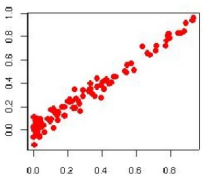
2.) Repetir para $i = 1, 2, 3, \ldots$ (hasta cierto criterio de paro):

- Calcular $S = \{h\%$ de las observaciones más cercanas a $\mu^{i-1}\}$
- Estimar μ^i y Σ^i con base en S.

$$A = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^T \;\Rightarrow\; \mathbf{x}^T A \mathbf{x} = \sum_i (\sigma_i \mathbf{x}^T \mathbf{u}_i)(\sigma_i \mathbf{u}_i^T \mathbf{x}) = \sum_i (\sigma_i \langle \mathbf{x}, \mathbf{u}_i \rangle)^2.$$

Extensiones no-lineales de PCA





Transformar los datos (similar a regresión no-lineal!)

E.g., definimos $\Phi(\mathbf{x}) = \Phi(x_1, x_2) = (x_1^2, x_2)$, y aplicamos PCA a los $\{\Phi(\mathbf{x}_i)\}$.

- Antes: En escalamiento multidimensional,
  $P_u(\mathbf{x}) = \sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle$, donde α_i depende sólo de $\mathbb{X}\mathbb{X}^T = [\langle \mathbf{x}_i, \mathbf{x}_j \rangle]_{i,j}$.
- Ahora: Transformamos $\mathbf{x}$ a $\Phi(\mathbf{x})$, y definimos $K_\Phi(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$.
  $P_u^\Phi(\mathbf{x}) = P_u(\Phi(\mathbf{x})) = \sum_i \alpha_i K_\Phi(\mathbf{x}_i, \mathbf{x})$, α_i depende de $\mathbb{K} = [K_\Phi(\mathbf{x}_i, \mathbf{x}_j)]_{i,j}$.

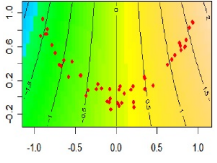
Kernel PCA

Definición

A la función K_Φ se le llama una ***función kernel*** *inducida por Φ. Típicamente, K_Φ debe ser simétrica y tal que $K_\Phi(\mathbf{x}, \mathbf{x}) \geq 0$. Además, se requiere que la matriz $\mathbb{K} = [K_\Phi(\mathbf{x}_i, \mathbf{x}_j)]_{i,j}$, sea definida positiva.*

En el metodo de Kernel PCA, definimos explícitamente $K(\mathbf{x}_i, \mathbf{x}_j)$, y sólo implícitamente Φ.

Problema: nuestra intuición no es buena para pensar en términos de productos puntos (contrario a distancias).



Kernel PCA

Ejemplo 1: Sea $\mathbf{z} = (z_1, z_2) \in \mathbb{R}^2$. Consideremos la transformación $\Phi : \mathbb{R}^2 \to \mathbb{R}^6$ dada por

$$\Phi(\mathbf{z}) = (1, \sqrt{2}z_1, \sqrt{2}z_2, z_1^2, \sqrt{2}z_1z_2, z_2^2).$$

Observe que

$$\begin{aligned} K_\Phi(\mathbf{x}_1, \mathbf{x}_2) &= \langle \Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2) \rangle \\ &= \langle (1, \sqrt{2}x_1, \sqrt{2}y_1, x_1^2, \sqrt{2}x_1y_1, y_1^2), (1, \sqrt{2}x_2, \sqrt{2}y_2, x_2^2, \sqrt{2}x_2y_2, y_2^2) \rangle \\ &= 1 + 2x_1x_2 + 2y_1y_2 + x_1^2x_2^2 + 2x_1x_2y_1y_2 + y_1^2y_2^2 \\ &= (1 + x_1x_2 + y_1y_2)^2 = \left(1 + \langle (x_1, y_1), (x_2, y_2) \rangle\right)^2 \\ &= (1 + \langle \mathbf{x}_1, \mathbf{x}_2 \rangle)^2. \end{aligned}$$

En general podemos definir $K_\Phi(\mathbf{x}, \mathbf{y}) = (1 + \langle \mathbf{x}, \mathbf{y} \rangle)^p$.

Kernel PCA

Ejemplo 2: cuando **x** no tiene una representación vectorial (natural).

Sean **x** y **y** dos cadenas de longitud d sobre el alfabeto $\mathcal{A}$, *i.e.* $\mathbf{x}, \mathbf{y} \in \mathcal{A}^d$.

Definimos $\Phi(\mathbf{x}) = \left(\Phi_s(\mathbf{x})\right)_{s \in \mathcal{A}^d}$, donde $\Phi_s(\mathbf{x})$ denota el número de veces que la subcadena s aparece en **x**.

Eso es más fácil de calcular que $\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$ directamente:

$$\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle = \sum_{s \in S(\mathbf{x}, \mathbf{y})} \Phi_s(\mathbf{x})\, \Phi_s(\mathbf{y}),$$

con $S(x, y)$ el conjunto de subcadenas comunes de **x** y **y**.

Kernel PCA

Ejemplo 3: cuando $\mathbf{x}$ no tiene una representación vectorial (natural).

Sea $\mathbb{P}()$ una distribución de probabilidad. Definimos

$$K_\Phi(\mathbf{x}, \mathbf{y}) = \mathbb{P}(\mathbf{x})\,\mathbb{P}(\mathbf{y}).$$

Interpretación usando la norma inducida por $||\cdot||$:

$$||\Phi(\mathbf{x}) - \Phi(\mathbf{y})||^2 = K_\Phi(\mathbf{x}, \mathbf{x}) - 2K_\Phi(\mathbf{x}, \mathbf{y}) + K_\Phi(\mathbf{y}, \mathbf{y}) = \big(\mathbb{P}(\mathbf{x}) - \mathbb{P}(\mathbf{y})\big)^2.$$

Este es un ejemplo de kernel generativo.

Kernel PCA

Ejemplo 4: trabajar con otras normas.

Una elección muy popular es $K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle = e^{-||\mathbf{x}-\mathbf{y}||^2/\sigma^2}$, un kernel de base radial.

$\Phi(\cdot)$ mapea datos a una hiperesfera en $\mathbb{R}^\infty$. La función de distancia correspondiente es

$$||\Phi(\mathbf{x}) - \Phi(\mathbf{y})||^2 = ke^{-||\mathbf{x}-\mathbf{y}||^2/\sigma^2}.$$