Clasificación lineal

J.D. Echeverry-Correa, Ph.D.
A. M. Alvarez-Meza, Ph.D.
jde@utp.edu.co;amalvarezme@unal.edu.co

Departamento de ingeniería eléctrica
Universidad Tecnológica de Pereria
Departamento de ingeniería eléctrica, electrónica y computación
Universidad Nacional de Colombia-sede Manizales

Contenido

Contenido

Hoja de ruta (1)

Las técnicas de Aprendizaje de Máquina aplicadas a la Ciencia de los Datos pueden ser:

- **De aprendizaje supervisado.**
  - Objetivo: Encontrar una función que permita relacionar unas entradas $\boldsymbol{X}$ con unas salidas $\boldsymbol{y}$, dado un set de pares de entradas-salidas $D = \{(\boldsymbol{x}_i, y_i)\}$
  - $\boldsymbol{X} \in \mathbb{R}^{N \times P}$
  - Las salidas $\boldsymbol{y}$ pueden ser: o bien valores reales $y_i \in \mathbb{R}^N$, o bien variables categóricas en donde $y_i \in \{1, ..., C\}$
  - En el primer caso hablaríamos de un *sistema de regresión* y en el segundo caso hablaríamos de un *sistema de clasificación*

Hoja de ruta (2)

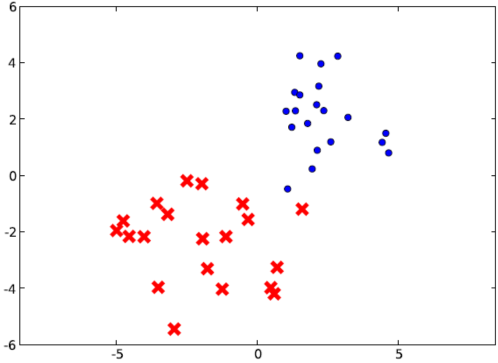
- **De aprendizaje no supervisado.**
  - Objetivo: Dadas unas entradas $\boldsymbol{X} \in \mathbb{R}^{N \times P}$, encontrar patrones de interés en los datos.
  - Inicialmente, no se conoce qué hay que buscar.
  - No hay métricas de error definidas (a diferencia del aprendizaje supervisado)

Contenido

Conceptos básicos (1)

Supuestos:

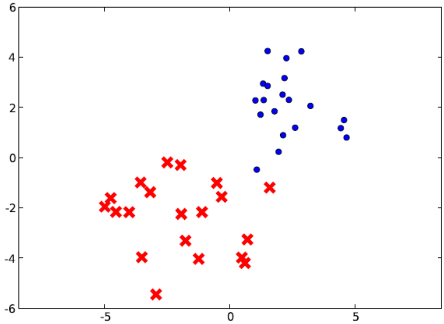
- Cada muestra corresponde a una única clase.
- Los datos conforman espacios linealmente separables.



- Empezaremos por analizar el caso de $K = 2$ clases.

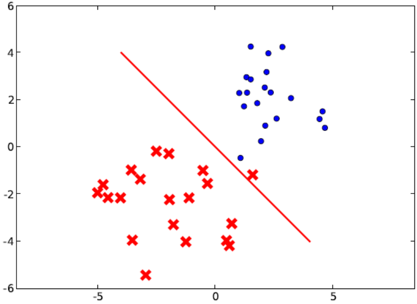
Conceptos básicos (2)

- Cada muestra (instancia) está descrita por un conjunto de números $\Rightarrow$ características
- Debemos escoger características que nos permitan discriminar entre las clases *ejemplos positivos*, *ejemplos negativos*
- En esta gráfica cada muestra está descrita por dos características



Conceptos básicos (3)

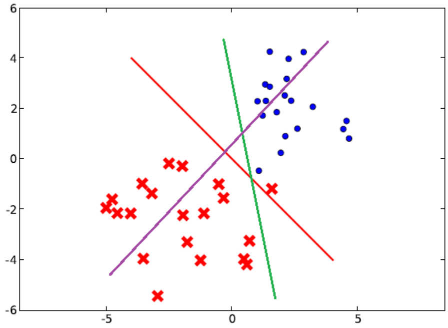
- El objetivo es dividir estos puntos con una línea recta
- Esto es lo que se conoce como **clasificación lineal**



- Si extendemos esta noción a múltiples características por cada una de las muestras, hablaremos entonces de planos (e hiperplanos) que separen entre los ejemplos positivos y los ejemplos negativos.

Conceptos básicos (4)

- No existe una única solución
- Pero sí se puede buscar la solución que satisfaga cierto criterio



Contenido

Función discriminante (1)

- Una función discriminante toma un vector de entradas $\boldsymbol{x}$ y lo asigna a una de K posibles clases:

$$y(\boldsymbol{x}) : \boldsymbol{x} \to k, \quad k \in \{1, ..., K\}$$

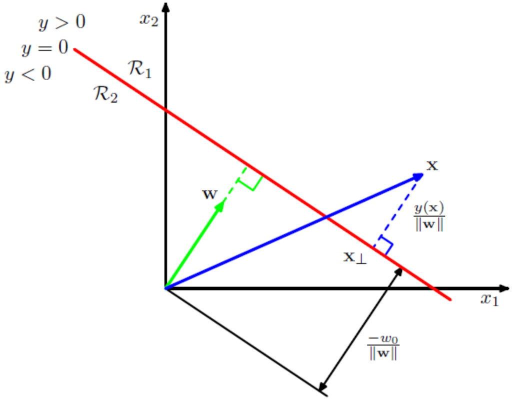
- La función discriminante está dada por:

$$y(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + w_0$$

- $\boldsymbol{w}$ es el vector de pesos y w_0 es el *bias* o tendencia.
- $\boldsymbol{x} \in \mathcal{C}_1$ si $y(\boldsymbol{x}) > 0$; de lo contrario $\boldsymbol{x} \in \mathcal{C}_2$.
- La línea de decisión o superficie de decisión es $y(\boldsymbol{x}) = 0$.
- El vector $\boldsymbol{w}$ es ortogonal a la superficie de decisión.

Función discriminante (2)

- Distancia de $y(\boldsymbol{x})$ al origen: $-w_0/\left\|\boldsymbol{w}\right\|$.
  Tarea: Demostrar.
- Distancia de un punto $\boldsymbol{x}$ a $y(\boldsymbol{x})$: $y(\boldsymbol{x})/\left\|\boldsymbol{w}\right\|$.
  Tarea: Demostrar.

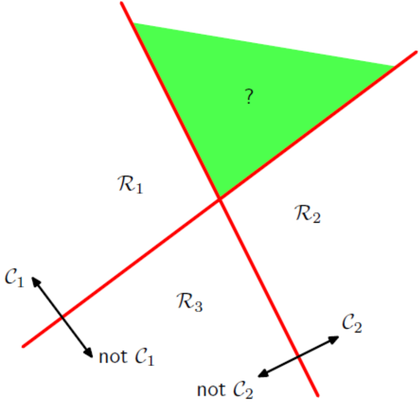


Función discriminante (3)

Con el objetivo de vectorizar las operaciones, se puede hacer $\widetilde{\boldsymbol{x}} = (x_0, \boldsymbol{x})$, $\widetilde{\boldsymbol{w}} = (w_0, \boldsymbol{w})$, por lo tanto $y(\boldsymbol{x}) = \widetilde{\boldsymbol{w}}^\top \widetilde{\boldsymbol{x}}$

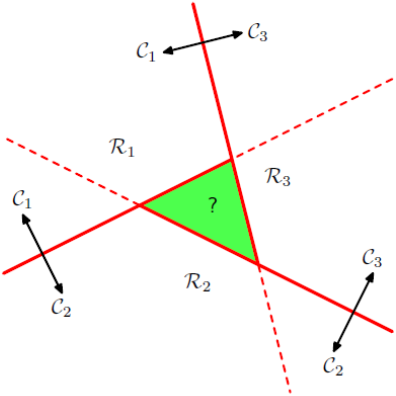
Múltiples clases (1)

- Consideremos ahora la extensión de las funciones discriminantes a $K > 2$ clases.
- Una posible opción: Considerar $K - 1$ discriminantes $\Rightarrow$ *One-versus-the-rest*:



Múltiples clases (2)

- Otra opción es considerar $K(K-1)/2$ discriminantes $\Rightarrow$ *One-versus-one*:

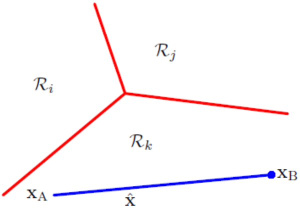


Múltiples clases (3)

- Solución: emplear discriminante de K clases con K funciones lineales

$$y_k(\boldsymbol{x}) = \boldsymbol{w}_k^\top \boldsymbol{x} + w_{k0}$$

- $\boldsymbol{x} \in \mathcal{C}_k$, si $y_k(\boldsymbol{x}) > y_j(\boldsymbol{x}), \quad k \neq j$



- El resultado serán hiperplanos de decisión de la forma

$$(\boldsymbol{w}_k - \boldsymbol{w}_j)^\top \boldsymbol{x} + (w_{k0} - w_{j0}) = 0$$

- Estos hiperplanos generarán regiones de decisión conectadas de manera simple y conexas.

Contenido

Mínimos cuadrados (1)

- En regresión lineal, ya vimos cómo ajustar un modelo mediante una relación lineal entre los datos de entrada y los parámetros.
- Consideremos K clases descritas por modelos lineales

$$y_k(\boldsymbol{x}) = \boldsymbol{w}_k^\top \boldsymbol{x} + w_{k0}, \quad k = 1, ..., K$$

- Podemos agrupar estos términos empleando notación vectorizada

$$\boldsymbol{y}(\boldsymbol{x}) = \widetilde{\boldsymbol{W}}^\top \widetilde{\boldsymbol{x}}$$

donde

$$\widetilde{\boldsymbol{W}} = [\widetilde{\boldsymbol{w}}_1, ..., \widetilde{\boldsymbol{w}}_k] = \begin{bmatrix} w_{10} & \cdots & w_{K0} \\ w_{11} & \cdots & w_{K1} \\ \vdots & \ddots & \vdots \\ w_{1D} & \cdots & w_{KD} \end{bmatrix}$$

- La salida será en la notación de 1-de-K y se podrá comparar entonces con los valores objetivo $\boldsymbol{t} = [t_1, ..., t_k]^\top$

Mínimos cuadrados (2)

- Supongamos entonces que tenemos un conjunto de entrenamiento $\{\boldsymbol{x}_n, \boldsymbol{t}_n\}_{n=1}^{N}$

$\boldsymbol{T}$ será una matriz con vectores fila $\boldsymbol{t}_n^\top$

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{t}_1^\top \\ \vdots \\ \boldsymbol{t}_N^\top \end{bmatrix}$$

$\widetilde{\boldsymbol{X}}$ será una matriz con vectores fila $\widetilde{\boldsymbol{x}}_n$

$$\widetilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{bmatrix}$$

- Para todo el set de entrenamiento tenemos entonces

$$\boldsymbol{Y}\left(\widetilde{\boldsymbol{X}}\right) = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{W}}$$

- El objetivo entonces es escoger $\widetilde{\boldsymbol{W}}$ que minimice

$$\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{W}} - \boldsymbol{T}$$

Mínimos cuadrados (3)

- Para esto, minimizaremos la función de error cuadrático

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\boldsymbol{w}_k^\top \boldsymbol{x}_n - t_{kn}\right)^2$$

En este punto nos conviene saber algo más de álgebra lineal

Propiedad de la traza de una matriz

$$\sum_{i,j} a_{ij}^2 = Tr\{\boldsymbol{A}^\top \boldsymbol{A}\}$$

- Entonces, el error cuadrático expresado en forma matricial será entonces:

$$E_D\left(\widetilde{\boldsymbol{W}}\right) = \frac{1}{2} Tr\left\{\left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{W}} - \boldsymbol{T}\right)^\top \left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{W}} - \boldsymbol{T}\right)\right\}$$

Mínimos cuadrados (4)

- Para minimizar esta expresión, derivamos

$$\frac{\partial}{\partial \widetilde{\boldsymbol{W}}} E_D\left(\widetilde{\boldsymbol{W}}\right) = \frac{1}{2}\frac{\partial}{\partial \widetilde{\boldsymbol{W}}} Tr\left\{\left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{W}} - \boldsymbol{T}\right)^{\top}\left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{W}} - \boldsymbol{T}\right)\right\}$$
$$= \widetilde{\boldsymbol{X}}^{\top}\left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{W}} - \boldsymbol{T}\right)$$

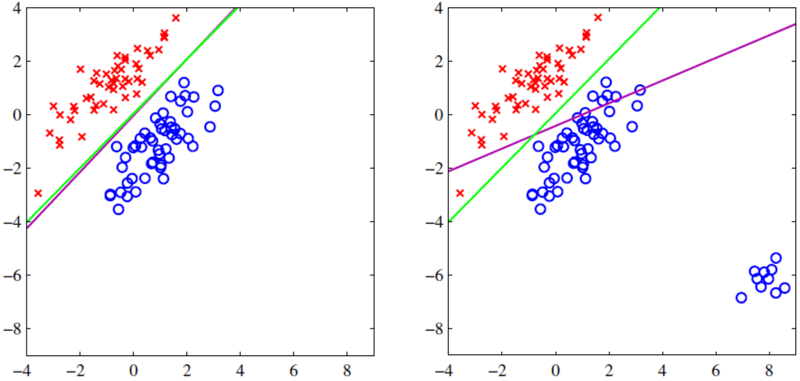
- E igualamos a cero

$$\widetilde{\boldsymbol{W}}_{MSE} = \left(\widetilde{\boldsymbol{X}}^{\top}\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}^{\top}\boldsymbol{T}$$
$$= \widetilde{\boldsymbol{X}}^{+}\boldsymbol{T} \qquad \text{donde } \widetilde{\boldsymbol{X}}^{+} \Rightarrow \text{pseudo-inversa de } \widetilde{\boldsymbol{X}}$$

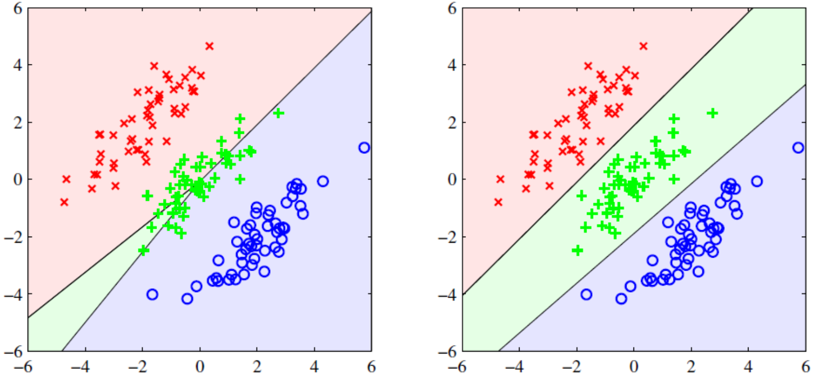
- La función discriminante está dada por

$$\boldsymbol{y}(\boldsymbol{x}) = \widetilde{\boldsymbol{W}}_{MSE}^{\top}\widetilde{\boldsymbol{x}} = \boldsymbol{T}^{\top}\left(\widetilde{\boldsymbol{X}}^{+}\right)^{\top}\widetilde{\boldsymbol{x}}$$

Algunos inconvenientes (1)



Algunos inconvenientes (2)



Análisis discriminante de Fisher (1)

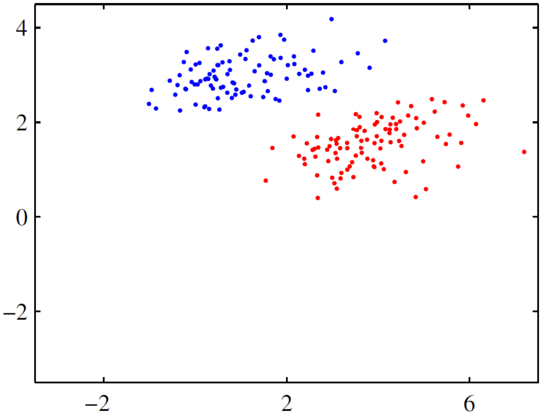
- El objetivo es proyectar los datos a un espacio de menor dimensionalidad donde la clasificación sea más sencilla.
- Sea $\boldsymbol{x} \in \mathbb{R}^D$.
- Se proyecta a una dimensión usando

$$y = \boldsymbol{w}^\top \boldsymbol{x}$$

- Se establece un umbral y_0, y se clasifica un nuevo punto como de la clase $\mathcal{C}_1$ si $y > y_0$, o de la clase $\mathcal{C}_2$ si sucede lo contrario.
- La idea es escoger $\boldsymbol{w}$ de manera que maximice la separabilidad de las clases.

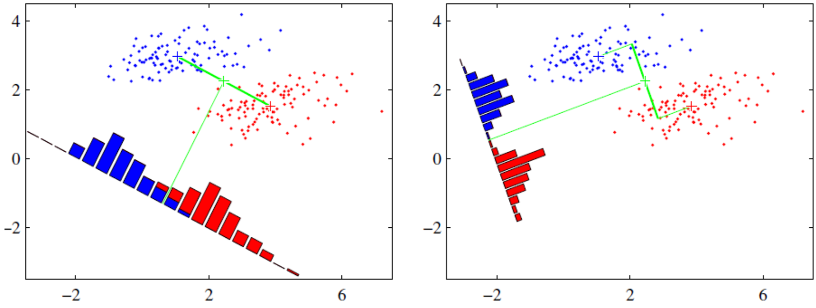
Análisis discriminante de Fisher (2)

Consideremos inicialmente un problema de dos clases



Intuición: Llevar el problema a una sola dimensión y buscar el umbral que separe ambas clases.

Podríamos separar las clases de diversas formas



Los puntos centrales son los vectores que representan a la media de cada grupo

$$\boldsymbol{m}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}\boldsymbol{x}_n, \qquad \boldsymbol{m}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}\boldsymbol{x}_n$$

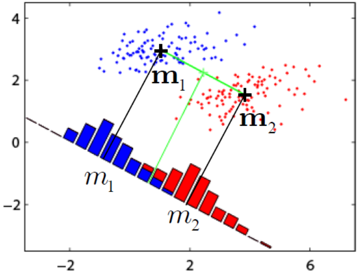
Una medida de separación entre las clases es

$$m_1 - m_2 = \boldsymbol{w}^\top(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

donde

$$m_k = \boldsymbol{w}^\top \boldsymbol{m}_k$$

Se debe escoger $\boldsymbol{w}$ de forma que se maximice la anterior expresión.



Análisis discriminante de Fisher (5)

- Se busca maximizar la distancia entre las medias y a la vez minimizar la variabilidad de las muestras en cada clase.
- La varianza intraclase se obtiene de los vectores transformados de la clase $\mathcal{C}_k$ como

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

  donde $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n$

- El criterio de Fisher se define como

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

  donde $(m_2 - m_1)^2 \Rightarrow$ varianza entre clases, y $s_1^2 + s_2^2 \Rightarrow$ varianza intraclase

- Haciendo los cambios necesarios para hacer la expresión dependiente de $\boldsymbol{w}$, tenemos

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}}$$

donde $\boldsymbol{S}_B$ es la matriz de covarianza entre clases, calculada como

$$\boldsymbol{S}_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top$$

y $\boldsymbol{S}_W$ es la matriz de covarianza intraclases, calculada como

$$\boldsymbol{S}_W = \sum_{n \in \mathcal{C}_1} (\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^\top + \sum_{n \in \mathcal{C}_2} (\boldsymbol{x}_n - \boldsymbol{m}_2)(\boldsymbol{x}_n - \boldsymbol{m}_2)^\top$$

Análisis discriminante de Fisher (7)

- Derivando

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}}$$

con respecto a $\boldsymbol{w}$, e igualando a cero, se tiene que $J(\boldsymbol{w})$ se maximiza cuando

$$(\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w})\boldsymbol{S}_W \boldsymbol{w} = (\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w})\boldsymbol{S}_B \boldsymbol{w}$$

Tarea: Demostrar.

Lo que importa de $\boldsymbol{w}$ es su dirección, no su magnitud.

$$\boldsymbol{w} \propto \boldsymbol{S}_W^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

Discriminante lineal de Fisher

Algoritmo del perceptrón (1)

- Es un algoritmo para problemas de clasificación de dos clases.
- El vector de entrada es transformado por medio de una función no lineal en un nuevo vector de características $\phi(\boldsymbol{x})$.
- Este vector luego es usado para construir una función lineal generalizada de la forma

$$y(\boldsymbol{x}) = f\left(\boldsymbol{w}^\top \phi(\boldsymbol{x})\right)$$

- La función $f(\cdot)$ es la función signo.

$$f(a) = \begin{cases} +1, & a \geq 0 \\ -1, & a < 0 \end{cases}$$

- En este algoritmo se asume $t = +1$ para $\mathcal{C}_1$, y $t = -1$ para $\mathcal{C}_2$

Algoritmo del perceptrón (2)

- La función a minimizar se conoce como el criterio del perceptrón

$$E_P(\boldsymbol{w}) = -\sum_{n \in \mathcal{M}} \boldsymbol{w}^\top \phi(\boldsymbol{x}_n) t_n$$

  donde $\mathcal{M}$ denota el conjunto de patrones incorrectamente clasificados.

- Aplicando el algoritmo de gradiente descendiente estocástico a esta función, se tiene

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \eta \nabla E_P(\boldsymbol{w}) = \boldsymbol{w}^{(\tau)} + \eta \phi(\boldsymbol{x}_n) t_n$$

  donde η se conoce como la razón de aprendizaje, y τ indexa los pasos del algoritmo.

Algoritmo del perceptrón (3)

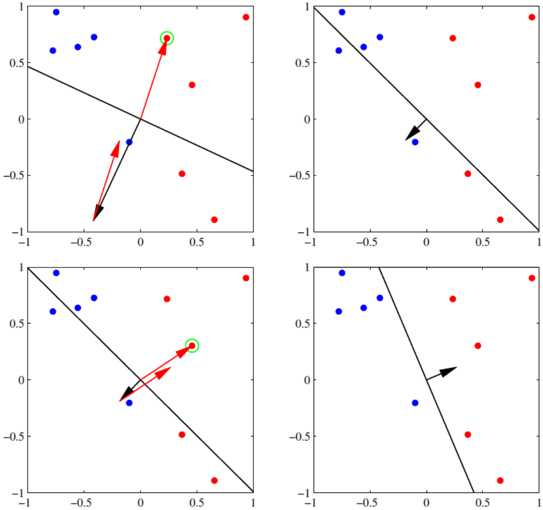
Básicamente

- Se empieza con un vector $\boldsymbol{w}$ inicial. (P.ej. $\boldsymbol{w} = [0, \cdots, 0]$)
- Se empieza a evaluar, uno por uno, cada uno de los datos del conjunto de entrenamiento.
- Si el vector $\boldsymbol{w}$ clasifica un dato de forma equivocada, se ajusta el vector $\boldsymbol{w}$ en la dirección "correcta".
- Si el vector $\boldsymbol{w}$ ha dejado de cambiar, se detiene el algoritmo.

Algoritmo del perceptrón (4)

- Si los datos son linealmente separables, el algoritmo converge.
- Sin embargo, la solución no es única.
- El algoritmo depende del orden en el que los datos son procesados.
- Separar los datos de entrenamiento no implica una separación de datos no vistos (de evaluación).

Algoritmo del perceptrón (5)



Ejercicios Laboratorio

En un cuaderno (notebook) de jupyter responda a las siguientes preguntas con ejemplos concretos de implementación sobre Python 3. Envie su notebook al correo `amalvarezme@unal.edu.co`.

- Consultar el funcionamiento (modelo matemático, función de costo y optimización) de los algoritmos de clasificación: `sklearn.naive_bayes.GaussianNB`, `sklearn.linear_model.SGDClassifier`, `sklearn.discriminant_analysis.LinearDiscriminantAnalysis` `sklearn.discriminant_analysis.QuadraticDiscriminantAnalysi` `sklearn.lda.LDA`, y `sklearn.neighbors.KNeighborsClassifier` según sus implementaciones en el paquete `Scikit-Learn` de Python.
- Utilizando la base de datos `LFW PEOPLE`, realice un análisis comparitivo de los métodos de clasificación mediante validación cruzada. Recuerde sintonizar los parámetros libres de cada algoritmo y calcular el acierto, la precisión, exahustividad, el F1 score y la matriz de confusión.

Referencias I

Murphy, K. (2012).
Machine Learning: A Probabilistic Perspective.
The MIT Press. 1st Edition. 2012

Bishop, C. (2006).
Pattern recognition.
Ed. Springer. 2006