Exponential Family and Generalized Linear Models

Piyush Rai
IIT Kanpur

Probabilistic Machine Learning (CS772A)

Jan 20, 2016

Generalized Linear Models

- Models we have seen so far..
  - (Probabilistic) Linear regression: when y is real-valued

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$$

  - Logistic regression: when y is binary (0/1)

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \text{Bernoulli}(\sigma(\boldsymbol{w}^\top \boldsymbol{x})) = [\sigma(\boldsymbol{w}^\top \boldsymbol{x})]^y [1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x})]^{1-y}$$

  where $\sigma(\boldsymbol{w}^\top \boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{w}^\top \boldsymbol{x})} = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}$

- In both, the model depends on the inputs $\boldsymbol{x}$ linearly via $\boldsymbol{w}^\top \boldsymbol{x}$
- Both are special cases of a more general class: Generalized Linear Models

$$p(y|\eta) = h(y) \exp(\eta y - A(\eta))$$

  .. a special type of exponential family distribution
- GLM can be used to also model responses that aren't reals/binary (can be any exponential family distribution in general)

Exponential Family Distributions

- An exponential family distribution is of the form

$$p(y|\eta) = h(y) \exp(\eta^\top T(y) - A(\eta))$$

- η is called the natural parameter
- $h(y)$ is usually a constant w.r.t. η
- $T(y)$ is the sufficient statistics: $p(y|\eta)$ depends on y only through $T(y)$
- $A(\eta)$: log partition function or cumulant function

$$A(\eta) = \log \int h(y) \exp(\eta^\top T(y)) dy$$

  .. can also be seen as the log of a normalization factor

Bernoulli as Exponential Family

- Bernoulli in the usual form:

$$\text{Bernoulli}(y|p) = p^y (1-p)^{1-y} = \exp\left(y \log\left(\frac{p}{1-p}\right) + \log(1-p)\right)$$

- Comparing it as $p(y|\eta) = h(y) \exp(\eta^\top T(y) - A(\eta))$, we have
  - $h(y) = 1$
  - $\eta = \log\left(\frac{p}{1-p}\right)$
  - $T(y) = y$
  - $A(\eta) = -\log(1-p)$

Gaussian as Exponential Family

- Gaussian in the usual form:

$$\mathcal{N}(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu}{\sigma^2}y - \frac{1}{2\sigma^2}y^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right)$$

- Comparing it as $p(y|\eta) = h(y)\exp(\eta^\top T(y) - A(\eta))$, we have
  - $h(y) = \frac{1}{\sqrt{2\pi}}$
  - $\eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T$
  - $T(y) = (y, y^2)^T$
  - $A(\eta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$

Some Useful Properties

- The log partition function $A(\eta)$ has several useful properties
- First derivative of $A(\eta)$ w.r.t. η is the expectation of the sufficient statistics

$$\frac{dA(\eta)}{d\eta} = \mathbb{E}[T(y)] \qquad \text{(proof done on board)}$$

- Second derivative of $A(\eta)$ w.r.t. η is the variance of sufficient statistics

$$\frac{d^2A(\eta)}{d\eta^2} = \text{var}[T(y)]$$

- Note: $A(\eta)$ is also convex (because second derivative is non-negative)

MLE for Exponential Family Distributions

- The log-likelihood is given by

$$L(\eta) = \log p(Y|\eta) = \log\prod_{n=1}^N p(y_n|\eta) = \log\prod_{n=1}^N h(y_n)\exp(\eta^\top T(y_n) - A(\eta))$$
$$= \log\prod_{n=1}^N h(y_n) + \eta^\top\left(\sum_{n=1}^N T(y_n)\right) - NA(\eta)$$

- Taking derivative w.r.t. η and setting it to zero

$$N\frac{dA(\eta)}{d\eta} = \sum_{n=1}^N T(y_n)$$

- Defining $\mu = \mathbb{E}[T(y)] = \frac{dA(\eta)}{d\eta}$, we get

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{n=1}^N T(y_n) \qquad \text{(can be used for parameter estimation via moment-matching)}$$

- Note that the estimate only depends on data via the sufficient statistics $T(y)$

Generalized Linear Models

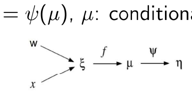
- An exp. fam. model for $\boldsymbol{x} \to y$ is a Generalized Linear Model if:
  1. Observed inputs $\boldsymbol{x}_n$ enter the model via linear combination $\boldsymbol{w}^\top\boldsymbol{x}_n$
  2. Conditional mean of response y_n depends on $\boldsymbol{w}^\top\boldsymbol{x}_n$ via a response function f

     $$\mu_n = \mathbb{E}[y_n] = f(\boldsymbol{w}^\top\boldsymbol{x}_n)$$

     - for linear regression $\mu_n = f(\boldsymbol{w}^\top\boldsymbol{x}_n) = \boldsymbol{w}^\top\boldsymbol{x}_n$,
     - for logistic regression $\mu_n = f(\boldsymbol{w}^\top\boldsymbol{x}_n) = \exp(\boldsymbol{w}^\top\boldsymbol{x}_n)/(1+\exp(\boldsymbol{w}^\top\boldsymbol{x}_n))$
  3. $T(y) = y$
- Form of a GLM

$$p(y|\eta) = h(y)\exp(\eta y - A(\eta))$$

where natural parameter $\eta = \psi(\mu)$, μ: conditional mean, ψ: link function



- Note: Some GLM can be represented as $p(y|\eta,\phi) = h(y,\phi)\exp(\frac{\eta y - A(\eta)}{\phi})$ where ϕ is a dispersion parameter (Gaussian/gamma GLMs use this rep.)

GLM with Canonical Response Function

- A GLM has a canonical response function f if $f = \psi^{-1}$
- For such a GLM, $\eta_n = \psi(\mu_n) = \psi(f(\boldsymbol{w}^\top \boldsymbol{x}_n)) = \boldsymbol{w}^\top \boldsymbol{x}_n$
- E.g., for logistic regression $\eta_n = \log \frac{\mu_n}{1-\mu_n} = \boldsymbol{w}^\top \boldsymbol{x}_n$ (exercise: verify by recalling the exponential family representation of Bernoulli distribution)
- Thus, for Canonical GLMs

$$\begin{aligned} p(y|\eta) &= h(y)\exp(\eta y - A(\eta)) \\ &= h(y)\exp(y\boldsymbol{w}^\top \boldsymbol{x} - A(\eta)) \end{aligned}$$

- Such design choices in the canonical GLM make parameter estimation easy

MLE for Generalized Linear Models

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n)\exp(y_n \boldsymbol{w}^\top x_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Taking derivative w.r.t. $\boldsymbol{w}$

$$\sum_{n=1}^{N}\left(y_n \boldsymbol{x}_n - A'(\eta_n)\frac{d\eta_n}{d\boldsymbol{w}}\right) = \sum_{n=1}^{N}(y_n \boldsymbol{x}_n - \mu_n \boldsymbol{x}_n) = \sum_{n=1}^{N}(y_n - \mu_n)\boldsymbol{x}_n$$

  where $\mu_n = f(\boldsymbol{w}^\top \boldsymbol{x}_n)$ and 'f' $(= \psi^{-1})$ depends on type of response y, e.g.,
  - Real-valued y (linear regression): f is identity, i.e., $\mu_n = \boldsymbol{w}^\top \boldsymbol{x}_n$
  - Binary y (logistic regression): f is logistic function, i.e., $\mu_n = \frac{\exp(\boldsymbol{w}^\top x_n)}{1+\exp(\boldsymbol{w}^\top x_n)}$
  - Count-valued y (Poisson regression): $\mu_n = \exp(\boldsymbol{w}^\top \boldsymbol{x}_n)$
  - Positive reals y (gamma regression): $\mu_n = -(\boldsymbol{w}^\top \boldsymbol{x}_n)^{-1}$
- To estimate $\boldsymbol{w}$, either set the derivative to zero or use iterative methods (e.g., gradient descent, iteratively reweighted least squares, etc.)

Next class:
Clustering via Gaussian Mixture Models