Generalized Linear Models and Exponential Families

David M. Blei

COS424
Princeton University

April 12, 2012
Generalized Linear Models

- Linear regression and logistic regression are both **linear models**. The coefficient $\beta$ enters the distribution of $y_n$ through a linear combination of $x_n$.

- Both are amenable to regularization via a Bayesian prior.

- Call $x_n$ the **input** and $y_n$ the **response**.
  - Linear regression: Real-valued response
  - Logistic regression: Binary response

- These ideas can be generalized to many kinds of response variables with **generalized linear models**.
  - E.g., categorical, positive real, positive integer, ordinal
A probability density in the exponential family has this form

\[ p(x \mid \eta) = h(x) \exp\{\eta^\top t(x) - a(\eta)\}, \]

where

- \( \eta \) is the natural parameter
- \( t(x) \) are sufficient statistics
- \( h(x) \) is the “underlying measure”, ensures \( x \) is in the right space
- \( a(\eta) \) is the log normalizer

Examples: Gaussian, Gamma, Poisson, Bernoulli, Multinomial

Distributions not in this family: Chi-Squared, Student-t
The log normalizer

\[ p(x \mid \eta) = h(x) \exp\{\eta^\top t(x) - a(\eta)\} \]

- The log normalizer ensures that the density integrates to 1,

\[ a(\eta) = \log \int h(x) \exp\{\eta^\top t(x)\} dx \]

- This is the negative logarithm of the normalizing constant.
Example: Bernoulli

The Bernoulli you are used to seeing is

\[ p(x | \pi) = \pi^x (1 - \pi)^{1-x} \quad x \in \{0, 1\} \]

In exponential family form

\[
\begin{align*}
 p(x | \pi) &= \exp\{\log \pi^x (1 - \pi)^{1-x}\} \\
 &= \exp\{x \log \pi + (1 - x) \log(1 - \pi)\} \\
 &= \exp\{x \log \pi - x \log(1 - \pi) + \log(1 - \pi)\} \\
 &= \exp\{x \log(\pi/(1 - \pi)) + \log(1 - \pi)\}
\end{align*}
\]
Example: Bernoulli (cont)

\[ p(x | \eta) = h(x) \exp{\eta^\top t(x) - a(\eta)} \]

This form reveals the exponential family

\[ p(x | \pi) = \exp\{x \log(\pi/(1 - \pi)) + \log(1 - \pi)\}, \]

where

- \( \eta = \log(\pi/(1 - \pi)) \)
- \( t(x) = x \)
- \( a(\eta) = -\log(1 - \pi) = \log(1 + e^\eta) \)
- \( h(x) = 1 \)
Log normalizer of the Bernoulli

- We express the log normalizer as a function of $\eta$.
- Recall that $\eta = \log(\pi/1 - \pi)$ and $a(\eta) = -\log(1 - \pi)$.

$$
\log(1 + e^\eta) = \log(1 + \pi/(1 - \pi))
\quad = \log((1 - \pi + \pi)/(1 - \pi))
\quad = \log(1/(1 - \pi))
\quad = -\log(1 - \pi)
$$

- The relationship between $\pi$ and $\eta$ is invertible

$$
\pi = 1/(1 + e^{-\eta})
$$

This is the **logistic function**.
Moments of the exponential family

Derivatives of \( a(\eta) \) give moments of the sufficient statistics.

\[
\nabla_\eta a = \nabla_\eta \left\{ \log \int \exp\{\eta^\top t(x)\} h(x) \, dx \right\} \\
= \frac{\nabla_\eta \int \exp\{\eta^\top t(x)\} h(x) \, dx}{\int \exp\{\eta^\top t(x)\} h(x) \, dx} \\
= \int t(x) \frac{\exp\{\eta^\top t(x)\} h(x)}{\int \exp\{\eta^\top t(x)\} h(x) \, dx} \, dx \\
= \mathbb{E}_\eta [t(X)]
\]

Higher order derivatives give higher order moments.
Mean parameters and natural parameters

- This expectation tells us that the mean parameter $E[t(X)]$ and natural parameter $\eta$ have a 1-1 relationship.

- We saw this with the logistic function, where note that $\pi = E[X]$ (because $X$ is an indicator).

- There is a 1-1 relationship between $E[t(X)]$ and $\eta$.
  - $\text{Var}(t(X)) = \nabla^2 a_\eta$ is positive.
  - $\rightarrow a(\eta)$ is convex.
  - $\rightarrow$ 1-1 relationship between its argument and first derivative

- Notation for later
  - The mean parameter is $\mu = E[t(X)]$.
  - The inverse map is $\psi(\mu)$, gives the $\eta$ such that $E[t(X)] = \mu$. 
The data are $\mathcal{D} = \{x_n\}_{n=1}^N$. We want to find the value of $\eta$ that maximizes the likelihood. The log likelihood is

$$
\mathcal{L} = \sum_{n=1}^N \log p(x_n | \eta)
= \sum_{n=1}^N \left( \log h(x_n) + \eta^\top t(x_n) - a(\eta) \right)
= \sum_{n=1}^N \log h(x_n) + \eta^\top \sum_{n=1}^N t(x_n) - N \cdot a(\eta)
$$

As a function of $\eta$, the log likelihood only depends on $\sum_{n=1}^N t(x_n)$.

- Has fixed dimension; no need to store the data.
- Is sufficient for $\eta$. 
Maximum likelihood estimation of an exponential family

\[ \mathcal{L} = \sum_{n=1}^{N} \log h(x_n) + \eta^\top \sum_{n=1}^{N} t(x_n) - a(\eta) \]

- Take the gradient and set to zero:
  \[ \nabla_\eta \mathcal{L} = \sum_{n=1}^{N} t(x_n) - N\nabla_\eta a(\eta) \]
- It’s easy to solve for the mean parameter:
  \[ \mu_{ML} = \frac{\sum_{n=1}^{N} t(x_n)}{N} \]
- The inverse map gives us the natural parameter:
  \[ \eta_{ML} = \psi(\mu_{ML}) \]
It's easy to solve for the mean parameter:

\[ \mu_{ML} = \frac{\sum_{n=1}^{N} t(x_n)}{N} \]

The inverse map gives us the natural parameter:

\[ \eta_{ML} = \psi(\mu_{ML}) \]

Consider the Bernoulli. \( \mu_{ML} \) is just the sample mean. The natural parameter is the corresponding log odds.
Idea behind logistic and linear regression: The conditional expectation of $y_n$ depends on $x_n$ through a function of a linear relationship,

$$E[y_n | x_n, \beta] = f(\beta^\top x_n) = \mu_n$$

- linear regression: $f$ is the identity.
- logistic regression: $f$ is the logistic.

Endow $y_n$ with a distribution that depends on $\mu_n$.
- linear regression: Gaussian
- logistic regression: Binary
Generalized linear models

\[ p(y_n \mid x_n) = h(y_n) \exp \{ \eta_n^T y_n - a(\eta_n) \} \]
\[ \eta_n = \psi(\mu_n) \]
\[ \mu_n = f(\beta^T x_n) \]

- Input \( x_n \) enters the model through \( \beta^T x_n \).
- The conditional mean \( \mu_n \) is a function \( f(\beta^T x_n) \) called the response function or link function.
- Response \( y_n \) has conditional mean \( \mu_n \).
- Its natural parameter is denoted \( \eta_n = \psi(\mu_n) \).
- Lets us build probabilistic predictors of many kinds of responses.
Generalized linear models

\[ p(y_n|x_n) = h(y_n) \exp\{\eta_n^\top t(y_n) - a(\eta_n)\} \]

\[ \eta_n = \psi(\mu_n) \]

\[ \mu_n = f(\beta^\top x_n) \]

- Two choices:
  1. Exponential family for response \( y_n \)
  2. Response function \( f(\beta^\top x_n) \)

- The family is usually determined by the form of \( y_n \).

- The response function:
  - Somewhat constrained—must give a mean in the right space
  - But also offers freedom, e.g., probit or logistic
The canonical response function

\[ p(y_n | x_n) = h(y_n) \exp\{\eta_n^\top t(y_n) - a(\eta_n)\} \]
\[ \eta_n = \psi(\mu_n) \]
\[ \mu_n = f(\beta^\top x_n) \]

- The **canonical response function** is \( f = \psi^{-1} \), which maps a natural parameter to the conditional mean that gives that natural parameter.
- Means that the natural parameter is \( \beta^\top x_n \),

\[ p(y_n | x_n) = h(y_n) \exp\{((\beta^\top x_n)^\top t(y_n) - a(\eta_n))\} \]
- Examples: logistic (binary) and identity (real)
Another important perspective

\[ p(y_n | x_n) = h(y_n) \exp\{\eta_n^\top t(y_n) - a(\eta_n)\} \]
\[ \eta_n = \psi(\mu_n) \]
\[ \mu_n = f(\beta^\top x_n) \]

- We can also think about this as

\[ y_n = f(\beta^\top x_n) + \epsilon_n, \]

where \( \epsilon_n \) is a zero-mean error term.

- \( \beta \) is the **systematic component**; \( \epsilon_n \) is the **random component**.
- Different response types lead to different error distributions.
Fitting a GLM

- The data are input/response pairs \( \{x_n, y_n\}_{n=1}^N \)
- The conditional likelihood is

\[
\mathcal{L}(\beta) = \sum_{n=1}^N h(y_n) + \eta_n^\top t(y_n) - a(\eta_n),
\]

and recall that \( \eta_n \) is a function of \( \beta \) and \( x_n \) (via \( f \) and \( \psi \)).
- Define each term to be \( \mathcal{L}_n \). The gradient is

\[
\nabla_\beta \mathcal{L} = \sum_{n=1}^N \nabla_{\eta_n} \mathcal{L}_n \nabla_\beta \eta_n
\]

\[
= \sum_{n=1}^N (t(y_n) - \nabla_{\eta_n} a(\eta_n)) \nabla_\beta \eta_n
\]

\[
= \sum_{n=1}^N (t(y_n) - \mathbb{E}[Y | x_n, \beta]) (\nabla_{\mu_n} \eta_n) (\nabla_{\theta_n} \mu_n) x_n
\]
In a canonical GLM, \( \eta_n = \beta^T x_n \) and

\[
\nabla_\beta \mathcal{L} = \sum_{n=1}^{N} (t(y_n) - \mathbb{E}[Y| x_n, \beta]) x_n
\]

Recall logistic and linear regression derivatives.