Generalized Linear Models

February 21, 2019

Extending Linear Models to Non-linear cases

- In L3, we discussed least squares solutions of non-linear models. One common type of nonlinear model is the Generalized Linear Model (GLM).
- Given an invertible function $g(\cdot)$, obersvations $\mathbf{x}$ with probability density function $p(\mathbf{x};\theta)$, a known matrix $\mathbf{H}$, and a vector of parameters θ, a GLM has two components:
  - A Linear Predictor: $\mathbf{H}\theta$
  - A Link Function: $g(\cdot)$, invertible, s.t.

$$\mathbb{E}[\mathbf{x}] = g^{-1}(\mathbf{H}\theta) = \mu$$

Extending Linear Models to Non-linear cases

- In L3, we discussed least squares solutions of non-linear models. One common type of nonlinear model is the Generalized Linear Model (GLM).
- Given an invertible function $g(\cdot)$, obersvations $\mathbf{x}$ with probability density function $p(\mathbf{x};\theta)$, a known matrix $\mathbf{H}$, and a vector of parameters θ, a GLM has two components:
  - A Linear Predictor: $\mathbf{H}\theta$
  - A Link Function: $g(\cdot)$, invertible, s.t.

$$\mathbb{E}[\mathbf{x}] = g^{-1}(\mathbf{H}\theta) = \mu$$

- Note: This is not the same as transforming $\mathbf{x}$ because $g(\mathbb{E}[\mathbf{x}]) \neq \mathbb{E}[g(\mathbf{x})]$

Exponential Family Distributions

- Let $p(\mathbf{x};\theta)$ belong to the exponential family, e.g.

$$p(\mathbf{x};\theta) = h(\mathbf{x})\exp(\eta^T\mathbf{T}(\mathbf{x}) - A(\eta))$$

- Then we know that $\mathbb{E}(\mathbf{T}(\mathbf{x})) = \frac{\partial A}{\partial \eta} = \mu$.
- The Canonical Link Function is

$$g_c(\mu) = \left(\frac{\partial A}{\partial \eta}\right)^{-1}$$

- Then the linear prediction is:

$$\mathbb{E}[\mathbf{x}] = \mu = \frac{\partial A}{\partial \eta}(\mathbf{H}\theta)$$

Exponential Family Continued

- To find θ, use the Mean Value Parameterization - change variables to μ.
- The pdf becomes:

$$p(\mathbf{x};\theta) = h(\mathbf{x})\exp(C(\mu)^T\mathbf{T}(\mathbf{x}) - D(\mu))$$

  where $C(\cdot)$ and $D(\cdot)$ are the functions resulting from the change of variables.
- Substituting $\mu = g^{-1}(\mathbf{H}\theta)$ we have

$$p(\mathbf{x};\theta) = h(\mathbf{x})\exp(C(g^{-1}(\mathbf{H}\theta))^T\mathbf{T}(\mathbf{x}) - D(g^{-1}(\mathbf{H}\theta)))$$

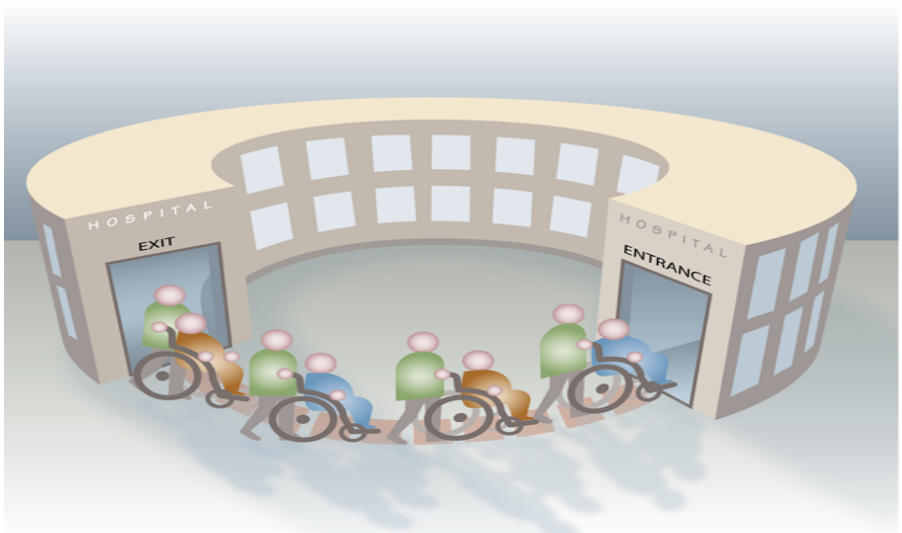
  and θ can be found via maximum likelihood estimation as normal.

Example Link Functions

- Binomial Distribution: Used for binary observations, $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$ - this is the case of Logistic Regression from your homework.
- Poisson Distribution: Used for integer observations, such as counting the clicks on a website or for arrival times at a train station.
  Link function is $g(\mu) = \ln(\mu)$.
- Exponential Distribution: Used for positive data, link is $g(\mu) = \frac{1}{\mu}$.
- GLM's are used when the assumptions for linear models don't hold, but you still want to find a "linear-ish" mapping between parameters and the information contained in $\mathbf{H}$.

Hospital Readmissions

- Hospital Readmission: patients are admitted to a hospital within 30 days of discharge - accrue significant additional hospital costs.
- Cost: The Agency for Healthcare Research and Quality reports that in 2011 an additional $41 billion dollars in hospital costs were caused by readmissions across the country.
- Overall Goal: Identify patients at high risk of readmission and reduce Barnes Jewish's annual readmissions rates to under 17.5%.



Graphic credit: Greg Cross/The Bulletin. Republished with permission.

Quick Data Overview

- Observations: $\mathbf{x}$ consists of the results of 776 patients who were either readmitted ($x[n] = 1$) or not-readmitted ($x[n] = 0$).
- Data: The matrix $\mathbf{H}$ has rows representing patients and columns representing:
  - LACE Score - current hospital risk metric
  - Whether the patient has diabetes
  - Principal Diagnoses - Heart Failure, Chronic Obstructory Pulminary Disease, Myocardial Infarction, or Pneumonia
  - Ethnicity
  - Gender
  - Month patient was treated
  - Length of Stay
  - Access to Primary Care Provider
  - Age
  - Discharge Status - back to home, to rehab, etc.
  - Zipcode - very rough proxy for income/demographic data

Parameters

- In a logistic regression model, the paramters $\theta_1, ..., \theta_{11}$ will represent the impact of each of these factors on the readmissions rate.
- The Odds Ratio gives the relationship between variables and the observation, i.e.:

$$\text{OR} = \frac{p(\text{readmission; patient has diabetes})}{p(\text{readmission; patient does not have diabetes})}$$

- In the case of Linear Models, the odds ratio for each variable is simply θ_i.
- For Logistic Regression, the odds ratio for each variable is $\exp(\theta_i)$.

Logistic Regression Results

Predictor	Odds ratio	95% Confidence Interval	p value
LACE	1.22	(1.14, 1.31)	<0.001
COPD (vs CHF)	0.21	(0.11, 0.45)	<0.001
MI (vs CHF)	0.66	(0.45, 0.99)	<0.001
Discharged to Skilled Nursing Facility (vs Home)	0.50	(0.29, 0.86)	0.01
Discharged with Home Health (vs Home)	2.34	(1.57, 3.49)	<0.001
Male (vs Female)	1.97	(1.36, 2.86)	<0.001
LOS	0.97	(0.95, 0.99)	0.03
Age70-74yrs (vs 65-69 yrs)	0.69	(0.49, 0.98)	0.04
Age75-79 (vs 65-69 yrs)	1.78	(1.2, 2.6)	<0.001
Has PCP (vs no PCP)	1.77	(1.15, 2.72)	0.01

Figure 1: Odds Ratios, with confidence intervals and p-values for each estimated parameter. Variables with high p-values, bad confidence intervals are assigned $\theta_i = 0$ as they don't help the model. Results were originally reported at the 2018 GSA conference.

Connection to Machine Learning: Predicting Readmissions

- Once θ_i have been estimated, we want to predict whether or not patients will be readmitted, so define a Decision Threshold:

$$D(\mathbf{x}, T) = \begin{cases} 1 & \text{if } p(\text{readmission}; \theta) > T \\ 0 & \text{otherwise} \end{cases}$$

- Then we choose T by trying to maximize:
  - True Positives: Correctly predicted readmissions.
  - True Negatives: Correctly predicted safe patients.

- and minimize:
  - False Positive: Predicting that a safe patient will be readmitted.
  - False Negative: Predicting that a readmitted patient would not return.

- This is a Binary Classification Problem, and there is usually a tradeoff between the metrics we want to find.

Comments

- The best logistic regression models with decision thresholds predict roughly %60 of patients correctly.
- This is ... non-ideal.
- We tested many different models and were able to come up with a model of the form:

$$p(x;\theta) = \sum_{i=1}^{k} a_i f_i(\mathbf{H}, \theta)$$

  where a_i are constants, f_i represents logistic regression models trained on cleverly selected sub-sets of $\mathbf{H}$, and θ is a vector containing all the parameters for each f_i.
- While we have better classification performance, this model loses the easy interpretation of the GLM approach.