Exponential Families and Friends: Learning the Parameters of the a Fully Observed BN

Kayhan Batmanghelich
Machine Learning

The **data** inspires the structures we want to predict.

Inference finds \{best structure, marginals, partition function\} for a new observation.

(\text{Inference} \text{ is usually called as a subroutine in learning})

Our **model** defines a score for each structure.

It also tells us what to optimize.

\textbf{Learning} tunes the parameters of the model.
Machine Learning

Data

Inference

(Inference is usually called as a subroutine in learning)

Model

Objective

Learning

1. Alice saw Bob on a hill with a telescope

2. Time flies like an arrow

3. Objective

$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5$

4. Alice saw Bob on a hill with a telescope

3
Today’s Lecture

\[ p(X_1, X_2, X_3, X_4, X_5) = \]
\[ p(X_5 | X_3)p(X_4 | X_2, X_3) \]
\[ p(X_3)p(X_2 | X_1)p(X_1) \]
Today’s Lecture

\[ p(X_1, X_2, X_3, X_4, X_5) = p(X_5|X_3)p(X_4|X_2, X_3) p(X_3)p(X_2|X_1)p(X_1) \]
Today’s Lecture

How do we define and learn these conditional and marginal distributions for a Bayes Net?
Today’s Lecture

1. Exponential Family Distributions
   A candidate for marginal distributions, $p(X_i)$

2. Generalized Linear Models
   Convenient form for conditional distributions,
   $p(X_j | X_i)$

3. Learning Fully Observed Bayes Nets
   Easy thanks to decomposability
A candidate for **marginal** distributions, $p(X_i)$

1. EXPONENTIAL FAMILY
Why the **Exponential Family**?

1. **Pitman-Koopman-Darmois theorem**: it is the only family of distributions with **sufficient statistics that do not grow** with the size of the dataset

2. Only family of distributions for which **conjugate priors exist** (see Murphy textbook for a description)

3. It is the distribution that is closest to uniform (i.e. **maximizes entropy**) – subject to moment matching constraints

4. Key to **Generalized Linear Models** (**next section**) 

5. Includes some of your favorite distributions!

Adapted from Murphy (2012) textbook
Whiteboard

• Definition of multivariate exponential family
Whiteboard

• Example 1: Categorical distribution
Whiteboard

• Example 2: Multivariate Gaussian distribution
Moments and the Partition Function

\[ p(x; \theta) = \exp \left[ x^T \theta - A(\theta) \right] h(x) \]
Moments and the Partition Function

\[ p(x; \theta) = \exp \left[ x^T \theta - A(\theta) \right] h(x) \]

\[ \nabla_\theta A(\theta) = \mathbb{E}[T(x)] \]

\[ \nabla^2_\theta A(\theta) = \mathbb{E}[T(x)T(x)^T] - \mathbb{E}[T(x)]\mathbb{E}[T(x)]^T \]
Sufficiency

• For \( p(x; \theta) \), \( T(x) \) is **sufficient** for \( \theta \) if there is no information in \( X \) regarding \( \theta \) beyond that in \( T(x) \).
  
  – We can throw away \( X \) for the purpose of inference w.r.t. \( \theta \).

  – **Bayesian view**
    \[
    p(\theta | T(x), x) = p(\theta | T(x))
    \]

  – **Frequentist view**
    \[
    p(x | T(x), \theta) = p(x | T(x))
    \]

  – The Neyman factorization theorem
    • \( T(x) \) is **sufficient** for \( \theta \) if
      \[
      p(x, T(x), \theta) = \psi_1(T(x), \theta)\psi_2(x, T(x))
      \Rightarrow p(x | \theta) = g(T(x), \theta)h(x, T(x))
      \]
Sufficiency

\[ p(x; \theta) = \exp \left[ x^T \theta - A(\theta) \right] h(x) \]

- Let’s assume \( x_i \overset{iid}{\sim} p(x; \theta) \)

\[ p(x_1, \cdots, x_n; \theta) = \left( \prod_{j=1}^{n} h(x_j) \right) \exp \left( \theta^T \sum_{j}^{n} T(x_j) - nA(\theta) \right) \]
MLE for Exponential Family

- For iid data, the log-likelihood is

\[
\ell (\eta; D) = \log \prod_n h(x_n) \exp \{ \eta^T T(x_n) - A(\eta) \} \\
= \sum_n \log h(x_n) + \left( \eta^T \sum_n T(x_n) \right) - NA(\eta)
\]

- Take derivatives and set to zero:

\[
\frac{\partial \ell}{\partial \eta} = \sum_n T(x_n) - N \frac{\partial A(\eta)}{\partial \eta} = 0 \\
\frac{\partial A(\eta)}{\partial \eta} = \frac{1}{N} \sum_n T(x_n) \\
\Rightarrow \bar{\mu}_{MLE} = \frac{1}{N} \sum_n T(x_n)
\]

- This amounts to moment matching.

- We can infer the canonical parameters using

\[
\hat{\eta}_{MLE} = \psi(\bar{\mu}_{MLE})
\]
Examples

- **Gaussian:**
  
  \[
  \eta = \left[ \Sigma^{-1}\mu; -\frac{1}{2} \text{vec}(\Sigma^{-1}) \right] \\
  T(x) = \left[ x; \text{vec}(xx^T) \right] \\
  A(\eta) = \frac{1}{2} \mu^T \Sigma^{-1} \mu + \frac{1}{2} \log |\Sigma| \\
  h(x) = (2\pi)^{-k/2} \\
  \Rightarrow \mu_{MLE} = \frac{1}{N} \sum_n T_1(x_n) = \frac{1}{N} \sum_n x_n
  \]

- **Multinomial:**
  
  \[
  \eta = \left[ \ln \left( \frac{\pi_k}{\pi_K} \right); 0 \right] \\
  T(x) = [x] \\
  A(\eta) = -\ln \left( 1 - \sum_{k=1}^{K-1} \pi_k \right) = \ln \left( \sum_{k=1}^{K} e^{\eta_k} \right) \\
  h(x) = 1 \\
  \Rightarrow \mu_{MLE} = \frac{1}{N} \sum_n x_n
  \]

- **Poisson:**
  
  \[
  \eta = \log \lambda \\
  T(x) = x \\
  A(\eta) = \lambda = e^{\eta} \\
  h(x) = \frac{1}{x!} \\
  \Rightarrow \mu_{MLE} = \frac{1}{N} \sum_n x_n
  \]
Whiteboard

• Bayesian estimation of exponential family
  \[ p(x|\theta) = \exp \left[ x^T \theta - A(\theta) \right] h(x) \]

• We have observed iid samples and we are interested in
  \[ p(\theta|\{x_1, \cdots, x_n\}) \]
Posterior Mean Under Conjugate Prior

\[ p(x|\theta) = \exp \left[ x^T \theta - A(\theta) \right] h(x) \]

\[ p(\theta; \tau, n_0) = \exp \left( \tau^T \theta - n_0 A(\theta) - \tilde{A}(\tau, n_0) \right) \]

\[ p(\theta|\mathcal{D}) = p(\theta; \tau + \sum_i T(x_i); n + n_0) \]

- Posterior mean of \( \theta \)

\[ \mathbb{E}[\theta|\mathcal{D}] = \frac{n}{n + n_0} \left( \sum_i \frac{T(x_i)}{n} \right) + \frac{n_0}{n_0 + n} \left( \frac{\tau}{n_0} \right) \]
2. GENERALIZED LINEAR MODELS

Convenient form for conditional distributions, $p(X_j | X_i)$
Why Generalized Linear Models? (GLIMs)

1. Generalization of linear regression, logistic regression, probit regression, etc.
2. Provides a framework for creating new conditional distributions that come with some convenient properties.
3. Special case: GLMs with canonical response functions are easy to train with MLE.
4. No Free Lunch: What about Bayesian estimation of GLMs? Unfortunately, we have to turn to approximation techniques since, in general, there isn't a closed form of the posterior.
Generalized Linear Models (GLMs)

- **GLM**
  - The observed input $x$ is assumed to enter into the model via a linear combination of its elements $\xi = \theta^T x$
  - The conditional mean $\mu$ is represented as a function $f(\xi)$ of $\xi$, where $f$ is known as the response function
  - The observed output $y$ is assumed to be characterized by an **exponential family distribution** with conditional mean $\mu$. 

© Eric Xing @ CMU, 2005-2015
Whiteboard

- Constructive definition of GLMs
- Definition of GLMs with canonical response functions
Examples of the canonical response functions

<table>
<thead>
<tr>
<th>Distrib.</th>
<th>Link $g(\mu)$</th>
<th>$\theta = \psi(\mu)$</th>
<th>$\mu = \psi^{-1}(\theta) = \mathbb{E}[y]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{N}(\mu, \sigma^2)$</td>
<td>identity</td>
<td>$\theta = \mu$</td>
<td>$\mu = \theta$</td>
</tr>
<tr>
<td>Bin($N, \mu$)</td>
<td>logit</td>
<td>$\theta = \log\left(\frac{\mu}{1-\mu}\right)$</td>
<td>$\mu = \text{sigm}(\theta)$</td>
</tr>
<tr>
<td>Poi($\mu$)</td>
<td>log</td>
<td>$\theta = \log(\mu)$</td>
<td>$\mu = e^\theta$</td>
</tr>
</tbody>
</table>

\[ w \quad \eta_i \quad g^{-1} \quad \mu_i \quad \Psi \quad \theta_i \]

\[ x_i \quad g \quad \Psi^{-1} \]
Whiteboard

• MLE with GLM with Canonical response
MLE for GLMs with canonical response

• Log-likelihood

$$\mathcal{L}(w) = \sum_i \log h(y_i) + \sum_i (y_i w^T x_i - A(\eta_i))$$

• Derivative of Log-likelihood

$$\nabla_w \mathcal{L}(w) = \sum_i \left( x_i y_i - \frac{dA(\eta_i)}{d\eta_i} \frac{d\eta_i}{\theta} \right)$$

$$= \sum_i (y_i - \mu_i) x_i$$

$$= X^T (y - \mu)$$

• Online learning for canonical GLMs

  – Stochastic gradient ascent = least mean squares (LMS) algorithm:

$$w^{t+1} = w^t + \rho (y_i - \mu_i^t) x_i$$

This is a function of $w$

Step length
Batch learning for canonical GLMs

• The Hessian matrix

\[
H = -\frac{1}{\sigma^2} \sum_{i=1}^{N} \frac{d\mu_i}{d\theta_i} x_i x_i^T = -\frac{1}{\sigma^2} X^T S X
\]

\[
S = \text{diag}(\frac{d\mu_1}{d\theta_1}, \ldots, \frac{d\mu_N}{d\theta_N})
\]

Involves the second derivative of \( A(\theta) \)
Iteratively Reweighted Least Squares (IRLS)

\[ \nabla_w \mathcal{L}(w) = X^T(y - \mu) \]

\[ H = -\frac{1}{\sigma^2} \sum_{i=1}^{N} \frac{d\mu_i}{d\theta_i} x_i x_i^T = -\frac{1}{\sigma^2} X^T S X \]

- Recall Newton-Raphson methods with cost function

\[ w^{t+1} = w^t + H^{-1}(w^t) \nabla \mathcal{L}(w^t) \]

\[ = (X^T S(w^t) X)^{-1} \left[ X^T S(w^t) X w^t + X^T (y - \mu) \right] \]

\[ = (X^T S(w^t) X)^{-1} X^T S(w^t) z^t \]

\[ z^t = X w^t + S(w^t)^{-1} (y - \mu^t) \]
Iteratively Reweighted Least Squares (IRLS)

\[ \nabla_w \mathcal{L}(w) = X^T(y - \mu) \]

\[ H = -\frac{1}{\sigma^2} \sum_{i=1}^{N} \frac{d\mu_i}{d\theta_i} x_i x_i^T = -\frac{1}{\sigma^2} X^T S X \]

- Recall Newton-Raphson methods with cost function

\[ w^{t+1} = w^t + H^{-1}(w^t) \nabla \mathcal{L}(w^t) \]

\[ = (X^T S(w^t) X)^{-1} \left[ X^T S(w^t) X w^t + X^T (y - \mu) \right] \]

\[ = (X^T S(w^t) X)^{-1} X^T S(w^t) z^t \]

\[ z^t = X w^t + S(w^t)^{-1} (y - \mu^t) \]

It looks like \((X^T X)^{-1} X^T y\)
Iteratively Reweighted Least Squares (IRLS)

\[ \nabla_w \mathcal{L}(w) = X^T (y - \mu) \]

\[ H = -\frac{1}{\sigma^2} \sum_{i=1}^{N} \frac{d\mu_i}{d\theta_i} x_i x_i^T = -\frac{1}{\sigma^2} X^T S X \]

- Recall Newton-Raphson methods with cost function

\[
\begin{align*}
w^{t+1} &= w^t + H^{-1}(w^t) \nabla \mathcal{L}(w^t) \\
           &= (X^T S(w^t) X)^{-1} [X^T S(w^t) X w^t + X^T (y - \mu)] \\
           &= (X^T S(w^t) X)^{-1} X^T S(w^t) z^t \\
           &= X w^t + S(w^t)^{-1} (y - \mu^t)
\end{align*}
\]

- This can be understood as solving the following "Iteratively reweighted least squares" problem

\[
w^{t+1} = \arg \max_w (z^t - Xw)^T S(w^t) (z^t - Xw)\]
Examples

$$\nabla_w \mathcal{L}(w) = X^T(y - \mu)$$

$$H = -\frac{1}{\sigma^2} \sum_{i=1}^{N} \frac{d \mu_i}{d \theta_i} x_i x_i^T = -\frac{1}{\sigma^2} X^T S X$$

• Recall Newton-Raphson methods with cost function

$$w^{t+1} = w^t + H^{-1}(w^t) \nabla \mathcal{L}(w^t)$$

$$= (X^T S(w^t) X)^{-1} \left[ X^T S(w^t) X w^t + X^T (y - \mu) \right]$$

$$= (X^T S(w^t) X)^{-1} X^T S(w^t) z^t$$

$$z^t = X w^t + S(w^t)^{-1} (y - \mu^t)$$

$$w^{t+1} = \arg \max_w (z^t - X w)^T S(w^t) (z^t - X w)$$
Practical Issues

• It is very common to use regularized maximum likelihood.

\[
p(y = \pm 1 | x, \theta) = \frac{1}{1 + e^{-y \theta^T x}} = \sigma(y \theta^T x)
\]

\[
p(\theta) \sim \text{Normal}(0, \lambda^{-1} I)
\]

\[
l(\theta) = \sum_n \log(\sigma(y_n \theta^T x_n)) - \frac{\lambda}{2} \theta^T \theta
\]

– IRLS takes \(O(N d^3)\) per iteration, where \(N\) = number of training cases and \(d\) = dimension of input \(x\).

– Quasi-Newton methods, that approximate the Hessian, work faster.

– Conjugate gradient takes \(O(Nd)\) per iteration, and usually works best in practice.

– Stochastic gradient descent can also be used if \(N\) is large c.f. perceptron rule.
Today’s Lecture

1. Exponential Family Distributions
   A candidate for marginal distributions, $p(X_i)$

2. Generalized Linear Models
   Convenient form for conditional distributions, $p(X_j \mid X_i)$

3. Learning Fully Observed Bayes Nets
   Easy thanks to decomposability
3. LEARNING FULLY OBSERVED BNS

Easy thanks to decomposability
Simple GMs are the building blocks of complex BNs

Density estimation
- Parametric and nonparametric methods

Regression
- Linear, conditional mixture, nonparametric

Classification
- Generative and discriminative approach
Decomposable likelihood of a BN

- Consider the distribution defined by the directed acyclic GM:

\[ p(x \mid \theta) = p(x_1 \mid \theta_1)p(x_2 \mid x_1, \theta_2)p(x_3 \mid x_1, \theta_3)p(x_4 \mid x_2, x_3, \theta_4) \]

- This is exactly like learning four separate small BNs, each of which consists of a node and its parents.
Learning Fully Observed BNs

\[ \mathbf{\theta}^* = \arg\max_{\theta} \log p(X_1, X_2, X_3, X_4, X_5) \]

\[ = \arg\max_{\theta} \log p(X_5|X_3, \theta_5) + \log p(X_4|X_2, X_3, \theta_4) \]

\[ + \log p(X_3|\theta_3) + \log p(X_2|X_1, \theta_2) \]

\[ + \log p(X_1|\theta_1) \]

\[ \theta_1^* = \arg\max_{\theta_1} \log p(X_1|\theta_1) \]

\[ \theta_2^* = \arg\max_{\theta_2} \log p(X_2|X_1, \theta_2) \]

\[ \theta_3^* = \arg\max_{\theta_3} \log p(X_3|\theta_3) \]

\[ \theta_4^* = \arg\max_{\theta_4} \log p(X_4|X_2, X_3, \theta_4) \]

\[ \theta_5^* = \arg\max_{\theta_5} \log p(X_5|X_3, \theta_5) \]
Summary

1. **Exponential Family Distributions**
   - A candidate for marginal distributions, $p(X_i)$
   - Examples: Multinomial, Dirichlet, Gaussian, Gamma, Poisson
   - MLE has closed form solution
   - Bayesian estimation easy with conjugate priors
   - Sufficient statistics by inspection

2. **Generalized Linear Models**
   - Convenient form for conditional distributions, $p(X_j | X_i)$
   - Special case: GLIMs with canonical response
     - Output $y$ follows an exponential family
     - Input $x$ introduced via a linear combination
   - MLE for GLIMs with canonical response by SGD
   - In general, Bayesian estimation relies on approximations

3. **Learning Fully Observed Bayes Nets**
   - Easy thanks to decomposability