Reducción de Dimensionalidad

Aprendizaje Automático INF-393 II-2018

Ricardo Ñanculef

UTFSM Campus San Joaquín

Table of contents

1. Introducción
2. Análisis de Componentes Principales (PCA)
3. Análisis de Discriminantes Lineales (LDA)

Introducción

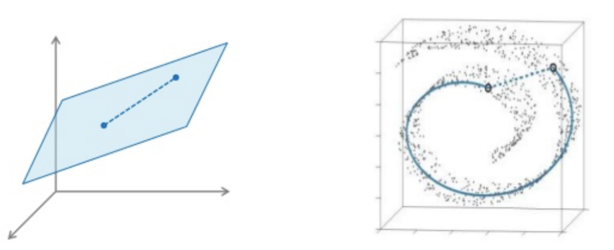
Objetivo

Reducción de Dimensionalidad Dada una representación de los datos como vectores $\mathbb{X} \subset \mathbb{R}^d$, se busca diseñar una función de la forma $\phi : \mathbb{R}^d \to \mathbb{R}^k$ con $k \ll d$, tal que $\phi(\mathbb{X})$ preserve propiedades de $\mathbb{X}$ que consideramos relevantes.

1. Esencialmente, se busca obtener una representación $\mathbb{Z} = \phi(\mathbb{X})$ con menos atributos, que permita:
    1.1 reducir el costo computacional de procesar esos datos (e.g. entrenar un modelo con ellos).
    1.2 reducir el riesgo de overfitting, mejorando la capacidad predictiva de un modelo que se quiere aprender a partir de ejemplos.
    1.3 reducir el impacto de la maldición de la dimensionalidad.

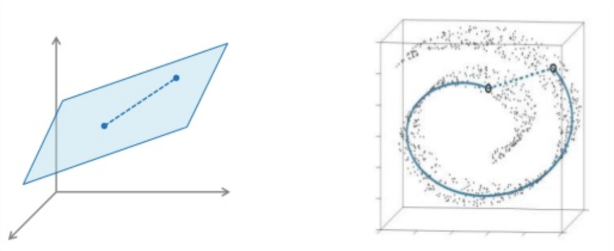
Estructura de los Datos

- Nuestro objetivo de "preservar las propiedades de $\mathbb{X}$" será más fácil de alcanzar si los datos se organizan "naturalmente" como una variedad de $\mathbb{R}^k$ con $k \ll d$, es decir forman una estructura de menor dimensionalidad que la sugiere su representación como vectores en $\mathbb{R}^d$.



Objetivo

- Un método de reducción de dimensionalidad puede verse entonces como un método para "descubrir" esa variedad o construir una variedad que aproxima la estructura general de los datos.



Análisis de Componentes Principales (PCA)

Elecciones Fundamentales

Forma de la Función ϕ. PCA implementa una función $\phi : \mathbb{R}^d \to \mathbb{R}^k$ como un mapa lineal de la forma $\phi(x) = Px$, donde $P \in \mathbb{R}^{kd}$.

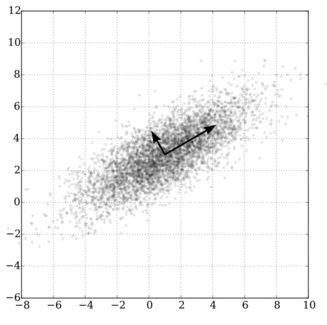
Criterio de Optimalidad. Se busca que la nueva representación preserve la varianza de las observaciones. Concretamente, se busca resolver el siguiente problema de optimización

$$\arg\max_P \frac{\mathbb{E}_x \|Px - P\mathbb{E}[x]\|^2}{\mathbb{E}\|x - \mathbb{E}[x]\|^2} = \arg\max_P \mathbb{E}\|Px - \mathbb{E}[Px]\|^2 . \quad (1)$$

- Si $z = Px$, $\mathbb{E}[z] = P\mathbb{E}[x]$. Por lo tanto, el objetivo anterior consiste en maximizar la varianza total del embedding, i.e., $\mathbb{E}\|z - \mathbb{E}[z]\|^2$.

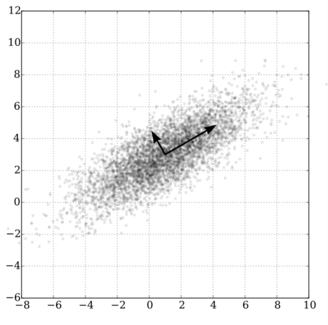
Elecciones Fundamentales

- Notemos que si $k = 1$, $Px = p^T x$ para algún $p \in \mathbb{R}^d$. ¿Cuál de las dos elecciones expuestas más abajo preserva mejor la varianza original de las observaciones (puntos grises)?



Elecciones Fundamentales

- Notemos que si $k = 2$, $z = Px = (z_1, z_2)^T$, con $z_1 = p_1^T x$ y $z_2 = p_2^T x$ para ciertos $p_1, p_2 \in \mathbb{R}^d$ (filas de P). De hecho, en el problema de más abajo, las dos elecciones expuestas preservan completamente la varianza original de las observaciones.



Observaciones Preliminares

- Notemos primero que el problema se puede simplificar asumiendo que $\mathbb{E}[x] = 0$ (basta centrar x). Debemos resolver

$$\max_P \mathbb{E} \|Px\|^2 . \tag{2}$$

- Deberíamos notar ahora que el problema planteado es degenerado. En efecto, si $P_1 = 2P_2$, $\|P_1 x\|^2 > \|P_2 x\|^2$. Es decir,

$$\max_P \mathbb{E} \|Px\|^2 = \infty . \tag{3}$$

- Para concentrarnos en elegir las direcciones correctas $p_1, p_2, \ldots, p_k$ (filas de P), necesitamos restringir su norma:

$$\max_P \mathbb{E} \|Px\|^2 \text{ s.t. } \|p_i\|^2 = \text{cte} \, \forall i . \tag{4}$$

Solución del Problema

- Para resolver

$$\mathcal{P}_1 : \ \max_P \mathbb{E}\, \|Px\|^2 \text{ s.t. } \|p_i\|^2 = \text{cte}\, \forall i\,, \tag{5}$$

podemos considerar la Lagrangiana

$$\mathcal{L}(P, \lambda) = \mathbb{E}\, \|Px\|^2 - \sum_i \lambda_i \left(\|p_i\|^2 - \text{cte} \right)\,. \tag{6}$$

- (KKT) Si P^* es la solución de $\mathcal{P}_1$, debe existir λ^* tal que

$$\frac{\partial \mathcal{L}(P^*, \lambda^*)}{\partial P} = 0\,. \tag{7}$$

Solución del Problema

- La Lagrangiana se puede escribir como

$$
\begin{aligned}
\mathcal{L}(P,\lambda) &= \mathbb{E}\,\|Px\|^2 - \sum\nolimits_i \lambda_i \left(\|p_i\|^2 - \mathrm{cte}\right) \\
&= \mathbb{E}\,(Px)^T(Px) - \sum\nolimits_i \lambda_i \left(p_i^T p_i - \mathrm{cte}\right) \\
&= \mathbb{E}\,\mathrm{tr}(x^T P^T P x) - \mathrm{tr}(\Lambda P^T P - \mathrm{cte}\,\Lambda I) \\
&= \mathbb{E}\,\mathrm{tr}(P x x^T P^T) - \mathrm{tr}(\Lambda P^T P - \mathrm{cte}\,\Lambda I) \\
&= \mathrm{tr}(P \Sigma P^T) - \mathrm{tr}(\Lambda P^T P - \mathrm{cte}\,\Lambda I)\,,
\end{aligned} \tag{8}
$$

con $\Sigma = \mathbb{E}(xx^T)$. La última igualdad la obtenemos de la invarianza cíclica de la traza. Recordando algunas otras (hermosas) propiedades de la traza

$$
\frac{\partial \mathrm{tr}(ABA^T C)}{\partial A} = CAB + C^T A B^T\,, \tag{9}
$$

obtenemos,

$$\frac{\partial \text{tr}(P\Sigma P^T)}{\partial A} = P\Sigma + P\Sigma^T = 2P\Sigma \tag{10}$$

$$\frac{\partial \text{tr}(\Lambda P^T P - \text{cte}\, I)}{\partial A} = P\Lambda + P\Lambda^T = 2P\Lambda\,.$$

- La condición de optimalidad es entonces,

$$P\Sigma = P\Lambda \;\Leftrightarrow\; \Sigma P^T = \Lambda P^T \tag{11}$$

$$\Leftrightarrow\; \Sigma p_i = \lambda_i p_i \,\forall i\,, \tag{12}$$

es decir, $\{p_i\}_{i=1}^k$ es un conjunto de vectores propios de la matriz Σ con valores propios $\{\lambda_i\}_{i=1}^k$.

Solución del Problema

- Reemplazando en la f.o de $\mathcal{P}_1$ obtenemos

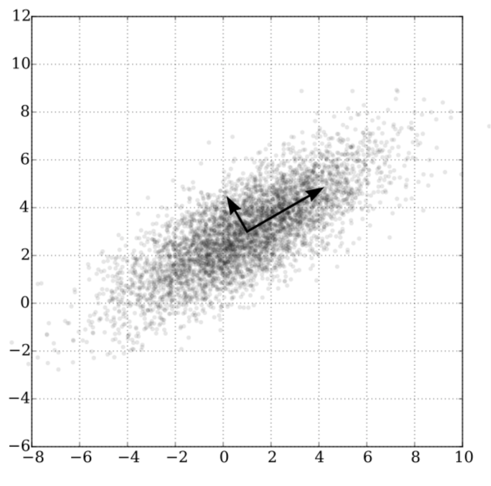
$$g(P) = \text{tr}(P\Sigma P^T) = \text{tr}(P\Lambda P^T) = \text{tr}(P\Lambda P^T) = \text{tr}(\Lambda P^T P) \tag{13}$$

- Como los vectores propios de la matriz Σ son ortogonales, obtenemos

$$g(P) = \text{cte} \sum_{i=1}^{k} \lambda_i \tag{14}$$

- Ahora, Como queremos maximizar la f.o. $g(P)$, se sigue que debemos elegir los vectores propios $\{p_i\}_{i=1}^{k}$ de Σ con valores propios $\{\lambda_i\}_{i=1}^{k}$ lo más grandes posible. Estos vectores propios definen las direcciones principales de los datos.

Direcciones Principales



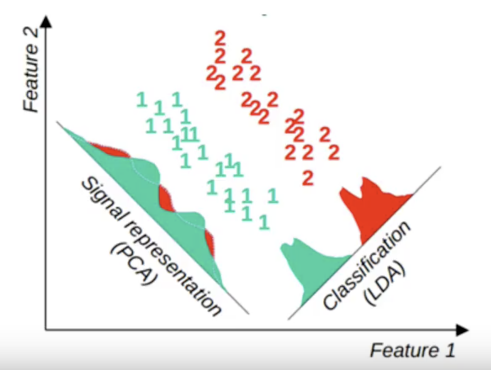
PCA

- Algoritmo:
  1. Estimar $\Sigma = \mathbb{E}(xx^T)$.
  2. Calcular de descomposición de valores propios de Σ, $\Sigma = U \Lambda U^T$.
  3. Ordenar las columnas de U en modo creciente según Λ_{ii}.
  4. $P = U_{1:k}^T$ (donde $U_{1:k}$ es la matriz U truncada a sus primeras k columnas).

Análisis de Discriminantes Lineales (LDA)

Problema de PCA

- En un problema de clasificación, debiésemos estar interesados en preservar la separación original de las clases, en vez de preservar la varianza total.



Descomposición de la Varianza

- Idea: Separar la varianza en una componente intra-clases y una componente inter-clases.
- Sea $m_y = \mathbb{E}_{x|y} x$ (media por clase) y veamos qué sucede con la varianza intra-clases

$$
\begin{aligned}
\mathbb{E}_y \mathbb{E}_{x|y} (x - m_y)^T (x - m_y) &= \mathbb{E}_y \mathbb{E}_{x|y} \, x^T x - m_y^T x - x^T m_y + m_y^T m_y \\
&= \mathbb{E}_y \mathbb{E}_{x|y} \, x^T x - 2 \, \mathbb{E}_y \mathbb{E}_{x|y} \, m_y^T x + \mathbb{E}_y \mathbb{E}_{x|y} \, m_y^T m_y \\
&= \mathbb{E}_x \, x^T x - 2 \, \mathbb{E}_y \, m_y^T m_y + \mathbb{E}_y \, m_y^T m_y \\
&= \mathbb{E}_x \left(x^T x \right) - \mathbb{E}_y \, \left(m_y^T m_y \right) . \qquad (15)
\end{aligned}
$$

Descomposición de la Varianza

- Recordando que $m = \mathbb{E}x = \mathbb{E}_y \mathbb{E}_{x|y} x = \mathbb{E}_y m_y = 0$

$$\begin{aligned} \mathbb{E}_x \left(x^T x \right) &= \mathbb{E}_x (x - m)^T (x - m) = \text{varianza total} \\ \mathbb{E}_y \left(m_y^T m_y \right) &= \mathbb{E}_y (m_y - m)^T (m_y - m) = \text{varianza inter} \\ \mathbb{E}_y \mathbb{E}_{x|y} (x - m_y)^T (x - m_y) &= \text{varianza intra} . \end{aligned} \tag{16}$$

- Por lo tanto

$$\text{varianza total} = \text{varianza intra} + \text{varianza inter} .$$

Reducción de Dimensionalidad vía LDA

Forma de la Función ϕ. LDA implementa la función $\phi : \mathbb{R}^d \to \mathbb{R}^k$ igual que PCA, es decir, como un mapa lineal de la forma $\phi(x) = Px$, donde $P \in \mathbb{R}^{kd}$.

Criterio de Optimalidad. Se busca que la nueva representación maximice la varianza inter después de la proyección, sin aumentar la varianza intra

$$\begin{aligned} \max_P \ & \mathbb{E}_y \left((Pm_y)^T (Pm_y) \right) \\ \text{s.t. } & \mathbb{E}_y \mathbb{E}_{x|y} \left(PxP - m_y \right)^T (Px - Pm_y) = \text{cte} \,, \end{aligned}$$

es decir,

$$\begin{aligned} \max_P \ & \mathbb{E}_y \ m_y^T P^T P m_y \\ \text{s.t. } & \mathbb{E}_y \mathbb{E}_{x|y} \left(x - m_y \right)^T P^T P (x - m_y) = \text{cte} \,. \end{aligned} \tag{17}$$

Reducción de Dimensionalidad vía LDA

- Notemos que si definimos $mP_y = \mathbb{E}_{x|y} Px$ y $mP = \mathbb{E}_x Px$, tenemos

$$
\begin{aligned}
mP_y &= \mathbb{E}_{x|y} Px = P\,\mathbb{E}_{x|y} x = Pm_y \\
mP &= \mathbb{E}_x Px = P\,\mathbb{E}_x x = Pm\,,
\end{aligned}
$$

por lo que efectivamente la f.o. en (17) es la varianza inter después de la proyección,

$$
\mathbb{E}_y\, m_y^T P^T P m_y = \mathbb{E}_y\,(mP_y - mP)^T (mP_y - mP)\,.
$$

- Ahora, si notamos que la f.o. (17) es de forma escalar, al igual que la restricción, podemos escribir (17) de manera más conveniente ...

Reducción de Dimensionalidad vía LDA

- En efecto (17) es equivalente a:

$$\begin{aligned} &\max_P \ \mathbb{E}_y \operatorname{tr}\left(m_y^T P^T P m_y\right) \\ &\text{s.t. } \mathbb{E}_y \mathbb{E}_{x|y} \operatorname{tr}\left((x - m_y)^T P^T P (x - m_y)\right) = \text{cte}\,. \end{aligned} \tag{18}$$

- Usando la linealidad de la traza, la linealidad del valor esperado y luego la invarianza de la traza frente a permutaciones cíclicas, obtenemos que (18) es equivalente a

$$\begin{aligned} &\max_P \ \operatorname{tr}\left(P\, \mathbb{E}_y(m_y m_y^T) P^T\right) \\ &\text{s.t. } \operatorname{tr}\left(P\, \mathbb{E}_y \mathbb{E}_{x|y}\left((x - m_y)(x - m_y)^T\right) P^T\right) = \text{cte}\,. \end{aligned} \tag{19}$$

Reducción de Dimensionalidad vía LDA

- Los nuevos términos involucrados tienen nombre propio:

$$\Sigma_B = \mathbb{E}_y(m_y m_y^T) = \text{matriz de covarianza inter} \tag{20}$$

$$\Sigma_I = \mathbb{E}_y \mathbb{E}_{x|y}\left((x - m_y)(x - m_y)^T\right) = \text{matriz de covarianza intra}$$

c• De este modo, nuestro problema se puede re-escribir como

$$\max_P \ \mathrm{tr}\left(P \Sigma_B P^T\right) \ \text{s.t.}\ \mathrm{tr}\left(P\, \Sigma_I P^T\right) = \mathrm{cte}\,. \tag{21}$$

- Escribiendo la Lagrangiana y usando las condiciones de KKT, obtenemos que P debe satisfacer la siguiente ecuación

$$\Sigma_B P^T = \Lambda \Sigma_I P^T \tag{22}$$

que corresponde a una ecuación generalizada de vectores propios.

Reducción de Dimensionalidad vía LDA

- Para que la restricción $\mathrm{tr}\left(P\,\Sigma_I P^T\right) = \mathrm{cte}$ involucre solo una elección sobre la norma de los vectores propios involucrados, podemos hacer la siguiente transformación

$$\tilde{P} = P\Sigma_I^{1/2} \Rightarrow P = \tilde{P}\Sigma_I^{-1/2}\,. \tag{23}$$

- De este modo, el problema de LDA se puede re-escribir como

$$\max_P\ \mathrm{tr}\left(\tilde{P}\Sigma_I^{-1/2}\Sigma_B\Sigma_I^{-1/2}\tilde{P}^T\right)\ \text{s.t. } \mathrm{tr}\left(\tilde{P}\tilde{P}^T\right) = \mathrm{cte}\,. \tag{24}$$

- Las condiciones de KKT llevan a la siguiente ecuación

$$\Sigma_I^{-1/2}\Sigma_B\Sigma_I^{-1/2}\tilde{P}^T = \Lambda\tilde{P}^T \tag{25}$$

que muestra que las filas de $\tilde{P}$ deben corresponder a vectores propios de la matriz $\Sigma_I^{-1/2}\Sigma_B\Sigma_I^{-1/2}$, con norma unitaria. Esto implica, en particular, que las filas de $\tilde{P}$ son ortogonales.

Reducción de Dimensionalidad vía LDA

- Reemplazando la condición en la f.o. notamos que

$$\operatorname{tr}\left(\tilde{P}\Sigma_I^{-1/2}\Sigma_B\Sigma_I^{-1/2}\tilde{P}^T\right) = \operatorname{tr}\left(\tilde{P}\Lambda\tilde{P}^T\right) = \operatorname{tr}\left(\Lambda\tilde{P}^T\tilde{P}\right) \quad = \operatorname{tr}(\Lambda) = \sum_{i=1}^{k}\lambda_i\,,$$

es decir, si queremos maximizar la f.o. de LDA conviene elegir los k vectores propios de la matriz $\Sigma_I^{-1/2}\Sigma_B\Sigma_I^{-1/2}$ que corresponden a los valores propios más grandes.

Reducción de Dimensionalidad vía LDA

- Algoritmo:

  1. Estimar la matriz de covarianza intra $\Sigma_B = \mathbb{E}_y(m_y m_y^T)$.
  2. Estimar la matriz de covarianza inter $\Sigma_I = \mathbb{E}_y \mathbb{E}_{x|y} \left((x - m_y)(x - m_y)^T\right)$.
  3. Calcular la matriz $M = \Sigma_I^{-1/2} \Sigma_B \Sigma_I^{-1/2}$.
  4. Calcular de descomposición de valores propios de M, $M = U \Lambda U^T$.
  5. Ordenar las columnas de U en modo creciente según Λ_{ii}.
  6. $P = U_{1:k}^T \Sigma_I^{-1/2}$ (donde $U_{1:k}$ es la matriz U truncada a sus primeras k columnas).