

Reducción de Dimensionalidad

Aprendizaje Automático INF-393 II-2018

Ricardo Nanculef

UTFSM Campus San Joaquín

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

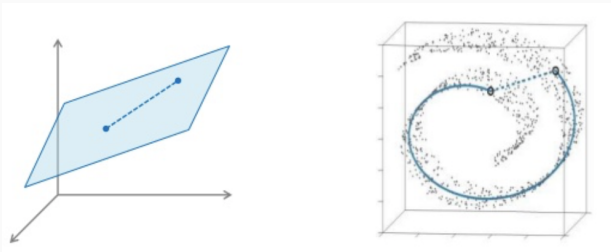
Introducción

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 $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^k$ 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.

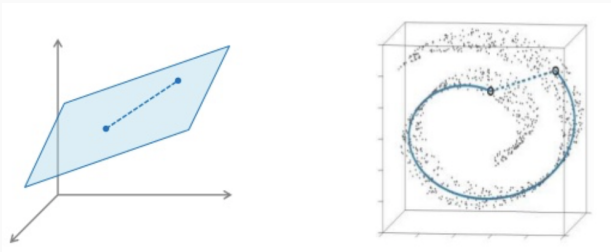
Estructura de los Datos

- 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)

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

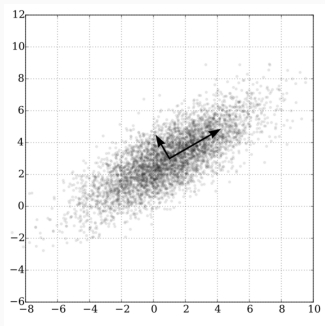
Criterio de Optimalidad. Se busca que la nueva representación **preserve la varianza** de las observaciones. Concretamente, se busca resolver el siguiente problema de optimización

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

- 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., $\mathbb{E} \|z - \mathbb{E}[z]\|^2$.

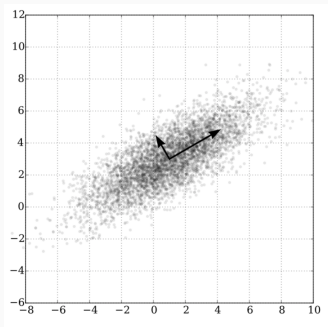
Elecciones Fundamentales

- 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 $m = \mathbb{E}[x] = 0$ (basta centrar x). Debemos resolver

$$\max_P \mathbb{E} \|P_X\|^2. \quad (2)$$

- 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,

$$\max_P \mathbb{E} \|P_X\|^2 = \infty. \quad (3)$$

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

$$\max_P \mathbb{E} \|P_X\|^2 \text{ s.t. } \|p_i\|^2 = \text{cte} \forall i. \quad (4)$$

Solución del Problema

- Para resolver

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

podemos considerar la Lagrangiana

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

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

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

Solución del Problema

- La Lagrangiana se puede escribir como

$$\begin{aligned}\mathcal{L}(P, \lambda) &= \mathbb{E} \|P\mathbf{x}\|^2 - \sum_i \lambda_i (\|p_i\|^2 - \text{cte}) \\ &= \mathbb{E} (P\mathbf{x})^T (P\mathbf{x}) - \sum_i \lambda_i (p_i^T p_i - \text{cte}) \\ &= \mathbb{E} \text{tr}(\mathbf{x}^T P^T P \mathbf{x}) - \text{tr}(\Lambda(P^T P - \text{cte} I)) \\ &= \mathbb{E} \text{tr}(P \mathbf{x} \mathbf{x}^T P^T) - \text{tr}(\Lambda P^T P - \text{cte} \Lambda) \\ &= \text{tr}(P \Sigma P^T) - \text{tr}(P \Lambda P^T) - \text{cte}',\end{aligned}\tag{8}$$

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

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

obtenemos,

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

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

- La condición de optimalidad es entonces,

$$P\Sigma = P\Lambda \Leftrightarrow \Sigma P^T = \Lambda P^T \quad (11)$$

$$\Leftrightarrow \Sigma p_i = \lambda_i p_i \forall i, \quad (12)$$

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

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

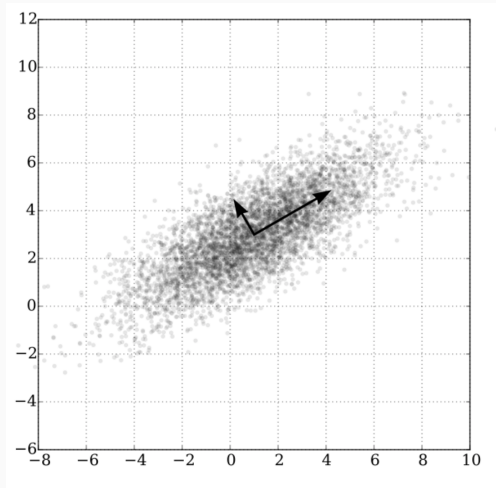
$$g(P) = \text{tr}(P\Sigma P^T) = \text{tr}(P\Lambda P^T) = \text{tr}(P\Lambda P^T) = \text{tr}(\Lambda P^T P) \quad (13)$$

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

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

- 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



- Algoritmo:

1. Estimar $\Sigma = \mathbb{E}(\mathbf{x}\mathbf{x}^T)$.
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).

Cómo estimar $\Sigma = \mathbb{E}(xx^T)$?

- Si tenemos un conjunto de ejemplos $\{x^{(\ell)}\}_{\ell=1}^n$ (no necesitamos etiquetas), y los datos están centrados, podemos estimar Σ como:

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n x^{(\ell)} x^{(\ell)T} \quad (15)$$

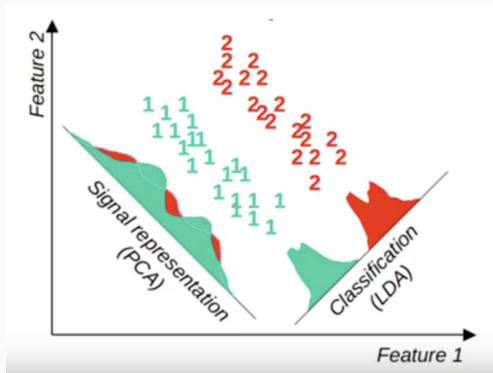
- Si los datos no están centrados, calculamos $\bar{x} = \frac{1}{n-1} \sum_{i=1}^n x^{(\ell)}$ y luego podemos estimar Σ como:

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n \left(x^{(\ell)} - \bar{x} \right) \left(x^{(\ell)} - \bar{x} \right)^T \quad (16)$$

Análisis de Discriminantes Lineales (LDA)

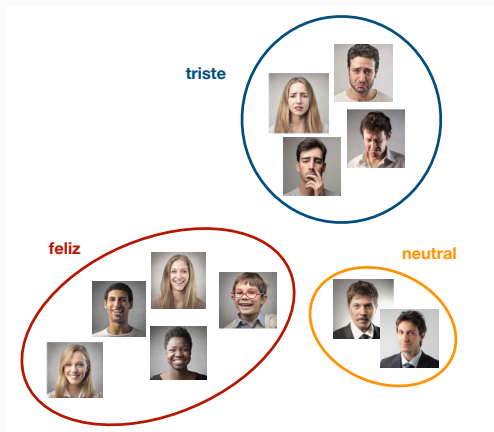
Problema de PCA

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



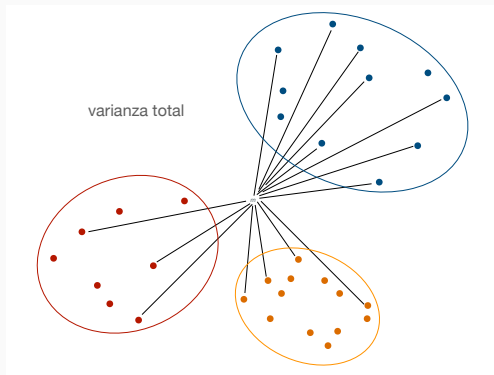
Descomposición de la Varianza

- **Idea:** Si tenemos etiquetas, tiene sentido separar la varianza en una componente **intra-clases** y una componente **inter-clases**.



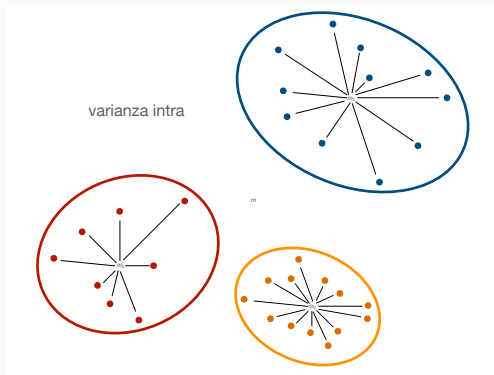
Descomposición de la Varianza

- **Idea:** Si tenemos etiquetas, tiene sentido separar la varianza en una componente **intra-clases** y una componente **inter-clases**.



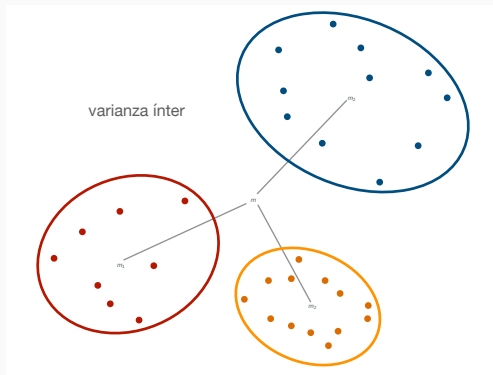
Descomposición de la Varianza

- **Idea:** Si tenemos etiquetas, tiene sentido separar la varianza en una componente **intra-clases** y una componente **inter-clases**.



Descomposición de la Varianza

- **Idea:** Si tenemos etiquetas, tiene sentido separar la varianza en una componente **intra-clases** y una componente **inter-clases**.



Descomposición de la Varianza

- **Idea:** Separar la varianza en una componente **intra-clases** y una componente **inter-clases**.

$$\begin{aligned}\text{varianza total} &= \mathbb{E}_x (x - m)^T (x - m) \\ \text{varianza inter} &= \mathbb{E}_y (m_y - m)^T (m_y - m) \\ \text{varianza intra} &= \mathbb{E}_y \mathbb{E}_{x|y} (x - m_y)^T (x - m_y).\end{aligned}\tag{17}$$

con $m = \mathbb{E}_x$, $m_y = \mathbb{E}_{x|y}$.

- Notemos que

$$\begin{aligned}\mathbb{E}_y \mathbb{E}_{x|y} (x - m_y)^T (x - m_y) &= \mathbb{E}_y \mathbb{E}_{x|y} x^T x - m_y^T x - x^T m_y + m_y^T m_y \\ &= \mathbb{E}_y \mathbb{E}_{x|y} x^T x - 2 \mathbb{E}_y \mathbb{E}_{x|y} m_y^T x + \mathbb{E}_y \mathbb{E}_{x|y} m_y^T m_y \\ &= \mathbb{E}_x x^T x - 2 \mathbb{E}_y m_y^T m_y + \mathbb{E}_y m_y^T m_y \\ &= \mathbb{E}_x (x^T x) - \mathbb{E}_y (m_y^T m_y).\end{aligned}\tag{18}$$

- Si tenemos que $m = \mathbb{E}x = \mathbb{E}_y \mathbb{E}_{x|y} x = \mathbb{E}_y m_y = 0$. Por lo tanto,

$$\begin{aligned}\mathbb{E}_x (x^T x) &= \mathbb{E}_x (x - m)^T (x - m) = \text{varianza total} \\ \mathbb{E}_y (m_y^T m_y) &= \mathbb{E}_y (m_y - m)^T (m_y - m) = \text{varianza inter}\end{aligned}\quad (19)$$

Por lo tanto,

$$\begin{aligned}\text{varianza intra} &= \text{varianza total} - \text{varianza inter} \\ \Leftrightarrow \text{varianza total} &= \text{varianza intra} + \text{varianza inter}.\end{aligned}\quad (20)$$

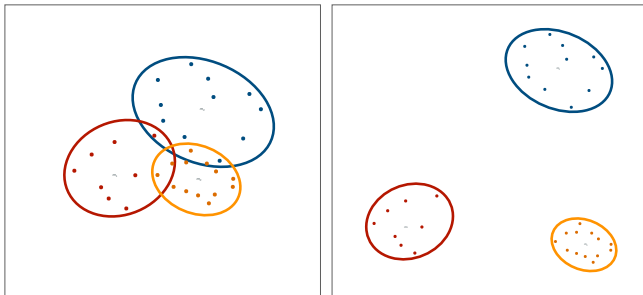
Forma de la Función ϕ . LDA implementa la función $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^k$ igual que PCA, es decir, como un mapa lineal de la forma $\phi(x) = Px$, donde $P \in \mathbb{R}^{kd}$.

Criterio de Optimalidad. Se busca que la nueva representación maximice la varianza inter (separación entre las clases) y minimice la varianza intra después de la proyección (concretamente, el denominado “cuociente de Rayleigh”):

$$\max_P \left(\frac{\text{varianza inter}}{\text{varianza intra}} \right) \quad \text{s.t. } \|p_i\|^2 = \text{cte } \forall i,$$

Reducción de Dimensionalidad vía LDA

Criterio de Optimalidad. Se busca que la nueva representación maximice la **varianza inter** (separación entre las clases) y minimice la **varianza intra** después de la proyección:



Criterio de Optimalidad:

$$\begin{aligned} \max_P \left(\frac{\text{varianza inter}}{\text{varianza intra}} \right) &= \frac{\mathbb{E}_y m_y^T P^T P m_y}{\mathbb{E}_y \mathbb{E}_{x|y} (x - m_y)^T P^T P (x - m_y)} \\ \text{s.t. } \|p_i\|^2 &= \text{cte} \forall i, \end{aligned} \quad (21)$$

- Notemos que si definimos $\tilde{m}_y = \mathbb{E}_{x|y} P x$ y $\tilde{m} = \mathbb{E}_x P x$, tenemos

$$\tilde{m}_y = \mathbb{E}_{x|y} P x = P \mathbb{E}_{x|y} x = P m_y$$

$$\tilde{m} = \mathbb{E}_x P x = P \mathbb{E}_x x = P m,$$

por lo que efectivamente los términos de la f.o. en (21) corresponden a las varianzas después de la proyección. Por ejemplo,

$$\mathbb{E}_y (\tilde{m}_y - \tilde{m})^T (\tilde{m}_y - \tilde{m}) = \mathbb{E}_y m_y^T P^T P m_y.$$

Reducción de Dimensionalidad vía LDA

- Ahora, si notamos que la f.o. (21) es de forma escalar, podemos escribir (21) de manera más conveniente ...

$$\begin{aligned} \max_P \left(\frac{\text{varianza inter}}{\text{varianza intra}} \right) &= \frac{\mathbb{E}_y \operatorname{tr} (m_y^T P^T P m_y)}{\mathbb{E}_y \mathbb{E}_{x|y} \operatorname{tr} ((x - m_y)^T P^T P (x - m_y))} \\ \text{s.t. } \|p_i\|^2 &= \text{cte } \forall i, \end{aligned} \quad (22)$$

Aprovechando las propiedades de la traza y la linealidad del valor esperado, obtenemos ...

$$\max_P \frac{\operatorname{tr} (P \mathbb{E}_y (m_y m_y^T) P^T)}{\operatorname{tr} (P \mathbb{E}_y \mathbb{E}_{x|y} ((x - m_y)(x - m_y)^T) P^T)} \quad \text{s.t. } \|p_i\|^2 = \text{cte } \forall i, \quad (23)$$

que se toma la forma clásica ...

$$\max_P \frac{\operatorname{tr} (P \Sigma_B P^T)}{\operatorname{tr} (P \Sigma_I P^T)} \quad \text{s.t. } \|p_i\|^2 = \text{cte } \forall i, \quad (24)$$

Reducción de Dimensionalidad vía LDA

- Los nuevos términos involucrados se denominan:

$$\Sigma_B = \mathbb{E}_y(m_y m_y^T) = \text{matriz de covarianza inter} \quad (25)$$

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

- Para resolver este problema, resulta útil hacer la transformación $\tilde{P} = P \Sigma_I^{-1/2}$ o bien $P = \tilde{P} \Sigma_I^{-1/2}$. Esto reduce el problema a

$$\max_{\tilde{P}} \frac{\text{tr}(\tilde{P} \Sigma_I^{-1/2} \Sigma_B \Sigma_I^{-1/2} \tilde{P}^T)}{\text{tr}(\tilde{P} \tilde{P}^T)} \quad \text{s.t. } \|\tilde{p}_i\|^2 = \text{cte} \forall i, \quad (26)$$

o más simple aún,

$$\max_{\tilde{P}} \text{tr}(\tilde{P} \Sigma_I^{-1/2} \Sigma_B \Sigma_I^{-1/2} \tilde{P}^T) \quad \text{s.t. } \|\tilde{p}_i\|^2 = \text{cte} \forall i, \quad (27)$$

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

$$\Sigma_I^{-1/2} \Sigma_B \Sigma_I^{-1/2} \tilde{P} = \Lambda \tilde{P} \quad (28)$$

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

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

$$\begin{aligned}\text{tr}\left(\tilde{P}\Sigma_I^{-1/2}\Sigma_B\Sigma_I^{-1/2}\tilde{P}^T\right) &= \text{tr}\left(\tilde{P}\Lambda\tilde{P}^T\right) = \text{tr}\left(\Lambda\tilde{P}^T\tilde{P}\right) \\ &= \sum_{i=1}^k \lambda_i \|\tilde{p}_i\|^2 = \text{cte} \sum_{i=1}^k \lambda_i ,\end{aligned}\tag{29}$$

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.

- 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
 $\Sigma_I = \mathbb{E}_y \mathbb{E}_{x|y} ((x - m_y)(x - m_y)^T)$.
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).