Discriminantes de Fisher

Lección 22

Dr. Pablo Alvarado Moya

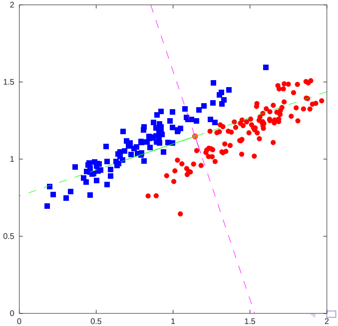
CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

Contenido

Introducción

- El Análisis de Componentes Principales (ACP, o PCA) permitió hacer una reducción de dimensiones **no** supervisada.
- PCA es en general **no** adecuado para pre-procesar tareas de clasificación
- Reducción no necesariamente favorecerá separabilidad de clases:

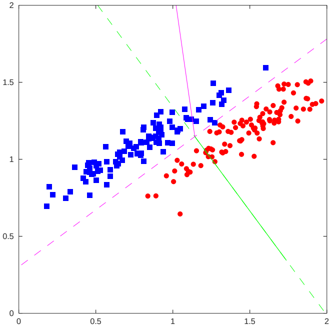


Discriminantes lineales

- Vimos métodos discriminantes y generativos para clasificación
- Varias estrategias separan clases linealmente:
  - Regresión logística
  - Perceptron
  - GDA
  - SVM con kernel lineal
- Queremos encontrar una transformación $y = \underline{\mathbf{w}}^T \underline{\mathbf{x}}$ tal que la clasificación se realice con un solo umbral.

Discriminantes de Fisher

- Presentamos aquí el Análisis de Discriminantes de Fisher, un tipo particular de análisis de discriminantes lineales.
- Esta es una tarea **supervisada**
- Objetivo es encontrar ejes que maximicen separabilidad de clases:



Planteamiento del problema

- Sea $(\underline{\mathbf{x}}^{(i)}, y^{(i)})$, $i = 1 \ldots m$ un conjunto de datos etiquetado con $y^{(i)} \in \{0,1\}$
- Tenemos m_0 datos para la clase 0 y m_1 datos para la clase 1

$$m_0 = \sum_{i=1}^{m} 1\left\{y^{(i)} = 0\right\} \qquad m_1 = \sum_{i=1}^{m} 1\left\{y^{(i)} = 1\right\}$$

- Todos los datos con $y^{(i)} = 0$ se agrupan en $\mathcal{C}_0$ y todos los datos en $y^{(i)} = 1$ en $\mathcal{C}_1$
- Además, para las medias de cada clase se cumple:

$$\underline{\boldsymbol{\mu}}_0 = \frac{1}{m_0} \sum_{i=1}^{m} 1\left\{y^{(i)} = 0\right\} \underline{\mathbf{x}}^{(i)} \qquad \underline{\boldsymbol{\mu}}_1 = \frac{1}{m_1} \sum_{i=1}^{m} 1\left\{y^{(i)} = 1\right\} \underline{\mathbf{x}}^{(i)}$$

Separación de clases

- La proyección de $\underline{\mathbf{x}}^{(i)}$ sobre un vector unitario $\underline{\mathbf{w}}$ es

$$y = \underline{\mathbf{w}}^T \underline{\mathbf{x}}^{(i)}$$

- Una primera idea: elijamos $\underline{\mathbf{w}}$ tal que maximice la distancia entre las medias

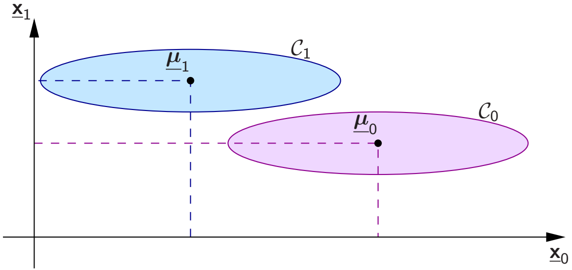
$$d_\mu = \underline{\mathbf{w}}^T (\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)$$

  sujeto a que $\|\underline{\mathbf{w}}\| = 1$

- A esta técnica se le llama PCA de medias

PCA de medias

- Luego de maximizar d_μ se obtiene que $\underline{\mathbf{w}} \propto (\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)$
- La siguiente figura ilusta el problema de PCA de medias:



- En el eje $\underline{\mathbf{x}}_0$ la separación de medias d_μ es mayor que la proyección en $\underline{\mathbf{x}}_1$
- Sin embargo, ¡la separación entre clases es mejor en eje $\underline{\mathbf{x}}_1$!
- El problema surge puesto que la **dispersión** intra-clase a lo largo de $\underline{\mathbf{x}}_0$ es grande.

Idea de Fisher

- Fisher propuso maximizar una función asociada a la diferencia entre medias normalizadas por una medida de dispersión intra-clase a lo largo de $\underline{\mathbf{w}}$
- La dispersión de los datos de la clase k proyectados por $\underline{\mathbf{w}}$ será:

$$s_k^2 = \sum_{i=1}^{m} 1\left\{y^{(i)} = k\right\}\left(\underline{\mathbf{w}}^T(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}_k)\right)^2$$

- La dispersión total intra-clase es $s_0^2 + s_1^2$.
- Notemos la similitud entre dispersión y varianza: solo se omite división por m o m_k.
- El criterio de Fisher es

$$J(\underline{\mathbf{w}}) = \frac{\left(\underline{\mathbf{w}}^T(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)\right)^2}{s_0^2 + s_1^2} = \frac{d_\mu^2}{s_0^2 + s_1^2}$$

Criterio de Fisher

- Puede demostrarse que lo anterior es equivalente a:

$$J(\underline{\mathbf{w}}) = \frac{\underline{\mathbf{w}}^T \mathbf{S}_B \underline{\mathbf{w}}}{\underline{\mathbf{w}}^T \mathbf{S}_W \underline{\mathbf{w}}}$$

  con $\mathbf{S}_B$ la matrix de dispersión inter-clase (*between-class scatter*):

$$\mathbf{S}_B = (\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)^T$$

  y $\mathbf{S}_W$ la dispersión total intra-clase (*within-class scatter*):

$$\mathbf{S}_W = \sum_{i=1}^{m} \left(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}_{y^{(i)}}\right)\left(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}_{y^{(i)}}\right)^T$$

- Calculando el gradiente e igualando a cero se encuentra máximo de $J(\underline{\mathbf{w}})$ si

$$\left(\underline{\mathbf{w}}^T \mathbf{S}_B \underline{\mathbf{w}}\right) \mathbf{S}_W \underline{\mathbf{w}} = \left(\underline{\mathbf{w}}^T \mathbf{S}_W \underline{\mathbf{w}}\right) \mathbf{S}_B \underline{\mathbf{w}}$$

Derivación de dirección de separación

- Con $\mathbf{S}_B = (\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)^T$ entonces $\mathbf{S}_B\underline{\mathbf{w}}$ siempre tiene la dirección de $(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)$, puesto que $(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)^T\underline{\mathbf{w}}$ es un escalar
- Solo estamos interesados en la dirección de $\underline{\mathbf{w}}$, por lo que considerando que los términos cuadráticos son escalares:

$$\begin{aligned}
\left(\underline{\mathbf{w}}^T\mathbf{S}_B\underline{\mathbf{w}}\right)\mathbf{S}_W\underline{\mathbf{w}} &= \left(\underline{\mathbf{w}}^T\mathbf{S}_W\underline{\mathbf{w}}\right)\mathbf{S}_B\underline{\mathbf{w}} \\
\alpha_B\mathbf{S}_W\underline{\mathbf{w}} &= \alpha_W\beta(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0) \\
\underline{\mathbf{w}} &\propto \mathbf{S}_W^{-1}(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)
\end{aligned}$$

- Esta dirección se conoce como **discriminante de Fisher**

Extensión a más de una dirección

- Solo hemos calculado una única dirección $\underline{\mathbf{w}}$ para discriminar.
- Nuestro interés es reducir dimensión de datos de entrada, pero posiblemente necesitemos más de una dirección.
- A partir $\underline{\mathbf{w}}$ buscamos una matriz de transformación que proyecte los datos $\underline{\mathbf{x}}^{(i)}$ a otra base, donde el primer vector de esa base debe ser $\underline{\mathbf{w}}$.
- Esa transformación no es única. Una selección común usa transformaciones de Householder, donde se modifica el signo de la última columna para convertirla de una reflexión a una rotación.
- Una vez proyectados los datos, eliminamos la primera dimensión y repetimos el proceso en el subespacio.
- Resultado **no** será necesariamente ortogonal

Extensión a múltiples clases

- Vamos a extender conceptos a más de dos clases.
- Supondremos que el número n de dimensiones de los datos es mayor que el número c de clases ($n > c$).
- Buscamos proyectar los datos a un nuevo espacio de $n' < n$ dimensiones, elegidas para facilitar la clasificación:

$$\hat{\underline{\mathbf{x}}}^{(i)} = \mathbf{W}\underline{\mathbf{x}}^{(i)}$$

- La generalización de la matriz de dispersión intra-clase es

$$\mathbf{S}_W = \sum_{i=1}^{m} (\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}_{y^{(i)}})(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}_{y^{(i)}})^T$$

con

$$\underline{\boldsymbol{\mu}}_k = \frac{1}{m_k} \sum_{i=1}^{m} 1\left\{y^{(i)} = k\right\} \underline{\mathbf{x}}^{(i)} \qquad m_k = \sum_{i=1}^{m} 1\left\{y^{(i)} = k\right\}$$

Matriz de dispersión interclase (1)

- La matriz de dispersión total es:

$$\mathbf{S}_T = \sum_{i=1}^{m} (\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}})(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}})^T$$

con

$$\underline{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i=1}^{m} \underline{\mathbf{x}}^{(i)} = \frac{1}{m}\sum_{i=1}^{c} m_k \underline{\boldsymbol{\mu}}_k \qquad m = \sum_{k=1}^{c} m_k$$

Matriz de dispersión interclase (2)

- Si asumimos que la matriz de dispersión total se descompone en las matrices de dispersión intra- e interclase:

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$$

entonces podemos despejar

$$\mathbf{S}_B = \sum_{k=1}^{c} m_k(\underline{\boldsymbol{\mu}}_k - \underline{\boldsymbol{\mu}})(\underline{\boldsymbol{\mu}}_k - \underline{\boldsymbol{\mu}})^T$$

Matrices en espacio de características

- La matriz de dispersión intra-clase en el espacio de características es

$$\mathbf{s}_W = \sum_{i=1}^{m} (\underline{\hat{\mathbf{x}}}^{(i)} - \underline{\boldsymbol{\mu}}_{y^{(i)}})(\underline{\hat{\mathbf{x}}}^{(i)} - \underline{\boldsymbol{\mu}}_{y^{(i)}})^T$$

y la matriz inter-clase en el espacio de características:

$$\mathbf{s}_B = \sum_{k=1}^{c} m_k (\underline{\hat{\boldsymbol{\mu}}}_k - \underline{\hat{\boldsymbol{\mu}}})(\underline{\hat{\boldsymbol{\mu}}}_k - \underline{\hat{\boldsymbol{\mu}}})^T$$

con

$$\underline{\hat{\boldsymbol{\mu}}}_k = \frac{1}{m_k} \sum_{i=1}^{m} 1\left\{ y^{(i)} = k \right\} \underline{\hat{\mathbf{x}}}^{(i)} \qquad \underline{\hat{\boldsymbol{\mu}}} = \frac{1}{m} \sum_{i=1}^{m} m_k \underline{\hat{\boldsymbol{\mu}}}_k$$

Solución analítica

- Queremos de nuevo encontrar ejes en donde la dispersión interclase es grande y la dispersión intraclase es pequeña.
- El criterio de Fukunaga define la función a optimizar como

$$J(\mathbf{W}) = \operatorname{tr}\{\mathbf{s}_W^{-1}\mathbf{s}_B\} = \operatorname{tr}\{(\mathbf{W}\mathbf{S}_W\mathbf{W}^T)^{-1}(\mathbf{W}\mathbf{S}_B\mathbf{W}^T)\}$$

- La solución analítica a dicho criterio está dada por los eigenvectores de $\mathbf{S}_W^{-1}\mathbf{S}_B$ correspondientes a los n' eigenvalores mayores.
- La matriz $\mathbf{S}_B$ tiene rango a lo sumo $c-1$, por lo que hay un máximo de $c-1$ eigenvalores no nulos.
- Esto implica que no es posible encontrar más de $c-1$ nuevas características con esta técnica.

Direcciones de Fisher (1)

- En general, la matriz $\mathbf{S}_W^{-1}\mathbf{S}_B$ **no** es simétrica, por lo que los eigenvectores no son ortogonales.
- Demostraremos que, pese a la asimetría, todos los eigenvalores son reales.
- Eso se debe a que $\mathbf{S}_B$ es simétrica, positiva definida, por lo que tiene eigenvalores reales positivos y se cumple:

$$\mathbf{S}_B = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

  con $\mathbf{U}$ los eigenvectores y $\mathbf{\Lambda}$ los eigenvalores (matriz diagonal).
- Si definimos $\mathbf{S}_B^{\frac{1}{2}} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}^T$, entonces

$$\mathbf{S}_B^{\frac{1}{2}}\mathbf{S}_B^{\frac{1}{2}} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{S}_B$$

Direcciones de Fisher (2)

- Si definimos ahora $\underline{\mathbf{v}} = \mathbf{S}_B^{\frac{1}{2}}\underline{\mathbf{w}}$ entonces

$$\begin{aligned}
\mathbf{S}_W^{-1}\mathbf{S}_B\underline{\mathbf{w}} &= \lambda\underline{\mathbf{w}} \\
\mathbf{S}_W^{-1}\mathbf{S}_B^{\frac{1}{2}}\underline{\mathbf{v}} &= \mathbf{S}_B^{-\frac{1}{2}}\lambda\underline{\mathbf{v}} \\
\mathbf{S}_B^{\frac{1}{2}}\mathbf{S}_W^{-1}\mathbf{S}_B^{\frac{1}{2}}\underline{\mathbf{v}} &= \lambda\underline{\mathbf{v}}
\end{aligned}$$

- La matriz $\mathbf{S}_B^{\frac{1}{2}}\mathbf{S}_W^{-1}\mathbf{S}_B^{\frac{1}{2}}$ es simétrica, positiva definida y por tanto tiene eigenvalores λ reales y eigenvectores ortogonales $\mathbf{v}$.
- Los eigenvectores de interés son $\underline{\mathbf{w}} = \mathbf{S}_B^{-\frac{1}{2}}\underline{\mathbf{v}}$

Resumen

Este documento ha sido elaborado con software libre incluyendo LATEX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux