# Clasificación

## Lección 08

Dr. Pablo Alvarado Moya

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

II Semestre, 2019

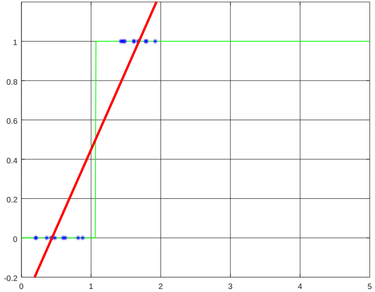
# Contenido

# Clasificación

- **Regresión**: $y = h_{\underline{\theta}}(\underline{\mathbf{x}})$ con $y \in \mathbb{R}$ continuo.
- **Clasificación**: $y = h_{\underline{\theta}}(\underline{\mathbf{x}})$ como antes, pero $y \in \mathcal{C}$ es discreto y posiblemente no ordenado
- Iniciaremos con problema de **clasificación binario** $y \in \{0,1\}$
- Ejemplos:
  - clasificación de correo-e como spam (1: clase positiva) o no-spam (0: clase negativa)
  - paciente tiene enfermedad o no tiene enfermedad
  - una máquina fallará o no fallará dadas entradas de sensores
- Dado $\underline{\mathbf{x}}^{(i)}$, el valor deseado correspondiente $y^{(i)}$ se llama **etiqueta**

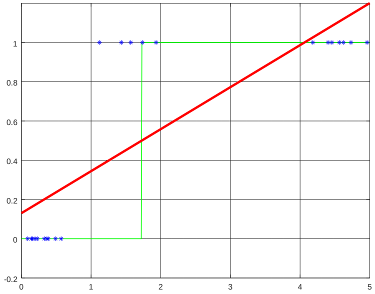
# Regresión lineal como clasificador

- Podríamos usar regresión, ignorando que $y$ es discreta:



# Regresión lineal como clasificador

- Podríamos usar regresión, ignorando que $y$ es discreta, **pero no funciona bien**:



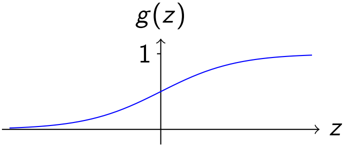
# Regresión logística

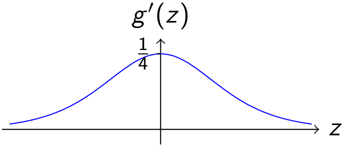
- Es mejor buscar una hipótesis $h(\underline{\mathbf{x}}) \in [0,1]$
- Elijamos

$$h_{\underline{\boldsymbol{\theta}}} = g(\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}) = \frac{1}{1 + e^{-\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}}}$$

con la función sigmoide / función logística

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



# Planteo probabilístico (1)

- Vamos a usar planteo probabilístico para derivar solución
- Supongamos que la hipótesis cumple

$$P(y = 1|\underline{\mathbf{x}}; \underline{\boldsymbol{\theta}}) = h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}})$$

y entonces además que

$$P(y = 0|\underline{\mathbf{x}}; \underline{\boldsymbol{\theta}}) = 1 - h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}})$$

que se pueden combinar (recuerdese Bernoulli)

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

# Planteo probabilístico (2)

- La verosimilitud, igual que caso de regresión, con i.i.d. es:

$$
\begin{aligned}
L(\underline{\boldsymbol{\theta}}) = P(\underline{\mathbf{y}}|\mathbf{X};\underline{\boldsymbol{\theta}}) &= \prod_{i=1}^{m} P(y^{(i)}|\underline{\mathbf{x}}^{(i)};\underline{\boldsymbol{\theta}}) \\
&= \prod_{i=1}^{m} h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)})^{y^{(i)}}(1 - h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)}))^{1-y^{(i)}}
\end{aligned}
$$

# Planteo probabilístico (3)

- Queremos maximizar esta verosimilitud, lo que es más fácil a través de la verosimilitud logarítmica

$$\begin{aligned}\ell(\underline{\boldsymbol{\theta}}) &= \ln L(\underline{\boldsymbol{\theta}}) \\ &= \sum_{i=1}^{m} \ln \left( h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)})^{y^{(i)}} (1 - h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)}))^{1-y^{(i)}} \right) \\ &= \sum_{i=1}^{m} y^{(i)} \ln(h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)})) + (1 - y^{(i)}) \ln(1 - h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)}))\end{aligned}$$

- Podemos maximizar esto utilizando ascenso de gradiente:

$$\underline{\boldsymbol{\theta}} \leftarrow \underline{\boldsymbol{\theta}} + \alpha \nabla_{\underline{\boldsymbol{\theta}}} \ell(\underline{\boldsymbol{\theta}})$$

# Planteo probabilístico (4)

- Si derivamos $\partial\ell(\underline{\boldsymbol{\theta}})/\partial\theta_j$ llegamos con $g'(z) = g(z)(1-g(z))$ a

$$\frac{\partial}{\partial\theta_j}\ell(\underline{\boldsymbol{\theta}}) = (y - h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}))x_j$$

- Con lo anterior obtenemos para el ascenso estocástico de gradiente:

$$\theta_j \leftarrow \theta_j + \alpha(y^{(i)} - h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)}))x_j^{(i)}$$

- ¡Es exactamente la misma expresión que para los mínimos cuadrados ordinarios (OLS)! (diferentes $h_{\underline{\boldsymbol{\theta}}}(\cdot)$)

# El perceptron (1)

- Primera propuesta de "red neuronal"
- 1957 Frank Rosenblatt, laboratorio aeronáutico de Cornell

# El perceptron (2)

- En regresión logística usamos $g(z)$ sigmoidal
- En el perceptron se usa $g(z) = u(z)$ (escalon unitario)
- En ambas $h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}) = g(\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}})$: hipótesis mapea a [0,1]
- La regla de actualización estocástica es similar al caso anterior:

  $$\theta_j \leftarrow \theta_j + \alpha(y^{(i)} - h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)}))x_j^{(i)}$$

  que difieren **mucho** en estilo de aprendizaje por la forma de $h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}})$
- Perceptron no tiene justificación probabilística

# Modelos lineales generalizados

# Modelos lineales generalizados

- Hemos visto dos tipos de problemas:
  - Regresión lineal con mínimos cuadrados: $y|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}} \sim \mathcal{N}(\underline{\boldsymbol{\mu}},\sigma^2)$
  - Regresión logística: $y|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}} \sim \text{Ber}(\phi)$
- ¿Por qué en ambos problemas llegamos a la misma regla de actualización de $\underline{\boldsymbol{\theta}}$:

$$\theta_j \leftarrow \theta_j + \alpha(y^{(i)} - h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)}))x_j^{(i)} \quad ?$$

- Ambos métodos pertenecen a los Modelos Lineales Generalizados (GLM)

# Familias de distribuciones

- Podemos ver las distribuciones con sus parámetros como **familias**:
  - $\mathrm{Ber}(\phi)$: $P(y=1;\phi)=\phi$
  - $\mathcal{N}(\mu,\sigma^2)$: $p(x;\mu,\sigma)=\dfrac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- Esto es: cada instancia de parámetros produce una distribución particular
- Mostraremos que estos son casos especiales de la **familia exponencial**

# La familia exponencial

- La familia exponencial incluye las distribuciones que se pueden expresar:

$$p(y; \underline{\boldsymbol{\eta}}) = b(y) \exp\left(\underline{\boldsymbol{\eta}}^T \underline{\mathbf{T}}(y) - a(\underline{\boldsymbol{\eta}})\right)$$

- $\underline{\boldsymbol{\eta}}$ es el **parámetro natural** (o canónico) de la distribución
- $\underline{\mathbf{T}}(y)$ es el **estadístico suficiente**
  En casos aquí, se cumple $\underline{\mathbf{T}}(y) = y$
- $a(\underline{\boldsymbol{\eta}})$ es la **función de partición logarítmica**
- Usualmente $e^{-a(\underline{\boldsymbol{\eta}})}$ es constante de normalización
- Elección fija de $\underline{\mathbf{T}}$, $a$ y $b$ define una **familia** (o conjunto) de distribuciones parametrizadas con $\underline{\boldsymbol{\eta}}$

# Caso de distribución de Bernoulli (1)

- Distribución de Bernoulli:

$$
\begin{aligned}
p(y;\phi) = \phi^y(1-\phi)^{1-y} &= \exp(\ln(\phi^y(1-\phi)^{1-y})) \\
&= \exp(y\ln\phi + (1-y)\ln(1-\phi)) \\
&= \exp\left(\underbrace{\left(\ln\left(\frac{\phi}{1-\phi}\right)\right)}_{\eta}\underbrace{y}_{\underline{\mathbf{T}}(y)} + \underbrace{\ln(1-\phi)}_{-a(\underline{\boldsymbol{\eta}})}\right)
\end{aligned}
$$

con lo que se deriva el parámetro natural

$$
\underline{\boldsymbol{\eta}} = \eta = \ln\left(\frac{\phi}{1-\phi}\right)
$$

# Caso de distribución de Bernoulli (2)

- Despejando $\phi$ en términos de $\underline{\boldsymbol{\eta}}$ resulta en:

$$\phi = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}}$$

¡que es la función logística que usamos anteriormente!

- Además,

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -\ln(1-\phi) = \ln(1+e^{\eta}) \\ b(y) &= 1 \end{aligned}$$

# Caso de distribución normal (1)

- Para la interpretación probabilística de regresión lineal, la varianza $\sigma^2$ no tuvo efecto
- Vamos a simplificar caso asumiendo varianza $\sigma^2 = 1$
- Para la distribución gaussiana entonces:

$$
\begin{aligned}
p(y;\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\mu)^2\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \cdot \exp\left(\mu y - \frac{1}{2}\mu^2\right)
\end{aligned}
$$

# Caso de distribución normal (2)

- El gaussiano entonces está en la familia exponencial con

$$
\begin{aligned}
\eta &= \mu \\
\underline{\mathbf{T}}(y) &= y \\
a(\eta) &= \frac{\mu^2}{2} = \frac{\eta^2}{2} \\
b(y) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)
\end{aligned}
$$

# Otras distribuciones en la familia exponencial

- Muchas otras distribuciones son parte de la familia exponencial:
  - Multinoulli
  - Poison
  - Gamma
  - Exponencial
  - Beta
  - Dirichlet
  - ...
- Si logramos hacer una derivación para la familia exponencial, entonces ¡cualquiera de las distribuciones anteriores podrá utilizarse con el método general!

# Modelos lineales generalizados

- Vamos a suponer:
  1. $y|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}} \sim \text{FamiliaExponencial}(\underline{\boldsymbol{\eta}})$
  2. Dado $\underline{\mathbf{x}}$, tarea es predecir $\mathrm{E}[\underline{\mathbf{T}}(y)|\underline{\mathbf{x}}]$

     Con frecuencia tomamos $\underline{\mathbf{T}}(y) = y$, con lo que $h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}) = \mathrm{E}[y|\underline{\mathbf{x}}]$
  3. (Criterio de diseño): parámetro natural $\underline{\boldsymbol{\eta}}$ relacionado **linealmente** con entrada $\underline{\mathbf{x}}$: $\eta = \underline{\boldsymbol{\theta}}^T\underline{\mathbf{x}}$

     Si $\underline{\boldsymbol{\eta}}$ es un vector entonces sus componentes $\eta_i = \underline{\boldsymbol{\theta}}_i^T\underline{\mathbf{x}}$.

Caso 1: Mínimos cuadrados ordinarios (OLS)

# Mínimos cuadrados ordinarios

- OLS (*ordinary least squares*) es un caso particular de GLM (*generalized linear models*)
- Considere la **variable de respuesta** continua $y$, con

$$y|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}} \sim \mathcal{N}(\underline{\boldsymbol{\mu}},\sigma^2)$$

- Como la gaussiana pertenece a la familia de distribuciones exponenciales

$$h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}) = \mathrm{E}[y|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}}] = \underline{\boldsymbol{\mu}} = \underline{\boldsymbol{\eta}} = \underline{\boldsymbol{\theta}}^T\underline{\mathbf{x}}$$

# Caso 2: Regresión logística (LR)

# Regresión logística

- La regresión logística hacemos clasificación binaria $y \in \{0,1\}$
- En este caso tiene sentido utilizar la familia de distribuciones de Bernoulli para caracterizar la distribución de $y|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}}$:

$$y|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}} \sim \mathrm{Ber}(\phi)$$

- Con eso entonces $\mathrm{E}[y|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}}] = p(y=1|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}}) = \phi$
- Con las derivaciones anteriores:

$$h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}) = \mathrm{E}[y|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}}] = \phi = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\underline{\boldsymbol{\theta}}^T\underline{\mathbf{x}}}}$$

- Nótese que esto es una justificación natural para haber seleccionado la función logística en la hipótesis de la regresión logística, consecuencia de utilizar distribución de Bernoulli.

# Funciones canónicas

- $g(\underline{\boldsymbol{\eta}}) = \mathsf{E}[\underline{\mathbf{T}}(y); \underline{\boldsymbol{\eta}}]$ se denomina **función canónica de respuesta**
- $g^{-1}(\underline{\boldsymbol{\eta}})$ se denomina **función canónica de enlace**
- Para la familia gaussiana la función canónica de respuesta es la identidad, pues $\mathsf{E}[y|\underline{\mathbf{x}}; \underline{\boldsymbol{\theta}}] = \eta$
- Para la familia Bernoulli la función canónica de respuesta es la función logística, pues $\mathsf{E}[y|\underline{\mathbf{x}}; \underline{\boldsymbol{\theta}}] = 1/\left(1 + e^{-\eta}\right)$

# Caso 3: Regresión Softmax

# Regresión Softmax (1)

- Consideremos la distribución multinomial: $y \in \{1, \ldots, k\}$
- Tenemos $k$ clases. Algoritmo aprende a separar esas clases
- Parámetros de distribución serán $\underline{\phi} = \left(\phi_1, \phi_2 \ldots \phi_k\right)$ tal que

$$P(y = i) = \phi_i$$

- Ya vimos que $\phi_k = 1 - (\phi_1 + \phi_2 + \ldots + \phi_{k-1})$ (esto es, solo hay $k - 1$ parámetros), pero por simplicidad de notación usaremos $\phi_k$

# Regresión Softmax (2)

- Para el planteo de la distribución multinomial como parte de la familia de distribuciones exponenciales, vamos a necesitar vectores unitarios $\underline{\mathbf{T}}(i) \in \mathbb{R}^{k-1}$ (¡aquí **no** usamos $\underline{\mathbf{T}}(y) = y$!):

$$\underline{\mathbf{T}}(1) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \underline{\mathbf{T}}(2) = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \quad \cdots \quad \underline{\mathbf{T}}(k-1) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \quad \underline{\mathbf{T}}(k) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

- Con la función indicadora $1\{a\} = 1$ si $a$ es verdadero o $1\{a\} = 0$ si $a$ es falso, observemos que

$$(\underline{\mathbf{T}}(y))_i = 1\{y = i\}$$

- Por lo tanto se cumple $\mathrm{E}[(\underline{\mathbf{T}}(y))_i] = P(y = i) = \phi_i$

# Regresión Softmax (3)

- Tenemos así que

$$
\begin{aligned}
p(y;\underline{\phi}) &= \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\dots\phi_k^{1\{y=k\}} \\
&= \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\dots\phi_k^{1-\sum_{i=1}^{k-1}1\{y=i\}} \\
&= \phi_1^{(T(y))_1}\phi_2^{(T(y))_2}\dots\phi_k^{1-\sum_{i=1}^{k-1}(T(y))_i} \\
&= \exp\left[(T(y))_1\ln(\phi_1)+(T(y))_2\ln(\phi_2)+\dots\right. \\
&\qquad \left.\dots+\left(1-\sum_{i=1}^{k-1}(T(y))_i\right)\ln(\phi_k)\right] \\
&= \exp\left((T(y))_1\ln\left(\frac{\phi_1}{\phi_k}\right)+(T(y))_2\ln\left(\frac{\phi_2}{\phi_k}\right)+\dots\right. \\
&\qquad \left.\dots+(T(y))_{k-1}\ln\left(\frac{\phi_{k-1}}{\phi_k}\right)+\ln(\phi_k)\right)
\end{aligned}
$$

# Multinoulli en la familia exponencial (1)

- Finalmente obtenemos los parámetros de la familia exponencial:

$$
\underline{\boldsymbol{\eta}} = \begin{bmatrix} \ln(\phi_1/\phi_k) \\ \ln(\phi_2/\phi_k) \\ \vdots \\ \ln(\phi_{k-1}/\phi_k) \end{bmatrix}
$$

$$
a(\underline{\boldsymbol{\eta}}) = -\ln(\phi_k)
$$

$$
b(y) = 1
$$

- La función de enlace está dada para $i = 1 \dots k$ con $\eta_i = \ln(\phi_i/\phi_k)$, con $\eta_k = 0$

# Multinoulli en la familia exponencial (2)

- La función de respuesta se deriva con

$$\begin{aligned} e^{\eta_i} &= \phi_i / \phi_k \\ \phi_k e^{\eta_i} &= \phi_i \\ \phi_k \sum_{i=1}^{k} e^{\eta_i} &= \sum_{i=1}^{k} \phi_i = 1 \\ \phi_k &= \frac{1}{\sum_{i=1}^{k} e^{\eta_i}} \\ \phi_i &= \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}} \end{aligned}$$

  que se conoce como la función **softmax**

- Si introducimos el criterio de que $\eta_i = \underline{\boldsymbol{\theta}}_i^T \underline{\mathbf{x}}$, $(i = 1 \ldots k-1)$ con $\underline{\boldsymbol{\theta}}_i \in \mathbb{R}^{n+1}$ y $\underline{\boldsymbol{\theta}}_k = \underline{\mathbf{0}}$ para que $\eta_k = \underline{\boldsymbol{\theta}}_k^T \underline{\mathbf{x}} = 0$

# Multinoulli en la familia exponencial (3)

- Nuestro modelo supone que

$$
\begin{aligned}
p(y = i|\underline{\mathbf{x}};\underline{\boldsymbol{\theta}}) &= \phi_i \\
&= \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}} \\
&= \frac{e^{\underline{\boldsymbol{\theta}}_i^T \underline{\mathbf{x}}}}{\sum_{j=1}^{k} e^{\underline{\boldsymbol{\theta}}_j^T \underline{\mathbf{x}}}}
\end{aligned}
$$

que se conoce como **regresión softmax**

# Multinoulli en la familia exponencial (4)

- La hipótesis es entonces

$$h_{\underline{\theta}}(\underline{\mathbf{x}}) = \mathrm{E}[\underline{\mathbf{T}}(y)|\underline{\mathbf{x}};\underline{\theta}]$$

$$= \mathrm{E}\begin{bmatrix} 1\{y=1\} \\ 1\{y=2\} \\ \vdots \\ 1\{y=k-1\} \end{bmatrix} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{bmatrix} = \begin{bmatrix} \frac{e^{\underline{\theta}_1^T\underline{\mathbf{x}}}}{\sum_{j=1}^k e^{\underline{\theta}_j^T\underline{\mathbf{x}}}} \\ \frac{e^{\underline{\theta}_2^T\underline{\mathbf{x}}}}{\sum_{j=1}^k e^{\underline{\theta}_j^T\underline{\mathbf{x}}}} \\ \vdots \\ \frac{e^{\underline{\theta}_{k-1}^T\underline{\mathbf{x}}}}{\sum_{j=1}^k e^{\underline{\theta}_j^T\underline{\mathbf{x}}}} \end{bmatrix}$$

# Ajuste de parámetros en regresión softmax

- Con $m$ ejemplos en el conjunto de entrenamiento $\{(\underline{\mathbf{x}}^{(i)}, y^{(i)}); i = 1 \ldots m\}$ queremos aprender $\underline{\boldsymbol{\theta}}$ que maximice la verosimilitud logarítmica

$$\ell(\underline{\boldsymbol{\theta}}) = \sum_{i=1}^{m} \ln P(y^{(i)} | \underline{\mathbf{x}}^{(i)}; \underline{\boldsymbol{\theta}})$$

$$= \sum_{i=1}^{m} \ln \prod_{l=1}^{k} \left( \frac{e^{\underline{\boldsymbol{\theta}}_l^T \underline{\mathbf{x}}}}{\sum_{j=1}^{k} e^{\underline{\boldsymbol{\theta}}_j^T \underline{\mathbf{x}}}} \right)^{1\left\{y^{(i)} = l\right\}}$$

que se maximiza por ascenso de gradiente o métodos similares

# Resumen

1. Clasificación
   - Clasificación
   - Regresión logística
2. Modelos lineales generalizados
   - Familia exponencial
   - Modelos lineales generalizados

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