Redes neuronales y clasificadores de margen máximo

Lección 10

Dr. Pablo Alvarado Moya

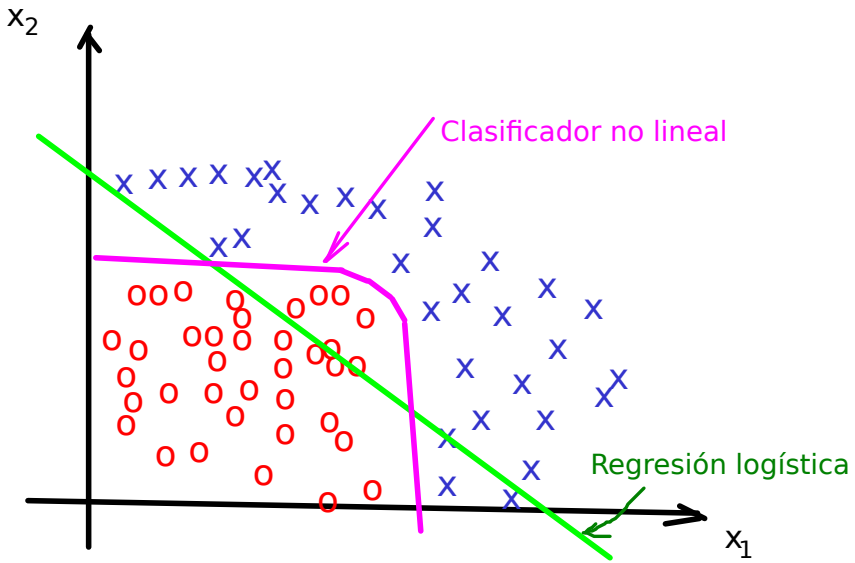
CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

Contenido

Fronteras de decisión no lineales

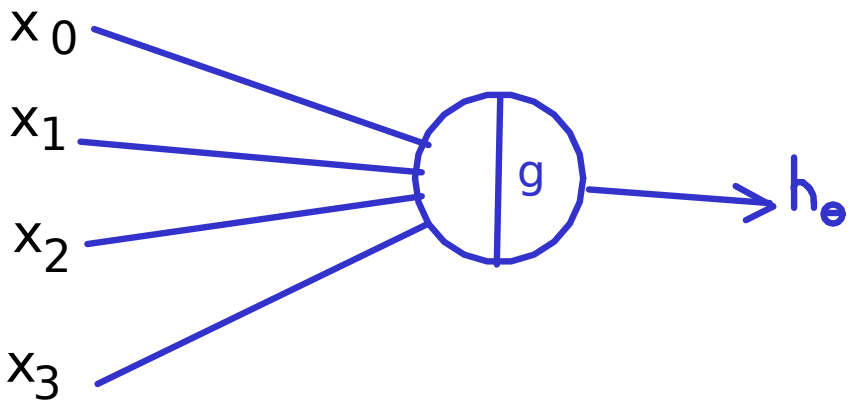
- Primer clasificador que vimos: regresión logística (RL)
- En RL hipótesis: $h_{\underline{\theta}}(\underline{\mathbf{x}}) = 1/(1 + e^{\underline{\theta}^T \underline{\mathbf{x}}})$
- Predicción usa $h_{\underline{\theta}}(\underline{\mathbf{x}}) > 1/2$ para predecir una determinada clase, es decir $\underline{\theta}^T \underline{\mathbf{x}} = 0$ es la **frontera de decisión**
- Esta hipótesis solo permite partir espacio de entrada en dos con una línea recta



- ¿Cómo podemos producir clasificadores con frontera de decisión no lineal?

Redes neuronales

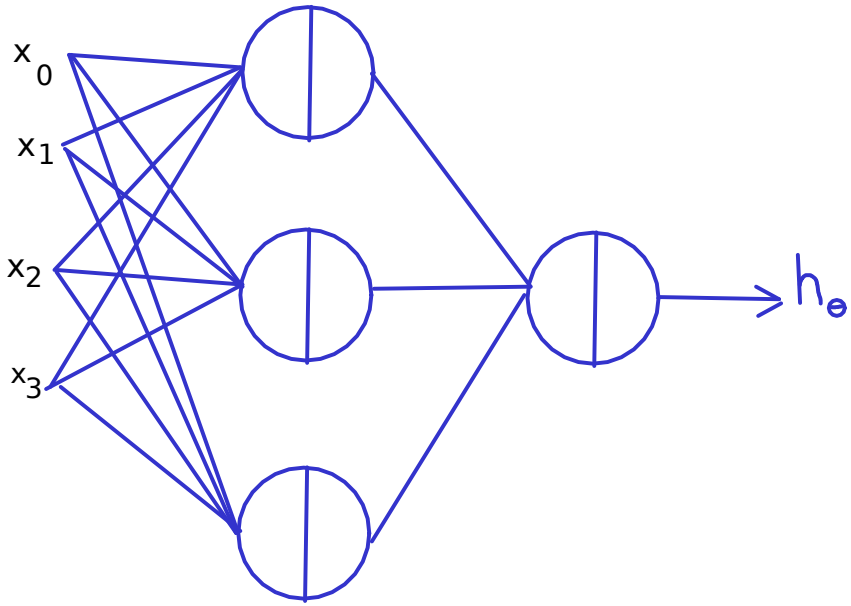
- Caso de regresión logística coincide con una red neuronal clásica de una capa y una salida:



- La salida de la neurona artificial es $h_{\underline{\theta}}(\underline{\mathbf{x}})$
- A $g(\cdot)$ se le denomina función de activación (en regresión logística la función sigmoide)
- La entrada de la neurona es $\underline{\mathbf{x}}$
- Los parámetros de la neurona son $\underline{\boldsymbol{\theta}}$
- La entrada a la función de activación es en este caso $\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}$.

No linealidad en red neuronal

- Una forma de alcanzar fronteras no lineales es armando una **red** neuronal



- La introducción de una capa **oculta** permite la no-linealidad de la frontera

Entrenamiento de la red neuronal

- De la capa de entrada, se tienen los valores intermedios

$$\underline{\mathbf{a}} = g(\mathbf{\Theta}\underline{\mathbf{x}})$$

donde las filas de $\mathbf{\Theta}$ corresponden a los parámetros de cada neurona por separado, y la función sigmoide se aplica a cada elemento del vector $\mathbf{\Theta}\underline{\mathbf{x}}$

- La salida de la red es entonces $h_{\mathbf{\Theta},\underline{\boldsymbol{\theta}}^{(2)}}(\underline{\mathbf{x}}) = g(\underline{\boldsymbol{\theta}}^{(2)^T}\underline{\mathbf{a}})$
- Para entrenar la red se minimiza

$$J(\mathbf{\Theta},\underline{\boldsymbol{\theta}}^{(2)}) = \frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - h_{\mathbf{\Theta},\underline{\boldsymbol{\theta}}^{(2)}}(\underline{\mathbf{x}}))^2$$

- La minimización por decenso de gradiente recibe en este contexto el nombre de: **algoritmo de retropropagación**

Convexidad de función de error

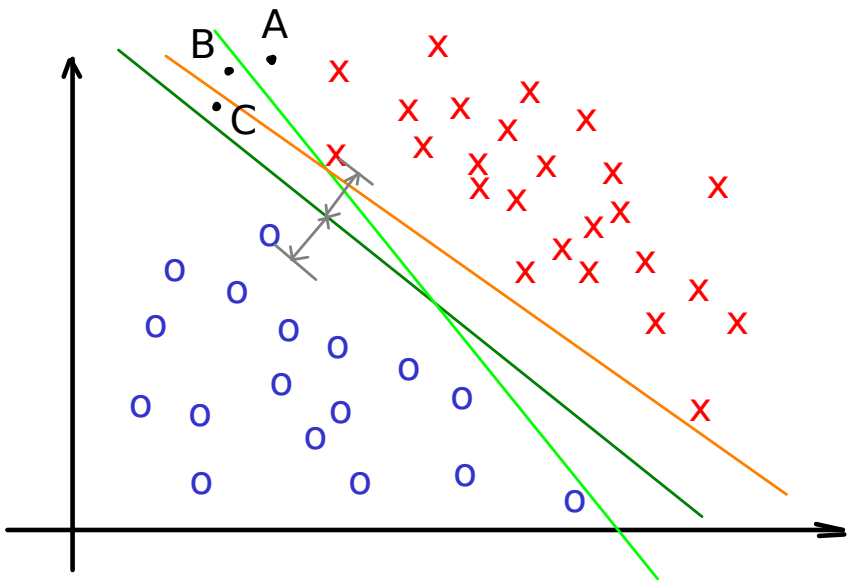
- Mientras que $J(\underline{\theta})$ en las regresiones lineal y logística era convexa cuadrática, en el caso de las redes neuronales por lo general no lo es.
- Esto implica que ninguno de los algoritmos utilizados para encontrar el mínimo puede asegurar convergencia a mínimo global
- El problema de optimización se torna complejo y requiere intervención manual
- En este sentido se prefieren las máquinas de soporte vectorial que veremos más adelante.
- Ver LeNet

Concepto para confiabilidad

- Regresión logística: Algoritmo que calcula $\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}$
  - Predice "1" sii $\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}} \geq 0$
  - Predice "0" sii $\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}} < 0$
- Si $\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}} \gg 0$ alta confiabilidad que $y = 1$
- Si $\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}} \ll 0$ alta confiabilidad que $y = 0$
- Deseable entonces que
  - $\forall i$ tales que $y^{(i)} = 1$, $\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}^{(i)}} \gg 0$
  - $\forall i$ tales que $y^{(i)} = 0$, $\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}^{(i)}} \ll 0$
- Este concepto se asocia al llamado **margen funcional**.

Concepto de discriminación lineal

- Supongamos un conjunto de entrenamiento separable linealmente
- Distintas lineas pueden separar correctamente el conjunto
- Buscamos entonces aquella que geométricamente tiene el mayor **margen geométrico** de separación.



Cambio de notación

- Necesitamos cambiar de notación para acoplarnos a la literatura de SVM
- Vamos a realizar clasificación binaria pero con $y \in \{-1, +1\}$ (en vez de $y \in \{0,1\}$)
- Vamos a parametrizar con $\underline{\mathbf{w}}$ y b en vez de $\underline{\boldsymbol{\theta}}$:

$$h_{\underline{\mathbf{w}},b}(\underline{\mathbf{x}}) = g(\underline{\mathbf{w}}^T \underline{\mathbf{x}} + b)$$

con

$$g(z) = \text{sgn}(z) = \begin{cases} 1 & z \geq 0 \\ -1 & z < 0 \end{cases}$$

- Nótese que ya no necesitamos $x_0 = 1$ y que $b = \theta_0$, $\underline{\mathbf{w}} = [\theta_1, \ldots, \theta_n]^T$

Margen funcional

- El **margen funcional** de un hiperplano $(\underline{\mathbf{w}}, b)$ con respecto al dato de entrenamiento $(\underline{\mathbf{x}}^{(i)}, y^{(i)})$ se **define** como:

$$\hat{\gamma}^{(i)} = y^{(i)}(\underline{\mathbf{w}}^T \underline{\mathbf{x}} + b)$$

- Para lograr que nuestra decisión sea confiable, queremos que:
  - Si $y^{(i)} = +1$, entonces $\underline{\mathbf{w}}^T \underline{\mathbf{x}} + b \gg 0$
  - Si $y^{(i)} = -1$, entonces $\underline{\mathbf{w}}^T \underline{\mathbf{x}} + b \ll 0$
- Si clasificamos correctamente entonces $y^{(i)}(\underline{\mathbf{w}}^T \underline{\mathbf{x}} + b) > 0$
- El margen funcional de **todo** el conjunto de entrenamiento es el peor de todos:

$$\hat{\gamma} = \min_i \hat{\gamma}^{(i)}$$

- Nótese que si se escalan $\underline{\mathbf{w}}$ y b por una constante α, entonces el margen funcional aumenta sin cambiar la frontera de decisión.
- Necesitaremos normalizar $\underline{\mathbf{w}}$ para evitar aumentos sin efectos.

Margen geométrico

- El margen **geométrico** es la distancia entre el punto de entrenamiento y la frontera de decisión:

$$\gamma^{(i)} = y^{(i)} \frac{\underline{\mathbf{w}}^T \underline{\mathbf{x}}^{(i)} + b}{\|\underline{\mathbf{w}}\|} = y^{(i)} \left[\left(\frac{\underline{\mathbf{w}}}{\|\underline{\mathbf{w}}\|} \right)^T \underline{\mathbf{x}}^{(i)} + \frac{b}{\|\underline{\mathbf{w}}\|} \right]$$

- Los márgenes geométrico y funcional se relacionan con:

$$\gamma^{(i)} = \frac{\hat{\gamma}^{(i)}}{\|\underline{\mathbf{w}}\|}$$

- Si $\|\underline{\mathbf{w}}\| = 1$ entonces $\gamma^{(i)} = \hat{\gamma}^{(i)}$
- El margen geométrico es invariante a escalamientos de $\underline{\mathbf{w}}$ y b
- El margen geométrico del conjunto de entrenamiento es

$$\gamma = \min_i \gamma^{(i)}$$

Clasificador de margen máximo
Maximum Margin Classifier

- Precursor de las máquinas de soporte vectorial (SVM)
- El clasificador de margen máximo optimiza

$$\begin{aligned} &\underset{\gamma,\underline{\mathbf{w}},b}{\text{máx}}\gamma \\ &\text{sujeto a } y^{(i)}(\underline{\mathbf{w}}^T\underline{\mathbf{x}}^{(i)}+b)\geq\gamma, \quad i=1,\ldots,m \\ &\|\underline{\mathbf{w}}\|=1 \end{aligned}$$

- La restricción $\|\underline{\mathbf{w}}\|=1$ asegura que los márgenes funcional y geométrico sean iguales
- Solución asegura que $(\underline{\mathbf{w}},b)$ produce el máximo margen geométrico respecto al conjunto de entrenamiento

Segundo planteo de clasificador de margen máximo

- Restricción $\|\underline{\mathbf{w}}\| = 1$ dificulta solución de problema (no es convexa), y planteo no solucionable con optimizadores "estándar"
- Necesitamos replantear el problema a optimizar:

$$\max_{\hat{\gamma},\underline{\mathbf{w}},b} \frac{\hat{\gamma}}{\|\underline{\mathbf{w}}\|}$$
$$\text{sujeto a } y^{(i)}(\underline{\mathbf{w}}^T \underline{\mathbf{x}}^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, m$$

- Con esto maximizamos $\hat{\gamma}/\|\underline{\mathbf{w}}\|$ sujeto a que todos los márgenes funcionales sean al menos $\hat{\gamma}$, sin la restricción $\|\underline{\mathbf{w}}\| = 1$
- La función objetivo $\frac{\hat{\gamma}}{\|\underline{\mathbf{w}}\|}$ es de nuevo no convexa, y tampoco existen métodos listos para resolver esto.

Tercer planteo de clasificador de margen máximo

- Anteriormente vimos que un escalamiento de $\underline{\mathbf{w}}$ y b no afectan la frontera
- Eligiremos el escalamiento que obligue:

$$\hat{\gamma} = 1 \qquad \text{ó equiv.} \qquad \min_i y^{(i)}(\underline{\mathbf{w}}^T \underline{\mathbf{x}}^{(i)} + b) = 1$$

- El problema de optimización es entonces equivalente a:

$$\min_{\gamma, \underline{\mathbf{w}}, b} \frac{1}{2} \|\underline{\mathbf{w}}\|^2$$
$$\text{sujeto a } y^{(i)}(\underline{\mathbf{w}}^T \underline{\mathbf{x}}^{(i)} + b) \geq 1, \quad i = 1, \ldots, m$$

- Este sí es un problema de optimización convexo
- Esto representa al **clasificador de margen óptimo**
- Se puede resolver con optimizadores cuadráticos genéricos

¿Qué sigue?

- Esto podría resolverse así. . .
- Sin embargo, forma **dual** permite algoritmos más eficientes
- Forma dual además permitirá introducir el "truco del kernel"

Resumen

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