Logistic Multiclass Classification

March 13, 2013

Exponential Family

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

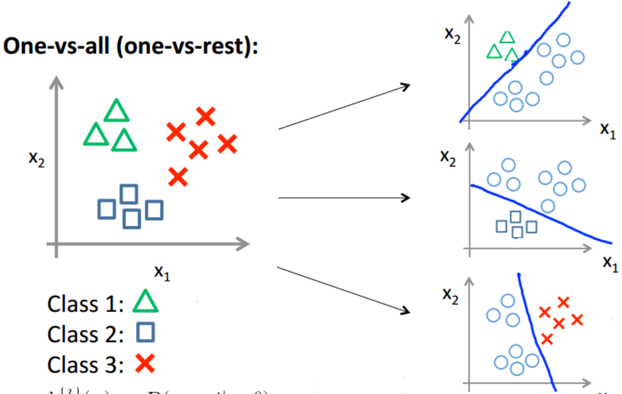
- η - natural parameter (actual parameter of distribution)
- $a(\eta), b(y)$ - specify the specific form of the distribution
- $T(y)$ - might be a vector

General Lenear Models

Assume

- $y|x; \theta$ ExpFamily(η)
- Given x, goal is to output $E[T(y)|x]$.
  - Want $h(x) = E[T(y)|x]$
- $\eta = \theta^T x$

Multiclass classification One vs All



$h_\theta^{(i)}(x) = P(y = i|x;\theta) \qquad (i = 1, 2, 3)$

Multiclass classification All vs All

Multiclass classification

- We whould like to classify instances into more than two classes

$$y \in \{1, 2, ..., k\}$$

- Lets derive GLM under assumption that $y|x$ is Multinomial
- The parameters of Multinomial distribution are $\phi_1, \phi_2, \ldots, \phi_k$, where

$$\phi_i = p(y = i; \phi)$$

$$p(y = k; \phi) = 1 - \sum_{i=1}^{k-1} \phi_i$$

Notation

To express the multinomial as an exponential family distribution, we will define $T(y) \in R^{k-1}$ as follows

$$T(1) = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix} \quad T(2) = \begin{bmatrix} 0 \\ 1 \\ \dots \\ 0 \end{bmatrix} \quad \dots \quad T(k-1) = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 1 \end{bmatrix} \quad T(k) = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}$$

- Let $(T(y))_i$ - ith coordinate of vector T(y)
- Let $1\{y = i\}$ - indicator function that returns 1 if expression in $\{\}$ is true and 0 otherwise
- $(T(y))_i = 1\{y = i\}$
- $E[(T(y))_i] = P(y = i) = \phi_i$

Multinomial is a Member of Exponential Family

$$
\begin{aligned}
p(y;\phi) &= \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\dots\phi_k^{1\{y=k\}} \\
&= \dots \\
&= \exp((T(y))_1\log(\phi_1/\phi_k) + (T(y))_2\log(\phi_2/\phi_k) + \\
&\quad \dots + (T(y))_{k-1}\log(\phi_{k-1}/\phi_k) + \log(\phi_k)) \\
&= b(y)\exp(\eta^T T(y) - a(\eta)),
\end{aligned}
$$

where

$$
\begin{aligned}
\eta &= \begin{bmatrix} \log(\phi_1/\phi_k) \\ \log(\phi_2/\phi_k) \\ \dots \\ \log(\phi_{k-1}/\phi_k) \end{bmatrix} \\
a(\eta) &= -\log(\phi_k) \\
b(y) &= 1
\end{aligned}
$$

Softmax Function

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \dots \\ \eta_{k-1} \\ \eta_k \end{bmatrix} = \begin{bmatrix} \log(\phi_1/\phi_k) \\ \log(\phi_2/\phi_k) \\ \dots \\ \log(\phi_{k-1}/\phi_k) \\ \log(\phi_k/\phi_k) \end{bmatrix}$$

Using that $\sum_{j=1}^{k} \phi_j = 1$ obtain

$$\phi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$$

$$\eta_k = \log(\phi_k/\phi_k) = 0$$

Softmax Regression

By assumption $\eta_i = \theta_i^T x$, where $\theta_1, \ldots, \theta_{k-1} \in R^{n+1}$, $\theta_k = 0$ so that $\theta_k^T x = 0$

$$p(y = i|x; \theta) = \phi_i = \frac{e^{\theta_i^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}}$$

$$h_\theta(x) = E[T(y)|x; \theta] = E \begin{bmatrix} \frac{e^{\theta_1^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}} \\ \frac{e^{\theta_2^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}} \\ \cdots \\ \frac{e^{\theta_{k-1}^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}} \end{bmatrix}$$

Generative Learning

March 13, 2013

Generative learning

- Earlier we consider $p(y|x)$ as $h_\theta(x) = g(\theta^T x)$ - discriminative learning algorithms
- Algotithms that estimates $p(x|y)$ and $p(y)$ - generative learning algorithms

$$\text{argmax}_y p(y|x) = \text{argmax}_y \frac{p(x|y)p(y)}{p(x)} = \text{argmax}_y p(x|y)p(y)$$

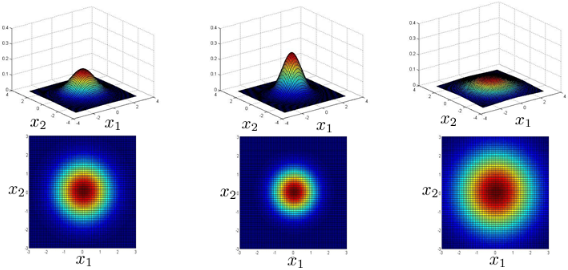
Multivariate Normal Distribution

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

$$E[X] = \int_x x p(x; \mu, \Sigma) dx = \mu$$

$$COV(X) = E[(X - E[X])(X - E[X])^T]$$

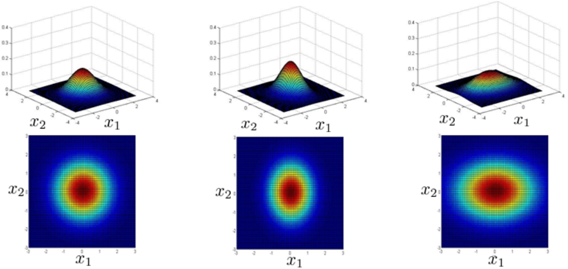
$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$



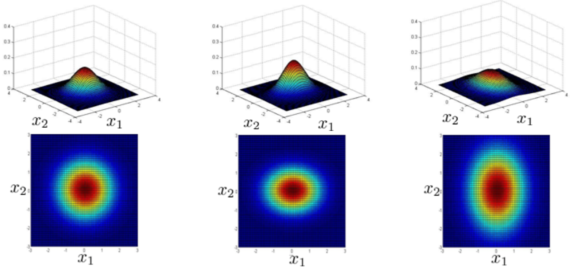
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$

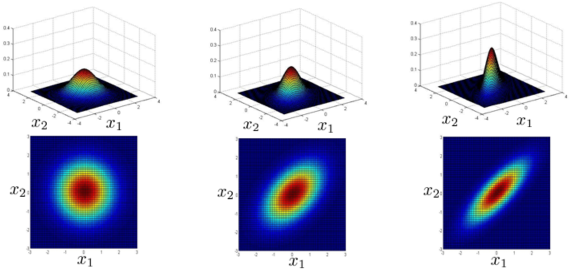
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$



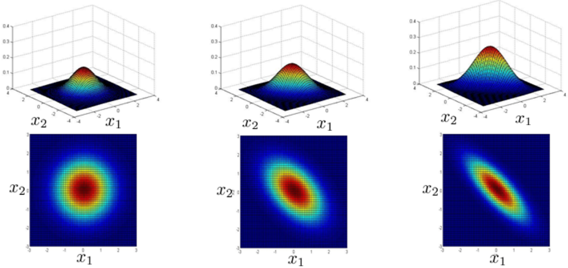
$\mu = \begin{bmatrix}0\\0\end{bmatrix} \Sigma = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ $\mu = \begin{bmatrix}0\\0\end{bmatrix} \Sigma = \begin{bmatrix}1 & 0\\0 & 0.6\end{bmatrix}$ $\mu = \begin{bmatrix}0\\0\end{bmatrix} \Sigma = \begin{bmatrix}1 & 0\\0 & 2\end{bmatrix}$



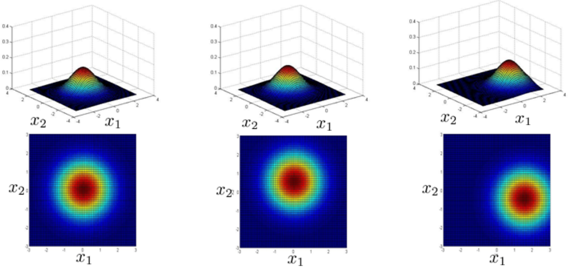
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\quad$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ $\quad$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Gaussian Discriminant Analysis Model

Intuition: predict $\text{argmax}_y p(y|x) = \text{argmax}_y \frac{p(x|y)p(y)}{p(x)}$ by modeling $p(x|y)$

- y : Bernoulli(ϕ)
- $x|y = 0$: (μ_0, Σ)
- $x|y = 1$: (μ_1, Σ)

$$p(y) = \phi^y(1-\phi)^{1-y}$$

$$p(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)$$

$$p(x|y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)$$

Gaussian Discriminant Analysis Model
Log-likelihood

$$\begin{aligned}\log L(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\ &= \log \prod p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi)\end{aligned}$$

$$\begin{aligned}\phi &= \frac{1}{m} \sum_{i=1}^{m} 1\{y^{(i)} = 1\} \\ \mu_0 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} \\ \mu_1 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} \\ \Sigma &= \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.\end{aligned}$$

Gaussian Discriminant Analysis
Example