COMS 4771 HW4 (Spring 2025)

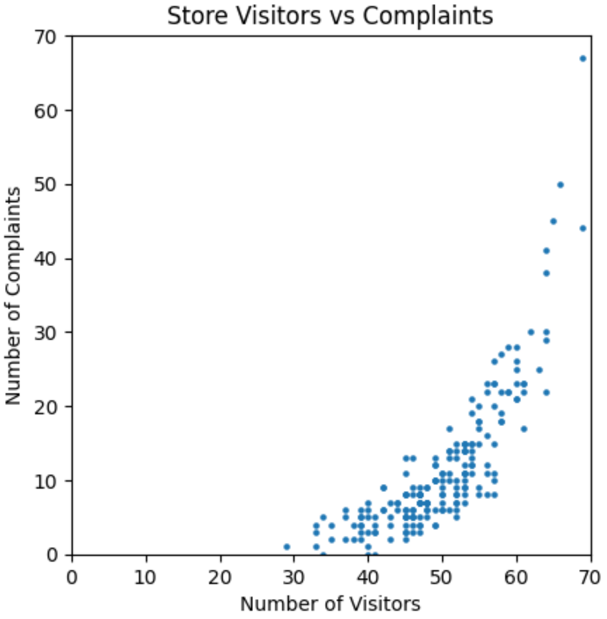
Brian Chen, Andrew Yang

1 Generalized Linear Models

In class, we have discussed linear regression, where we assume the output variable Y is some linear combination of some observed features $X \in \mathbb{R}^d$, and logistic regression, where the *log-odds* of the output variable are a linear combination of our features. In this problem, we will generalize this concept further in what are called **generalized linear models**.

1.1 Motivations for GLMs

OLS works great if our output variable is a linear combination of observed features. However, this assumption may not always hold in practice. Consider the following example data set comparing the number of visitors to a store to the number of complaints received:



(i) List some reasons why OLS is not suitable for this kind of data.

(ii) This is an example of 'count data', e.g. when the output variable is interpreted as the count of something (and therefore cannot be negative). Give some other example categories of output variables that are not suited to ordinary OLS.

1.2 Exponential family densities

In generalized linear models, we assume $Y|X$ has a probability density of the form

$$f(\mathbf{y}; \boldsymbol{\theta}) = \exp\left(\mathbf{T}(\mathbf{y}) \cdot \mathbf{m}(\boldsymbol{\theta}) - b(\boldsymbol{\theta}) + k(\mathbf{y})\right) \tag{1}$$

where $\mathbf{T}, \mathbf{m}$ are vector-valued functions, b, s, k are scalar-valued functions and $\boldsymbol{\theta}$ is a vector of parameters. Probability densities of this form are called 'exponential family densities', and $\boldsymbol{\theta}$ is called the 'parameter' of the family.

(iii) Show that the normal, Poisson, Bernoulli, and binomial distributions all belong to the exponential family. That is, specify what $\mathbf{T}, \mathbf{m}, b, k$ and $\boldsymbol{\theta}$ are for each distribution when written in the form of eq. 1. (Note for the binomial distribution we need to fix the number of trials n.) For reference, here are the normal distribution

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}},$$

the Poisson distribution

$$f(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!},$$

the Bernoulli distribution

$$f(y;p) = p^y(1-p)^{1-y} \tag{2}$$

and the Binomial distribution with fixed number of trials n

$$f_n(y;p) = \binom{n}{y} p^y (1-p)^{n-y}.$$

Do you notice anything familiar from your rewritings of these distributions? *Hint: one of them should contain terms that look like logistic regression!*

(iv) Argue why the Laplace distribution

$$f(y;\mu,b) = \frac{1}{2b} e^{-\frac{|y-\mu|}{b}}$$

is not a exponential family density.

(v) It is undesirable for the support of eq. 1 to change based on the parameter θ. As such, we also require the support of exponential family densities to be independent of θ. Name one class of densities excluded by this requirement.

(vi) Show that if T, m, y, θ are all scalar-valued,

$$\mathbb{E}[T(y)] = \frac{b'(\theta)}{m'(\theta)}$$

for a general exponential family density with compact support. (Hint: First consider the log-likelihood $L(y;\theta) = \ln(f(y;\theta))$. Then, directly compute $\mathbb{E}\left[\frac{\partial L}{\partial \theta}(y;\theta)\right]$ and use Leibniz's integral rule to argue the expectation is zero. Finally, relate $\mathbb{E}\left[\frac{\partial L}{\partial \theta}(y;\theta)\right]$ and $\mathbb{E}[T(y)]$ by linearity of expectation.)

In fact, one can always compute all moments of an exponential family density through its moment generating function. Not all densities have one, making this family particularly nice to work with.

1.3 A conjugate prior for the Bernoulli distribution

Exponential families have conjugate priors. This means for a likelihood $p(x|\theta)$, the prior $p(\theta)$ and posterior $p(\theta|x)$ are part of the same probability distribution family. Since the posterior can be computed

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{3}$$

$$= \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}, \tag{4}$$

having a conjugate prior often means the integral in the denominator of eq. 4 has an analytical solution, allowing the posterior to have a closed-form expression.

(vii) Show that if $p(x|\theta)$ and $p(\theta)$ are both exponential family densities, $p(x|\theta)p(\theta)$ is as well. This is good as integrals of exponential family densities often have nice analytical expressions.

(viii) Let your likelihood follow the Bernoulli distribution given in eq. 2 and $\theta \in [0,1]$. Let $\alpha, \beta \in \mathbb{N}$. Define our prior

$$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

through the Gamma function

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

Let $\mathbf{x} = (x_1, \cdots, x_n)^T$, where $x_i,\ i \in [n]$ are drawn i.i.d. from a Bernoulli distribution with parameter θ. Find functions $A(\mathbf{x})$ and $B(\mathbf{x})$ such that $p(\theta|\mathbf{x}) = p(\theta; A(\mathbf{x}), B(\mathbf{x}))$. It may be helpful to note

$$\int z^{\alpha-1}(1-z)^{\beta-1} dz = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

(ix) Given your answer to the previous part, show that if we observe a new x_{n+1} also drawn i.i.d. from the Bernoulli distribution with parameter θ, then we can write $A\left((\mathbf{x}, x_{n+1})^T\right) = A(\mathbf{x}) + a(x_{n+1})$ and $B\left((\mathbf{x}, x_{n+1})^T\right) = B(\mathbf{x}) + b(x_{n+1})$ for some functions a, b. This means that when new data comes in, we can easily update the parameters of the posterior. Furthermore, you do not need to keep track of any of the previous data $\mathbf{x}$ as all the information necessary is contained in the parameters $A(\mathbf{x}), B(\mathbf{x})$. We call $(A(\mathbf{x}), B(\mathbf{x}))^T$ a sufficient statistic as its value is sufficient to completely determine the posterior distribution.

All exponential family densities have a sufficient statistic $\mathbf{S}(\mathbf{x})$ of the easy-to-update form $\sum_{i=1}^n \mathbf{S}(x_i)$.

1.4 Defining GLMs

In this section we will motivate and define GLMs. We often want to estimate conditional probabilities given X, e.g. we want to estimate $\Pr[Y \mid X_i]$ for our possible input values for X_i. We could do this by just assuming all of these distributions $\Pr[Y \mid X_i]$ are independent of each other; this would amount to estimating some random distribution for every possible value of X_i. However, X might be related to the distribution $\Pr[Y \mid X_i]$ in some systematic way (indeed, this is likely the case!). GLMs give us a way to express this relationship.

In a generalized linear model, we make the following assumptions about the relationship between X and Y.

1. $Y \mid X$ is distributed according to some exponential family distribution, e.g. $Y \mid X \sim \mathcal{N}(\mu, \sigma^2)$ for OLS. This is called the **random component**.
2. The values of Y are related in some way to a linear combination of the input variables $\{X_i\}$. We'll call this linear combination $\eta \equiv \beta^T X$. This is called the **systemic component**.
3. The linear combination η can be mapped to $\mathbb{E}[Y \mid X] \equiv \mu$ through a smooth, invertible linking function

$$\eta = l(\mu)$$

   This is called the **link function**.

(x) With GLMs, we are interested in estimating $\mathbb{E}(y_i \mid x_i) \equiv \mu_i$. Suppose we are given i.i.d samples (x_i, y_i) for $i = 1, \ldots, n$ and each $y_i \mid x_i$ distribution is from the exponential family density (Eq. 1). Using the fact that $l(\mu_i) = \eta_i \equiv x_i^T \beta$, where l is the canonical link function, derive the log-likelihood function for β. Note: unlike in the Gaussian case, there is generally not a closed form for the maximizing β. In practice, we can use Newton's method to find the maximizing value.

(xi) Using your result from part (x), show that the likelihood function for the Poisson model is

$$l(\beta) = \sum_i y_i x_i^T \beta - \exp\left(x_i^T \beta\right)$$

This is called Poisson regression!

(xii) Let's return to the count data example from part (i) with our new approach to regression. The data csv can be found at [this google drive link]. Generate the Poisson regression fit to the given count data and compare it with the OLS solution. Does it avoid the issues described in part (i)?