COS 424: Interacting with Data

Lecturer: David Blei
Scribe: Güngör Polatkan

Lecture #19 Part 2
April 15, 2008

1 MLE is illuminating (continued)

Let $D = \{x_n\}_{n=1}^N$

$$\begin{aligned}\log p(x_{1:N}|\eta) &= \sum_{n=1}^N \log p(x_n|\eta) \\ &= \sum_{n=1}^N (\log h(x_n) + \eta^T t(x_n) - a(\eta)) \\ &= \sum_{n=1}^N \log h(x_n) + \eta^T \sum_{n=1}^N t(x_n) - N a(\eta).\end{aligned}$$

1.1 Notes

Note that $\sum_{n=1}^N t(x_n)$ is sufficient for η.

$$\nabla_\eta L = \sum_{n=1}^N t(x_n) - N \nabla_\eta a(\eta)$$

$$\nabla_\eta a(\eta) = \frac{\sum_{n=1}^N t(x_n)}{N} = \mathrm{E}[t(\mathbf{x})]$$

1.2 Back to Linear Models

The idea behind both linear and logistic regression is as the following:

$$\mathrm{E}[\mathbf{y}|\mathbf{x}] = f(\beta^T \mathbf{x}) \triangleq \mu$$

- At linear regression $f(a) = a$
- At logistic regression $f(a) =$logistic(a)

y is endowed with a distribution that depends on μ

- At linear regression $y \sim N(\mu, \sigma^2)$
- At logistic regression $y \sim Bernoulli(\mu)$

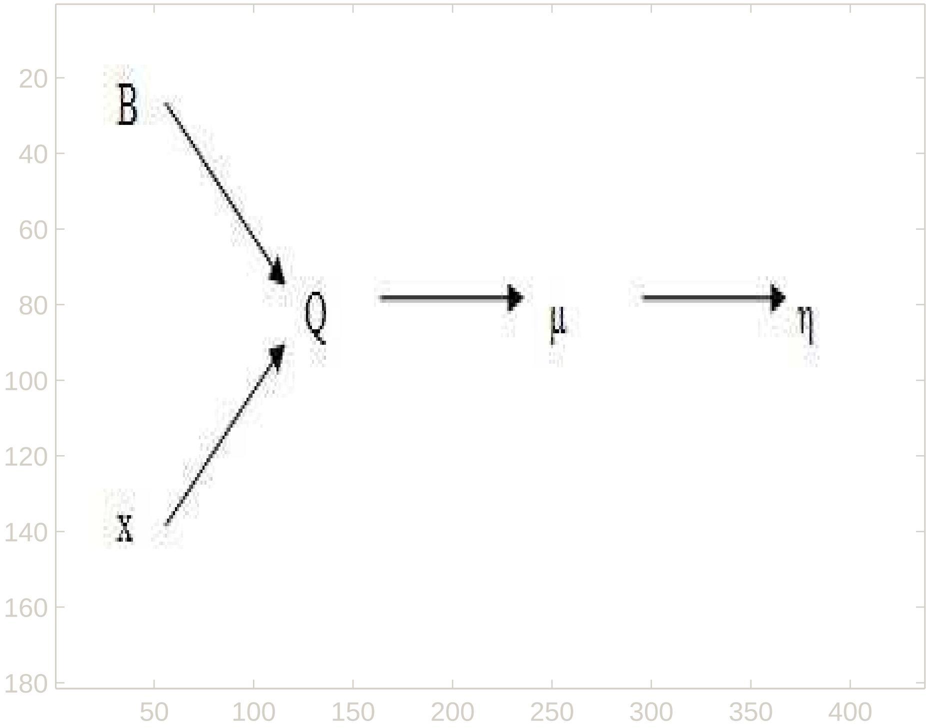

Figure 1: Relation between variables

1.3 Generalized Linear Model

- Input enters the model via $\beta^T x \triangleq Q$
- Conditional mean, $\mathrm{E}[\mathbf{y}|\mathbf{x}] \triangleq \mu$, is a function of Q called a *response function* or *link function.*
- Y comes from an exponential family with parameter μ.

Now let us model the diversity of response variables.
Choices:

- We need to decide which exponential distribution family to use for the response. (this is determined by the data type of y.
- We need to specify the response function f which is constrained but offers more freedom.

We will consider the *canonical response function.*