# Chapter 7

# Generalized Linear Models

In previous sections, we saw that the statistical framework provided valuable insights into linear regression, especially with respect to explicitly stating most of the assumptions in the system. These assumptions were necessary to rigorously estimate parameters of the model, which could then be subsequently used for prediction on previously unseen data. In this section, we introduce generalized linear models (GLMs) which extend ordinary least-squares regression beyond Gaussian probability distributions and linear dependencies between the features and the target. This generalization will also introduce you to a broader range of loss functions, called Bregman divergences.

We shall first revisit the main points of the ordinary least-squares regression. There, we assumed that a set of i.i.d. data points with their targets $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ were drawn according to some distribution $p(\mathbf{x}, y)$. We also assumed that an underlying relationship between the features and the target was linear; i.e.,

$$Y = \sum_{j=0}^{d} \omega_j X_j + \varepsilon,$$

where $\boldsymbol{\omega}$ was a set of unknown weights and $\varepsilon$ was a zero-mean normally distributed random variable with variance $\sigma^2$. To simplify generalization, we will slightly reformulate this model. In particular, it will be useful to separate the underlying linear relationship between the features and the target from the fact that $Y$ was normally distributed. That is, we will write that

1. $\mathbb{E}[y|\mathbf{x}] = \boldsymbol{\omega}^\top \mathbf{x}$
2. $p(y|\mathbf{x}) = \mathcal{N}(\mu, \sigma^2)$

with $\mu = \boldsymbol{\omega}^\top \mathbf{x}$ connecting the two expressions. This way of formulating linear regression will allow us (*i*) to generalize the framework to non-linear relationships between the features and the target as well as (*ii*) to use the error distributions other than Gaussian.

## 7.1 Exponential transfer and the Poisson distribution

We will start first with an example of a GLM, before moving on to the general class and general definition. Assume that data points correspond to cities in the world—described by some numerical features—and that the target variable is the number of sunny days observed in a particular year. The target variable $y$ may look like a Poisson distribution, given features $\mathbf{x}$. It would be more natural, therefore, to model $p(y|\mathbf{x}) = \text{Poisson}(\lambda)$, where $\lambda > 0$ is the parameter (mean) of the Poisson distribution: $\mathbb{E}[y|\mathbf{x}] = \lambda$. However, because

$\lambda \in \mathbb{R}^+$, it would not be appropriate to model $\lambda$ with $\boldsymbol{\omega}^\top \mathbf{x} \in \mathbb{R}$. Rather, we would like to transfer our linear prediction with some function $f$ to adjusts the range of the linear combination of features to the domain of the parameters of the probability distribution.

We can do so by introducing an exponential transfer for this Poisson distribution, and more generally, any invertible transfer function $f$. If we can instead estimate $\boldsymbol{\omega}$ such that $\lambda = e^{\boldsymbol{\omega}^\top \mathbf{x}}$, then we can guarantee our estimates are in the correct range. Alternatively, one can consider that we are learning a linear weighting of features to learn a transformed parameter, $\log(\lambda) = \boldsymbol{\omega}^\top \mathbf{x}$. This simple modification is why these models are called *generalized* linear models, because the key component is a still a linear weighting. We formalize the types of distributions and transfers that can be considered in the below sections, but first finish off this example with Poisson regression to provide a concrete example.

To establish the GLM model for Poisson regression, we assume (1) an exponential transfer between the expectation of the target and linear combination of features, and (2) the Poisson distribution for the target variable.

1. $\mathbb{E}[y|\mathbf{x}] = \exp\left(\boldsymbol{\omega}^\top \mathbf{x}\right)$ or $\log(\mathbb{E}[y|\mathbf{x}]) = \boldsymbol{\omega}^\top \mathbf{x}$
2. $p(y|\mathbf{x}) = \text{Poisson}(\lambda)$

Exploiting the fact that $E\left[y|\mathbf{x}\right] = \lambda$, we connect the two formulas using $\lambda = e^{\boldsymbol{\omega}^\top \mathbf{x}}$. The resulting probability distribution is

$$p(y|\mathbf{x}) = \frac{e^{\boldsymbol{\omega}^\top \mathbf{x} y} \cdot e^{-e^{\boldsymbol{\omega}^\top \mathbf{x}}}}{y!}$$

for any $y \in \mathbb{N}$.

We can use maximum likelihood estimation to find the parameters of the regression model. The log-likelihood function has the form

$$\begin{aligned} ll(\mathbf{w}) &= \sum_{i=1}^{n} ll_i(\mathbf{w}) \\ ll_i(\mathbf{w}) &= \mathbf{w}^\top \mathbf{x}_i y_i - e^{\mathbf{w}^\top \mathbf{x}_i} - \ln y_i! \end{aligned}$$

Our goal now is to maximize the log-likelihood function. It is easy to see that $\nabla ll(\mathbf{w}) = \mathbf{0}$ does not have a closed-form solution. Therefore, unlike in ordinary least-squares regression, we will have to use gradient descent. We could choose to use first-order or second-order gradient descent, and batch or stochastic gradient descent. The key step in any of these is to first compute the gradient for one example. We start by deriving the partial derivative of the log-likelihood

$$\begin{aligned} \frac{\partial ll_i(\mathbf{w})}{\partial w_j} &= x_{ij} y_i - e^{\mathbf{w}^\top \mathbf{x}_i} x_{ij} \\ &= x_{ij} \left(y_i - e^{\mathbf{w}^\top \mathbf{x}_i}\right). \end{aligned}$$

The gradient for one sample is

$$\nabla ll_i(\mathbf{w}) = \mathbf{x}_i \cdot (y_i - p_i)$$

where $p_i = e^{\mathbf{w}^\top \mathbf{x}_i}$ is the prediction. Notice that $y_i - p_i$ corresponds to a prediction error, for the $i^{\text{th}}$ example. The batch gradient is

$$\begin{aligned} -\nabla ll(\mathbf{w}) &= \sum_{i=1}^{n} \nabla ll_i(\mathbf{w}) \\ &= \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - p_i\right) \\ &= \mathbf{X}^\top \left(\mathbf{y} - \mathbf{p}\right) \end{aligned} \tag{7.1}$$

where $\mathbf{p}$ is a vector with elements $p_i = e^{\mathbf{w}^\top \mathbf{x}_i}$ and $\mathbf{p} - \mathbf{y}$ is an error vector.

Commonly, one would now just do stochastic or batch gradient descent. For stochastic gradient descent, each step consists of using the gradient for one example and for batch gradient descent, each step consists of using the gradient for all examples. We can additionally consider the Hessian matrix, both to evaluate the properties of the stationary points as well as to allow for second-order gradient descent—though it is likely too expensive if $d$ is large. The second partial derivative of the log-likelihood function for one example is

$$\begin{aligned} \frac{\partial^2 ll_i(\mathbf{w})}{\partial w_j \partial w_k} &= -x_{ij} e^{\mathbf{w}^\top \mathbf{x}_i} x_{ik} \\ &= -x_{ij} p_i x_{ik} \end{aligned}$$

with

$$\frac{\partial^2 ll(\mathbf{w})}{\partial w_j \partial w_k} = \sum_{i=1}^{n} \frac{\partial^2 ll_i(\mathbf{w})}{\partial w_j \partial w_k}.$$

For $\mathbf{P}$ an $n \times n$ diagonal matrix with $p_i$ on the diagonal, the Hessian matrix is therefore

$$H_{ll(\mathbf{w})} = -\mathbf{X}^\top \mathbf{P} \mathbf{X}. \tag{7.2}$$

This matrix is negative definite if $\mathbf{X}$ is not low-rank, which would mean there is only one stationary point and that it is the global maximum. In fact, we know that the objective for Poisson regression is concave, even if $\mathbf{X}$ is not full rank, and so all stationary points are global maxima.

## 7.2 Exponential family distributions

In the previous section, we used a specific example to illustrate how to generalize beyond Gaussian distributions. The approach more generally extends to any exponential family distribution. For simplicity, here we focus on the natural exponential family, which is sufficient for most generalized linear models. The natural exponential family is a class of probability distributions with the following form

$$p(x|\theta) = \exp\left(\theta x - a(\theta) + b(x)\right)$$

where $\theta \in \mathbb{R}$ is the parameter to the distribution, $a : \mathbb{R} \to \mathbb{R}$ is a log-normalizer function and $b : \mathbb{R} \to \mathbb{R}$ is a function of only $x$ that will typically be ignored in our optimization because it is not a function of $\theta$. Many of the often encountered (families of) distributions

are members of the exponential family; e.g. exponential, Gaussian, Gamma, Poisson, or the binomial distributions. Therefore, it is useful to generically study the exponential family to better understand commonalities and differences between individual member functions.

**Example 17:** The Poisson distribution can be expressed as

$$p(x|\lambda) = \exp\left(x \log \lambda - \lambda - \log x!\right),$$

where $\lambda \in \mathbb{R}^+$ and $\mathcal{X} = \mathbb{N}_0$. Thus, $\theta = \log \lambda$, $a(\theta) = e^\theta$, and $b(x) = -\log x!$. □

Now let us get some further insight into the properties of the exponential family parameters and why this class is convenient for estimation. The function $a(\theta)$ is typically called the log-partitioning function or simply a log-normalizer. It is called this because

$$a(\theta) = \log \int_{\mathcal{X}} \exp\left(\theta x + b(x)\right) dx$$

and so plays the role of ensuring that we have a valid density: $\int_{\mathcal{X}} p(x)dx = 1$. Importantly, for many common GLMs, the derivative of $a$ corresponds to the inverse of the link function. For example, for Poisson regression, the link function $g(\theta) = \log(\theta)$, and the derivative of $a$ is $e^\theta$, which is the inverse of $g$. Therefore, as we discuss below, the log-normalizer for an exponential family informs what link $g$ should be used (or correspondingly the transfer $f = g^{-1}$).

The properties of this log-normalizer are also key for estimation of generalized linear models. It can be derived that

$$\begin{aligned} \frac{\partial a(\theta)}{\partial \theta} &= \mathbb{E}\left[X\right] \\ \frac{\partial^2 a(\theta)}{\partial \theta^2} &= \mathrm{V}[X] \end{aligned}$$

## 7.3 Formalizing generalized linear models

We shall now formalize the generalized linear models. The two key components of GLMs can be expressed as

1. $\mathbb{E}[y|\mathbf{x}] = f(\boldsymbol{\omega}^\top \mathbf{x})$ or $g(\mathbb{E}[y|\mathbf{x}]) = \boldsymbol{\omega}^\top \mathbf{x}$ where $g = f^{-1}$
2. $p(y|\mathbf{x})$ is an Exponential Family distribution

The function $f$ is called the transfer function and $g$ is called the link function. For Poisson regression, $f$ is the exponential function, and as we shall see for logistic regression, $f$ is the sigmoid function. The transfer function adjusts the range of $\boldsymbol{\omega}^\top \mathbf{x}$ to the domain of $Y$; because of this relationship, link functions are usually not selected independently of the distribution for $Y$. The generalization to the exponential family from the Gaussian distribution used in ordinary least-squares regression, allows us to model a much wider range of target functions. GLMs include three widely used models, linear regression, Poisson regression and logistic regression, which we will talk about in the next chapter about classification.

To relate these more clearly to exponential family distributions, we have to consider conditional distributions. Each $p(y|\mathbf{x})$ is an exponential family distribution, with parameter

$\theta = \mathbf{x}^\top \mathbf{w}$. When learning $\mathbf{w}$—by maximizing likelihood—we are learning the parameter $\theta_i$ for each sample $(\mathbf{x}_i, y_i)$. The general log-likelihood is

$$\begin{aligned} ll(\mathbf{w}) &= \log \prod_{i=1}^{n} e^{\theta_i y_i - a(\theta) + b(y_i)} \\ &= \sum_i \left(\theta_i y_i - a(\theta) + b(y_i)\right) \\ &= \sum_i ll_i(\mathbf{w}) \end{aligned}$$

with gradients

$$\begin{aligned} \frac{\partial ll_i(\mathbf{w})}{\partial w_j} &= \frac{\partial \theta_i}{\partial w_j} y_i - \frac{\partial a(\theta_i)}{\partial w_j} \\ &= \frac{\partial \theta_i}{\partial w_j} y_i - \frac{\partial a(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial w_j} \\ &= \left(y_i - \frac{\partial a(\theta_i)}{\partial \theta_i}\right) \frac{\partial \theta_i}{\partial w_j}. \end{aligned}$$

As was clear for Poisson regression, there is no guarantee of a closed-form solution for $\mathbf{w}$. Therefore, GLM formulations usually use iterative techniques such as gradient descent. Hence, a single mechanism can be used for a wide range of link functions and probability distributions, using these above gradients.

This update can be made more concrete, using the most common setting for GLMs. Importantly, this setting only requires knowledge of the transfer function $f$, without explicitly needing to know the log-normalizer $a$. This simplification arises from the connection between the transfer $f$ and the log-normalizer $a$ alluded to above. We have discussed that the transfer function $f$ is chosen to reflect the range of the output variable $y$. However, the choice should have other properties as well. In particular, we would like to ensure that the $g$ provides a smooth, concave log-likelihood, to simplify optimization. Usefully, the parameter $a$ of the exponential family distribution provides us with just such a choice: $f = \nabla a$. Because $\frac{\partial \theta_i}{\partial w_j} = x_{ij}$ for $\theta_i = \mathbf{x}_i^\top \mathbf{w}$, we get that

$$\begin{aligned} \frac{\partial ll_i(\mathbf{w})}{\partial w_j} &= \left(y_i - \frac{\partial a(\theta_i)}{\partial \theta_i}\right) \frac{\partial \theta_i}{\partial w_j} \\ &= \left(y_i - f(\theta_i)\right) x_{ij} \\ &= \left(y_i - f(\mathbf{x}_i^\top \mathbf{w})\right) x_{ij} \end{aligned}$$

Therefore, given the appropriate transfer $f$ for the desired exponential family distribution, the stochastic gradient descent update is simply

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \left(y_i - f(\mathbf{x}_i^\top \mathbf{w}_t)\right) \mathbf{x}_i$$

and the batch gradient descent update is

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{w}_t - \eta_t \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}_i^\top \mathbf{w}_t)\right) \mathbf{x}_i \\ &= \mathbf{w}_t - \eta_t \mathbf{X}^\top (\mathbf{y} - \mathbf{p}) \end{aligned}$$

where $\mathbf{p}_i = f(\mathbf{x}_i^\top \mathbf{w}_t)$. To examine the Hessian, the second partial derivative of the log likelihood function for one example is

$$\frac{\partial^2 ll_i(\mathbf{w})}{\partial w_j \partial w_k} = -x_{ij} \frac{\partial f(\theta_i)}{\partial \theta_i} x_{ik}.$$

For $\mathbf{D}$ an $n \times n$ diagonal matrix with $\frac{\partial f(\theta_i)}{\partial \theta_i}$ on the diagonal, the Hessian matrix is therefore

$$H_{ll(\mathbf{w})} = -\mathbf{X}^\top \mathbf{D} \mathbf{X}. \tag{7.3}$$

As in Poisson regression, this matrix is guaranteed to be negative semi-definite, and further negative definite if $\mathbf{X}$ is not low-rank.

**Remark:** The common setting of $f = \nabla a$ for GLMs has a connection to widely used objectives called *Bregman divergences*. These divergences are written as $D_a(\hat{y}||y)$, indicating the difference between $\hat{y}$ and $y$, where the divergence is parametrized by $a$. The minimization of this Bregman divergence corresponds to the minimization of the negative log-likelihood of the corresponding natural exponential family:

$$\operatorname*{argmin}_\theta D_a(x||g(\theta)) = \operatorname*{argmin}_\theta - \ln p(x|\theta).$$

See [18, Section 2.2] and [1] for more details about this relationship.

Note that the chosen link does not necessarily have to correspond to the derivative of $a$. Rather, this provides a mechanism for ensuring a nice loss function, since Bregman divergences have nice properties, including being convex in the first argument. However, this does not mean that any other link will necessarily result in an undesirable loss function.