GLMs - Exercises

Wilker Aziz

February 12, 2026

1 Exercises

Problem 1 (Collectables)**.** In this exercise we will model the market value of collectables (e.g., LPs) based on textual data attached to them (e.g., description by seller, opinions of people who own the same item, etc.).

Data. We have a collection of items and their selling prices in an online platform. For each item, we have textual context x and a vector $\mathbf{y}$ storing the selling prices for the last 20 times the item was sold. See Figure 1 for some examples.

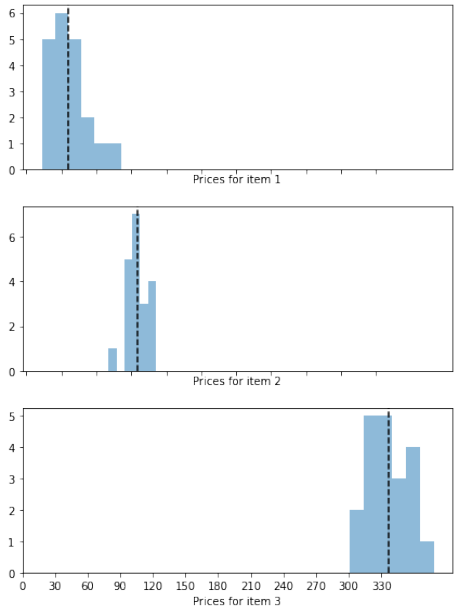

Figure 1: Histograms of selling price for 3 of the items in the collection. The dashed line is the mean selling price per item.

Task. Use the textual information x to predict the distribution of selling price for the item. Assume that for this task, we have already designed a good feature function $\mathbf{h}(x) \in \mathbb{R}^D$.

Question 1.1 ($\star$)**.** A first-year data analyst suggested that, for each item $(x, \mathbf{y})$, we take the average of the 20 measurements $\bar{y} = \frac{1}{S}\sum_{s=1}^{20} y_s$ and fit a linear regressor $g(x; \mathbf{w}, b) = \mathbf{w}^\top \mathbf{h}(x) + b$ with $\mathbf{w} \in \mathbb{R}^D$ and $b \in \mathbb{R}$. Explain at least 2 shortcomings of this idea. A good answer will likely ground the argument to observations about Figure 1.

A second-year data analyst, who has already taken NTMI, suggested a generalised linear model for textual data. She believes she has identified candidate distributions for a conditional model, her plan is to choose one of those candidates based on properties of the data, once she has decided, she intends to use the feature vector $\mathbf{h}(x)$ of an item to predict the parameter of that distribution's pdf (or pmf), then, she will assume the 20 measurements were each independently drawn from the conditional distribution prescribed by that pdf (or pmf). These are the candidates she chose:

1. Gamma distribution
2. Normal distribution

Moreover, one of her colleagues had suggested the Geometric distribution, which she discarded without running an experiment.

Question 1.2 ($\star$)**.** Explain why the analyst discarded the Geometric without running an experiment.

Question 1.3 ($\star$)**.** The analyst decided for the Gamma distribution. Reproduce what arguments she might have had for the Gamma and *against* the normal.

Let's design her Gamma GLM:

$$Y_s | X = x \sim \text{Gamma}(\alpha(x; \mathbf{w}, b), \beta(x; \mathbf{m}, c)) \tag{1}$$

$$\alpha(x; \mathbf{w}, b) = a(\mathbf{w}^\top \mathbf{h}(x) + b) \tag{2}$$

$$\beta(x; \mathbf{m}, c) = a(\mathbf{m}^\top \mathbf{h}(x) + c) \tag{3}$$

$$\tag{4}$$

where $\alpha(\cdot)$ predicts the Gamma's shape (strictly positive) and $\beta(\cdot)$ predicts the Gamma's rate (strictly positive). Each of these functions has their own parameters ($\mathbf{w}, b$ and $\mathbf{m}, c$, respectively).

Question 1.4 ($\star$)**.** State the shapes of the parameters of the GLM, and suggest an activation function $a(\cdot)$ that correctly constrains the linear predictors to valid Gamma parameters. Does this choice work for both the shape and the rate?

Question 1.5 ($\star$)**.** Use the pdf of the Gamma to state the log-likelihood function given a single item $(x, \mathbf{y})$ as a function of the *linear predictors* used in this model.

Question 1.6 ($\star$)**.** Use the log-likelihood function stated in the previous exercise, and assume that partial derivatives $\frac{\partial}{\partial w_d}\mathcal{L}_{x,\mathbf{y}}(\mathbf{w}, b, \mathbf{m}, \mathbf{c})$ for every $d = 1, \ldots, D$, $\frac{\partial}{\partial b}\mathcal{L}_{x,\mathbf{y}}(\mathbf{w}, b, \mathbf{m}, \mathbf{c})$, and similarly for $\mathbf{m}$ and c, are available to you without the need for manually computing them.

State the algorithmic steps necessary to go from the observation $(x, \mathbf{y})$ and an initial set of parameter values $\mathbf{w}^{(0)}, b^{(0)}\mathbf{m}^{(0)}, c^{(0)}$ to better parameter values

$\mathbf{w}^{(1)}, b^{(1)}, \mathbf{m}^{(1)}, c^{(1)}$ in an attempt to maximise the log-likelihood function of the model. You can develop your argument for a single pair $(x, \mathbf{y})$.

Question 1.7 ($\star$). Suppose you evaluate the model log-likelihood using M observed pairs $(x, \mathbf{y})$. Express the time complexity of this operation in units of time as a function of M and D. You may assume that

- assessing the Gamma pdf for a certain outcome, once the shape and rate parameters are known, takes one unit of time $\mathcal{O}(1)$;
- computing the feature vector $\mathbf{h}(x)$ takes D units of time.

Hint: it's easier if you use big-O notation.

2 Solutions

1.1 Nothing prevents the linear regressor from producing *negative* prices, after all, depending on the values $\mathbf{w}$ and b, the linear combination $\mathbf{w}^\top \mathbf{h}(x) + b$ could be *any* real value. Our measurements are always positive (they are prices), thus linear regression could make impossible predictions.

Another reason: by averaging the measurements, we loose information about the *spread* of the data points. For example, in Figure 1 we can see that the average (dashed line) cannot tell us that the prices for the first item spread quite a bit (from under EUR 30 to near EUR 120).

1.2 The Geometric mode is always 0 (or 1, depending on the version), while for different items in the plot, we can see histograms with modes that vary along the positive real line. The analyst probably knows the Geometric distribution well, and can already see, by comparison to the plots, that it is not flexible enough to model the kinds of distributions we need.

1.3 The Normal distribution is defined over the entire real line, thus outcomes sampled from the Normal distribution can be positive and/or negative, while our data (prices) are clearly always positive. Besides, the Normal distribution is symmetric about its mean (which is also its mode), and we can see in the plots that some histograms are not quite symmetric (for example, the first one is stretches more to the right than to the left). That is also intuitive: as prices are positive quantities, small prices (close to 0) could never distribute symmetrically, as symmetry would force some values to distribute over the negative line. The Gamma distribution is defined over the strictly positive real line, thus its outcomes are more compatible with prices. Besides, the Gamma distribution can take non-symmetric shapes.

1.4 As this GLM uses a D-dimensional feature function, we have $\mathbf{w} \in \mathbb{R}^D$, $b \in \mathbb{R}$, $\mathbf{m} \in \mathbb{R}^D$, and $c \in \mathbb{R}$. That is, we have $2D + 2$ parameters, each an unconstrained real value.

As the Gamma distribution requires strictly positive shape and rate parameters, we need an activation function that can correctly constrain each of the linear predictors to the strictly positive real line $\mathbb{R}_{>0}$. The exponential function can do that for us, thus $a(s) = \exp(s)$, which works for both of our linear predictors.

1.5 The Gamma distribution with shape α and rate β has pdf:

$$\text{Gamma}(y|\alpha, \beta) = \frac{\beta^\alpha y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)} \tag{5}$$

where the quantity in the denominator is the Gamma-function (a generalisation of the factorisal function that works for positive real numbers).

So, for an observation $(x, \mathbf{y})$, where $\mathbf{y} \in \mathbb{R}^{20}_{>0}$ are 20 price measurements,

we have the following log-likelihood function:

$$\mathcal{L}_{x,\mathbf{y}}(\mathbf{w}, b, \mathbf{m}, \mathbf{c}) = \sum_{s=1}^{20} \log \text{Gamma}(y_s | \alpha(x; \mathbf{w}, b), \beta(x; \mathbf{m}, c)) \tag{6}$$

$$= \sum_{s=1}^{20} \log \frac{\beta(x; \mathbf{m}, c)^{\alpha(x;\mathbf{w},b)} y_s^{\alpha(x;\mathbf{w},b)-1} e^{-\beta(x;\mathbf{m},c)y_s}}{\Gamma(\alpha(x; \mathbf{w}, b))} \tag{7}$$

$$= \sum_{s=1}^{20} \alpha(x; \mathbf{w}, b) \log \beta(x; \mathbf{m}, c) \tag{8}$$

$$+ (\alpha(x; \mathbf{w}, b) - 1) \log y_s \tag{9}$$

$$- \beta(x; \mathbf{m}, c) y_s - \log \Gamma(\alpha(x; \mathbf{w}, b)) \tag{10}$$

$$= \sum_{s=1}^{20} \exp\Big(\mathbf{w}^\top \mathbf{h}(x) + b\Big)(\mathbf{m}^\top \mathbf{h}(x) + c) \tag{11}$$

$$+ \Big(\exp\Big(\mathbf{w}^\top \mathbf{h}(x) + b\Big) - 1\Big) \log y_s \tag{12}$$

$$- \exp\Big(\mathbf{m}^\top \mathbf{h}(x) + c\Big) y_s - \log \Gamma\left(\exp\Big(\mathbf{w}^\top \mathbf{h}(x) + b\Big)\right) \tag{13}$$

where the first step is true because the measurements are independent of one another given the Gamma parameters $\alpha(x; \mathbf{w}, b)$ and $\beta(x; \mathbf{m}, c)$, which the GLM predicts from x; the second step is just applying the definition of the Gamma pdf and evaluating it for each of measurements y_s, the third step applies properties of logarithm to simplify the expression a bit; the last step rewrites the Gamma parameters as a function of the linear predictors $\mathbf{w}^\top \mathbf{h}(x) + b$ and $\mathbf{m}^\top \mathbf{h}(x) + c$ for this document.

1.6

1. We compute the feature vector for x, that is, $\mathbf{h}(x)$;
2. We then compute the linear predictors $s_1 = \mathbf{w}^{(0)\top} \mathbf{h}(x) + b^{(0)}$ and $s_2 = \mathbf{m}^{(0)\top} \mathbf{h}(x) + c^{(0)}$;
3. We then apply the exponential activation function to obtain the Gamma parameters: $\alpha(x; \mathbf{w}^{(0)}, b^{(0)}) = \exp(s_1)$ and $\beta(x; \mathbf{m}^{(0)}, c^{(0)}) = \exp(s_2)$;
4. Next we assess the log-likelihood function $\mathcal{L}_{x,\mathbf{y}}(\mathbf{w}^{(0)}, b^{(0)}, \mathbf{m}^{(0)}, c^{(0)})$, which requires assessing the logarithm of the Gamma pdf for each of the 20 measurements and summing those values:
   $L = \sum_{s=1}^{20} \log \text{Gamma}(y_s | \exp(s_1), \exp(s_2))$;
5. We then obtain gradients of this quantity with respect to the parameters of the GLM, each coordinate of the gradient vector is a partial derivative (pdv). We have pdvs for each of the weights $\frac{\partial}{\partial w_d^{(0)}} L$ and each of the weights $\frac{\partial}{\partial m_d^{(0)}} L$ as well as for each of the biases $\frac{\partial}{\partial b^{(0)}} L$ and $\frac{\partial}{\partial c^{(0)}} L$;
6. With those pdvs, we can compute parameter updates:

$$
\begin{aligned}
w_d^{(1)} &= w_d^{(0)} + \gamma \frac{\partial}{\partial w_d^{(0)}} L && \text{for } d = 1, \ldots, D \\
b_d^{(1)} &= b_d^{(0)} + \gamma \frac{\partial}{\partial b^{(0)}} L \\
m_d^{(1)} &= m_d^{(0)} + \gamma \frac{\partial}{\partial m_d^{(0)}} L && \text{for } d = 1, \ldots, D \\
c_d^{(1)} &= c_d^{(0)} + \gamma \frac{\partial}{\partial c^{(0)}} L
\end{aligned}
$$

where $\gamma > 0$ is a learning rate.

1.7 Let's start with predicting the Gamma parameters for 1 document:

- To assess the Gamma pdf of an observation, we first need the Gamma parameters, which the GLM predicts for a document. Here's what we need to do: featurise the document, which takes time $\mathcal{O}(D)$; then take dot product and add bias (for the first parameter), then do it again for the second parameter, which takes time $\mathcal{O}(2D)$. Altogether we have $\mathcal{O}(D + 2D) = \mathcal{O}(D)$.

Once we have this Gamma, we can assess the pdf of all 20 measurements we have for it, the exercise said the pdf can be assessed in unit time once we have the parameters so we have $\mathcal{O}(D + 20) = \mathcal{O}(D)$.

We have just concluded that for 1 observed pair $(x, \mathbf{y})$ it takes time $\mathcal{O}(D)$ to assess the quantities needed for the log-likelihood function. Thus, for M observed pairs we repeat all that M times, which gives us $\mathcal{O}(M \times D)$.