Regresión Logística

M.Sc. William Caicedo Torres

Universidad Tecnológica de Bolívar

caicedo77@gmail.com

19 de octubre de 2016

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Machine Learning: Clasificación

- Clasificación: $y \in \{c_1, c_2, \ldots, c_n\}$

Machine Learning: Clasificación

- Clasificación: $y \in \{c_1, c_2, \ldots, c_n\}$
  - La salida del modelo es un valor discreto perteneciente a un conjunto finito.
  - La salida se interpreta como la clase a la que pertenecen las entradas.
  - Ejemplo: Dada una imagen, predecir si la imagen es de un perro, gato, silla, etc.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Machine Learning: Clasificación

- El problema canónico de clasificación involucra 2 clases, es decir que $y \in \{0, 1\}$, donde:
  0: Clase Negativa (ej., tumor benigno)
  1: Clase Positiva (ej., tumor maligno)
- La idea de un modelo de clasificación basado en Machine Learning es encontrar la forma de separar los ejemplos de la clase positiva de los pertenecientes a la clase negativa, de forma automática.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Cómo clasificar los datos automáticamente?

- Consideremos por ejemplo, la función

$$f(x_1, x_2) = -2 - 3x_1 + x_2$$

- Dicha función la podemos expresar de forma más succinta. Si tenemos que

$$\theta = \begin{bmatrix} -2 \\ -3 \\ 1 \end{bmatrix} \quad x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

entonces

$$f(x) = \theta^t x$$

- $f(x) = 0$ es una recta en el plano, y $f(x) > 0$ para todo x **por encima** de la recta, mientras que $f(x) < 0$ para todo x ubicado **por debajo** de ella.

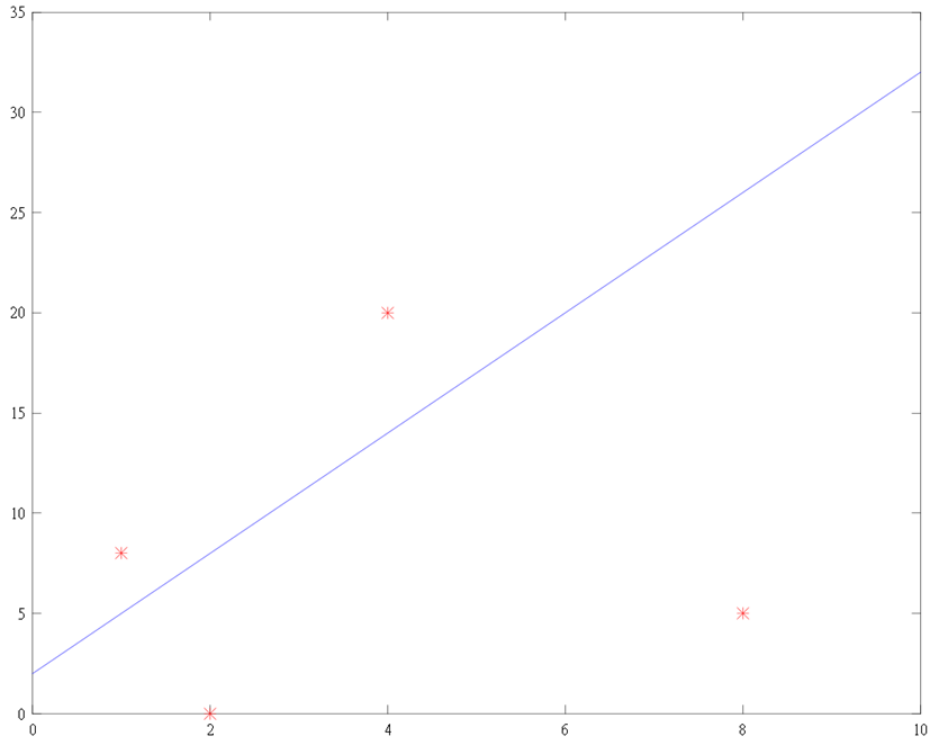
Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Cómo clasificar los datos automáticamente?

$$f(x_1, x_2) = -2 - 3x_1 + x_2$$

x_1	x_2	$f(x_1, x_2)$
1	8	3
2	0	-8
4	20	6
8	5	-21

Observe el signo de $f(x)$ de acuerdo a si x está por encima o por debajo de la recta.



Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

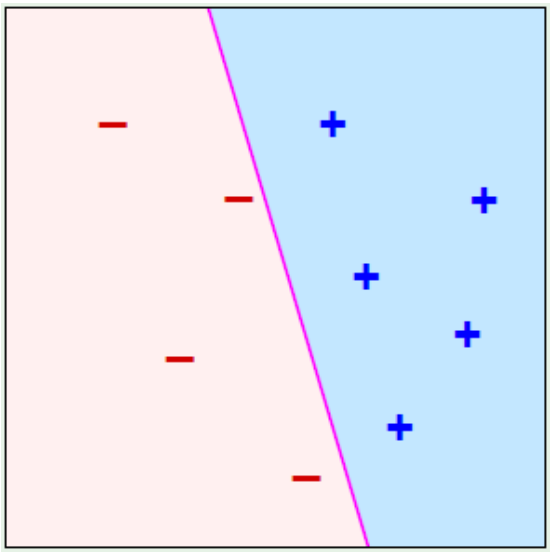
Cómo clasificar los datos automáticamente?

- Vemos que podemos separar el plano en 2 regiones, una donde la función produce una salida positiva, otra donde dicha salida es negativa.
- Esta es la base de un algoritmo de clasificación: Al ubicar cuidadosamente la función, podemos discriminar x de acuerdo a sus características. A la hipótesis $f(x)$ se le llama **Frontera de Decisión**.

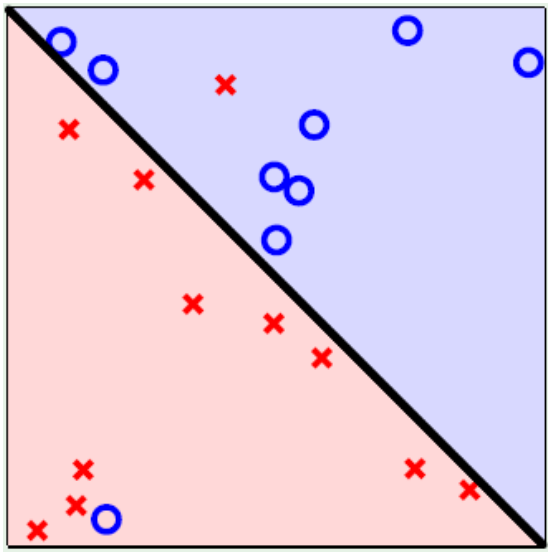
Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Separabilidad Lineal

- Se dice que un problema de clasificación es Linealmente Separable cuando es posible clasificar correctamente **TODAS** las entradas, usando una hipótesis lineal como frontera de decisión.



Problema lineálmente separable.



Problema no lineálmente separable.

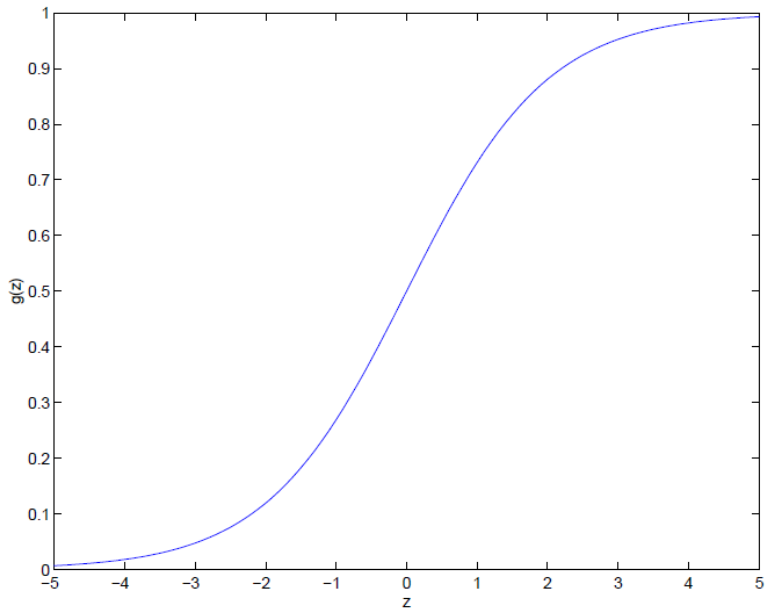
Función Logística Sigmoide

- Como vimos anteriormente, nuestra hipótesis lineal separa el plano en 2 clases.
- Sin embargo, la salida de $f(x)$ está en $(-\infty, +\infty)$.
- Pero el problema canónico de clasificación tiene que $y \in \{0, 1\}$.
- Cómo restringimos la salida de nuestro modelo de clasificación al intervalo $y \in (0, 1)$?.
- La respuesta: Usamos la función Logística Sigmoide.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Función Logística Sigmoide

- $g(z) = \frac{1}{1+e^{-z}}$
- La salida de $g(z)$ siempre estará en $(0, 1)$
- Si hacemos $z = \theta^t x$, entonces la salida de nuestro modelo de clasificación estará restringida efectivamente a $(0, 1)$
- Un plus: Podemos interpretar la salida del modelo como la probabilidad de que la entrada pertenezca a la clase positiva.



Función Logistica Sigmoide

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Regresión Logística

- Presentamos entonces el algoritmo de Regresión Logística.
- Para una entrada x, la probabilidad de que dicha entrada pertenezca a la clase positiva esta dada por:

$$P(y = 1|x; \theta) = \frac{1}{1 + e^{-\theta^t x}}$$

- La pregunta como siempre es, cómo hallamos el vector θ que nos de la mejor clasificación posible, teniendo en cuenta que la frontera de decisión es lineal?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Regresión Logística

- Presentamos entonces el algoritmo de Regresión Logística.
- Para una entrada x, la probabilidad de que dicha entrada pertenezca a la clase positiva esta dada por:

$$P(y = 1|x; \theta) = \frac{1}{1 + e^{-\theta^t x}}$$

- La pregunta como siempre es, cómo hallamos el vector θ que nos de la mejor clasificación posible, teniendo en cuenta que la frontera de decisión es lineal?
- Las técnicas de optimización acuden a nuestro rescate!

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Maximum Likelihood Estimation

- Vamos a asumir que

$$P(y = 1|x; \theta) = h_\theta(x)$$

$$P(y = 0|x; \theta) = 1 - h_\theta(x)$$

Maximum Likelihood Estimation

- Vamos a asumir que

$$P(y = 1|x; \theta) = h_\theta(x)$$
$$P(y = 0|x; \theta) = 1 - h_\theta(x)$$

- De forma más compacta:

$$P(y|x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Maximum Likelihood Estimation

- Vamos a asumir que

$$P(y = 1|x; \theta) = h_\theta(x)$$
$$P(y = 0|x; \theta) = 1 - h_\theta(x)$$

- De forma más compacta:

$$P(y|x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}$$

- Asumiendo m ejemplos de entrenamiento independientes, la función de probabilidad (Likelihood) asociada sería:

$$L(\theta) = p(\vec{y}|x; \theta) = \prod_{i=1}^{m} p(y_i|x_i; \theta)$$
$$= \prod_{i=1}^{m} (h_\theta(x_i))^{y_i} (1 - h_\theta(x_i))^{1-y_i}$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Maximum Likelihood Estimation

- Meta: Queremos maximizar la probabilidad de que los ejemplos de entrenamiento caigan del lado correcto de la frontera de decisión.
- Por lo anterior, queremos maximizar nuestra función de probabilidad $L(\theta)$ usando el criterio de la primera derivada.
- Para simplificar la derivación, vamos a maximizar el logaritmo de la función de probabilidad (Log-Likelihood). El logaritmo introduce una propiedad interesante en este caso: Concavidad!
- Entonces:

$$\ell = \log L(\theta)$$

$$\ell = \sum_{i=1}^{m} y_i(\log h(x_i)) + (1 - y_i) log(1 - h(x_i))$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Cross-Entropy Error

- Lastimosamente no existe una solución en forma cerrada para este problema de optimización.
- Debemos recurrir a un método iterativo: Gradiente Ascendente.
- Pero antes de proceder a calcular el gradiente, consideremos lo siguiente: Si anteponemos un signo negativo y la fracción $\frac{1}{m}$, el problema pasaría a ser un problema de minimización:

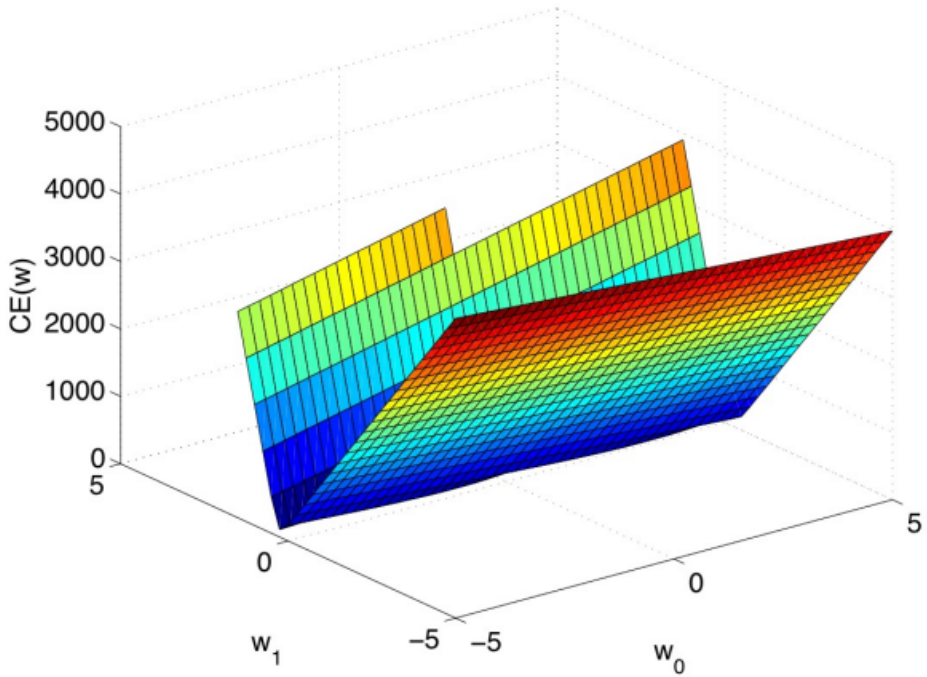
$$min_\theta \ J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y_i(\log h(x_i)) + (1-y_i)log(1-h(x_i))$$

- A la función objetivo anterior se le conoce como la función de error de entropía cruzada (Cross-Entropy Error).

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Cross-Entropy Error

- La función de entropía cruzada se usa para cuantificar el error en el que incurre un clasificador.
- Tiene una propiedad bastante afortunada: Es convexa!
- Penaliza al clasificador de manera proporcional a cada error cometido.



Cross-Entropy Loss

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Gradiente Descendente aplicado a la Regresión Logística

- La aplicación del gradiente descendente al entrenamiento en la regresión logística es sorprendentemente similar al caso de la regresión lineal.
- El gradiente del error está dado por:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)x_{ij}$$

- Por lo que la regla de actualización de θ en cada iteración del método sería:

$$\theta j = \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)x_{ij}$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Derivación del gradiente

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y_i(\log h(x_i)) + (1-y_i)log(1-h(x_i))$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m} y_i \frac{1}{h(x_i)}\frac{\partial h(x_i)}{\partial \theta_i} + (1-y_i)\frac{1}{1-h(x_i)}\frac{\partial (1-h(x_i))}{\partial \theta_i}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m} y_i \frac{1}{h(x_i)}\frac{\partial h(x_i)}{\partial \theta_i} - (1-y_i)\frac{1}{1-h(x_i)}\frac{\partial (h(x_i))}{\partial \theta_i}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m} \left(y_i \frac{1}{h(x_i)} - (1-y_i)\frac{1}{1-h(x_i)} \right) \frac{\partial (h(x_i))}{\partial \theta_i}$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Derivada de la función logística sigmoide

Ahora veremos como podemos beneficiarnos de la particular estructura de la derivada del la función logística sigmoide.

$$g(z) = \frac{1}{1+e^{-z}} = (1+e^{-z})^{-1}$$

$$\frac{dg}{dz} = e^{-z}(1+e^{-z})^{-2}$$

$$\frac{dg}{dz} = \frac{e^{-z}}{(1+e^{-z})^2}$$

$$\frac{dg}{dz} = \frac{1}{(1+e^{-z})}\frac{e^{-z}}{(1+e^{-z})}$$

$$\frac{dg}{dz} = g(z).\frac{-1+1+e^{-z}}{(1+e^{-z})}$$

$$\frac{dg}{dz} = g(z)\left(\frac{1+e^{-z}}{(1+e^{-z})} - \frac{1}{(1+e^{-z})}\right)$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Derivada de la función logística sigmoide

$$\frac{dg}{dz} = g(z)\left(1 - \frac{1}{(1+e^{-z})}\right)$$

$$\frac{dg}{dz} = g(z)(1 - g(z))$$

Es muy conveniente que podamos calcular la derivada de la función logística sigmoide en términos de la propia función logistica sigmoide, desde el punto de vista del desempeño computacional.

Derivación del gradiente

Entonces, en nuestro caso tenemos que:

$$h(x) = g(\theta^t x) = \frac{1}{1 + e^{-\theta^t x}}$$

$$\frac{\partial g(\theta^t x)}{\partial \theta_j} = g(\theta^t x)(1 - g(\theta^t x))\frac{\partial(\theta^t x)}{\partial \theta_j}$$

Reemplazando en nuestro gradiente tenemos:

$$\frac{\partial J(\theta)}{\partial \theta_j} =$$

$$-\frac{1}{m}\sum_{i=1}^{m}\left(y_i \frac{1}{g(\theta^t x_i)} - (1 - y_i)\frac{1}{1 - g(\theta^t x_i)}\right) g(\theta^t x_i)(1-g(\theta^t x_i))\frac{\partial(\theta^t x_i)}{\partial \theta_j}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left(y_i(1 - g(\theta^t x_i)) - (1 - y_i)g(\theta^t x_i)\right)\frac{\partial(\theta^t x_i)}{\partial \theta_j}$$

Derivación del gradiente

Por tanto:

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i - y_i g(\theta^t x_i) - g(\theta^t x_i) + y_i g(\theta^t x_i)\right)\frac{\partial(\theta^t x_i)}{\partial \theta_j}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i - g(\theta^t x_i)\right)\frac{\partial(\theta^t x_i)}{\partial \theta_j}$$

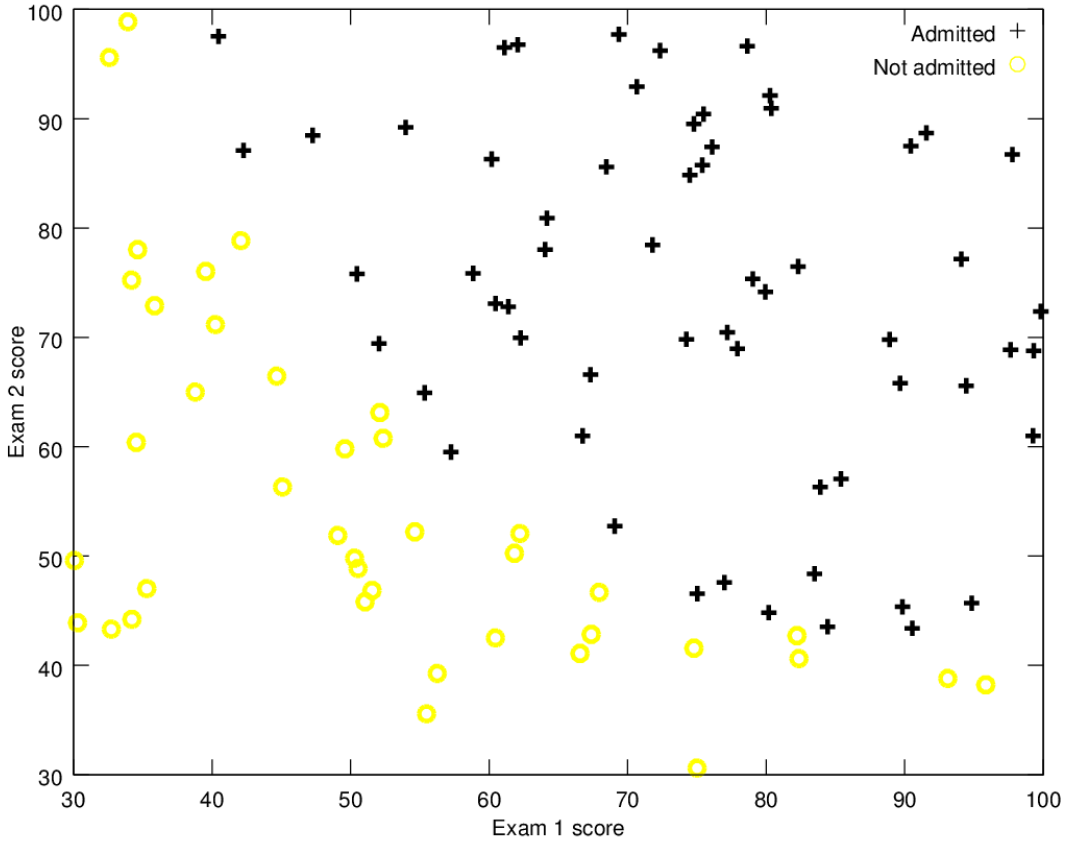
$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)x_{ij}$$

Ejemplo Regresión Logística

- Tenemos las calificaciones de m estudiantes en 2 exámenes.
- Queremos predecir la probabilidad de admisión de un estudiante a una universidad.
- La idea es desarrollar un modelo de Regresión Logística para tal fin.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Ejemplo Regresión Logística



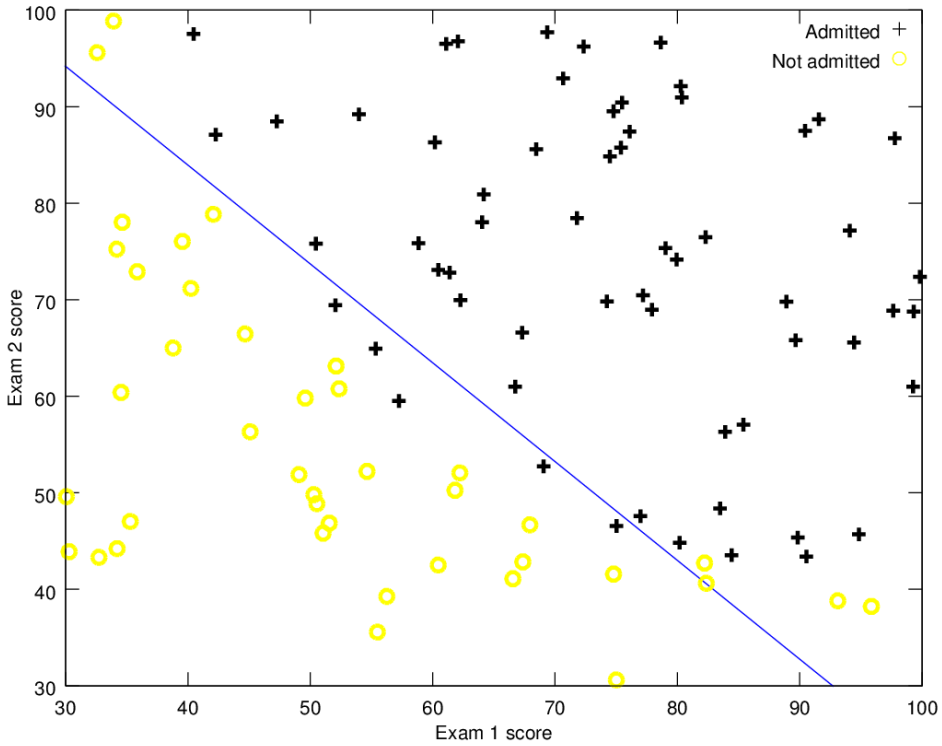
Datos de los estudiantes (notas en los exámenes contra aceptación)

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Ejemplo Regresión Logística

- El resultado el entrenamiento es:

$$\theta = \begin{bmatrix} -25,1612 \\ 0,2062 \\ 0,2014 \end{bmatrix}$$



Datos de los estudiantes (notas en los exámenes contra aceptación)

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Ejemplo Regresión Logística

- Una vez tenemos el modelo de clasificación entrenado, podemos usarlo para predecir la probabilidad de aceptación para un nuevo estudiante, dadas sus notas.
- Ejemplo: Si la nota 1 del estudiante es 45, y la nota 2 es 85, la probabilidad de aceptación para dicho estudiate es:

$$\theta^t x = \begin{bmatrix} -25,1612, 0,2062, 0,2014 \end{bmatrix} * \begin{bmatrix} 1 \\ 45 \\ 85 \end{bmatrix}$$

$$y = 0,776289$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Muchas gracias!

Preguntas?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS