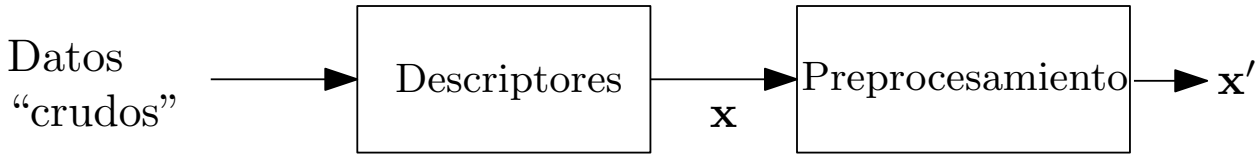
Preprocesamiento

Fernando Lozano

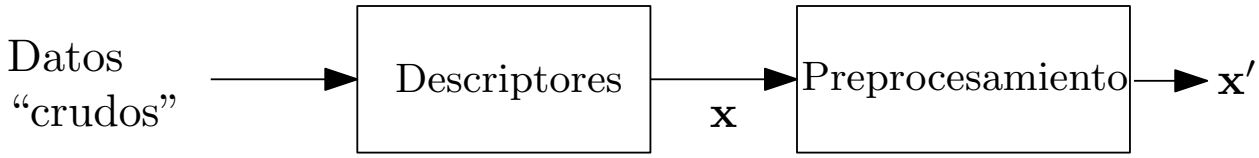
Universidad de los Andes

24 de octubre de 2014

Preprocesamiento

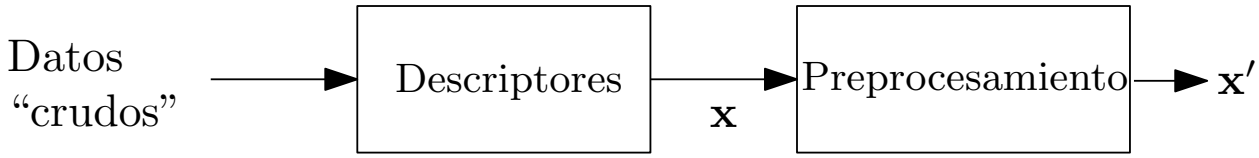


Preprocesamiento



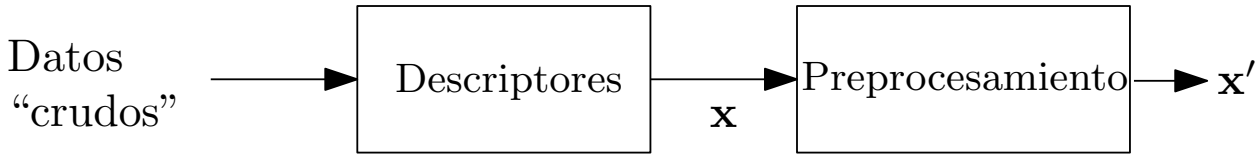
- Usualmente aprendizaje no es posible con datos “crudos”.

Preprocesamiento



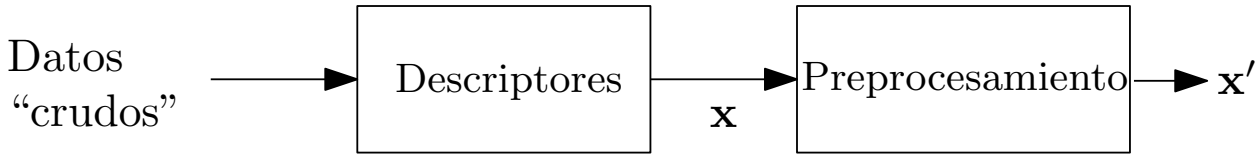
- Usualmente aprendizaje no es posible con datos “crudos”.
- Preprocesamiento es muchas veces la fase más importante.

Preprocesamiento



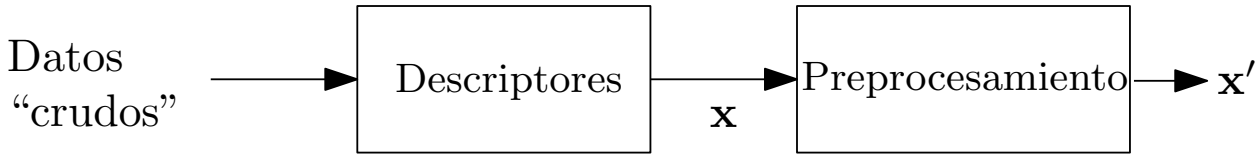
- Usualmente aprendizaje no es posible con datos “crudos”.
- Preprocesamiento es muchas veces la fase más importante.
- Extracción de descriptores (features) :

Preprocesamiento



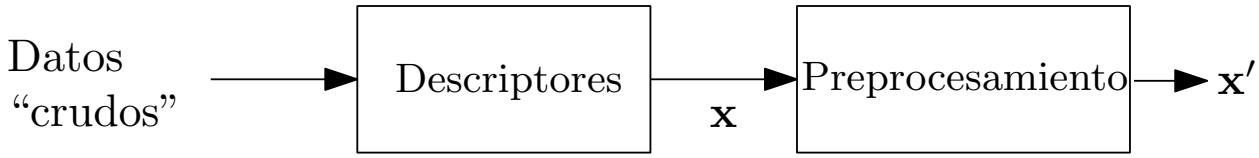
- Usualmente aprendizaje no es posible con datos “crudos”.
- Preprocesamiento es muchas veces la fase más importante.
- Extracción de descriptores (features) :
  - Conocimiento previo del problema

Preprocesamiento



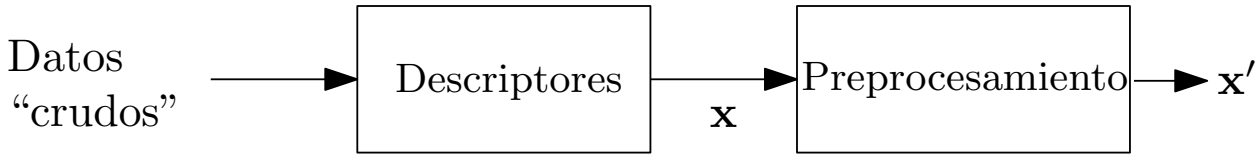
- Usualmente aprendizaje no es posible con datos “crudos”.
- Preprocesamiento es muchas veces la fase más importante.
- Extracción de descriptores (features) :
  - Conocimiento previo del problema
  - Descriptores similares entre datos de la misma clase, diferentes entre datos de clases diferentes.

Preprocesamiento



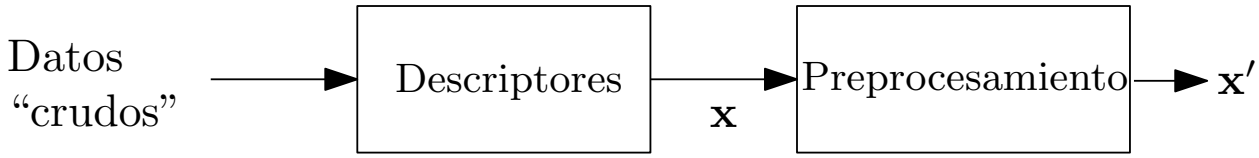
- Usualmente aprendizaje no es posible con datos “crudos”.
- Preprocesamiento es muchas veces la fase más importante.
- Extracción de descriptores (features) :
  - Conocimiento previo del problema
  - Descriptores similares entre datos de la misma clase, diferentes entre datos de clases diferentes.
- Preprocesamiento:

Preprocesamiento



- Usualmente aprendizaje no es posible con datos “crudos”.
- Preprocesamiento es muchas veces la fase más importante.
- Extracción de descriptores (features) :
  - Conocimiento previo del problema
  - Descriptores similares entre datos de la misma clase, diferentes entre datos de clases diferentes.
- Preprocesamiento:
  - Adecuación de los datos.

Preprocesamiento



- Usualmente aprendizaje no es posible con datos “crudos”.
- Preprocesamiento es muchas veces la fase más importante.
- Extracción de descriptores (features) :
  - Conocimiento previo del problema
  - Descriptores similares entre datos de la misma clase, diferentes entre datos de clases diferentes.
- Preprocesamiento:
  - Adecuación de los datos.
  - Reducción de dimensionalidad: $\mathbf{x}' = f(\mathbf{x})$, con $\dim(\mathbf{x}') \ll \dim(\mathbf{x})$

Normalización

- Media:

$$\mathbf{x}' = \mathbf{x} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

Normalización

- Media:

$$\mathbf{x}' = \mathbf{x} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

- Escala (varianza):

$$\mathbf{x}'' = \frac{\mathbf{x}'}{\sigma^2}, \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i')^2$$

Normalización

- Media:

$$\mathbf{x}' = \mathbf{x} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

- Escala (varianza):

$$\mathbf{x}'' = \frac{\mathbf{x}'}{\sigma^2}, \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}'_i)^2$$

- Datos centrados, en la misma escala.

Blanqueo

- Suponga datos centrados $\mathbf{x}$.

Blanqueo

- Suponga datos centrados $\mathbf{x}$.
- Matriz de covarianza: $\mathbf{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$.

Blanqueo

- Suponga datos centrados $\mathbf{x}$.
- Matriz de covarianza: $\mathbf{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$.
- $\mathbf{\Sigma}$ tiene vectores propios $\mathbf{u}_i$ con valores propios λ_i, y

$$\mathbf{U} = [\mathbf{u}_1 | \mathbf{u}_2 | \dots | \mathbf{u}_n] \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

Blanqueo

- Suponga datos centrados $\mathbf{x}$.
- Matriz de covarianza: $\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$.
- $\mathbf{\Sigma}$ tiene vectores propios $\mathbf{u}_i$ con valores propios λ_i, y

$$\mathbf{U} = [\mathbf{u}_1|\mathbf{u}_2|\ldots|\mathbf{u}_n] \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

- Transformación lineal: $\mathbf{y} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{x}$.

Blanqueo

- Suponga datos centrados $\mathbf{x}$.
- Matriz de covarianza: $\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$.
- $\mathbf{\Sigma}$ tiene vectores propios $\mathbf{u}_i$ con valores propios λ_i, y

$$\mathbf{U} = [\mathbf{u}_1 | \mathbf{u}_2 | \ldots | \mathbf{u}_n] \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

- Transformación lineal: $\mathbf{y} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{x}$.

$$\mathbf{\Sigma}_\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^T$$

Blanqueo

- Suponga datos centrados $\mathbf{x}$.
- Matriz de covarianza: $\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$.
- $\mathbf{\Sigma}$ tiene vectores propios $\mathbf{u}_i$ con valores propios λ_i, y

$$\mathbf{U} = [\mathbf{u}_1 | \mathbf{u}_2 | \ldots | \mathbf{u}_n] \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

- Transformación lineal: $\mathbf{y} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{x}$.

$$\mathbf{\Sigma}_\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^T = \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Lambda}^{-1/2}\mathbf{U}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{U}\mathbf{\Lambda}^{-1/2}$$

Blanqueo

- Suponga datos centrados $\mathbf{x}$.
- Matriz de covarianza: $\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$.
- $\mathbf{\Sigma}$ tiene vectores propios $\mathbf{u}_i$ con valores propios λ_i, y

$$\mathbf{U} = [\mathbf{u}_1 | \mathbf{u}_2 | \dots | \mathbf{u}_n] \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

- Transformación lineal: $\mathbf{y} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{x}$.

$$\begin{aligned} \mathbf{\Sigma}_\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^T &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Lambda}^{-1/2}\mathbf{U}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{U}\mathbf{\Lambda}^{-1/2} \\ &= \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{\Sigma}\mathbf{U}\mathbf{\Lambda}^{-1/2} \end{aligned}$$

Blanqueo

- Suponga datos centrados $\mathbf{x}$.
- Matriz de covarianza: $\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$.
- $\mathbf{\Sigma}$ tiene vectores propios $\mathbf{u}_i$ con valores propios λ_i, y

$$\mathbf{U} = [\mathbf{u}_1|\mathbf{u}_2|\ldots|\mathbf{u}_n] \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

- Transformación lineal: $\mathbf{y} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{x}$.

$$\begin{aligned} \mathbf{\Sigma}_\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^T &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Lambda}^{-1/2}\mathbf{U}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{U}\mathbf{\Lambda}^{-1/2} \\ &= \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{\Sigma}\mathbf{U}\mathbf{\Lambda}^{-1/2} = \mathbf{I} \end{aligned}$$

Selección de características

- Reducir dimensionalidad seleccionando un subconjunto de los descriptores:

Selección de características

- Reducir dimensionalidad seleccionando un subconjunto de los descriptores:
  - Eliminar descriptores redundates.

Selección de características

- Reducir dimensionalidad seleccionando un subconjunto de los descriptores:
  - Eliminar descriptores redundates.
  - Prevenir overfitting.

Selección de características

- Reducir dimensionalidad seleccionando un subconjunto de los descriptores:
  - Eliminar descriptores redundates.
  - Prevenir overfitting.
  - Reducir tiempo de entrenamiento.

Selección de características

- Reducir dimensionalidad seleccionando un subconjunto de los descriptores:
  - Eliminar descriptores redundates.
  - Prevenir overfitting.
  - Reducir tiempo de entrenamiento.
- Hay si $\dim(\mathbf{x}) = d$ hay $d!$ posibles subconjuntos. Si fijamos un tamaño $\tilde{d}$, hay $\binom{d}{\tilde{d}}$ posibles subconjuntos.

Selección de características

- Reducir dimensionalidad seleccionando un subconjunto de los descriptores:
  - Eliminar descriptores redundates.
  - Prevenir overfitting.
  - Reducir tiempo de entrenamiento.
- Hay si $\dim(\mathbf{x}) = d$ hay $d!$ posibles subconjuntos. Si fijamos un tamaño $\tilde{d}$, hay $\binom{d}{\tilde{d}}$ posibles subconjuntos.
- Elementos:

Selección de características

- Reducir dimensionalidad seleccionando un subconjunto de los descriptores:
  - Eliminar descriptores redundates.
  - Prevenir overfitting.
  - Reducir tiempo de entrenamiento.
- Hay si $\dim(\mathbf{x}) = d$ hay $d!$ posibles subconjuntos. Si fijamos un tamaño $\tilde{d}$, hay $\binom{d}{\tilde{d}}$ posibles subconjuntos.
- Elementos:
  - Criterio de evaluación.

Selección de características

- Reducir dimensionalidad seleccionando un subconjunto de los descriptores:
  - Eliminar descriptores redundates.
  - Prevenir overfitting.
  - Reducir tiempo de entrenamiento.
- Hay si $\dim(\mathbf{x}) = d$ hay $d!$ posibles subconjuntos. Si fijamos un tamaño $\tilde{d}$, hay $\binom{d}{\tilde{d}}$ posibles subconjuntos.
- Elementos:
  - Criterio de evaluación.
  - Método de búsqueda.

Criterio de evaluación

- Entrenar modelo en subconjunto de descriptores

Criterio de evaluación

- Entrenar modelo en subconjunto de descriptores (costoso computacionalmente).

Criterio de evaluación

- Entrenar modelo en subconjunto de descriptores (costoso computacionalmente).
- Entrenar modelo sencillo (e.g.. LMS)

Criterio de evaluación

- Entrenar modelo en subconjunto de descriptores (costoso computacionalmente).
- Entrenar modelo sencillo (e.g.. LMS)
- Criterio de separabilidad de los datos.

Criterio de Fisher

- Proyección $y = \mathbf{w}^T \mathbf{x}$
- Sean $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ las medias de los datos de cada clase y $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ sus covarianzas.

Criterio de Fisher

- Proyección $y = \mathbf{w}^T \mathbf{x}$
- Sean $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ las medias de los datos de cada clase y $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ sus covarianzas.
- Para una proyección proyección $y = \mathbf{w}^T \mathbf{x}$, el criterio de Fisher es:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \boldsymbol{\mu}_1 - \mathbf{w}^T \boldsymbol{\mu}_2}{\sigma_1^2 + \sigma_2^2}$$

donde σ_1^2, σ_2^2 son las varianzas de las proyecciones de cada clase.

Criterio de Fisher

- Proyección $y = \mathbf{w}^T \mathbf{x}$
- Sean $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ las medias de los datos de cada clase y $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ sus covarianzas.
- Para una proyección proyección $y = \mathbf{w}^T \mathbf{x}$, el criterio de Fisher es:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \boldsymbol{\mu}_1 - \mathbf{w}^T \boldsymbol{\mu}_2}{\sigma_1^2 + \sigma_2^2}$$

  donde σ_1^2, σ_2^2 son las varianzas de las proyecciones de cada clase.
- El vector que maximiza $J(\mathbf{w})$ satisface:

$$\mathbf{w} \propto (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

Cota de Chernoff

- Error de Bayes:

$$\epsilon = \int \text{mín}[P_1 p_1(\mathbf{x}), P_2 p_2(\mathbf{x})] d\mathbf{x}$$

Cota de Chernoff

- Error de Bayes:

$$\epsilon = \int \text{mín}[P_1 p_1(\mathbf{x}), P_2 p_2(\mathbf{x})] d\mathbf{x}$$

- Usando $\text{mín}[a, b] \leq a^s b^{1-s}$ para $0 \leq s \leq 1$:

Cota de Chernoff

- Error de Bayes:

$$\epsilon = \int \text{mín}[P_1 p_1(\mathbf{x}), P_2 p_2(\mathbf{x})] d\mathbf{x}$$

- Usando $\text{mín}[a, b] \leq a^s b^{1-s}$ para $0 \leq s \leq 1$:

$$\epsilon \leq P_1^s P_2^{1-s} \int p_1^s(\mathbf{x}) p_2^{1-s}(\mathbf{x}) d\mathbf{x}$$

Cota de Chernoff

- Error de Bayes:

$$\epsilon = \int \text{mín}[P_1 p_1(\mathbf{x}), P_2 p_2(\mathbf{x})] d\mathbf{x}$$

- Usando $\text{mín}[a, b] \leq a^s b^{1-s}$ para $0 \leq s \leq 1$:

$$\epsilon \leq P_1^s P_2^{1-s} \int p_1^s(\mathbf{x}) p_2^{1-s}(\mathbf{x}) d\mathbf{x}$$

- Para clases distribuídas normalmente:

$$\int p_1^s(\mathbf{x}) p_2^{1-s}(\mathbf{x}) d\mathbf{x} = e^{-\mu(s)}$$

donde

$$\mu(s) = \frac{s(1-s)}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left(s\boldsymbol{\Sigma}_1 + (1-s)\boldsymbol{\Sigma}_2\right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \ln \left(\frac{|s\boldsymbol{\Sigma}_1 + (1-s)\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|^s |\boldsymbol{\Sigma}_2|^{1-s}} \right)$$

Cota de Bhattacharyya

- Cota de Chernoff con $s = \frac{1}{2}$:

Cota de Bhattacharyya

- Cota de Chernoff con $s = \frac{1}{2}$:

$$\epsilon \leq \sqrt{P_1 P_2} e^{-\mu(1/2)}$$

Cota de Bhattacharyya

- Cota de Chernoff con $s = \frac{1}{2}$:

$$\epsilon \leq \sqrt{P_1 P_2} e^{-\mu(1/2)}$$

donde

$$\mu(1/2) = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left[\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right]^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \ln \frac{\left|\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2}}$$

Cota de Bhattacharyya

- Cota de Chernoff con $s = \frac{1}{2}$:

$$\epsilon \leq \sqrt{P_1 P_2} e^{-\mu(1/2)}$$

donde

$$\mu(1/2) = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left[\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right]^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \ln \frac{\left|\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2}}$$

- Equivale a cota de Chernoff óptima cuando $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$

Método de búsqueda

Método de búsqueda

- Exahustiva.

Método de búsqueda

- Exahustiva.
- Secuencial hacia adelante.

Método de búsqueda

- Exahustiva.
- Secuencial hacia adelante.
- Secuencial hacia atrás.

Método de búsqueda

- Exahustiva.
- Secuencial hacia adelante.
- Secuencial hacia atrás.
- Branch and Bound.

Método de búsqueda

- Exahustiva.
- Secuencial hacia adelante.
- Secuencial hacia atrás.
- Branch and Bound.
- Otros métodos heurísticos.

Branch and Bound

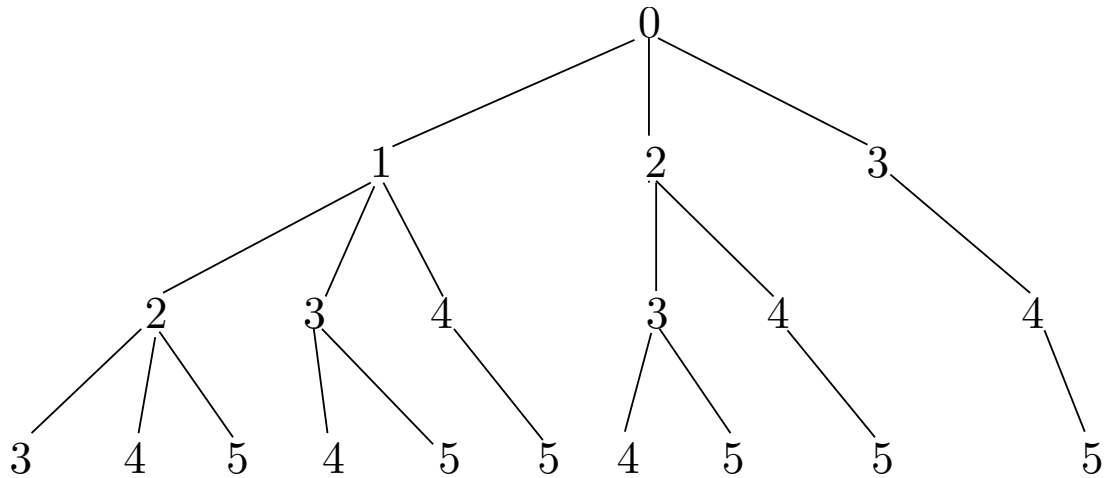
- Descartar m descriptores de n posibles.

Branch and Bound

- Descartar m descriptores de n posibles.
- Enumeración: $(x_1, \ldots, x_m)$. Basta considerar $x_1 < x_2 > \ldots x_m$

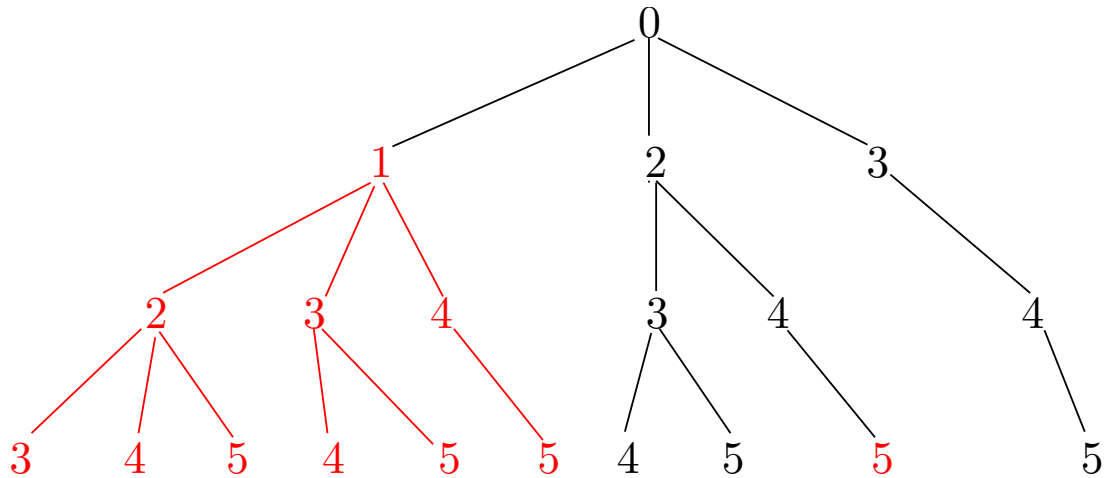
Branch and Bound

- Descartar m descriptores de n posibles.
- Enumeración: $(x_1, \ldots, x_m)$. Basta considerar $x_1 < x_2 > \ldots x_m$
- Supone criterio de evaluación monótono J: si A, B son conjuntos de descriptores $A \subseteq B \Rightarrow J(A) \leq J(B)$



Branch and Bound

- Descartar m descriptores de n posibles.
- Enumeración: $(x_1, \ldots, x_m)$. Basta considerar $x_1 < x_2 > \ldots x_m$
- Supone criterio de evaluación monótono J: si A, B son conjuntos de descriptores $A \subseteq B \Rightarrow J(A) \leq J(B)$



Análisis de componentes principales (PCA)

- Suponga que dos componentes de $\mathbf{x}$ están perfectamente correlacionados: $x_i = \alpha x_j$.

Análisis de componentes principales (PCA)

- Suponga que dos componentes de $\mathbf{x}$ están perfectamente correlacionados: $x_i = \alpha x_j$.
  - Información redundante.

Análisis de componentes principales (PCA)

- Suponga que dos componentes de $\mathbf{x}$ están perfectamente correlacionados: $x_i = \alpha x_j$.
  - Información redundante.
  - Podemos descartar una variable y reducir dimensión de los datos por 1.

Análisis de componentes principales (PCA)

- Suponga que dos componentes de $\mathbf{x}$ están perfectamente correlacionados: $x_i = \alpha x_j$.
  - Información redundante.
  - Podemos descartar una variable y reducir dimensión de los datos por 1.
- En general, correlación no es perfecta.

Análisis de componentes principales (PCA)

- Suponga que dos componentes de $\mathbf{x}$ están perfectamente correlacionados: $x_i = \alpha x_j$.
  - Información redundante.
  - Podemos descartar una variable y reducir dimensión de los datos por 1.
- En general, correlación no es perfecta.
- Podemos tener conjuntos de más de 2 variables correlacionadas entre sí.

Análisis de componentes principales (PCA)

- Suponga que dos componentes de $\mathbf{x}$ están perfectamente correlacionados: $x_i = \alpha x_j$.
  - Información redundante.
  - Podemos descartar una variable y reducir dimensión de los datos por 1.
- En general, correlación no es perfecta.
- Podemos tener conjuntos de más de 2 variables correlacionadas entre sí.
- Suponemos datos normalizados:

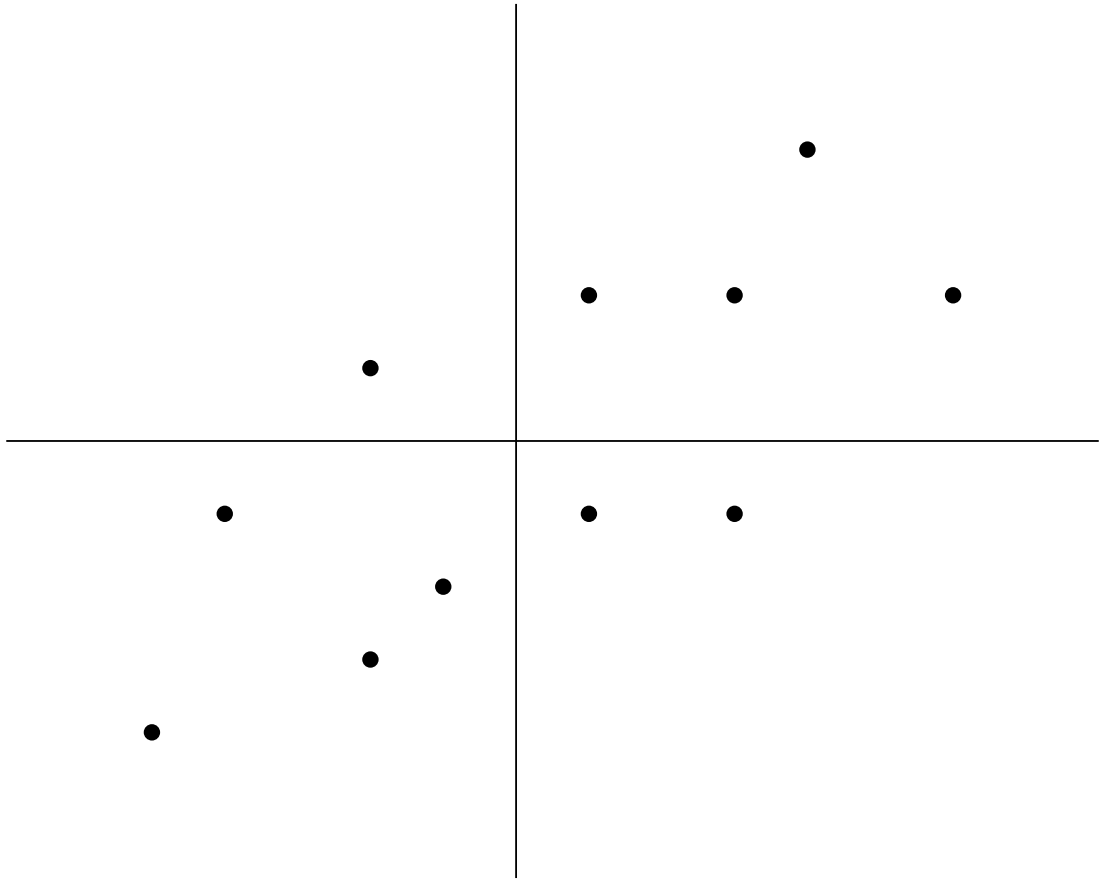
Análisis de componentes principales (PCA)

- Suponga que dos componentes de $\mathbf{x}$ están perfectamente correlacionados: $x_i = \alpha x_j$.
  - Información redundante.
  - Podemos descartar una variable y reducir dimensión de los datos por 1.
- En general, correlación no es perfecta.
- Podemos tener conjuntos de más de 2 variables correlacionadas entre sí.
- Suponemos datos normalizados:
  1. $\sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}$

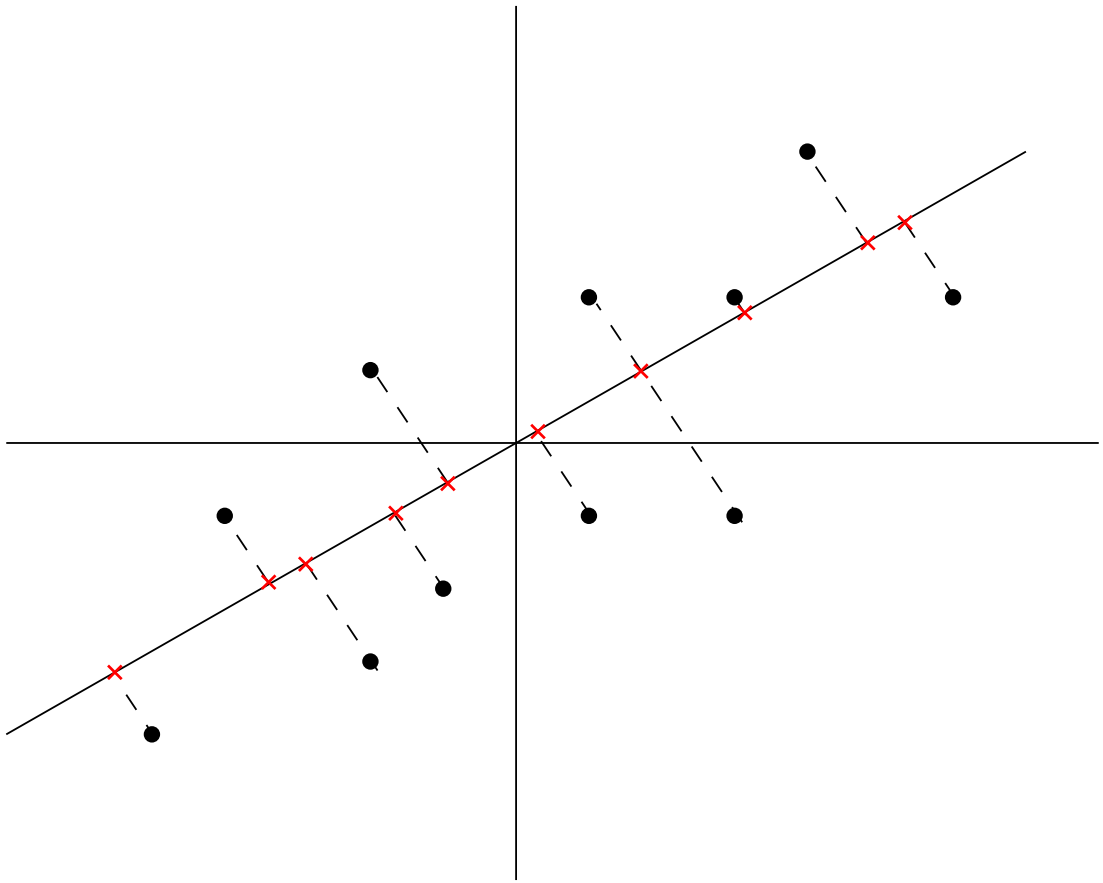
Análisis de componentes principales (PCA)

- Suponga que dos componentes de $\mathbf{x}$ están perfectamente correlacionados: $x_i = \alpha x_j$.
  - Información redundante.
  - Podemos descartar una variable y reducir dimensión de los datos por 1.
- En general, correlación no es perfecta.
- Podemos tener conjuntos de más de 2 variables correlacionadas entre sí.
- Suponemos datos normalizados:
  1. $\sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}$
  2. $\sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} x_j^{(i)} = 1, \quad j = 1, \ldots, d$

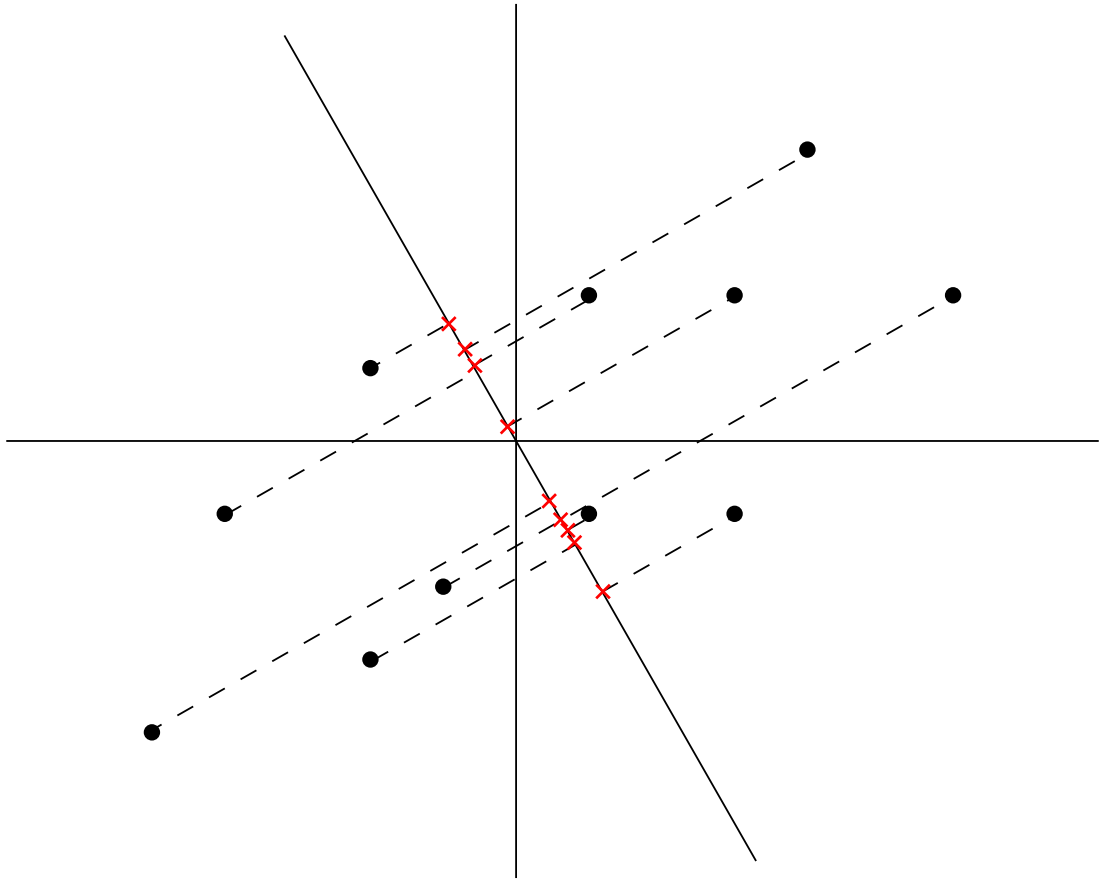
Ejemplo para $d = 2$



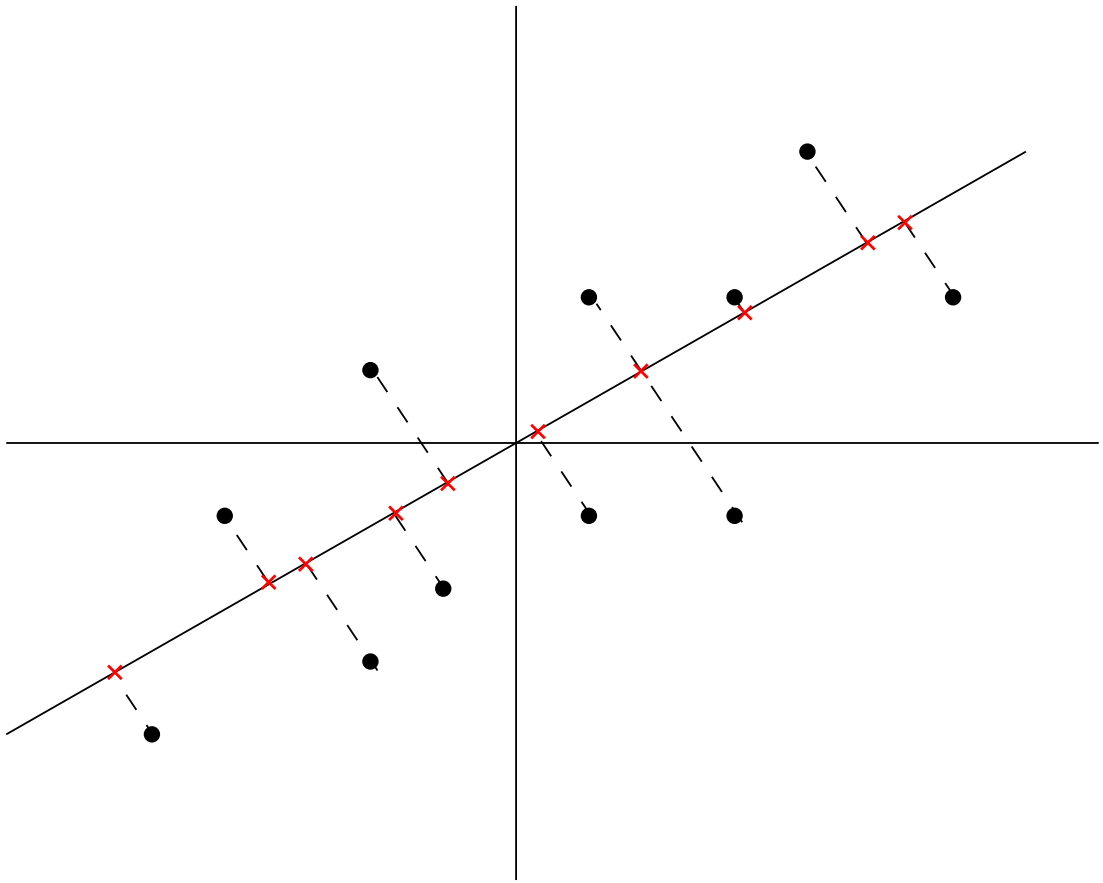
Ejemplo para $d = 2$



Ejemplo para $d = 2$



Ejemplo para $d = 2$



- Queremos proyección con máxima varianza.

- Proyección ortogonal sobre línea $\mathbf{u}^T\mathbf{x} = 0$, con $\|\mathbf{u}\| = 1$.

- Proyección ortogonal sobre línea $\mathbf{u}^T\mathbf{x} = 0$, con $\|\mathbf{u}\| = 1$.
- Queremos escoger $\mathbf{u}$ que maximice la varianza:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{u})$$

- Proyección ortogonal sobre línea $\mathbf{u}^T\mathbf{x} = 0$, con $\|\mathbf{u}\| = 1$.
- Queremos escoger $\mathbf{u}$ que maximice la varianza:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{u}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{u})\end{aligned}$$

- Proyección ortogonal sobre línea $\mathbf{u}^T\mathbf{x} = 0$, con $\|\mathbf{u}\| = 1$.
- Queremos escoger $\mathbf{u}$ que maximice la varianza:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{u}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{u}) \\ &= \frac{1}{n}\mathbf{u}^T\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{u}\end{aligned}$$

- Proyección ortogonal sobre línea $\mathbf{u}^T\mathbf{x} = 0$, con $\|\mathbf{u}\| = 1$.
- Queremos escoger $\mathbf{u}$ que maximice la varianza:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{u}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{u}) \\
&= \frac{1}{n}\mathbf{u}^T\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{u} \\
&= \frac{1}{n}\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}
\end{aligned}$$

- Proyección ortogonal sobre línea $\mathbf{u}^T\mathbf{x} = 0$, con $\|\mathbf{u}\| = 1$.
- Queremos escoger $\mathbf{u}$ que maximice la varianza:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{u}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{u}) \\ &= \frac{1}{n}\mathbf{u}^T\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{u} \\ &= \frac{1}{n}\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}\end{aligned}$$

donde $\boldsymbol{\Sigma} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$ es la matriz de covarianza de los datos $\{\mathbf{x}_i\}_{i=1}^{n}$.

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} \\ \text{sujeto a} \quad & \|\mathbf{u}\|^2 = 1 \end{aligned}$$

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & -\mathbf{u}^T \mathbf{\Sigma} \mathbf{u} \\ \text{sujeto a} \quad & \|\mathbf{u}\|^2 = 1 \end{aligned}$$

- El lagrangiano:

$$L(\mathbf{u}, \lambda) = -\mathbf{u}^T \mathbf{\Sigma} \mathbf{u} + \lambda(\mathbf{u}^T \mathbf{u} - 1)$$

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} \\ \text{sujeto a} \quad & \|\mathbf{u}\|^2 = 1 \end{aligned}$$

- El lagrangiano:

$$L(\mathbf{u}, \lambda) = -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} + \lambda(\mathbf{u}^T \mathbf{u} - 1)$$

- Derivando con respecto a $\mathbf{u}$:

$$\nabla_{\mathbf{u}} L(\mathbf{u}, \lambda) = -2\boldsymbol{\Sigma}\mathbf{u} + 2\lambda\mathbf{u} = \mathbf{0}$$

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} \\ \text{sujeto a} \quad & \|\mathbf{u}\|^2 = 1 \end{aligned}$$

- El lagrangiano:

$$L(\mathbf{u}, \lambda) = -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} + \lambda(\mathbf{u}^T \mathbf{u} - 1)$$

- Derivando con respecto a $\mathbf{u}$:

$$\nabla_{\mathbf{u}} L(\mathbf{u}, \lambda) = -2\boldsymbol{\Sigma}\mathbf{u} + 2\lambda\mathbf{u} = \mathbf{0} \Rightarrow \boldsymbol{\Sigma}\mathbf{u} = \lambda\mathbf{u}$$

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} \\ \text{sujeto a} \quad & \|\mathbf{u}\|^2 = 1 \end{aligned}$$

- El lagrangiano:

$$L(\mathbf{u}, \lambda) = -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} + \lambda(\mathbf{u}^T \mathbf{u} - 1)$$

- Derivando con respecto a $\mathbf{u}$:

$$\nabla_{\mathbf{u}} L(\mathbf{u}, \lambda) = -2\boldsymbol{\Sigma}\mathbf{u} + 2\lambda\mathbf{u} = \mathbf{0} \Rightarrow \boldsymbol{\Sigma}\mathbf{u} = \lambda\mathbf{u}$$

- Es decir, $\mathbf{u}$ es un vector propio de $\boldsymbol{\Sigma}$ con valor propio λ.

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} \\ \text{sujeto a} \quad & \|\mathbf{u}\|^2 = 1 \end{aligned}$$

- El lagrangiano:

$$L(\mathbf{u}, \lambda) = -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} + \lambda(\mathbf{u}^T \mathbf{u} - 1)$$

- Derivando con respecto a $\mathbf{u}$:

$$\nabla_{\mathbf{u}} L(\mathbf{u}, \lambda) = -2\boldsymbol{\Sigma}\mathbf{u} + 2\lambda\mathbf{u} = \mathbf{0} \Rightarrow \boldsymbol{\Sigma}\mathbf{u} = \lambda\mathbf{u}$$

- Es decir, $\mathbf{u}$ es un vector propio de $\boldsymbol{\Sigma}$ con valor propio λ.
- Más aún $\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} = \lambda \|\mathbf{u}\|^2 = \lambda$,

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} \\ \text{sujeto a} \quad & \|\mathbf{u}\|^2 = 1 \end{aligned}$$

- El lagrangiano:

$$L(\mathbf{u}, \lambda) = -\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} + \lambda(\mathbf{u}^T \mathbf{u} - 1)$$

- Derivando con respecto a $\mathbf{u}$:

$$\nabla_{\mathbf{u}} L(\mathbf{u}, \lambda) = -2\boldsymbol{\Sigma}\mathbf{u} + 2\lambda\mathbf{u} = \mathbf{0} \Rightarrow \boldsymbol{\Sigma}\mathbf{u} = \lambda\mathbf{u}$$

- Es decir, $\mathbf{u}$ es un vector propio de $\boldsymbol{\Sigma}$ con valor propio λ.
- Más aún $\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} = \lambda \|\mathbf{u}\|^2 = \lambda$, es decir, $\mathbf{u}$ es el vector propio de $\boldsymbol{\Sigma}$ correspondiente al máximo valor propio

- $\mathbf{u}_1 = \mathbf{u}$ es el primer componente principal.

- $\mathbf{u}_1 = \mathbf{u}$ es el primer componente principal.
- Suponga que queremos escoger un segundo componente principal $\mathbf{u}_2 \perp \mathbf{u}_1$ de manera que la varianza del residuo sea máxima:

- $\mathbf{u}_1 = \mathbf{u}$ es el primer componente principal.
- Suponga que queremos escoger un segundo componente principal $\mathbf{u}_2 \perp \mathbf{u}_1$ de manera que la varianza del residuo sea máxima:

$$\frac{1}{n} = \sum_{i=1}^{n} \left(\mathbf{u}_2^T(\mathbf{x}_i - (\mathbf{u}_1^T\mathbf{x}_i)\mathbf{u}_1)\right)^2$$

- $\mathbf{u}_1 = \mathbf{u}$ es el primer componente principal.
- Suponga que queremos escoger un segundo componente principal $\mathbf{u}_2 \perp \mathbf{u}_1$ de manera que la varianza del residuo sea máxima:

$$\frac{1}{n} = \sum_{i=1}^{n} \left(\mathbf{u}_2^T(\mathbf{x}_i - (\mathbf{u}_1^T\mathbf{x}_i)\mathbf{u}_1)\right)^2 = \sum_{i=1}^{n} \left(\mathbf{u}_2^T\mathbf{x}_i\right)^2$$

- $\mathbf{u}_1 = \mathbf{u}$ es el primer componente principal.
- Suponga que queremos escoger un segundo componente principal $\mathbf{u}_2 \perp \mathbf{u}_1$ de manera que la varianza del residuo sea máxima:

$$\frac{1}{n} = \sum_{i=1}^{n} \left(\mathbf{u}_2^T(\mathbf{x}_i - (\mathbf{u}_1^T\mathbf{x}_i)\mathbf{u}_1)\right)^2 = \sum_{i=1}^{n} \left(\mathbf{u}_2^T\mathbf{x}_i\right)^2$$

- El vector $\mathbf{u}_2$ con $\|\mathbf{u}_2\| = 1$ que maximiza la varianza del residuo es

- $\mathbf{u}_1 = \mathbf{u}$ es el primer componente principal.
- Suponga que queremos escoger un segundo componente principal $\mathbf{u}_2 \perp \mathbf{u}_1$ de manera que la varianza del residuo sea máxima:

$$\frac{1}{n} = \sum_{i=1}^{n} \left(\mathbf{u}_2^T(\mathbf{x}_i - (\mathbf{u}_1^T\mathbf{x}_i)\mathbf{u}_1)\right)^2 = \sum_{i=1}^{n} \left(\mathbf{u}_2^T\mathbf{x}_i\right)^2$$

- El vector $\mathbf{u}_2$ con $\|\mathbf{u}_2\| = 1$ que maximiza la varianza del residuo es el vector propio de $\mathbf{\Sigma}$ correspondiente al segundo valor propio más grande.

- $\mathbf{u}_1 = \mathbf{u}$ es el primer componente principal.
- Suponga que queremos escoger un segundo componente principal $\mathbf{u}_2 \perp \mathbf{u}_1$ de manera que la varianza del residuo sea máxima:

$$\frac{1}{n} = \sum_{i=1}^{n} \left(\mathbf{u}_2^T(\mathbf{x}_i - (\mathbf{u}_1^T\mathbf{x}_i)\mathbf{u}_1)\right)^2 = \sum_{i=1}^{n} \left(\mathbf{u}_2^T\mathbf{x}_i\right)^2$$

- El vector $\mathbf{u}_2$ con $\|\mathbf{u}_2\| = 1$ que maximiza la varianza del residuo es el vector propio de $\mathbf{\Sigma}$ correspondiente al segundo valor propio más grande.
- La varianza total en los componentes $\mathbf{u}_1, \mathbf{u}_2$ es

- $\mathbf{u}_1 = \mathbf{u}$ es el primer componente principal.
- Suponga que queremos escoger un segundo componente principal $\mathbf{u}_2 \perp \mathbf{u}_1$ de manera que la varianza del residuo sea máxima:

$$\frac{1}{n} = \sum_{i=1}^{n} \left(\mathbf{u}_2^T(\mathbf{x}_i - (\mathbf{u}_1^T\mathbf{x}_i)\mathbf{u}_1)\right)^2 = \sum_{i=1}^{n} \left(\mathbf{u}_2^T\mathbf{x}_i\right)^2$$

- El vector $\mathbf{u}_2$ con $\|\mathbf{u}_2\| = 1$ que maximiza la varianza del residuo es el vector propio de $\mathbf{\Sigma}$ correspondiente al segundo valor propio más grande.
- La varianza total en los componentes $\mathbf{u}_1, \mathbf{u}_2$ es $\lambda_1 + \lambda_2$

- En general, la proyección en un espacio de $M \leq d$ dimensiones en el que la varianza es máxima está dada por los M vectores propios $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M$ de $\mathbf{\Sigma}$ correspondientes a los valores propios $\lambda_1, \lambda_2, \ldots, \lambda_n$ más grandes.

- En general, la proyección en un espacio de $M \leq d$ dimensiones en el que la varianza es máxima está dada por los M vectores propios $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M$ de $\mathbf{\Sigma}$ correspondientes a los valores propios $\lambda_1, \lambda_2, \ldots, \lambda_n$ más grandes.
- La varianza total es $\lambda_1 + \lambda_2 + \cdots + \lambda_M$

- Note que los vectores propios $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_d$ son (o pueden escogerse) ortonormales.

- Note que los vectores propios $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_d$ son (o pueden escogerse) ortonormales.
- $y = \mathbf{U}^T\mathbf{x}$ es un cambio de base ortonormal (rotación).

- Note que los vectores propios $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_d$ son (o pueden escogerse) ortonormales.
- $y = \mathbf{U}^T \mathbf{x}$ es un cambio de base ortonormal (rotación).
- PCA escoge los componentes de la base para los cuales la varianza es máxima.

- Note que los vectores propios $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_d$ son (o pueden escogerse) ortonormales.
- $y = \mathbf{U}^T\mathbf{x}$ es un cambio de base ortonormal (rotación).
- PCA escoge los componentes de la base para los cuales la varianza es máxima.
- PCA puede derivarse como la proyección a un subconjunto de una base ortonormal óptima.

PCA en altas dimensiones

- En muchos casos $d \gg n$

PCA en altas dimensiones

- En muchos casos $d \gg n$
- $d - n + 1$ valores propios son cero.

PCA en altas dimensiones

- En muchos casos $d \gg n$
- $d - n + 1$ valores propios son cero.
- Cálculo de $\mathbf{u}_i, \lambda_i$ es $O(d^3)$.

PCA en altas dimensiones

- En muchos casos $d \gg n$
- $d - n + 1$ valores propios son cero.
- Cálculo de $\mathbf{u}_i, \lambda_i$ es $O(d^3)$.
- Sea $\mathbf{X}$ la matriz $n \times d$ donde cada fila es un dato:

PCA en altas dimensiones

- En muchos casos $d \gg n$
- $d - n + 1$ valores propios son cero.
- Cálculo de $\mathbf{u}_i, \lambda_i$ es $O(d^3)$.
- Sea $\mathbf{X}$ la matriz $n \times d$ donde cada fila es un dato:

$$\frac{1}{n}\mathbf{X}^T\mathbf{X}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

PCA en altas dimensiones

- En muchos casos $d \gg n$
- $d - n + 1$ valores propios son cero.
- Cálculo de $\mathbf{u}_i, \lambda_i$ es $O(d^3)$.
- Sea $\mathbf{X}$ la matriz $n \times d$ donde cada fila es un dato:

$$\frac{1}{n}\mathbf{X}^T\mathbf{X}\mathbf{u}_i = \lambda_i\mathbf{u}_i$$

$$\frac{1}{n}\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{u}_i) = \lambda_i(\mathbf{X}\mathbf{u}_i)$$

PCA en altas dimensiones

- En muchos casos $d \gg n$
- $d - n + 1$ valores propios son cero.
- Cálculo de $\mathbf{u}_i, \lambda_i$ es $O(d^3)$.
- Sea $\mathbf{X}$ la matriz $n \times d$ donde cada fila es un dato:

$$\frac{1}{n}\mathbf{X}^T\mathbf{X}\mathbf{u}_i = \lambda_i\mathbf{u}_i$$

$$\frac{1}{n}\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{u}_i) = \lambda_i(\mathbf{X}\mathbf{u}_i)$$

$$\frac{1}{n}\mathbf{X}\mathbf{v}_i = \lambda_i\mathbf{v}_i$$

PCA en altas dimensiones

- En muchos casos $d \gg n$
- $d - n + 1$ valores propios son cero.
- Cálculo de $\mathbf{u}_i, \lambda_i$ es $O(d^3)$.
- Sea $\mathbf{X}$ la matriz $n \times d$ donde cada fila es un dato:

$$\frac{1}{n}\mathbf{X}^T\mathbf{X}\mathbf{u}_i = \lambda_i\mathbf{u}_i$$

$$\frac{1}{n}\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{u}_i) = \lambda_i(\mathbf{X}\mathbf{u}_i)$$

$$\frac{1}{n}\mathbf{X}\mathbf{v}_i = \lambda_i\mathbf{v}_i$$

- Cálculo en n dimensiones.