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.

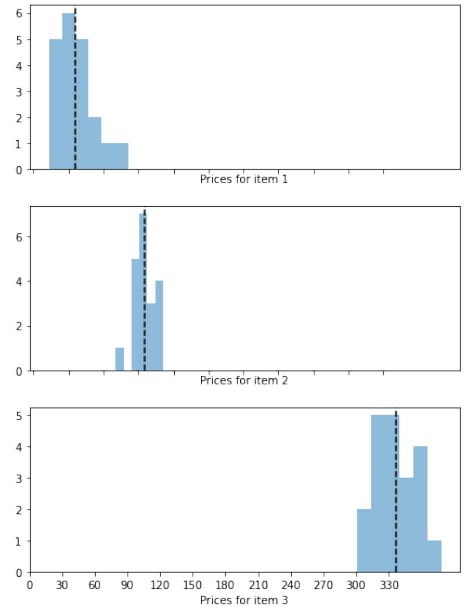

Figure 1: Histograms of selling price for 3 of the items in the collection. The dashed line is the mean selling price per item.

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:

$$\begin{aligned}
Y_s|X = x &\sim \mathrm{Gamma}(\alpha(x; \mathbf{w}, b), \beta(x; \mathbf{m}, c)) && (1)\\
\alpha(x; \mathbf{w}, b) &= a(\mathbf{w}^\top \mathbf{h}(x) + b) && (2)\\
\beta(x; \mathbf{m}, c) &= a(\mathbf{m}^\top \mathbf{h}(x) + c) && (3)\\
& && (4)
\end{aligned}$$

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 $d = 1, \dots, D$, $\frac{\partial}{\partial b}\mathcal{L}_{x,\mathbf{y}}(\mathbf{w}, b, \mathbf{m}, \mathbf{c})$, 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

Available after class.