Clase 15 - Aprendizaje no supervisado: reducción de dimensionalidad

Aprendizaje de Máquinas - MA5204

Felipe Tobar

Department of Mathematical Engineering &
Center for Mathematical Modelling
Universidad de Chile

14 de marzo de 2021

Reducción de dimensionalidad

El problema de reducción de dimensionalidad consiste con construir una representación de dimensión estrictamente menor que los datos originales con la finalidad de interpretar de mejor forma la información contenida en nuestros datos así como también disminuir el costo computacional en el entrenamiento.

En este curso, se trabajarán dos técnicas de reducción de dimensionalidad:

- Análisis de componentes principales (PCA).
- Discriminante lineal de Fisher.

Reducción de dimensionalidad

El problema de reducción de dimensionalidad consiste con construir una representación de dimensión estrictamente menor que los datos originales con la finalidad de interpretar de mejor forma la información contenida en nuestros datos así como también disminuir el costo computacional en el entrenamiento.

En este curso, se trabajarán dos técnicas de reducción de dimensionalidad:

- Análisis de componentes principales (PCA).
- Discriminante lineal de Fisher.

Análisis de componentes principales (PCA): idea general

Consideremos un conjunto de observaciones de $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^M$, se denotará por x_{ij} al valor del atributo j para la observación i.

Cada observación puede descomponerse en la base canónica $\{\mathbf{e}_i\}_{i=1}^M$ de $\mathbb{R}^M$ de la forma

$$\mathbf{x}_i = x_{i1}\mathbf{e}_1 + x_{i2}\mathbf{e}_2 + \cdots + x_{iM}\mathbf{e}_M$$

Notemos que es posible aproximar cada observación $\mathbf{x}_i$ mediante una cantidad $M' < M$ de términos, truncando la representación anterior, es decir,

$$\mathbf{x}_i \approx \sum_{j=1}^{M'} x_{i\sigma(j)}\mathbf{e}_{\sigma(j)}$$

donde $\sigma : \{1, 2, \ldots, M\} \mapsto \{1, 2, \ldots, M\}$ es una permutación que prioriza las coordenadas más representativas de los datos. Dichas aproximaciones son una versión de baja dimensión de las observaciones $\{\mathbf{x}_i\}_{i=1}^N$.

Análisis de componentes principales (PCA): idea general

Consideremos un conjunto de observaciones de $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^M$, se denotará por x_{ij} al valor del atributo j para la observación i.

Cada observación puede descomponerse en la base canónica $\{\mathbf{e}_i\}_{i=1}^M$ de $\mathbb{R}^M$ de la forma

$$\mathbf{x}_i = x_{i1}\mathbf{e}_1 + x_{i2}\mathbf{e}_2 + \cdots + x_{iM}\mathbf{e}_M$$

Notemos que es posible aproximar cada observación $\mathbf{x}_i$ mediante una cantidad $M' < M$ de términos, truncando la representación anterior, es decir,

$$\mathbf{x}_i \approx \sum_{j=1}^{M'} x_{i\sigma(j)}\mathbf{e}_{\sigma(j)}$$

donde $\sigma : \{1, 2, \ldots, M\} \mapsto \{1, 2, \ldots, M\}$ es una permutación que prioriza las coordenadas más representativas de los datos. Dichas aproximaciones son una versión de baja dimensión de las observaciones $\{\mathbf{x}_i\}_{i=1}^N$.

Análisis de componentes principales (PCA): idea general

Consideremos un conjunto de observaciones de $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^M$, se denotará por x_{ij} al valor del atributo j para la observación i.

Cada observación puede descomponerse en la base canónica $\{\mathbf{e}_i\}_{i=1}^M$ de $\mathbb{R}^M$ de la forma

$$\mathbf{x}_i = x_{i1}\mathbf{e}_1 + x_{i2}\mathbf{e}_2 + \cdots + x_{iM}\mathbf{e}_M$$

Notemos que es posible aproximar cada observación $\mathbf{x}_i$ mediante una cantidad $M' < M$ de términos, truncando la representación anterior, es decir,

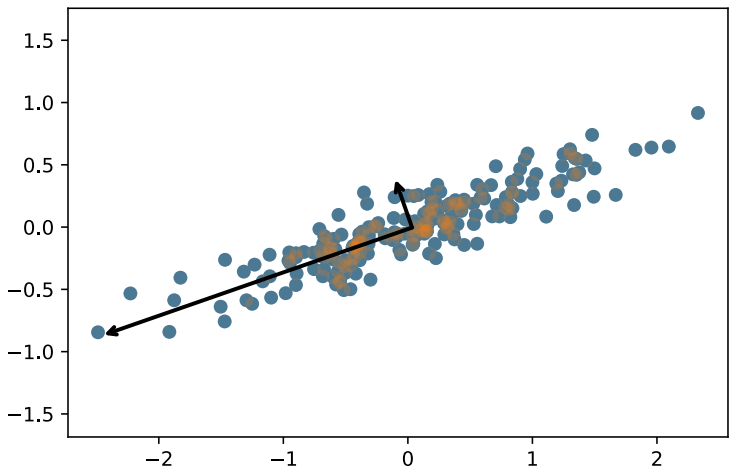
$$\mathbf{x}_i \approx \sum_{j=1}^{M'} x_{i\sigma(j)}\mathbf{e}_{\sigma(j)}$$

donde $\sigma : \{1, 2, \ldots, M\} \mapsto \{1, 2, \ldots, M\}$ es una permutación que prioriza las coordenadas más representativas de los datos. Dichas aproximaciones son una versión de baja dimensión de las observaciones $\{\mathbf{x}_i\}_{i=1}^N$.

Análisis de componentes principales (PCA): criterio

Dada una dimensión $M' < M$:

- ¿es efectivamente un subconjunto de los vectores canónicos la mejor base para descomponer las observaciones?



- ¿cómo encontramos la *mejor* base?

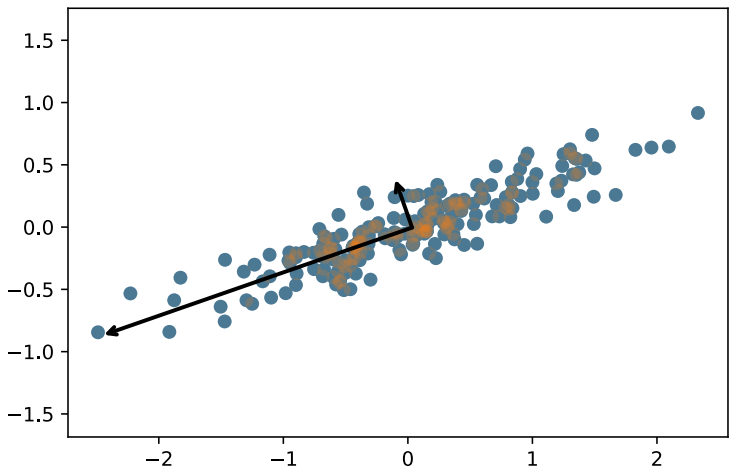
  Lo primero que se requiere es definir qué se entiende por *mejor*. Nos enfocaremos en determinar una base cuyos componentes **ordenados** $\mathbf{c}_1, \mathbf{c}_2, \ldots$ capturan las M' direcciones ortogonales de máxima variabilidad de nuestros datos.

Este criterio es conocido como **análisis de componentes principales (PCA)**.

Análisis de componentes principales (PCA): criterio

Dada una dimensión $M' < M$:

- ¿es efectivamente un subconjunto de los vectores canónicos la mejor base para descomponer las observaciones?



- ¿cómo encontramos la *mejor* base?

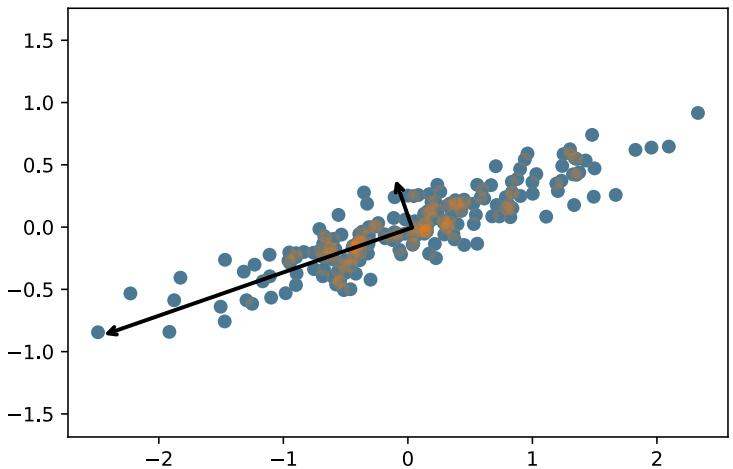
  Lo primero que se requiere es definir qué se entiende por *mejor*. Nos enfocaremos en determinar una base cuyos componentes **ordenados** $\mathbf{c}_1, \mathbf{c}_2, \ldots$ capturan las M' direcciones ortogonales de máxima variabilidad de nuestros datos.

Este criterio es conocido como **análisis de componentes principales (PCA)**.

Análisis de componentes principales (PCA): criterio

Dada una dimensión $M' < M$:

- ¿es efectivamente un subconjunto de los vectores canónicos la mejor base para descomponer las observaciones?



- ¿cómo encontramos la *mejor* base?

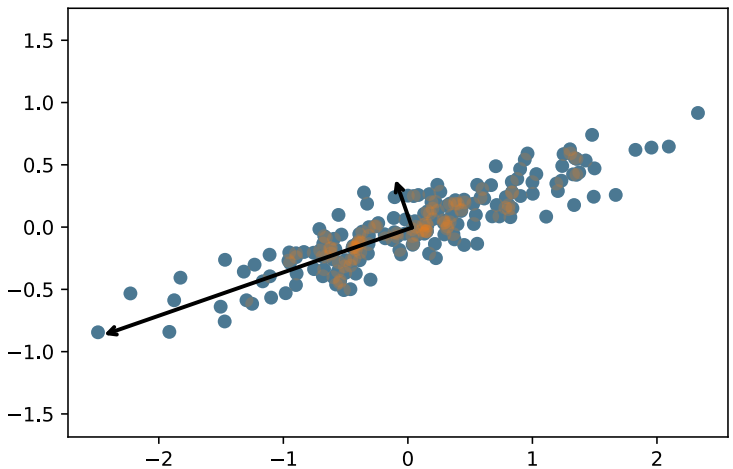
Lo primero que se requiere es definir qué se entiende por *mejor*. Nos enfocaremos en determinar una base cuyos componentes **ordenados** $\mathbf{c}_1, \mathbf{c}_2, \ldots$ capturan las M' direcciones ortogonales de máxima variabilidad de nuestros datos.

Este criterio es conocido como **análisis de componentes principales (PCA)**.

Análisis de componentes principales (PCA): criterio

Dada una dimensión $M' < M$:

- ¿es efectivamente un subconjunto de los vectores canónicos la mejor base para descomponer las observaciones?



- ¿cómo encontramos la *mejor* base?

  Lo primero que se requiere es definir qué se entiende por *mejor*. Nos enfocaremos en determinar una base cuyos componentes **ordenados** $\mathbf{c}_1, \mathbf{c}_2, \ldots$ capturan las M' direcciones ortogonales de máxima variabilidad de nuestros datos.

Este criterio es conocido como **análisis de componentes principales (PCA)**.

Análisis de componentes principales (PCA): criterio

De esta forma, dado que $\langle \mathbf{c}, \mathbf{x} \rangle$ representa la proyección ortogonal de $\mathbf{x}$ sobre $\mathbf{c}$, el primer elemento de la nueva base estará dado por

$$\mathbf{c}_1 = \underset{||\mathbf{c}||=1}{\arg\max} \langle \mathbf{c}, \mathbf{x} \rangle$$

- la restricción $||\mathbf{c}_1|| = 1$ es necesaria ya que $\langle \lambda \mathbf{c}_1, \mathbf{x} \rangle = \lambda \langle \mathbf{c}_1, \mathbf{x} \rangle$ por lo que $\langle \mathbf{c}_1, \mathbf{x} \rangle$ puede crecer indefinidamente si no se fija una restricción sobre la norma de $\mathbf{c}$.
- Además, es importante estandarizar los datos:

1. **Características de media nula:** la matriz X debe tener columnas con media 0. El objetivo de este ajuste es poder centrar los datos.

2. **Varianzas marginales unitarias:** si una dimensión tiene una varianza marginal mayor que el resto, esta será más importante en la determinación de la dirección de máxima varianza solo por su magnitud y no por la relación entre variables.

Análisis de componentes principales (PCA): criterio

De esta forma, dado que $\langle \mathbf{c}, \mathbf{x} \rangle$ representa la proyección ortogonal de $\mathbf{x}$ sobre $\mathbf{c}$, el primer elemento de la nueva base estará dado por

$$\mathbf{c}_1 = \underset{||\mathbf{c}||=1}{\arg\max} \langle \mathbf{c}, \mathbf{x} \rangle$$

- la restricción $||\mathbf{c}_1|| = 1$ es necesaria ya que $\langle \lambda \mathbf{c}_1, \mathbf{x} \rangle = \lambda \langle \mathbf{c}_1, \mathbf{x} \rangle$ por lo que $\langle \mathbf{c}_1, \mathbf{x} \rangle$ puede crecer indefinidamente si no se fija una restricción sobre la norma de $\mathbf{c}$.
- Además, es importante estandarizar los datos:

1. **Características de media nula:** la matriz X debe tener columnas con media 0. El objetivo de este ajuste es poder centrar los datos.

2. **Varianzas marginales unitarias:** si una dimensión tiene una varianza marginal mayor que el resto, esta será más importante en la determinación de la dirección de máxima varianza solo por su magnitud y no por la relación entre variables.

Análisis de componentes principales (PCA): criterio

De esta forma, dado que $\langle \mathbf{c}, \mathbf{x} \rangle$ representa la proyección ortogonal de $\mathbf{x}$ sobre $\mathbf{c}$, el primer elemento de la nueva base estará dado por

$$\mathbf{c}_1 = \underset{||\mathbf{c}||=1}{\arg\max} \langle \mathbf{c}, \mathbf{x} \rangle$$

- la restricción $||\mathbf{c}_1|| = 1$ es necesaria ya que $\langle \lambda \mathbf{c}_1, \mathbf{x} \rangle = \lambda \langle \mathbf{c}_1, \mathbf{x} \rangle$ por lo que $\langle \mathbf{c}_1, \mathbf{x} \rangle$ puede crecer indefinidamente si no se fija una restricción sobre la norma de $\mathbf{c}$.
- Además, es importante estandarizar los datos:

1. **Características de media nula:** la matriz X debe tener columnas con media 0. El objetivo de este ajuste es poder centrar los datos.

2. **Varianzas marginales unitarias:** si una dimensión tiene una varianza marginal mayor que el resto, esta será más importante en la determinación de la dirección de máxima varianza solo por su magnitud y no por la relación entre variables.

Análisis de componentes principales (PCA): criterio

De esta forma, dado que $\langle \mathbf{c}, \mathbf{x} \rangle$ representa la proyección ortogonal de $\mathbf{x}$ sobre $\mathbf{c}$, el primer elemento de la nueva base estará dado por

$$\mathbf{c}_1 = \underset{||\mathbf{c}||=1}{\arg\max} \langle \mathbf{c}, \mathbf{x} \rangle$$

- la restricción $||\mathbf{c}_1|| = 1$ es necesaria ya que $\langle \lambda \mathbf{c}_1, \mathbf{x} \rangle = \lambda \langle \mathbf{c}_1, \mathbf{x} \rangle$ por lo que $\langle \mathbf{c}_1, \mathbf{x} \rangle$ puede crecer indefinidamente si no se fija una restricción sobre la norma de $\mathbf{c}$.
- Además, es importante estandarizar los datos:

1. **Características de media nula:** la matriz X debe tener columnas con media 0. El objetivo de este ajuste es poder centrar los datos.

2. **Varianzas marginales unitarias:** si una dimensión tiene una varianza marginal mayor que el resto, esta será más importante en la determinación de la dirección de máxima varianza solo por su magnitud y no por la relación entre variables.

Análisis de componentes principales (PCA): criterio

De esta forma, dado que $\langle \mathbf{c}, \mathbf{x} \rangle$ representa la proyección ortogonal de $\mathbf{x}$ sobre $\mathbf{c}$, el primer elemento de la nueva base estará dado por

$$\mathbf{c}_1 = \underset{||\mathbf{c}||=1}{\arg\max} \langle \mathbf{c}, \mathbf{x} \rangle$$

- la restricción $||\mathbf{c}_1|| = 1$ es necesaria ya que $\langle \lambda \mathbf{c}_1, \mathbf{x} \rangle = \lambda \langle \mathbf{c}_1, \mathbf{x} \rangle$ por lo que $\langle \mathbf{c}_1, \mathbf{x} \rangle$ puede crecer indefinidamente si no se fija una restricción sobre la norma de $\mathbf{c}$.
- Además, es importante estandarizar los datos:

1. **Características de media nula:** la matriz X debe tener columnas con media 0. El objetivo de este ajuste es poder centrar los datos.

2. **Varianzas marginales unitarias:** si una dimensión tiene una varianza marginal mayor que el resto, esta será más importante en la determinación de la dirección de máxima varianza solo por su magnitud y no por la relación entre variables.

Análisis de componentes principales (PCA): formulación

En general no contamos con la distribución de las observaciones $p(\mathbf{x})$, por lo que se puede considerar una aproximación muestral de la varianza y resolver

$$\mathbf{c}_1 = \arg\max_{||\mathbf{c}||=1} \sum_{i=1}^{N} \langle \mathbf{c}, \mathbf{x}_i \rangle^2.$$

Usando la siguiente notación

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NM} \end{bmatrix}$$

Se puede reescribir el problema como

$$\mathbf{c}_1 = \arg\max_{||\mathbf{c}||=1} ||X\mathbf{c}||^2 = \arg\max_{||\mathbf{c}||=1} \mathbf{c}^\top X^\top X \mathbf{c} = \arg\max_{\mathbf{c}} \frac{\mathbf{c}^\top X^\top X \mathbf{c}}{\mathbf{c}^\top \mathbf{c}}$$

Análisis de componentes principales (PCA): formulación

En general no contamos con la distribución de las observaciones $p(\mathbf{x})$, por lo que se puede considerar una aproximación muestral de la varianza y resolver

$$\mathbf{c}_1 = \arg\max_{||\mathbf{c}||=1} \sum_{i=1}^{N} \langle \mathbf{c}, \mathbf{x}_i \rangle^2.$$

Usando la siguiente notación

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NM} \end{bmatrix}$$

Se puede reescribir el problema como

$$\mathbf{c}_1 = \arg\max_{||\mathbf{c}||=1} ||X\mathbf{c}||^2 = \arg\max_{||\mathbf{c}||=1} \mathbf{c}^\top X^\top X \mathbf{c} = \arg\max_{\mathbf{c}} \frac{\mathbf{c}^\top X^\top X \mathbf{c}}{\mathbf{c}^\top \mathbf{c}}$$

Análisis de componentes principales (PCA): formulación

En general no contamos con la distribución de las observaciones $p(\mathbf{x})$, por lo que se puede considerar una aproximación muestral de la varianza y resolver

$$\mathbf{c}_1 = \arg\max_{||\mathbf{c}||=1} \sum_{i=1}^{N} \langle \mathbf{c}, \mathbf{x}_i \rangle^2.$$

Usando la siguiente notación

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NM} \end{bmatrix}$$

Se puede reescribir el problema como

$$\mathbf{c}_1 = \arg\max_{||\mathbf{c}||=1} ||X\mathbf{c}||^2 = \arg\max_{||\mathbf{c}||=1} \mathbf{c}^\top X^\top X \mathbf{c} = \arg\max_{\mathbf{c}} \frac{\mathbf{c}^\top X^\top X \mathbf{c}}{\mathbf{c}^\top \mathbf{c}}$$

Análisis de componentes principales (PCA): cociente de Rayleigh

Para el problema anterior, se tiene la siguiente propiedad:

Lemma (minimización del cociente de Rayleigh)

Sea $M \in \mathcal{M}_{nn}(\mathbb{R})$ matriz cuadrada simétrica, entonces, para el cociente de Rayleigh

$$R(M, x) := \frac{x^\top M x}{x^\top x}$$

Su valor mínimo corresponde al menor valor propio de M, y es alcanzado en su vector propio asociado.

De esta forma, dado que $X^\top X$ es simétrica, su cociente de Rayleigh es maximizado en el vector propio asociado al valor propio máximo de $X^\top X$.

Consecuentemente, la proyección de una observación $\mathbf{x}_i$ en la dirección de máxima varianza, o bien la *primera componente principal*, está dada por

$$\mathbf{x}_i^{(1)} = \langle \mathbf{x}_i, \mathbf{c}_1 \rangle$$

donde $\mathbf{c}_1$ es el vector propio asociado al mayor valor propio de la matriz de covarianza muestral $XX^\top$.

Análisis de componentes principales (PCA): cociente de Rayleigh

Para el problema anterior, se tiene la siguiente propiedad:

Lemma (minimización del cociente de Rayleigh)

Sea $M \in \mathcal{M}_{nn}(\mathbb{R})$ matriz cuadrada simétrica, entonces, para el cociente de Rayleigh

$$R(M, x) := \frac{x^\top M x}{x^\top x}$$

Su valor mínimo corresponde al menor valor propio de M, y es alcanzado en su vector propio asociado.

De esta forma, dado que $X^\top X$ es simétrica, su cociente de Rayleigh es maximizado en el vector propio asociado al valor propio máximo de $X^\top X$.

Consecuentemente, la proyección de una observación $\mathbf{x}_i$ en la dirección de máxima varianza, o bien la *primera componente principal*, está dada por

$$\mathbf{x}_i^{(1)} = \langle \mathbf{x}_i, \mathbf{c}_1 \rangle$$

donde $\mathbf{c}_1$ es el vector propio asociado al mayor valor propio de la matriz de covarianza muestral $XX^\top$.

Análisis de componentes principales (PCA): cociente de Rayleigh

Para el problema anterior, se tiene la siguiente propiedad:

Lemma (minimización del cociente de Rayleigh)

Sea $M \in \mathcal{M}_{nn}(\mathbb{R})$ matriz cuadrada simétrica, entonces, para el cociente de Rayleigh

$$R(M, x) := \frac{x^\top M x}{x^\top x}$$

Su valor mínimo corresponde al menor valor propio de M, y es alcanzado en su vector propio asociado.

De esta forma, dado que $X^\top X$ es simétrica, su cociente de Rayleigh es maximizado en el vector propio asociado al valor propio máximo de $X^\top X$.

Consecuentemente, la proyección de una observación $\mathbf{x}_i$ en la dirección de máxima varianza, o bien la *primera componente principal*, está dada por

$$\mathbf{x}_i^{(1)} = \langle \mathbf{x}_i, \mathbf{c}_1 \rangle$$

donde $\mathbf{c}_1$ es el vector propio asociado al mayor valor propio de la matriz de covarianza muestral $XX^\top$.

Kernel PCA

Es posible utilizar el truco del kernel en PCA. En ese sentido, en vez de calcular la matriz de covarianza empírica $X^\top X$, se utiliza la matriz de Gram dada por un kernel K donde

$$K_{ij} = K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

Luego, se realiza PCA utilizando dicha matriz.

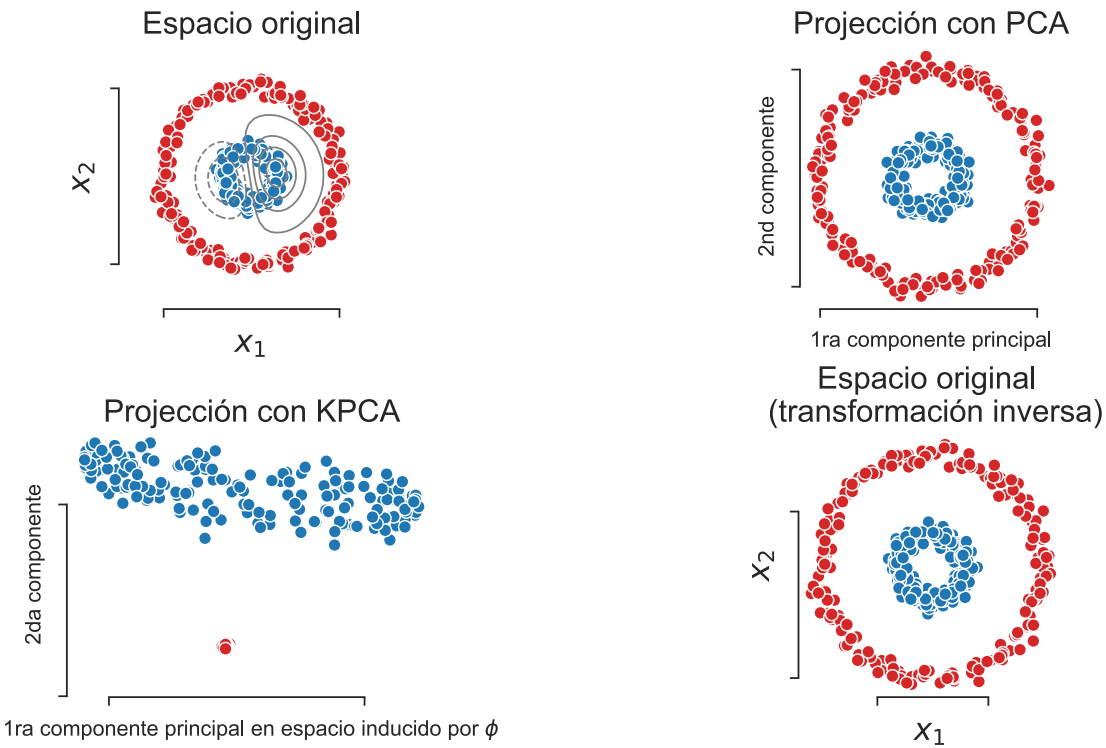


Fig.. Ejemplo de KPCA sobre un conjunto de datos que no es linealmente separable.

Kernel PCA

Es posible utilizar el truco del kernel en PCA. En ese sentido, en vez de calcular la matriz de covarianza empírica $X^\top X$, se utiliza la matriz de Gram dada por un kernel K donde

$$K_{ij} = K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

Luego, se realiza PCA utilizando dicha matriz.

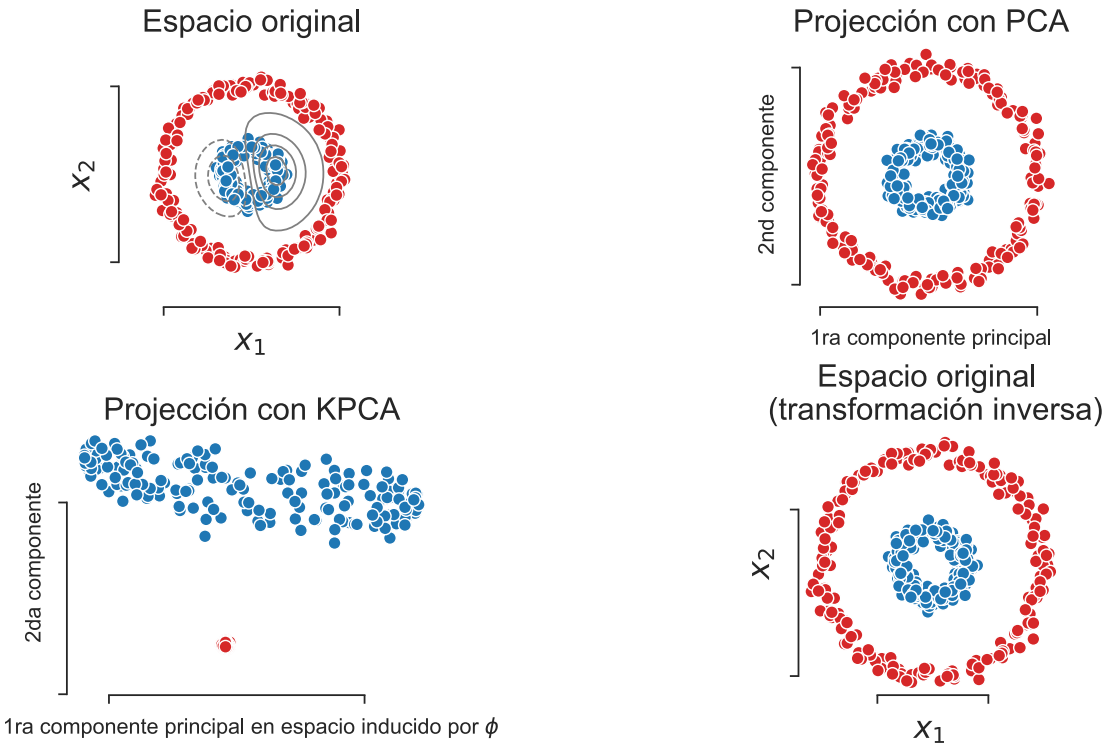


Fig.. Ejemplo de KPCA sobre un conjunto de datos que no es linealmente separable.

Discriminante lineal de Fisher: idea general

Es posible interpretar el problema de clasificación como uno de *reducción de dimensionalidad*:

- La reducción consiste representar nuestros datos en solo una dimensión, la cual representa su grado de pertenencia a una clase.
- Al proyectar un objeto M-dimensional en un espacio 1-dimensional, se pierde gran parte de la información, por lo que clases claramente separadas en el espacio M-dimensional puedan traslaparse al ser proyectadas a 1 dimensión.
- Sin embargo, es posible ajustar el vector a con la finalidad de obtener una proyección de x que maximice el grado de separación entre clases.
- Para el problema de clasificación binaria de $x \in \mathbb{R}^M$, se proyecta x en un espacio **unidimensional** con respecto a un vector $a \in \mathbb{R}^M$ de acuerdo a:

$$y = a^\top x$$

donde se puede definir un umbral b para asignar x a $\mathcal{C}_1$ si $y + b \geq 0$ y x a $\mathcal{C}_2$ en caso contrario. Esto recupera el modelo lineal para clasificación.

Discriminante lineal de Fisher: idea general

Es posible interpretar el problema de clasificación como uno de *reducción de dimensionalidad*:

- La reducción consiste representar nuestros datos en solo una dimensión, la cual representa su grado de pertenencia a una clase.
- Al proyectar un objeto M-dimensional en un espacio 1-dimensional, se pierde gran parte de la información, por lo que clases claramente separadas en el espacio M-dimensional puedan traslaparse al ser proyectadas a 1 dimensión.
- Sin embargo, es posible ajustar el vector a con la finalidad de obtener una proyección de x que maximice el grado de separación entre clases.
- Para el problema de clasificación binaria de $x \in \mathbb{R}^M$, se proyecta x en un espacio **unidimensional** con respecto a un vector $a \in \mathbb{R}^M$ de acuerdo a:

$$y = a^\top x$$

donde se puede definir un umbral b para asignar x a $\mathcal{C}_1$ si $y + b \geq 0$ y x a $\mathcal{C}_2$ en caso contrario. Esto recupera el modelo lineal para clasificación.

Discriminante lineal de Fisher: idea general

Es posible interpretar el problema de clasificación como uno de *reducción de dimensionalidad*:

- La reducción consiste representar nuestros datos en solo una dimensión, la cual representa su grado de pertenencia a una clase.
- Al proyectar un objeto M-dimensional en un espacio 1-dimensional, se pierde gran parte de la información, por lo que clases claramente separadas en el espacio M-dimensional puedan traslaparse al ser proyectadas a 1 dimensión.
- Sin embargo, es posible ajustar el vector a con la finalidad de obtener una proyección de x que maximice el grado de separación entre clases.
- Para el problema de clasificación binaria de $x \in \mathbb{R}^M$, se proyecta x en un espacio **unidimensional** con respecto a un vector $a \in \mathbb{R}^M$ de acuerdo a:

$$y = a^\top x$$

donde se puede definir un umbral b para asignar x a $\mathcal{C}_1$ si $y + b \geq 0$ y x a $\mathcal{C}_2$ en caso contrario. Esto recupera el modelo lineal para clasificación.

Discriminante lineal de Fisher: idea general

Es posible interpretar el problema de clasificación como uno de *reducción de dimensionalidad*:

- La reducción consiste representar nuestros datos en solo una dimensión, la cual representa su grado de pertenencia a una clase.
- Al proyectar un objeto M-dimensional en un espacio 1-dimensional, se pierde gran parte de la información, por lo que clases claramente separadas en el espacio M-dimensional puedan traslaparse al ser proyectadas a 1 dimensión.
- Sin embargo, es posible ajustar el vector a con la finalidad de obtener una proyección de x que maximice el grado de separación entre clases.
- Para el problema de clasificación binaria de $x \in \mathbb{R}^M$, se proyecta x en un espacio **unidimensional** con respecto a un vector $a \in \mathbb{R}^M$ de acuerdo a:

$$y = a^\top x$$

donde se puede definir un umbral b para asignar x a $\mathcal{C}_1$ si $y + b \geq 0$ y x a $\mathcal{C}_2$ en caso contrario. Esto recupera el modelo lineal para clasificación.

Discriminante lineal de Fisher: idea general

Es posible interpretar el problema de clasificación como uno de *reducción de dimensionalidad*:

- La reducción consiste representar nuestros datos en solo una dimensión, la cual representa su grado de pertenencia a una clase.
- Al proyectar un objeto M-dimensional en un espacio 1-dimensional, se pierde gran parte de la información, por lo que clases claramente separadas en el espacio M-dimensional puedan traslaparse al ser proyectadas a 1 dimensión.
- Sin embargo, es posible ajustar el vector a con la finalidad de obtener una proyección de x que maximice el grado de separación entre clases.
- Para el problema de clasificación binaria de $x \in \mathbb{R}^M$, se proyecta x en un espacio **unidimensional** con respecto a un vector $a \in \mathbb{R}^M$ de acuerdo a:

$$y = a^\top x$$

donde se puede definir un umbral b para asignar x a $\mathcal{C}_1$ si $y + b \geq 0$ y x a $\mathcal{C}_2$ en caso contrario. Esto recupera el modelo lineal para clasificación.

Discriminante lineal de Fisher: primera formulación

Se buscará un vector a con la finalidad de obtener una proyección de x que maximice el grado de separación entre clases. Se propone el siguiente esquema:

- Cardinales de clase: $N_1 = |\{x \in \mathcal{D} : x \in \mathcal{C}_1\}|$ y $N_2 = |\{x \in \mathcal{D} : x \in \mathcal{C}_2\}|$.
- Estos permiten calcular los promedios muestrales (centros de masa) de cada clase:

$$\mu_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} x_n \qquad \mu_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} x_n$$

- De esta forma, la medida más simple de separación entre las proyecciones de las clases sobre a es la distancia entre las medias de sus proyecciones:

$$m_1 - m_2 = a^\top (\mu_1 - \mu_2)$$

donde $m_k = a^\top \mu_k$ corresponde al promedio de los elementos de la clase $\mathcal{C}_k$ proyectado sobre el vector a (centro de masa sobre la recta).

Discriminante lineal de Fisher: primera formulación

Se buscará un vector a con la finalidad de obtener una proyección de x que maximice el grado de separación entre clases. Se propone el siguiente esquema:

- Cardinales de clase: $N_1 = |\{x \in \mathcal{D} : x \in \mathcal{C}_1\}|$ y $N_2 = |\{x \in \mathcal{D} : x \in \mathcal{C}_2\}|$.
- Estos permiten calcular los promedios muestrales (centros de masa) de cada clase:

$$\mu_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} x_n \qquad \mu_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} x_n$$

- De esta forma, la medida más simple de separación entre las proyecciones de las clases sobre a es la distancia entre las medias de sus proyecciones:

$$m_1 - m_2 = a^\top (\mu_1 - \mu_2)$$

donde $m_k = a^\top \mu_k$ corresponde al promedio de los elementos de la clase $\mathcal{C}_k$ proyectado sobre el vector a (centro de masa sobre la recta).

Discriminante lineal de Fisher: primera formulación

Se buscará un vector a con la finalidad de obtener una proyección de x que maximice el grado de separación entre clases. Se propone el siguiente esquema:

- Cardinales de clase: $N_1 = |\{x \in \mathcal{D} : x \in \mathcal{C}_1\}|$ y $N_2 = |\{x \in \mathcal{D} : x \in \mathcal{C}_2\}|$.
- Estos permiten calcular los promedios muestrales (centros de masa) de cada clase:

$$\mu_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} x_n \qquad \mu_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} x_n$$

- De esta forma, la medida más simple de separación entre las proyecciones de las clases sobre a es la distancia entre las medias de sus proyecciones:

$$m_1 - m_2 = a^\top (\mu_1 - \mu_2)$$

donde $m_k = a^\top \mu_k$ corresponde al promedio de los elementos de la clase $\mathcal{C}_k$ proyectado sobre el vector a (centro de masa sobre la recta).

Discriminante lineal de Fisher: primera formulación

Se buscará un vector a con la finalidad de obtener una proyección de x que maximice el grado de separación entre clases. Se propone el siguiente esquema:

- Cardinales de clase: $N_1 = |\{x \in \mathcal{D} : x \in \mathcal{C}_1\}|$ y $N_2 = |\{x \in \mathcal{D} : x \in \mathcal{C}_2\}|$.
- Estos permiten calcular los promedios muestrales (centros de masa) de cada clase:

$$\mu_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} x_n \qquad \mu_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} x_n$$

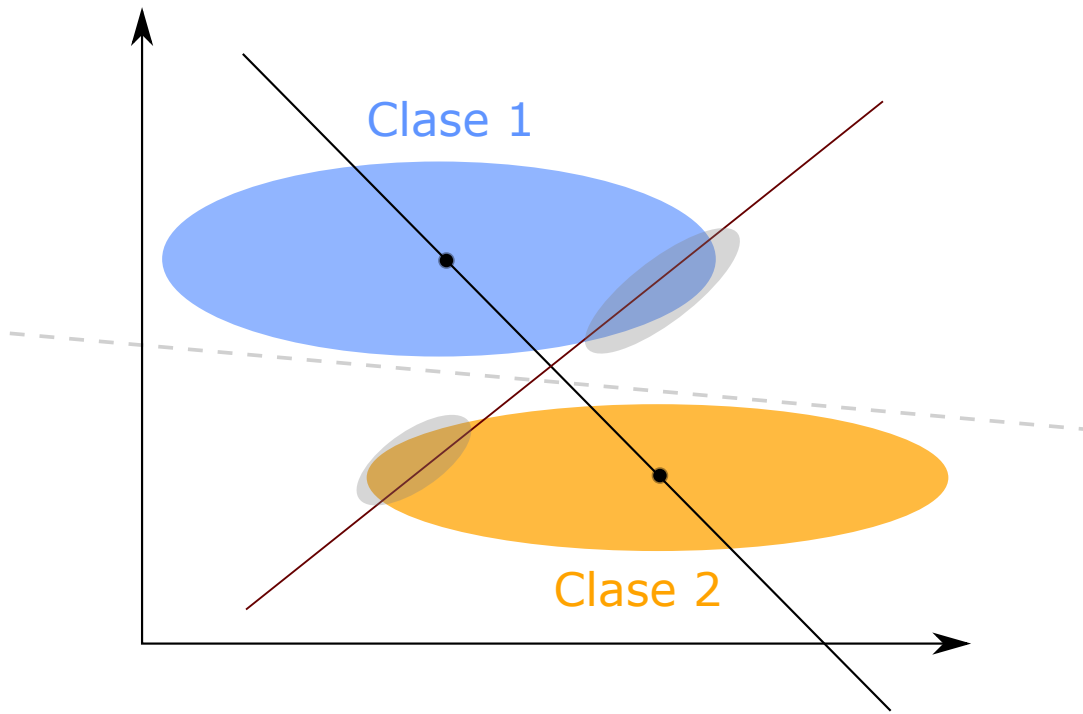
- De esta forma, la medida más simple de separación entre las proyecciones de las clases sobre a es la distancia entre las medias de sus proyecciones:

$$m_1 - m_2 = a^\top (\mu_1 - \mu_2)$$

donde $m_k = a^\top \mu_k$ corresponde al promedio de los elementos de la clase $\mathcal{C}_k$ proyectado sobre el vector a (centro de masa sobre la recta).

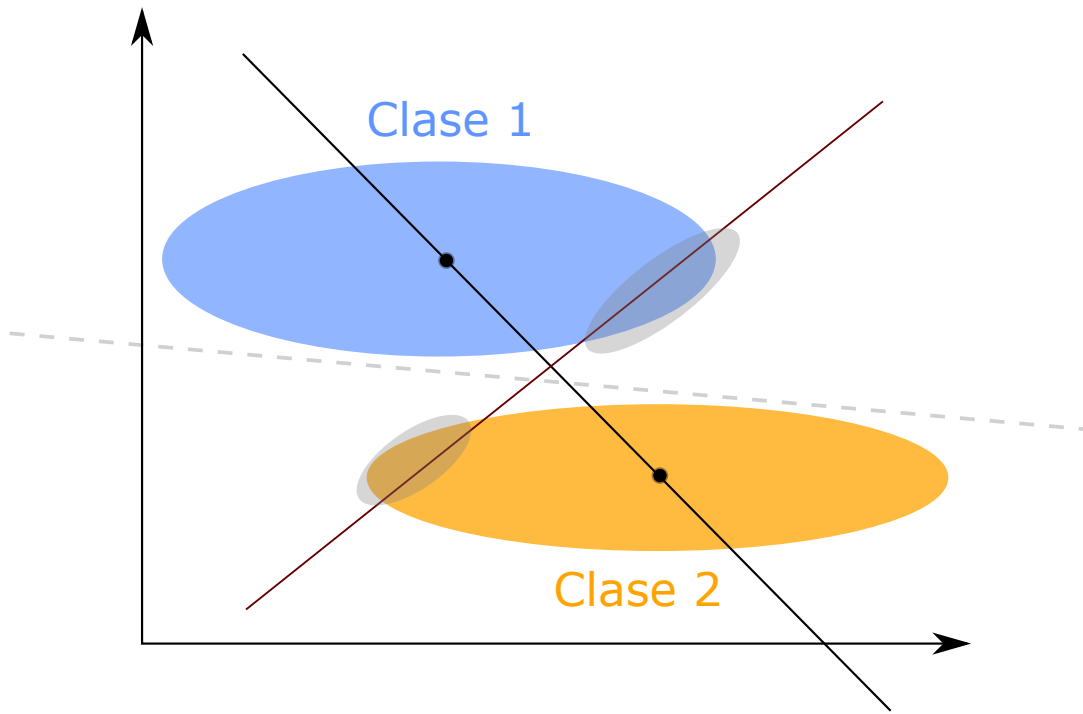
Discriminante lineal de Fisher: primera formulación (desventaja)

El problema de este enfoque es que pueden existir 2 clases bien separadas en el espacio M-dimensional, pero que al proyectar los datos sobre la recta que une sus promedios, las proyecciones de cada clase se traslapen.



Discriminante lineal de Fisher: primera formulación (desventaja)

El problema de este enfoque es que pueden existir 2 clases bien separadas en el espacio M-dimensional, pero que al proyectar los datos sobre la recta que une sus promedios, las proyecciones de cada clase se traslapen.



Discriminante lineal de Fisher: segunda formulación

Para resolver este problema, Fisher propuso el siguiente esquema:

- maximizar la distancia entre las medias de las clases proyectadas (primera formulación).
- Adicionalmente, minimizar la dispersión de los elementos de una misma clase con el objetivo de disminuir el traslape entre las proyecciones de las clases.

Como medida de dispersión, se define la varianza muestral proyectada de los elementos de la clase $\mathcal{C}_k$ mediante

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (a^\top (x_n - \mu_k))^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

Donde el factor de correción $\frac{1}{N_k - 1}$ fue omitido ya que de lo contrario, todas las clases pesarían lo mismo sin importar la cantidad de elementos de la clase.

Lo anterior permite definir la siguiente función objetivo:

$$J(a) = \frac{m_1 - m_2}{s_1^2 + s_2^2}$$

Discriminante lineal de Fisher: segunda formulación

Para resolver este problema, Fisher propuso el siguiente esquema:

- maximizar la distancia entre las medias de las clases proyectadas (primera formulación).
- Adicionalmente, minimizar la dispersión de los elementos de una misma clase con el objetivo de disminuir el traslape entre las proyecciones de las clases.

Como medida de dispersión, se define la varianza muestral proyectada de los elementos de la clase $\mathcal{C}_k$ mediante

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (a^\top (x_n - \mu_k))^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

Donde el factor de correción $\frac{1}{N_k - 1}$ fue omitido ya que de lo contrario, todas las clases pesarían lo mismo sin importar la cantidad de elementos de la clase.

Lo anterior permite definir la siguiente función objetivo:

$$J(a) = \frac{m_1 - m_2}{s_1^2 + s_2^2}$$

Discriminante lineal de Fisher: segunda formulación

Para resolver este problema, Fisher propuso el siguiente esquema:

- maximizar la distancia entre las medias de las clases proyectadas (primera formulación).
- Adicionalmente, minimizar la dispersión de los elementos de una misma clase con el objetivo de disminuir el traslape entre las proyecciones de las clases.

Como medida de dispersión, se define la varianza muestral proyectada de los elementos de la clase $\mathcal{C}_k$ mediante

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (a^\top (x_n - \mu_k))^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

Donde el factor de correción $\frac{1}{N_k - 1}$ fue omitido ya que de lo contrario, todas las clases pesarían lo mismo sin importar la cantidad de elementos de la clase.

Lo anterior permite definir la siguiente función objetivo:

$$J(a) = \frac{m_1 - m_2}{s_1^2 + s_2^2}$$

Discriminante lineal de Fisher: segunda formulación

Para resolver este problema, Fisher propuso el siguiente esquema:

- maximizar la distancia entre las medias de las clases proyectadas (primera formulación).
- Adicionalmente, minimizar la dispersión de los elementos de una misma clase con el objetivo de disminuir el traslape entre las proyecciones de las clases.

Como medida de dispersión, se define la varianza muestral proyectada de los elementos de la clase $\mathcal{C}_k$ mediante

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (a^\top (x_n - \mu_k))^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

Donde el factor de correción $\frac{1}{N_k - 1}$ fue omitido ya que de lo contrario, todas las clases pesarían lo mismo sin importar la cantidad de elementos de la clase.

Lo anterior permite definir la siguiente función objetivo:

$$J(a) = \frac{m_1 - m_2}{s_1^2 + s_2^2}$$

Discriminante lineal de Fisher: segunda formulación

Se puede expresar este costo directamente como función del vector de proyección a:

$$J(a) = \frac{m_1 - m_2}{s_1^2 + s_2^2} = \frac{a^\top S_B a}{a^\top S_W a}$$

donde la matriz de covarianza entre clases S_B y matriz total de covarianza dentro de clases S_W están dadas por

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top$$
$$S_W = \sum_{n \in \mathcal{C}_1} (x_n - \mu_1)(x_n - \mu_1)^\top + \sum_{n \in \mathcal{C}_2} (x_n - \mu_2)(x_n - \mu_2)^\top.$$

Discriminante lineal de Fisher: optimización

Aplicando la condición de primer orden para $J(a) = \frac{a^\top S_B a}{a^\top S_W a}$, obtenemos que el vector a óptimo debe cumplir

$$(a^\top S_B a) S_W a = (a^\top S_W a) S_B a.$$

- La norma del vector a es irrelevante (solo interesa su orientación), con lo que ignorando los escalares $a^\top S_B a$ y $a^\top S_W a$ tenemos que la relación de optimalidad es $S_W a \propto S_B a$.
- por la definición de S_B, sabemos que $S_B a \propto (\mu_1 - \mu_2)$, con lo que la relación de optimalidad se convierte en es $S_W a \propto (\mu_1 - \mu_2)$.

Consecuentemente, el vector optimo a en el criterio de Fisher debe cumplir

$$a \propto S_W^{-1}(\mu_1 - \mu_2).$$

Discriminante lineal de Fisher: optimización

Aplicando la condición de primer orden para $J(a) = \frac{a^\top S_B a}{a^\top S_W a}$, obtenemos que el vector a óptimo debe cumplir

$$(a^\top S_B a) S_W a = (a^\top S_W a) S_B a.$$

- La norma del vector a es irrelevante (solo interesa su orientación), con lo que ignorando los escalares $a^\top S_B a$ y $a^\top S_W a$ tenemos que la relación de optimalidad es $S_W a \propto S_B a$.
- por la definición de S_B, sabemos que $S_B a \propto (\mu_1 - \mu_2)$, con lo que la relación de optimalidad se convierte en es $S_W a \propto (\mu_1 - \mu_2)$.

Consecuentemente, el vector optimo a en el criterio de Fisher debe cumplir

$$a \propto S_W^{-1}(\mu_1 - \mu_2).$$

Discriminante lineal de Fisher: optimización

Aplicando la condición de primer orden para $J(a) = \frac{a^\top S_B a}{a^\top S_W a}$, obtenemos que el vector a óptimo debe cumplir

$$(a^\top S_B a) S_W a = (a^\top S_W a) S_B a.$$

- La norma del vector a es irrelevante (solo interesa su orientación), con lo que ignorando los escalares $a^\top S_B a$ y $a^\top S_W a$ tenemos que la relación de optimalidad es $S_W a \propto S_B a$.
- por la definición de S_B, sabemos que $S_B a \propto (\mu_1 - \mu_2)$, con lo que la relación de optimalidad se convierte en es $S_W a \propto (\mu_1 - \mu_2)$.

Consecuentemente, el vector optimo a en el criterio de Fisher debe cumplir

$$a \propto S_W^{-1}(\mu_1 - \mu_2).$$

Discriminante lineal de Fisher: optimización

Aplicando la condición de primer orden para $J(a) = \frac{a^\top S_B a}{a^\top S_W a}$, obtenemos que el vector a óptimo debe cumplir

$$(a^\top S_B a) S_W a = (a^\top S_W a) S_B a.$$

- La norma del vector a es irrelevante (solo interesa su orientación), con lo que ignorando los escalares $a^\top S_B a$ y $a^\top S_W a$ tenemos que la relación de optimalidad es $S_W a \propto S_B a$.
- por la definición de S_B, sabemos que $S_B a \propto (\mu_1 - \mu_2)$, con lo que la relación de optimalidad se convierte en es $S_W a \propto (\mu_1 - \mu_2)$.

Consecuentemente, el vector optimo a en el criterio de Fisher debe cumplir

$$a \propto S_W^{-1}(\mu_1 - \mu_2).$$

Discriminante lineal de Fisher: ejemplo

En la siguiente figura se observa el discriminador lineal que solo considera los promedios (izquierda) y la corrección de Fisher (derecha).