COS513 Scribe Note: Generalized linear models (II)

Donghun Lee

November 10, 2010

1 Generalized Linear Model, cont'd

Schematically, GLM is:

$$\beta^\top x_n \xrightarrow{f} \mathbb{E}\left[Y_n|X_n\right] = \mu_n \xrightarrow{\psi} \eta_n$$

And, corresponding form of conditional probability density is:

$$\begin{aligned}\mathbb{P}\left[Y_n|X_n\right] &= h\left(Y_n\right)\exp\left(\psi\left(f\left(\beta^\top x_n\right)\right)\right)\cdot y - a\left(\psi\left(f\left(\beta^\top x_n\right)\right)\right)\\ &= h\left(Y_n\right)\exp\left(\psi\left(\mu_n\right)\right)\cdot y - a\left(\psi\left(\mu_n\right)\right)\\ &= h\left(Y_n\right)\exp\left(\eta_n\right)\cdot y - a\left(\eta_n\right)\end{aligned}$$

When using canonical link function, the inverse link function f is defined such that $\eta_n = \beta^\top x_n$, which results in $f \overset{\triangle}{=} \psi^{-1}$. This sounds like an arbitrary condition, but in the case of logistic regression, this happens naturally.

Logistic regression and probit regression:

	Logistic	Probit
f	logit function i.e. $\log\left(\frac{\beta^\top x}{1-\beta^\top x}\right)$	tail prob. of $N(0,1)$ i.e. $1-\Phi\left(\beta^\top x\right)$
map	$\psi \circ f : \mathbb{R} \mapsto [0,1]$	
ψ	$f = \psi^{-1}$ when $Y \sim$ Bernoulli	ψ more complicated

Softmax regression is a multivariate extension of logistic:

	Logistic	Softmax
response Y	$Y \in \{0,1\}$	$Y \in \{1,2,\ldots,K\}$
ψ	logistic function	softmax function

In summary, GLM gives

- Flexibility: because a user can do model selection by choosing f (and its domain, codomain as well).
- Robustness: because ψ is determined from the choice of probabilistic model, which is from a well-defined “exponential” family of models.

2 Fitting GLM with Maximum Likelihood setup

The idea is:

$$\text{If given } \hat{\beta}_{MLE}\text{, we predict } Y \text{ with probability: } \mathbb{P}\left[\hat{Y}|X\right] = f\left(\hat{\beta}_{MLE}, X\right)$$

Denote:

- $\mathcal{D} = \{(x_n, y_n)\} \quad (n = 1 \ldots N)$
- $\eta_n := \psi\left(f\left(\beta^\top x_n\right)\right)$: per-observation natural parameter
- $\mu_n := f\left(\beta^\top x_n\right)$: per-observation mean parameter
- $l\left(\beta; \mathcal{D}\right)$: log-likelihood of data $\mathcal{D}$ given β, such that:

$$\begin{aligned} l\left(\beta; \mathcal{D}\right) &:= \log \prod_{n=1}^{N} h\left(y_n\right) \exp\left(\eta_n^\top y_n - a\left(\eta_n\right)\right) \\ &= \sum_{n=1}^{N} \log\left(h\left(y_n\right)\right) + \sum_{n=1}^{N}\left(\eta_n^\top y_n - a\left(\eta_n\right)\right) \end{aligned}$$

Now, finding MLE requires the following steps:

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \sum_{n=1}^{N} \frac{\partial l}{\partial \eta_n} \frac{\partial \eta_n}{\partial \beta} \\ &= \sum_{n=1}^{N} \left(y_n - a'\left(\eta_n\right)\right) \frac{\partial \eta_n}{\partial \beta} \qquad (1) \end{aligned}$$

Note the following two equalities that can be applied to Equation 1:

- $\frac{\partial \eta_n}{\partial \beta} = \psi'\left(f\left(\beta^\top x_n\right)\right) f'\left(\beta^\top x_n\right)$ from definition of η_n
- $a'\left(\eta_n\right) = \mu_n$ from exponential family's property

From this viewpoint, equation 1 can be seem as a weighted sum of gradients for each observation x_n, where each weight is the difference of observed response variable y_n and its mean μ_n.

When f is the canonical response function (i.e. $f = \psi^{-1}$), then $\eta_n = \beta^\top x_n$. Using this, equation 1 can be further simplified:

$$\begin{aligned}\nabla_\beta l\left(\beta;\mathcal{D}\right) &= \frac{\partial}{\partial\beta}\sum_{n=1}^{N}\left(\left(\beta^\top x_n\right)^\top y_n - a\left(\beta^\top x_n\right)\right)\\ &= \sum_{n=1}^{N}\left(x_n^\top y_n - a'\left(\beta^\top x_n\right)x_n\right)\\ &= \sum_{n=1}^{N}\left(y_n-\mu_n\right)x_n\end{aligned}$$

Now, this can be seen as an iterative update rule, where $y_n - \mu_n$ is the signed residual for each observation and x_n sets the direction to update β using the observation.

3 Dimension Reduction

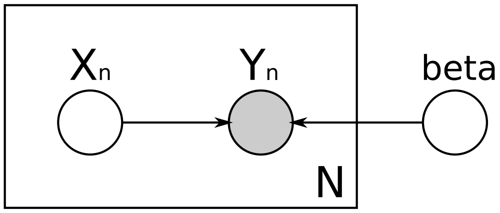

Figure 1: A graphical model for dimensionality reduction

Note that now we do *not* see X_n, but sees Y_n. Idea is to guess simpler distribution of data (X_n in Figure 1) from the observed data (Y_n in Figure 1). The simpler distribution should preserve the original distribution to some degree (in statistical sense).

Now let's use conventional notation that we observe X_n. To reduce the dimension of the data means to transform $(x_1, \ldots, x_p)$ into $(x_1, \ldots, x_q)$ where $q < p$. For example, clustering data into k numbers of cluser transforms each observed data vector X_n into its cluster assignment (e.g. one of k integers), where each cluster has a probabilistic model which uses β as its parameter. We focus only on linear cases, and we limit our projection space to $\mathbb{R}^q$.

3.1 Principal Component Analysis (PCA) & Factor Analysis (FA)

Dating back to Pearson (1901) and Hotelling (1933), there were many reinvention of this technique. Idea is to project the original data onto a lower dimensional manifold in the original data space. The free parameters define the manifold, and we present three different perspectives on how the objective is formulated.

1. Maximize the variance of the projection along each of the q components (Hotelling 1933). This criterion has identifiability issue (that there can be more than one subspace that satisfies the criterion), so in practice, the original data is normalized to have intercept at 0.
2. Minimize the reconstruction error (Pearson 1901). In other words, minimize the distance in the original space between the original data and its reconstruction from the projection values.
3. PCA and FA solutions are the MLE's of corresponding probabilistic model (Bishop 1996, Rowes 1998, Schapire 2001). This is discussed more in detail below:

Let's define a probabilistic model as follows, where $\vec{Z}$ is $q-$dimensional random variable, and p-dimensional random variable $\vec{X}$ corresponds to the observed data:

- $\vec{Z} \sim N_q\left(\vec{0}, I\right)$. N_q is q-dimensional multivariate Gaussian distribution, and I is a diagonal identity matrix.
- $\vec{X} \sim N_p\left(\vec{\mu} + \Lambda Z, \Sigma\right)$. When $\Sigma = I\sigma^2$, then this corresponds to PCA (where all components have the same variance); when Σ is set to have different variances for each component, then this corresponds to FA.

In this setup, the followings are true:

- Projection of data X_n is equivalent to $\mathbb{E}\left[Z_n|X_n\right]$.
- Fitting Λ with MLE is equivalent to finding a subspace that satisfies criteria 1 and 2.
- MLE Λ can be found using EM algorithm (this will be covered in next class).