Generalized Linear Model

Mohammad Emtiyaz Khan
EPFL

Oct 15, 2015

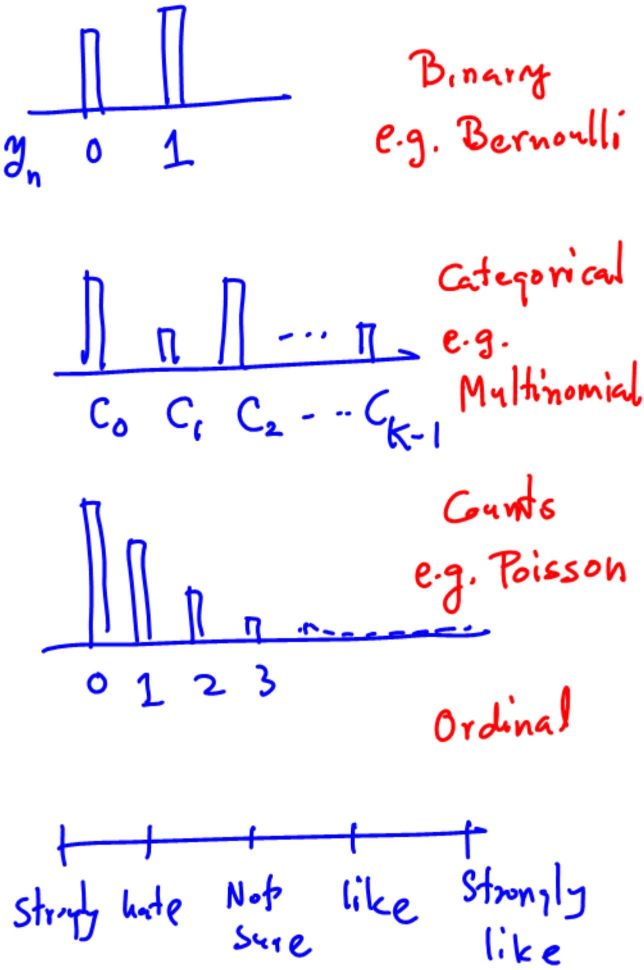
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

Logistic regression extends linear regression to binary outputs. Can we generalize this procedure to other kind of outputs, e.g. categorical, ordinal, counts, etc?

Specifically, can we generalize the maximum-likelihood procedure to a generic distribution $p(y_n|\widetilde{\mathbf{x}}_n^T\boldsymbol{\beta})$?

The answer is yes, we can do this for exponential family distributions.



Logistic regression revisited

In logistic regression, we used the following distribution: $\eta_n = \tilde{x}_n^T \beta$

$$p(y_n|\eta_n) := \frac{\exp(y_n\eta_n)}{1+e^{\eta_n}} = \frac{e^{y_n\eta_n}}{e^{\log(1+e^{\eta_n})}} = e^{y_n\eta_n - \log(1+e^{\eta_n})}$$

$$= \exp\left[\underbrace{y_n\eta_n}_{\text{linear}} - \underbrace{\log(1+e^{\eta_n})}_{A(\eta_n)}\right].$$

This specific form of the distribution is not coincidental.

In general, we require the following form of the distribution:

$$p(y_n|\eta_n) := \exp\left[\underbrace{y_n\eta_n}_{\text{linear}} - \underbrace{A(\eta_n)}_{\text{function of } \eta_n}\right] \overbrace{h(y_n)}^{\swarrow},$$

where h does not depend on $\underline{\eta_n}$.

Exponential family distribution

$$Z = \int h(y) \exp[\eta^T \phi(y) - A(\eta)]\, dy$$

Exponential family distribution is a general class of distributions defined as shown below,

$$p(y|\boldsymbol{\eta}) := \frac{\overbrace{h(y)}}{\underbrace{Z}_{\text{partition function}}} \exp\left[\overset{\text{natural parameter}}{\boldsymbol{\eta}}^T \underbrace{\boldsymbol{\phi}(y)}_{\text{Sufficient statistics}} - \overset{\text{log-partition function}}{A(\boldsymbol{\eta})}\right].$$

This definition is in the *canonical* form and assumes a scalar-valued random variable y. See Wikipedia for a more general definition.

Bernoulli distribution

$$\left.\begin{aligned} p(y=1/\mu) &= \mu \\ p(y=0/\mu) &= 1-\mu, \end{aligned}\right\}$$

$\mu \in (0,1)$

Example: Bernoulli distribution is in exponential family.

$$p(y|\mu) := \underline{\mu^y(1-\mu)^{1-y}}, \text{ where } \mu \in (0,1)$$

$$= \exp\left[\underbrace{y}_{\phi(y)} \underbrace{\log\frac{\mu}{1-\mu}}_{\eta} + \underbrace{\log(1-\mu)}_{-A(\eta)}\right]$$

$$\mu^y(1-\mu)^{1-y}$$
$$= \exp\left[\log\left\{\mu^y(1-\mu)^{1-y}\right\}\right]$$
$$= \exp\left[y\log\mu + (1-y)\log(1-\mu)\right]$$

There is a relationship between η and μ through the link function,

$$\underbrace{\eta = \log\frac{\mu}{1-\mu}}_{g(\mu)} \iff \mu = \frac{e^\eta}{1+e^\eta}.$$

Note that $\boxed{\mu = \mathbb{E}(\phi(y))}$ is the mean parameter of y (and y is the *sufficient* statistics to describe the distribution).

$$\begin{aligned} A(\eta) &= -\log(1-\mu) \\ &= -\log\left(1-\frac{e^\eta}{1+e^\eta}\right) \\ &= +\log(1+e^\eta) \end{aligned}$$

Useful properties of exponential family

In general, there is a relationship between the mean $\boldsymbol{\mu} := \mathbb{E}(\boldsymbol{\phi}(y))$ and $\boldsymbol{\eta}$ defined using a link function g.

$$\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\mu}) \iff \boldsymbol{\mu} = \mathbf{g}^{-1}(\boldsymbol{\eta}).$$

$$\begin{tikzcd} p(y/.) \arrow[r] & \underline{\phi}(y) \arrow[d, "\mathbb{E}"] \\ \underline{\eta} \arrow[u] & \underline{\mu} \arrow[l, "g"] \end{tikzcd}$$

See the table of link functions and many other examples of exponential family in the KPM book chapter on "Generalized Linear Model".

Derive this for Gaussian and multinomial distribution!

Another useful property is that the first and second derivatives of $A(\eta)$ are related to the mean and the variance of the sufficient statistics.

$$\frac{\partial A(\eta)}{\partial \eta} = \mathbb{E}[\underline{\phi}(\eta)] \ , \ \frac{\partial^2 A(\eta)}{\partial \eta^2} = \mathrm{Var}[\underline{\phi}(\eta)].$$

Therefore, A is convex.

Using this property, we can generalize the normal equation.

The maximum likelihood estimate

The generalized maximum likelihood cost to minimize is,

$$\underbrace{p(y_n/\eta_n)} \qquad \eta_n := \underline{x}_n^T\underline{\beta}$$

$$\min \mathcal{L}(\boldsymbol{\beta}) := -\sum_{n=1}^{N} \log p(y_n|\widetilde{\mathbf{x}}_n^T\boldsymbol{\beta}),$$

Find a $\underline{\beta}$ that is a "good-fit" to my data $(y_n, \underline{x}_n)$

where $p(y_n|\eta_n)$ is an exponential family distribution with $\eta_n = \widetilde{\mathbf{x}}_n^T\boldsymbol{\beta}$ as the natural parameters.

Assume η. is a scaler

$$\mathcal{L}_n(\underline{\beta}) := -\log p(y_n/\tilde{\underline{x}}_n^T\underline{\beta})$$

$$= -\sum_{n=1}^{N}\left[\eta_n\,\phi(y_n) - A(\eta_n)\right] + \text{cnst}$$

The maximum likelihood solution can be obtained by solving the following normal equation:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \widetilde{\mathbf{X}}^T\left[\mathbf{g}^{-1}(\boldsymbol{\eta}) - \underline{\phi}(\underline{y})\right]$$

$$\frac{\partial \mathcal{L}_n}{\partial \eta_n} = -\phi(y_n) + \frac{\partial}{\partial \eta_n}A(\eta_n)$$

$$= -\phi(y_n) + \mu_n$$

$$= -\phi(y_n) + g^{-1}(\eta_n)$$

where $\boldsymbol{\eta}$ is a vector of all η_n and the function $\mathbf{g}^{-1}(\boldsymbol{\eta})$ is a vector of $\mathbf{g}^{-1}(\eta_n)$.

$$\frac{\partial \mathcal{L}}{\partial \underline{\beta}} = \tilde{X}^T\left[g^{-1}(\underline{\eta}) - \phi(\underline{y})\right] \quad \Big| \quad \frac{\partial}{\partial \underline{\beta}} = \frac{\partial \eta_n}{\partial \underline{\beta}}\frac{\partial \mathcal{L}_n}{\partial \eta_n} = \tilde{\underline{x}}_n \frac{\partial \mathcal{L}_n}{\partial \eta_n}$$

vectorize

To do

1. Read the following sections in the KPM book: Section 9.1.1 to 9.1.4, Section 9.3.1 to 9.3.2.
2. Derive exponential family form for Gaussian distribution and multinomial distributions.
3. Derive generalized linear model for regression with Poisson distribution for count data.

Hessian

↳ $\dfrac{\partial^2 \mathcal{L}}{\partial \underline{\beta}\,\partial \underline{\beta}^T} = X^T S X$, where S is a diagonal matrix with $S_{nn} = \dfrac{\partial^2 A(\eta_n)}{\partial \eta_n^2}$

↳ always pos. semi-definite | with $\eta_n := \tilde{\underline{x}}_n^T\underline{\beta}$