

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 Pereira

Departamento de ingeniería eléctrica, electrónica y computación

Universidad Nacional de Colombia-sede Manizales



- 1 Hoja de ruta
- 2 Conceptos y definiciones
- 3 Función discriminante
- 4 Estimación de parámetros de funciones discriminantes

- 1 Hoja de ruta
- 2 Conceptos y definiciones
- 3 Función discriminante
- 4 Estimación de parámetros de funciones discriminantes

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 \mathbf{X} con unas salidas \mathbf{y} , dado un set de pares de entradas-salidas $D = \{(\mathbf{x}_i, y_i)\}$
- $\mathbf{X} \in \mathbb{R}^{N \times P}$
- Las salidas \mathbf{y} pueden ser: o bien valores reales $y_i \in \mathbb{R}^N$, o bien variables categóricas en donde $y_i \in \{1, \dots, C\}$
- En el primer caso hablaríamos de un *sistema de regresión* y en el segundo caso hablaríamos de un *sistema de clasificación*

- **De aprendizaje no supervisado.**

- Objetivo: Dadas unas entradas $\mathbf{X} \in \mathbb{R}^{N \times P}$, 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)

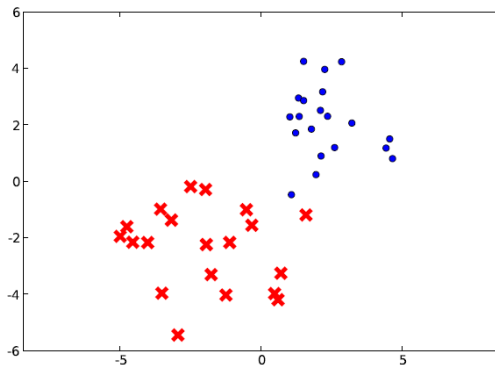
Contenido

- 1 Hoja de ruta
- 2 Conceptos y definiciones**
- 3 Función discriminante
- 4 Estimación de parámetros de funciones discriminantes

Conceptos básicos (1)

Supuestos:

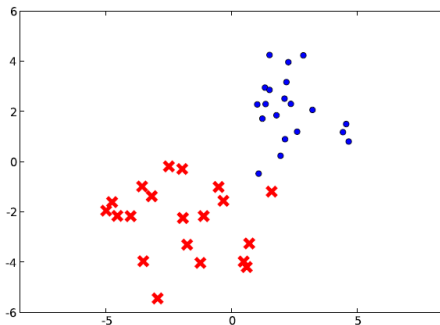
- Cada muestra corresponde a una única clase.
- Los datos conforman espacios linealmente separables.



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

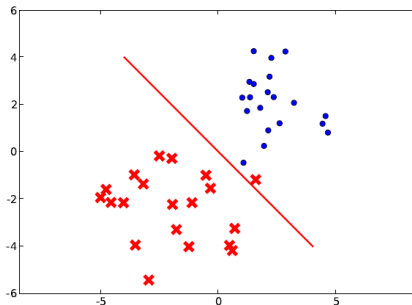
Conceptos básicos (2)

- Cada muestra (instancia) está descrita por un conjunto de números
⇒ 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)

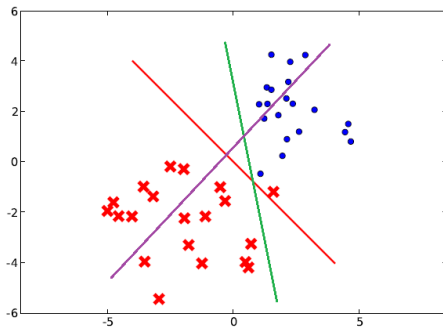
- El objetivo es dividir estos puntos con una línea recta
- Esto es lo que se conoce como **clasificación lineal**



- 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

- 1 Hoja de ruta
- 2 Conceptos y definiciones
- 3 Función discriminante**
- 4 Estimación de parámetros de funciones discriminantes

Función discriminante (1)

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

$$y(\mathbf{x}) : \mathbf{x} \rightarrow k, \quad k \in \{1, \dots, K\}$$

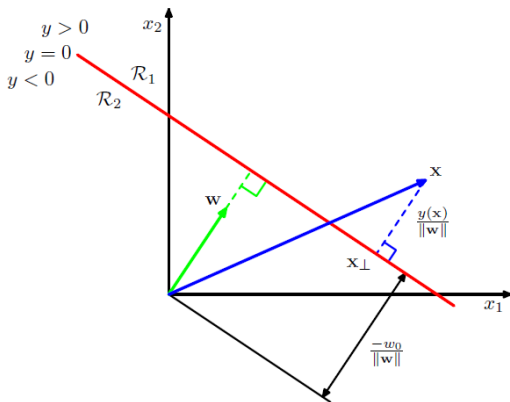
- La función discriminante está dada por:

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

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

Función discriminante (2)

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

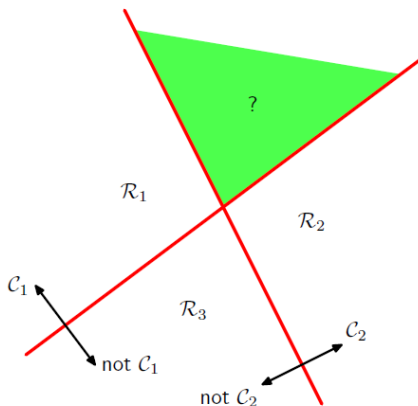


Función discriminante (3)

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

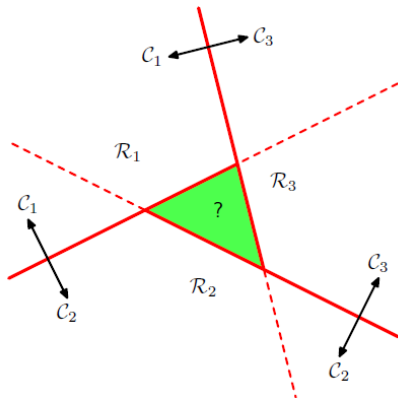
Múltiples clases (1)

- 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:

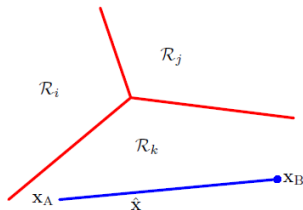


Múltiples clases (3)

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

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

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



- El resultado serán hiperplanos de decisión de la forma

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

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

- 1 Hoja de ruta
- 2 Conceptos y definiciones
- 3 Función discriminante
- 4 Estimación de parámetros de funciones discriminantes

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

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

- Podemos agrupar estos términos empleando notación vectorizada

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

donde

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

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

Mínimos cuadrados (2)

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

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

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

$\tilde{\mathbf{X}}$ será una matriz con vectores fila $\tilde{\mathbf{x}}_n$

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix}$$

- Para todo el set de entrenamiento tenemos entonces

$$\mathbf{Y}(\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}\tilde{\mathbf{W}}$$

- El objetivo entonces es escoger $\tilde{\mathbf{W}}$ que minimice

$$\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T}$$

Mínimos cuadrados (3)

- Para esto, minimizaremos la función de error cuadrático

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

En este punto nos conviene saber algo más de álgebra lineal

Propiedad de la traza de una matriz

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

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

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

Mínimos cuadrados (4)

- Para minimizar esta expresión, derivamos

$$\begin{aligned}\frac{\partial}{\partial \widetilde{\mathbf{W}}} E_D(\widetilde{\mathbf{W}}) &= \frac{1}{2} \frac{\partial}{\partial \widetilde{\mathbf{W}}} \text{Tr} \left\{ \left(\widetilde{\mathbf{X}} \widetilde{\mathbf{W}} - \mathbf{T} \right)^\top \left(\widetilde{\mathbf{X}} \widetilde{\mathbf{W}} - \mathbf{T} \right) \right\} \\ &= \widetilde{\mathbf{X}}^\top \left(\widetilde{\mathbf{X}} \widetilde{\mathbf{W}} - \mathbf{T} \right)\end{aligned}$$

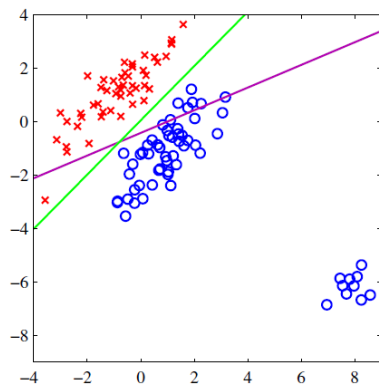
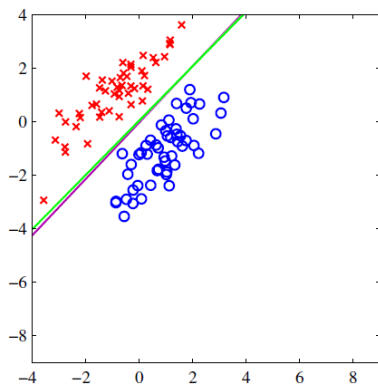
- E igualamos a cero

$$\begin{aligned}\widetilde{\mathbf{W}}_{MSE} &= \left(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} \right)^{-1} \widetilde{\mathbf{X}}^\top \mathbf{T} \\ &= \widetilde{\mathbf{X}}^+ \mathbf{T} \quad \text{donde } \widetilde{\mathbf{X}}^+ \Rightarrow \text{pseudo-inversa de } \widetilde{\mathbf{X}}\end{aligned}$$

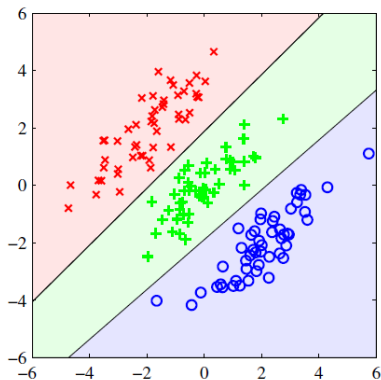
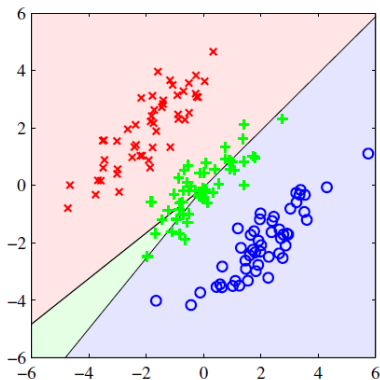
- La función discriminante está dada por

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

Algunos inconvenientes (1)



Algunos inconvenientes (2)



Análisis discriminante de Fisher (1)

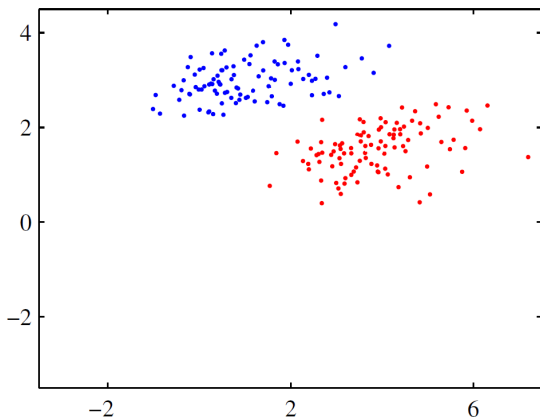
- El objetivo es proyectar los datos a un espacio de menor dimensionalidad donde la clasificación sea más sencilla.
- Sea $\mathbf{x} \in \mathbb{R}^D$.
- Se proyecta a una dimensión usando

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

- 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 \mathbf{w} de manera que maximice la separabilidad de las clases.

Análisis discriminante de Fisher (2)

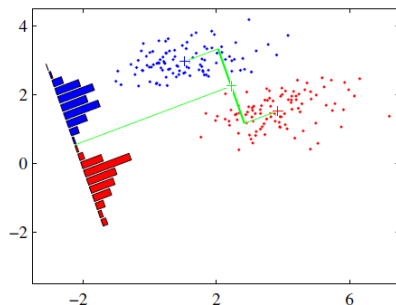
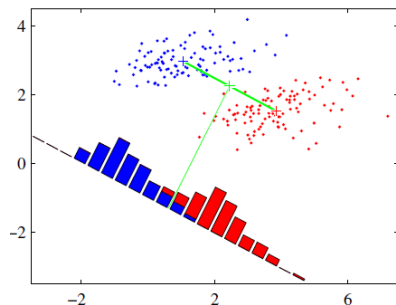
Consideremos inicialmente un problema de dos clases



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

Análisis discriminante de Fisher (3)

Podríamos separar las clases de diversas formas



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

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

Análisis discriminante de Fisher (4)

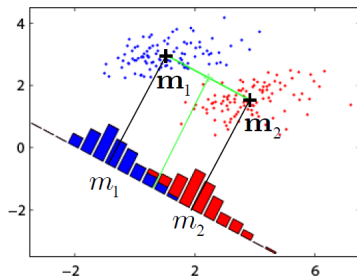
Una medida de separación entre las clases es

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

donde

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

Se debe escoger \mathbf{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 $y_n = \mathbf{w}^\top \mathbf{x}_n$

- El criterio de Fisher se define como

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

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

Análisis discriminante de Fisher (6)

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

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

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

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

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

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

Análisis discriminante de Fisher (7)

- Derivando

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

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

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

Tarea: Demostrar.

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

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

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 $\phi(\mathbf{x})$.
- Este vector luego es usado para construir una función lineal generalizada de la forma

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

- 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

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

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

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

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

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

Algoritmo del perceptrón (3)

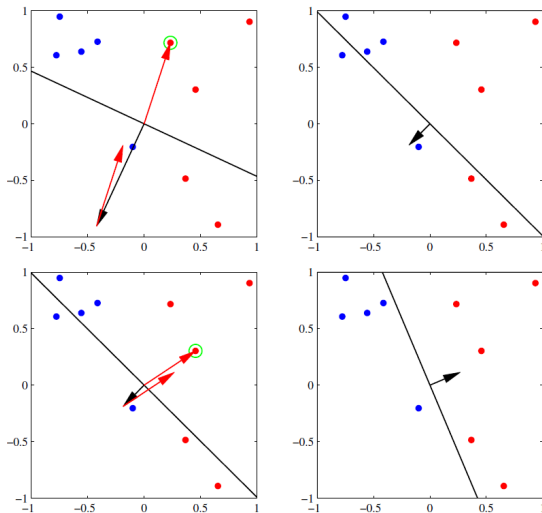
Básicamente

- Se empieza con un vector \mathbf{w} inicial. (P.ej. $\mathbf{w} = [0, \dots, 0]$)
- Se empieza a evaluar, uno por uno, cada uno de los datos del conjunto de entrenamiento.
- Si el vector \mathbf{w} clasifica un dato de forma equivocada, se ajusta el vector \mathbf{w} en la dirección “correcta”.
- Si el vector \mathbf{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. Envíe 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.QuadraticDiscriminantAnalysis`
`sklearn.lda.LDA`, y `sklearn.neighbors.KNeighborsClassifier`
según sus implementaciones en el paquete Scikit-Learn de Python.
- Utilizando la base de datos MNIST, realice un análisis comparativo de los métodos de clasificación mediante validación cruzada. Recuerde sintonizar los parámetros libres de cada algoritmo y calcular el acierto, la precisión, exhaustividad y el F1 score.

Referencias I



Murphy, K. (2012).

Machine Learning: A Probabilistic Perspective.

The MIT Press. 1st Edition. 2012



Bishop, C. (2006).

Pattern recognition.

Ed. Springer. 2006