Tools & Models for Data Science

Generalized Linear Models

Chris Jermaine & Risa Myers

Rice University

Last Class: Classical Linear Regression

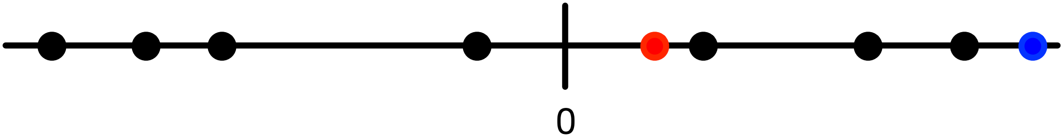
- LR in closed form

$$\hat{r} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- LR using Gradient Descent
  - Using the Mean Squared Error Loss function:

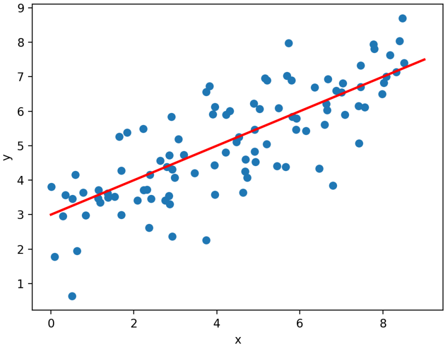
$$\frac{\sum_i (y_i - x_i \cdot r)^2}{n}$$

- Introduction to issues with using LR to handle categorical data

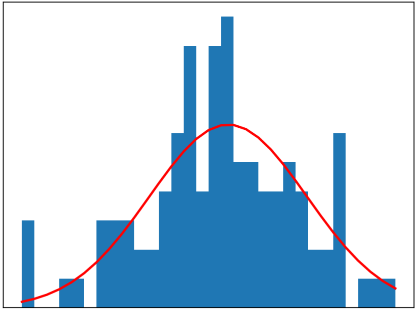


Linear Regression: Generative Statistical Model with Normal Error

Data and LR line

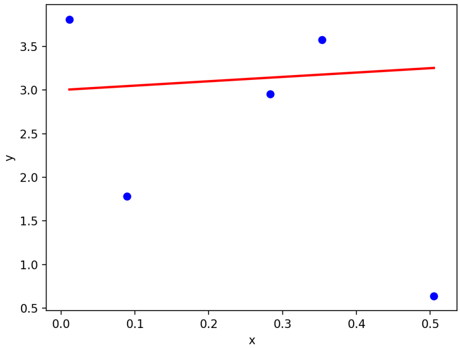


Histogram of error



Where is the Error?

Residuals



Probabilistic Interpretation of Classic LR

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

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

Probabilistic Interpretation of Classic LR

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

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

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

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

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

- Same loss function as LR, when divided by n
- This looks a lot like minimizing the squared loss
- But: note the negative sign!
  - Because we are maximizing instead of minimizing, so invert the function

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”

Which Distributions are in the Exponential Family?

?

Which Distributions are in the Exponential Family?

- Normal
- Bernoulli
- Exponential
- Chi-squared
- Dirichlet
- Poisson
- ...

? What determines if a distribution is 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 of θ

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 $\boldsymbol{y}$ from an input vector $\boldsymbol{x}$
2. It is natural to assume randomness/error/uncertainty on $\boldsymbol{y}$ is produced by some exponential family
3. The exponential family parameter θ is **linearly related** to x

$$\theta = X\boldsymbol{r} = \begin{bmatrix} \text{———} & \boldsymbol{x_1} & \text{———} \\ & \vdots & \\ \text{———} & \boldsymbol{x_i} & \text{———} \\ & \vdots & \\ \text{———} & \boldsymbol{x_n} & \text{———} \end{bmatrix} \times \begin{bmatrix} \\ \\ \boldsymbol{r} \\ \\ \\ \end{bmatrix} = \begin{bmatrix} x_1 r_1 \\ \vdots \\ x_i r_i \\ \vdots \\ x_n r_n \end{bmatrix}$$

- 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:

$$\Pi Pr(y_i|\theta) = \prod_i b(y_i)\exp(\theta_i T(y_i) - f(\theta_i))$$

- Where θ_i is produced by the dot product of the feature vector and the regression coefficients

$$\prod_i b(y_i)\exp\left(T(y_i)\left(X_i\cdot\boldsymbol{r}\right) - f\left(X_i\cdot\boldsymbol{r}\right)\right)$$

- Take the log to get the LLH:

$$\sum_i \left(\log b(y_i) + T(y_i)X_i\cdot\boldsymbol{r}\right) - f\left(X_i\cdot\boldsymbol{r}\right)\Big)\Big)$$

Then, For Any Member of Exponential Family...

- We have

$$LLH = \sum_i \left(\log b(y_i) + T(y_i) X_i \cdot \boldsymbol{r} \right) - f\left(X_i \cdot \boldsymbol{r} \right)))$$

- Maximize to learn the model
  - Take the derivative and set to 0, or
  - Use Gradient Descent to determine $\boldsymbol{r}$
- Let's look at another example:

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 in exponential form 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, for Bernoulli, we have:
  - θ is $(\log p - \log(1-p))$ = $\log(\frac{p}{1-p})$
  - $f(\theta)$ = $-\log(1-p) = \log(1+e^{\theta})$
  - $T(y)$ is y
  - $b(y)$ is 1
- Here θ is the “natural parameter” of the distribution

Example: Logistic Regression

- Plugging in the Bernoulli expression into the LLH
- LLH for GLM is:

$$\sum_i \left(\log b(y_i) + T(y_i) X_i \cdot \boldsymbol{r}\right) - f\left(X_i \cdot \boldsymbol{r}\right)\Big)\Big)$$

- 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 in these values (and letting $\theta_i = X_i \cdot \boldsymbol{r}$)

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

Example: Logistic Regression

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

- Dropping the $\log 1$ and maximizing wrt $\boldsymbol{r}$ gives us logistic regression
- ? Why can we drop the $\log 1$
- How to maximize?
  - Use any method we've discussed
  - Typically using gradient **ascent**
    - Ascent, not descent because we are typically using an MLE, which is a maximization problem

Example: Logistic Regression

$$LLH = \sum_i \log 1 + y_i(X_i \cdot \boldsymbol{r}) - \log(1 + e^{X_i \cdot \boldsymbol{r}})$$

- How to predict?
  - Given $\boldsymbol{r}, X_i$ make a prediction for unknown y_i, choose y_i to maximize LLH
  - That is, choose y_i to match sign of $X_i \cdot \boldsymbol{r}$
  - Note that no closed form exists

Prediction using Logistic Regression

$$LLH = \sum_i y_i(X_i \cdot \boldsymbol{r}) - \log(1 + e^{X_i \cdot \boldsymbol{r}})$$

- Given $\boldsymbol{r}, X_i$ make a prediction for unknown y_i, choose y_i to maximize LLH
- That is, choose y_i to match sign of $x_i \cdot r$

- Example
- Look at a lesion. Is it breast cancer or not?
- Learn $\boldsymbol{r}$
- At application time, given $\boldsymbol{r}$ and the new data $\boldsymbol{x}$, predict $\hat{y}$
- Answer 0 (no breast cancer) or 1 (breast cancer)
- Plug $\boldsymbol{x}$ and $\boldsymbol{r}$ into the equation
- Assign the label based on the sign of the computation

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
  - GLMs are meaningful if you have a single natural parameter
    - Normal(μ,1) vs. Poisson(λ)

How do You Choose the Distribution?

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

Why Bother with GLMs?

- Least squares or mean square error may not make sense for our application
  - Classification
  - Or predicting a duration (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?

- What do we know now that we didn't know before?

- How can we use what we learned today?