COMP 543: Tools & Models for Data Science

Generalized Linear Models

Chris Jermaine & Risa Myers

Rice University

Classical LR

- Last class
  - LR in closed form
  - LR using Gradient Descent
  - Discussion of issues with using LR to handle categorical data
- LR can be viewed as a generative statistical model with Normal error
- How?

Probabilistic Interpretation of Classic LR

- Given x_i, let $y_i \sim \text{Normal}(r \cdot x_i, \sigma^2)$
- Where we treat $r \cdot x_i$ as the expected value of the regression coefficients and the features of x
- Then, assuming iid data, the likelihood of data set is $\prod_i \text{Normal}(y_i | r \cdot x_i, \sigma^2)$
- We can replace the Normal function with its PDF

$$\prod_i \sigma^{-1} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(y_i - r \cdot x_i)^2 \sigma^{-2}}$$

Probabilistic Interpretation of Classic LR

$$\prod_i \sigma^{-1}(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(y_i - r \cdot x_i)^2 \sigma^{-2}}$$

- Take the log of this function to get the Log likelihood

$$LLH \propto \sum_i -\frac{1}{2}(y_i - r \cdot x_i)^2 \sigma^{-2}$$

- And an MLE over r is going to try to maximize

$$-\sum_i (y_i - r \cdot x_i)^2$$

- Same loss function as LR!!
- This looks a lot like minimizing the squared loss
- But: note the negative sign!

People Noticed This a Long Time Ago

- And wondered:
- Can I use other error models (besides Normal error) with LR?
- Answer, naturally, is yes!

Generalized Linear Models (GLM)

- Generalization of LR
- Allows error to be generated by a wide variety of distributions
- In particular, any in the "exponential family"

When is a Distribution in the Exponential Family?

- Any probability distribution that can be written in this canonical form:

$$p(y|\theta) = b(y)\exp(\theta T(y) - f(\theta))$$

- θ are the natural parameters
- y is the output
- b and T are some arbitrary functions
- f is some function function of θ

Example: Bernoulli

- Recall the Bernoulli distribution, which models a coin flip
- {Tails, Heads} = {0, 1}
- First, write Bernoulli as:

$$
\begin{aligned}
p(y|p) &= p^y \times (1-p)^{(1-y)} \\
&= \exp(y \log p + (1-y)\log(1-p)) \\
&= \exp((\log p - \log(1-p))y + \log(1-p))
\end{aligned}
$$

- p is the natural parameter for Bernoulli

Example: Bernoulli

- First, write Bernoulli as:

$$\begin{aligned} p(y|p) =& p^y \times (1-p)^{(1-y)} \\ =& \exp(y \log p + (1-y)\log(1-p)) \\ =& \exp((\log p - \log(1-p))y + \log(1-p)) \end{aligned}$$

- Recall, exponential family distribution that can be written as:

$$p(y|\theta) = b(y)\exp(\theta T(y) - f(\theta))$$

- So we have:
  - θ is $(\log p - \log(1-p))$ or $\log(p/(1-p))$
  - $f(\theta)$ = $-\log(1-p) = \log(1+e^{\theta})$
  - $T(y)$ is y
  - $b(y)$ is 1
- Here θ is the “natural parameter” of the distribution

Example: Normal

- Assume the variance is 1 (for simplicity):

$$
\begin{aligned}
p(y|\mu) &= \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}(y-\mu)^2) \\
&= \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}y^2 + y\mu - \frac{1}{2}\mu^2) \\
&= \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}y^2) \exp(\mu y - \frac{1}{2}\mu^2)
\end{aligned}
$$

- Which is the Normal distribution in canonical form

Example: Normal

- If variance is 1 (for simplicity):

$$p(y|\mu) = \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}y^2) \exp(\mu y - \frac{1}{2}\mu^2)$$

- Recall, exponential family distribution that can be written as:

$$p(y|\theta) = b(y) \exp(\theta T(y) - f(\theta))$$

- So we have:
  - θ is μ
  - $f(\theta) = \frac{1}{2}\theta^2$
  - $T(y)$ is y
  - $b(y)$ is $\frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}y^2)$

This Brings Us to GLMs

- Say we have a prediction problem where:

1. We want to predict output y from an input vector x
2. It is natural to assume randomness/error/uncertainty on y is produced by some exponential family
3. The exponential family parameter θ is **linearly related** to x: that is, assuming x is vector-valued:

$$\theta = \sum_j x_j \times r_j$$

- Then this is known as an instance of a "generalized linear model"
- E.g.: We might use a Poisson distribution, to predict an arrival time with some error or uncertainty

LLH Of Data Produced by GLM

- From GLM definition, likelihood of the data set is:

$$\prod_i b(y_i) \exp(\theta_i T(y_i) - f(\theta_i))$$

- Where θ_i is produced by the dot product of the feature vector and the regression coefficients
- Substituting $x_{i,j} \times r_{i,j}$ for θ_i, we have:

$$\prod_i b(y_i) \exp\left(T(y_i) \sum_j \left(x_{i,j} \times r_{i,j}\right) - f\left(\sum_j x_{i,j} \times r_{i,j}\right)\right)$$

- Take the log to get the LLH:

$$\sum_i \left(\log b(y_i) + T(y_i) \sum_j \left(x_{i,j} \times r_{i,j}\right) - f\left(\sum_j \left(x_{i,j} \times r_{i,j}\right)\right)\right)$$

Then, For Any Member of Exponential Family...

- Just substitute and then maximize to learn the model!
- Choose the r vector to maximize the log likelihood

Example: Logistic Regression

- LLH for GLM is:

$$\sum_i \left(\log b(y_i) + T(y_i) \sum_j \left(x_{i,j} \times r_{i,j} \right) - f\left(\sum_j \left(x_{i,j} \times r_{i,j} \right) \right) \right)$$

- For Bernoulli data have:
  - θ is $(\log p - \log(1-p))$ or $\log(p/(1-p))$
  - $f(\theta)$ = $-\log(1-p) = \log(1+e^{\theta})$
  - $T(y)$ is y
  - $b(y)$ is 1
- Substituting (and letting $\theta_i = x_i \cdot r$)

$$\sum_i \log 1 + y_i(x_i \cdot r) - \log(1 + e^{x_i \cdot r})$$

Example: Logistic Regression

$$\sum_i \log 1 + y_i(x_i \cdot r) - \log(1 + e^{x_i \cdot r})$$

- Dropping the $\log 1$ and maximizing wrt r gives us logistic regression
- How to maximize?
  - Use any method we've discussed
  - Typically using gradient **ascent**
- How to predict?
  - Given r, x_i make a prediction for unknown y_i, choose y_i to max LLH
  - That is, choose y_i to match sign of $x_i \cdot r$

Example: Linear Regression

- LLH for GLM is:

$$\sum_i \left(\log b(y_i) + T(y_i) \sum_j \left(x_{i,j} \times r_{i,j} \right) - f\left(\sum_j \left(x_{i,j} \times r_{i,j} \right) \right) \right)$$

- For Normal data have:
  - θ is μ
  - $f(\theta) = \frac{1}{2}\mu^2$
  - $T(y)$ is y
  - $b(y)$ is $\frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}y^2)$
- Plug in the values ...

Example: Linear Regression

- Substituting (and letting $\theta_i = x_i \cdot r$)
- Notice: the natural parameter $x_i \cdot r$ is a linear function of the feature vector

$$
\begin{aligned}
&\sum_i \log\left(\frac{1}{\sqrt{(2\pi)}}\exp(-\frac{1}{2}y_i^2)\right) + y_i(x_i \cdot r) - \frac{1}{2}(x_i \cdot r)^2 \\
&= \sum_i \log\frac{1}{\sqrt{(2\pi)}} - \frac{1}{2}y_i^2 + y_i(x_i \cdot r) - \frac{1}{2}(x_i \cdot r)^2 \\
&= \sum_i \log\frac{1}{\sqrt{(2\pi)}} - \frac{1}{2}(y_i^2 - 2y_i(x_i \cdot r) + (x_i \cdot r)^2) \\
&= \sum_i \log\frac{1}{\sqrt{(2\pi)}} - \frac{1}{2}(y_i - (x_i \cdot r))^2
\end{aligned}
$$

- Maximizing this LLH wrt r gives us linear regression

Some Thoughts on Linear Regression

$$\sum_i \log \frac{1}{\sqrt{(2\pi)}} - \frac{1}{2}(y_i - (x_i \cdot r))^2$$

- First term has no bearing on the maximization
- Second term is the negation of the squared error

Some Thoughts on GLM

- Key points
  - For the exponential family of distributions
  - Which is pretty much everything (except uniform)
  - θ is the natural parameter
  - θ is a **linear** function of the features
  - θ can be vector, but is often a single parameter
  - Sometimes you learn multiple models using the different exponential distributions and choose the best
  - This if meaningful if you have a single natural parameter
    - Normal(μ,1) vs. Poisson(λ)

How do You Choose the Distribution?

- Part black magic
- Part experience
- Part math
- Keep in mind the common uses for the distributions
  - Poisson - arrival times, time to completion
  - Bernoulli - coin flip
  - ...

More Thoughts on GLM

- Why bother?
  - Least squares may not make sense for our application
  - E.g. Classification
  - Or predicting a during (non-negative value)
  - Or choosing 1 of N categories
- GLM gives us a way to extend linear regression to other distributions

Other Common GLMs

- Poisson Regression
- Multinomial Regression
- Binomial Regression
- ...

Questions?