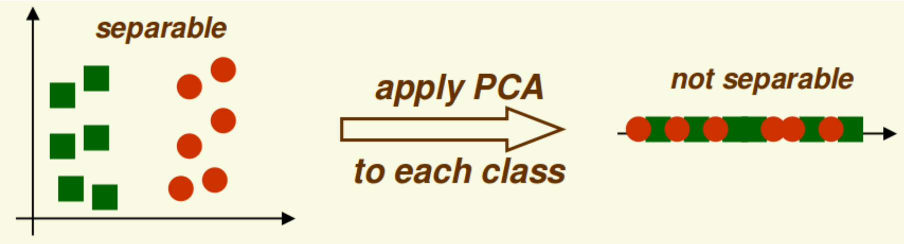
Análisis Discriminante de Fisher (o Lineal)

Aunque PCA encuentra componentes que son útiles para representar datos, no existe ninguna razón para asumir que dichas componentes deben ser útiles para discriminar entre conjuntos de datos en diferentes clases.

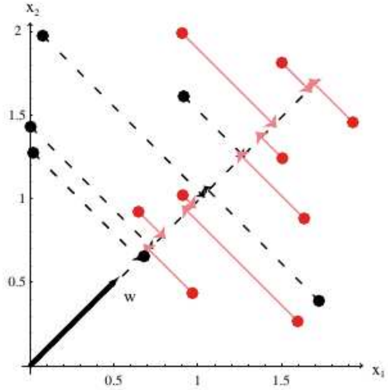


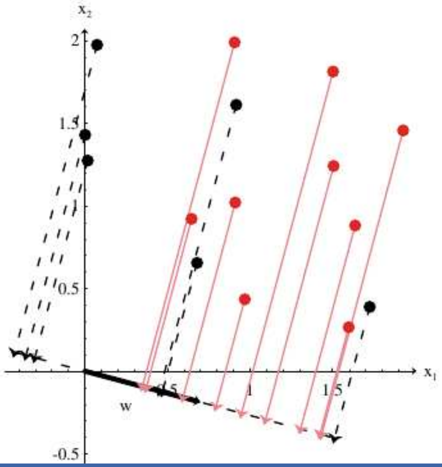
Análisis Discriminante de Fisher

Mientras PCA busca direcciones de proyección que son eficientes para representar los datos, el **análisis discriminante** busca direcciones que son eficientes para discriminación, es decir que permiten una mejor separación de las clases en el espacio de menor dimensión.

El análisis discriminate de Fisher proyecta los datos a una línea (o hiperplano), que preserva la(s) direcciones útiles para la **clasificación de datos**.

Proyecciones Discriminantes



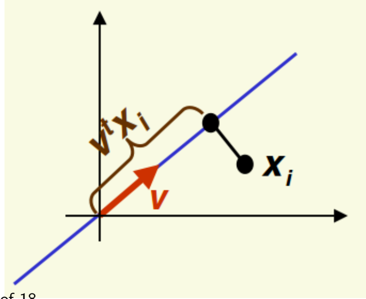


Formulación de LDA

Supongamos que tenemos un problema de clasificación de 2 clases, cada muestra $\mathbf{x}_i$ es d-dimensional.

- n_1 muestras clase 1
- n_2 muestras clase 2

Consideremos la proyección de un punto (muestra) con respecto una dirección representada por un vector unitario $\mathbf{v}$



El escalar $\mathbf{v}^T\mathbf{x}_i$ corresponde a la distancia de la proyección de $\mathbf{x}_i$ desde el origen.

Formulación de LDA

¿Cómo medir la separación de las proyecciones de diferentes clases?

- Sea $\widetilde{\mu}_1$ y $\widetilde{\mu}_2$ las medias de las proyecciones de las clases 1 y 2 respectivamente.
- Sea μ_1 y μ_2 las medias de las clases 1 y 2 en el espacio original.

Una buena medida de separación entre clases podría ser: $|\widetilde{\mu}_1 - \widetilde{\mu}_2|$.

$$\widetilde{\mu}_1 = \frac{1}{n_1} \sum_{\mathbf{x}_i \in C_1}^{n_1} \mathbf{v}^T \mathbf{x}_i = \mathbf{v}^T \left(\frac{1}{n_1} \sum_{\mathbf{x}_i \in C_1}^{n_1} \mathbf{x}_i \right) = \mathbf{v}^T \mu_1$$

de manera similar $\widetilde{\mu}_2 = \mathbf{v}^T \mu_2$

Formulación de LDA

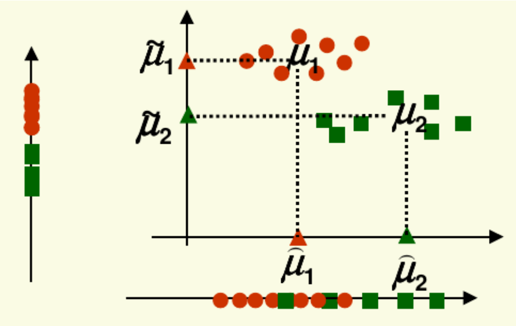
¿Que tan buena es $|\widetilde{\mu}_1 - \widetilde{\mu}_2|$ como medida de separación?

¿Entre mayor sea $|\widetilde{\mu}_1 - \widetilde{\mu}_2|$ mejor será la separación esperada?

Formulación de LDA

¿Que tan buena es $|\widetilde{\mu}_1 - \widetilde{\mu}_2|$ como medida de separación?

¿Entre mayor sea $|\widetilde{\mu}_1 - \widetilde{\mu}_2|$ mejor será la separación esperada?



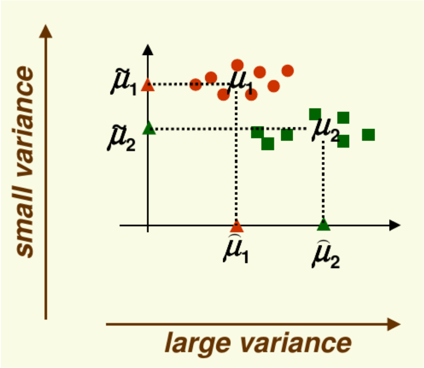
Formulación de LDA

- En la figura anterior el eje vertical es una mejor dirección que el eje horizontal para proyectar los datos, teniendo en cuenta la separabilidad entre clases.
- Sin embargo $|\hat{\mu}_1 - \hat{\mu}_2| > |\widetilde{\mu}_1 - \widetilde{\mu}_2|$

Formulación de LDA

- En la figura anterior el eje vertical es una mejor dirección que el eje horizontal para proyectar los datos, teniendo en cuenta la separabilidad entre clases.
- Sin embargo $|\hat{\mu}_1 - \hat{\mu}_2| > |\widetilde{\mu}_1 - \widetilde{\mu}_2|$

El **problema** con $|\widetilde{\mu}_1 - \widetilde{\mu}_2|$ es que no considera la varianza de las clases.

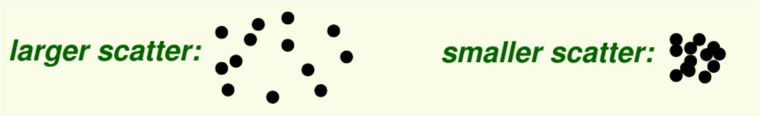


Formulación de LDA

Es necesario entonces normalizar $|\widetilde{\mu}_1 - \widetilde{\mu}_2|$ con respecto a un factor que sea proporcional a la varianza.

Definamos entonces la dispersión **S** de un conjunto de muestras como la varianza multiplicada por n, es decir:

$$\mathbf{S} = \sum_{i=1}^{n} (\mathbf{x}_i - \mu)^2$$



Formulación de LDA

La solución proporcionada por el análisis discriminante de Fisher es entonces normalizar la distancia entre las medias por la dispersión.

- Sea $\mathbf{y}_i = \mathbf{v}^T \mathbf{x}_i$, es decir, $\mathbf{y}_i$'s son las muestras proyectadas.
- La dispersión de las muestras proyectadas de la clase 1 es

$$\widetilde{\mathbf{S}}_1^2 = \sum_{\mathbf{y}_i \in C_1} (\mathbf{y}_i - \widetilde{\mu}_1)^2$$

- La dispersión de las muestras proyectadas de la clase 2 es

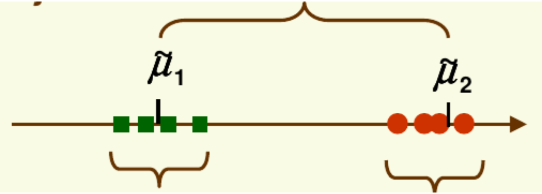
$$\widetilde{\mathbf{S}}_2^2 = \sum_{\mathbf{y}_i \in C_2} (\mathbf{y}_i - \widetilde{\mu}_2)^2$$

Formulación de LDA

Es necesario normalizar con respecto a ambas dispersiones, por lo tanto el análisis discriminante de Fisher proyecta los datos en la dirección **v** que maximiza el criterio:

$$J(\mathbf{v}) = \frac{(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2}{\widetilde{\mathbf{S}}_1^2 + \widetilde{\mathbf{S}}_2^2}$$

Si encontramos un vector **v** que haga a $J(\mathbf{v})$ grande, garantizaremos que las clases están bien separadas.



Formulación de LDA

Lo que se necesita es entonces expresar $J(\mathbf{v})$ en términos de $\mathbf{v}$ y maximizar. Para es necesario definir las matrices de dispersión de las muestras en el espacio original:

$$\mathbf{S}_1 = \sum_{\mathbf{x}_i \in C_1} (\mathbf{x}_i - \mu_1)(\mathbf{x}_i - \mu_1)^T$$

$$\mathbf{S}_2 = \sum_{\mathbf{x}_i \in C_2} (\mathbf{x}_i - \mu_2)(\mathbf{x}_i - \mu_2)^T$$

Y definamos la matriz de dispersión **intra** clase como: $\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$.

Recordando los resultados sobre la media de las muestras proyectadas, se puede llegar a la conclusión que: $\widetilde{\mathbf{S}}_1^2 = \mathbf{v}^T \mathbf{S}_1 \mathbf{v}$ y de manera similar para la clase 2, $\widetilde{\mathbf{S}}_2^2 = \mathbf{v}^T \mathbf{S}_2 \mathbf{v}$

Formulación de LDA

Entonces

$$\widetilde{\mathbf{S}}_1^2 + \widetilde{\mathbf{S}}_2^2 = \mathbf{v}^T \mathbf{S}_1 \mathbf{v} + \mathbf{v}^T \mathbf{S}_2 \mathbf{v} = \mathbf{v}^T \mathbf{S}_W \mathbf{v}$$

Formulación de LDA

Entonces

$$\widetilde{\mathbf{S}}_1^2 + \widetilde{\mathbf{S}}_2^2 = \mathbf{v}^T \mathbf{S}_1 \mathbf{v} + \mathbf{v}^T \mathbf{S}_2 \mathbf{v} = \mathbf{v}^T \mathbf{S}_W \mathbf{v}$$

Se define la matriz de dispersión **entre** clases como:

$$\mathbf{S}_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

la cual mide la separación entre las medias de dos clases (antes de la proyección). Utilizando nuevamente la definición de las medias proyectadas desarrollada antes es posible obtener:

$$(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2 = \mathbf{v}^T \mathbf{S}_B \mathbf{v}$$

Formulación de LDA

Por consiguiente el criterio $J(\mathbf{v})$ se puede expresar como:

$$J(\mathbf{v}) = \frac{(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2}{\widetilde{\mathbf{S}}_1^2 + \widetilde{\mathbf{S}}_2^2} = \frac{\mathbf{v}^T \mathbf{S}_B \mathbf{v}}{\mathbf{v}^T \mathbf{S}_W \mathbf{v}}$$

Para minimizar derivamos $J(\mathbf{v})$ con respecto a $\mathbf{v}$ e igualamos a cero. Después de algunas operaciones matriciales se llega a:

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{v} = \lambda \mathbf{v}$$

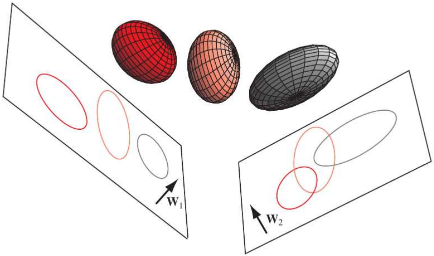
Que corresponde nuevamente a un problema de cálculo de valores propios pero está vez de la matriz $\mathbf{S}_W^{-1}\mathbf{S}_B$. Es posible también encontrar que para un caso bi-clase, la proyección que maximiza el criterio de Fisher corresponde al vector $\mathbf{v} = \mathbf{S}_W^{-1}(\mu_1 - \mu_2)$.

Análisis Discriminante Múltiple

El análisis discriminante de Fisher se puede generalizar para múltiples clases, en cuyo caso el criterio a optimizar es:

$$J(\mathbf{V}) = \frac{\det\left(\mathbf{V}^T \mathbf{S}_B \mathbf{V}\right)}{\det\left(\mathbf{V}^T \mathbf{S}_W \mathbf{V}\right)}$$

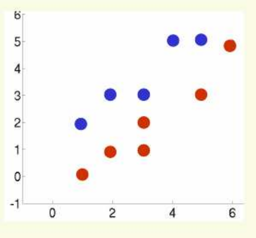
donde **V** es llamada la matriz de proyección.



Ejercicio

Considere un problema de clasificación de 2 clases, con el siguiente conjunto de muestras:

- La clase 1 tiene 5 muestras: $C_1 = [(1,2),(2,3),(3,3),(4,5),(5,5)]$
- La clase 2 tiene 6 muestras: $C_2 = [(1,0),(2,1),(3,1),(3,2),(5,3),(6,5)]$



- ¿Cuál sería la proyección de PCA?
- Calcule los puntos proyectados por LDA.

Fin del curso

"Any fool can know. The point is to understand."
– Albert Einstein

Referencias

[1] Velten K., *Mathematical Modeling and Simulation*, WILEY-VCH, 2009.

[2] Murphy K.P., *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[3] Duda R.O., Hart P.E., Stork D.G., *Pattern Classification*. 2ed, WILEY-INTERSCIENCE, 2001.

[4] Bishop, C.M. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] Webb, A.R. *Statistical Pattern Recognition*, 2nd Revised edition, John Wiley & Sons Ltd, 2002.

[6]Peña, D. *Análisis de datos multivariantes*. Mc Graw Hill, Madrid, España, 2002.