Logistic Regression & Information Theory

Matthew Jörke

Logistic regression is one of the most widely used statistical and machine learning models, but many are unaware of its rich connections to information theory. In this note, I present the mathematical derivation of logistic regression as a *maximum entropy model*, drawing from concepts in information theory and optimization theory. Understanding logistic regression from this lens yields insight into why logistic regression takes its particular form and why the optimal parameters are determined using maximum likelihood estimation.

0 Introduction

Logistic regression is a discriminative regression model that is used to estimate probabilities. The output space of a logistic regression model is restricted to the range $[0, 1]$, essentially modelling the disitribution of a Bernoulli response variable Y conditioned on some (multivariate) input variable $\mathbf{X}$. Suppose we are given the following labelled data $\mathcal{D}$

- $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^m$ (datapoints)
- $y_1, \ldots, y_n \in \{0, 1\}$ (labels)

where $y_i = 1$ if $\mathbf{x}_i$ corresponds to a 'positive' example in the dataset. For example, Y can represent the presence of a particular disease, whether a customer purchased a certain product, or whether a credit card transaction is fraudulent. $\mathbf{X}$ represents features of the domain you wish to predict from, e.g. a patient's health records, a customer's past purchases, or a credit card's transaction history.

Given the training data, the goal of logistic regression is to determine a function $f : \mathbf{R}^m \to [0, 1]$ such that $f(\mathbf{x})$ approximates $\Pr(Y = 1 \mid \mathbf{X} = \mathbf{x})$. There are many ways to define such a function, but in logistic regression the function is parameterized by $\theta \in \mathbf{R}^{m+1}$ and takes the form

$$f(\mathbf{x}; \theta) = \frac{\exp\left(\theta^\top \mathbf{x}\right)}{1 + \exp\left(\theta^\top \mathbf{x}\right)} = \frac{1}{1 + \exp\left(-\theta^\top \mathbf{x}\right)}^{1} \tag{1}$$

1 The Standard Derivation

Traditionally, logistic regression is thought of as a linear model of the log-odds function (this is also how logistic regression was originally defined in Cox, 1958). To demonstrate this, let us define the *logit function*, an abbreviation of **log**istic un**it**, as

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1 - p}\right) \tag{2}$$

[1] Note that $\theta^\top \mathbf{x} = \theta_0 + \sum_{i=1}^m \theta_i x_i$ is taken to include a bias term θ_0, hence the mismatched dimensionality of $\mathbf{x}$ and θ. The bias term is cumbersome for our analysis but incredibly important when implementing logistic regression in practice. With a simple trick, this is of no concern; we can append a constant 1 to each $\mathbf{x}_i$ (now in $\mathbf{R}^{m+1}$) to mimic the bias term.

The logit function is also known as the *log-odds*: the logarithm of the odds ratio. When $p \in [0, 1]$ is a probability, the logit function maps p to the full range of real numbers—a useful property. This transform allows us to define the log-odds of our model as linear function of $\mathbf{x}$.

$$\ln\left(\frac{\Pr(Y = 1 \mid \mathbf{x})}{\Pr(Y = 0 \mid \mathbf{x})}\right) = \theta^\top \mathbf{x}$$

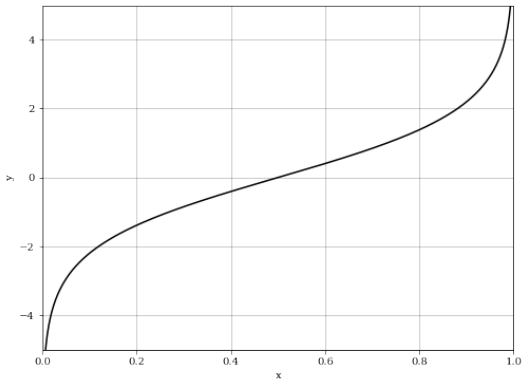
$$\frac{\Pr(Y = 1 \mid \mathbf{x})}{1 - \Pr(Y = 1 \mid \mathbf{x})} = e^{\theta^\top \mathbf{x}}$$

$$\Pr(Y = 1 \mid \mathbf{x}) = \frac{e^{\theta^\top \mathbf{x}}}{1 + e^{\theta^\top \mathbf{x}}}$$

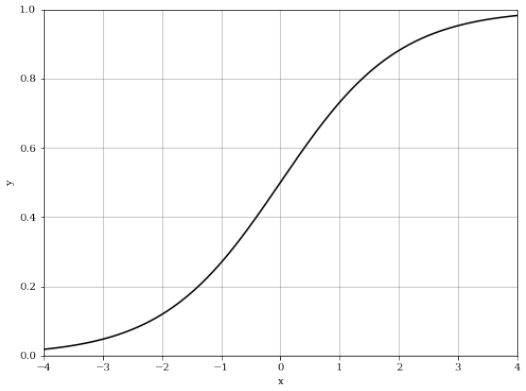
After solving for $\Pr(Y = 1 \mid \mathbf{x})$, we recognize the exact same equation for logistic regression as defined in (1). Here, the logit function is a so-called link function to the linear model. We could have chosen any function that maps $[0, 1]$ to $\mathbf{R}$, such as the inverse Gaussian CDF (this is known as a *probit* model). However, in Section 4 we will see that the logit function uniquely yields some desirable properties.

Figure 1

(a) Logit Function (2)



(b) Sigmoid Function (3)



The inverse of the logit function is the *logistic function*, commonly referred to as the *sigmoid function* (due to the characteristic *S*-shaped curve). As its inverse, the sigmoid function maps $\mathbf{R}$ back to $[0, 1]$.

$$\sigma(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}} \tag{3}$$

More compactly, we can now write equation (1) as

$$f(\mathbf{x}; \theta) = \sigma\left(\theta^\top \mathbf{x}\right)$$

The Multiclass Case

Binary logistic regression can easily be extended to the multiclass case, known as *multinomial logistic regression*. Now, $Y \in \{1, \ldots, k\}$ is a categorical random variable and f is a mapping $f : \mathbf{R}^m \to \mathbf{R}^k$, where $f(\mathbf{x})$ outputs a probability distribution over the k possible outcomes. Treating outcome k as a pivot, for each $i = 1, \ldots, k-1$ we can model the log-odds of outcome i vs outcome k as a linear function of $\mathbf{x}$.

$$\text{for } i = 1, \ldots, k-1: \quad \ln\left(\frac{\Pr(Y = i \mid \mathbf{x})}{\Pr(Y = k \mid \mathbf{x})}\right) = \theta_i^\top \mathbf{x}$$

$$\Pr(Y = i \mid \mathbf{x}) = e^{\theta_i^\top \mathbf{x}} * \Pr(Y = k \mid \mathbf{x})$$

The probabilities over the k outcomes must sum to 1, which allows us to determine $\Pr(Y = k \mid \mathbf{x})$.

$$\sum_{i=1}^{k} \Pr(Y = i \mid \mathbf{x}) = \Pr(Y = k \mid \mathbf{x})\left(1 + \sum_{i=1}^{k-1} e^{\theta_i^\top \mathbf{x}}\right) = 1$$

$$\Pr(Y = k \mid \mathbf{x}) = \frac{1}{1 + \sum_{i=1}^{k-1} e^{\theta_i^\top \mathbf{x}}}$$

If we fix $\theta_k = 0$, this yields the standard formula for multinomial logistic regression.

$$\Pr(Y = i \mid \mathbf{X} = \mathbf{x}) = \frac{e^{\theta_i^\top \mathbf{x}}}{\sum_{j=1}^{k} e^{\theta_j^\top \mathbf{x}}}, \quad \text{for } i = 1, \ldots, k \tag{4}$$

In fact, the relative probabilities do not change even if $\theta_k \neq 0$.[2] Generalizing the 1-dimensional sigmoid (3), we define the multi-dimensional *softmax function* as

$$\sigma : \mathbf{R}^n \to \Delta^{n-1}, \quad \sigma(\mathbf{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

where $\Delta^{n-1} = \{\mathbf{x} \in \mathbf{R}^n : \sum_{i=1}^{n} x_i = 1\}$ is called the *probability simplex*[3] in $\mathbf{R}^n$. Just as the sigmoid function maps any real number to a valid probability, the softmax function maps any n-dimensional vector $\mathbf{x}$ to a valid probability distribution. Defining $\Theta \in \mathbf{R}^{m \times k}$ as $\Theta = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_k \end{bmatrix}$, we can compactly express (4) as

$$f(\mathbf{x}; \Theta) = \sigma(\Theta^\top \mathbf{x})$$

For clarity, we will use the notation $\hat{\Pr}_\theta(Y \mid \mathbf{X})$ in place of $f(\mathbf{x}; \Theta)$ for the remainder of this note. This notation highlights the probabilistic interpretation of logistic regression, where $\hat{\Pr}_\theta(Y \mid \mathbf{X})$ is an empirical approximation of the true distribution $\Pr(Y \mid \mathbf{X})$.

2 Maximum Likelihood & Minimum Cross Entropy

Thus far, we have been ignoring a glaringly important issue—how do we determine the optimal weights for $\theta_1, ..., \theta_k$? This is commonly achieved by maximizing the *likelihood* of the data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ under our model.

$$\mathcal{L}(\theta_1, \ldots, \theta_k \mid \mathcal{D}) = \prod_{i=1}^{n} \hat{\Pr}_\theta(Y = y_i \mid \mathbf{X} = \mathbf{x}_i)$$

$$\theta_1^*, \ldots, \theta_k^* = \underset{\theta_1, \ldots, \theta_k}{\operatorname{argmax}} \mathcal{L}(\theta_1, \ldots, \theta_k \mid \mathcal{D})$$

Typically, we maximize the average log-likelihood, which does not change the maximizing θ_i^*. The logarithm transforms products intro more tractable summations.

$$\frac{1}{n} \log \mathcal{L}(\theta_1, \ldots, \theta_k \mid \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\hat{\Pr}_\theta(Y = y_i \mid \mathbf{X} = \mathbf{x}_i)\right)$$

[2] To see this, note that $\exp\left(\theta_i^\top \mathbf{x}\right) / \sum_{j=1}^{k} \exp\left(\theta_j^\top \mathbf{x}\right) = \exp\left((\theta_i + \epsilon)^\top \mathbf{x}\right) / \sum_{j=1}^{k} \exp\left((\theta_j + \epsilon)^\top \mathbf{x}\right)$ for any constant vector ϵ. Constraining $\theta_k = 0$ uniquely determines the remaining $k - 1$ weights, but in practice solvers might not always enforce this (the constraint makes the search space smaller, which can make it more difficult to find a solution). One can always set $\theta_k' = 0$, $\theta_i' = \theta_i - \theta_k$ to convert between one form and the other.

[3] No, this is not a typo: the simplex in $\mathbf{R}^n$ is denoted by Δ^{n-1}. This is because the n-simplex is uniquely determined by $n - 1$ parameters—it is an $n - 1$ dimensional hyperplane embedded in $\mathbf{R}^n$

In this section, we will see an equivalence between maximizing log-likelihood and minimizing cross-entropy, an information theoretical quantity. Before we proceed: if your information theory is rusty, I highly recommend this excellent article by Chris Olah. Information theory can be quite a mind-boggling subject and I won't spend too much time on the details.

Information Theory Primer

The most fundamental concept in information theory is *entropy*, a measure of a probability distribution's average information content.

$$H(p) = \sum_x p(x) \log \left(\frac{1}{p(x)} \right) = -\sum_x p(x) \log \big(p(x) \big)$$

Suppose your are transmitting messages drawn from p over a channel. Shannon's random coding theorem states that it is impossible to encode your messages such that the expected message length (in bits) is lower than the entropy of p. The entropy provides a strict lower bound on the lossless compression limit of p; if you achieve anything lower, you are certainly losing information.

Entropy can also be interpreted as a measure of *surprise*: the less predictable (i.e. more surprising) a distribution is, the more information it contains. A deterministic distribution is perfectly unsurprising and contains zero information. The uniform distribution is perfectly unpredictable and contains maximal information (more on this later). Entropy is used as a measure of information content because a distribution which requires more bits to compress is inherently less predictable, thus containing more information.

Suppose that you are transmitting messages over a channel again. You think think that you are transmitting messages drawn from q and compute optimal compression codes for your messages. However, unbeknownst to you, the messages are actually drawn from p. The *cross entropy* between probability distributions p and q is given by

$$H(p, q) = -\sum_x p(x) \log \big(q(x) \big)$$

Using the same random coding interpretation as above, the cross entropy measures the expected codeword length when we assume we are coding for q but the actual distribution is p. Moreover, the cross-entropy can be factored into

$$H(p, q) = H(p) + D_{KL}(p \| q)$$

where $D_{KL}(p \| q)$ is the so-called *Kullback-Leibler (KL) divergence* from q to p. The KL-divergence measures the number of additional bits required to code for samples from p when q was assumed.

$$D_{KL}(p \| q) = -\sum_x p(x) \log \left(\frac{q(x)}{p(x)} \right)$$

Strictly speaking, the KL-divergence is not a distance metric—is not symmetric nor does it satisfy the triangle inequality—but it is often used as a proxy for the similarity between two distributions. It can be interpreted as the *information gain* achieved when replacing an approximate distribution q (e.g. a statistical model's output) with the true distribution p (e.g. the distribution of the data).

Principle of Minimum Cross Entropy

In the case of logistic regression, p is the unknown, theoretical distribution $\Pr(Y \mid \mathbf{X})$, while q is our model's output $\hat{\Pr}_\theta(Y \mid \mathbf{X})$. When designing a model, we would like q to be as close to p as possible; the information gain when switching from q to p should be minimal. It seems natural to minimize the cross entropy over q, equivalent to minimizing the KL-divergence from q to p.

$$\begin{aligned}\min_q H(p,q) &= \min_q H(p) + D_{KL}(p\|q) \\ &= \min_q D_{KL}(p\|q)\end{aligned}$$

This is known as the *Principle of Minimum Cross Entropy* (or the *Principle of Minimum Discrimination Information*), first outlined in 1957 by Kullback in Information Theory & Statistics (pg. 36-37). It is closely related to the Principle of Maximum Entropy, which will be covered in Section 4.

It would be desirable to minimize $H(p,q)$ directly, but since $\Pr(Y \mid \mathbf{X})$ is unknown, we must use an empirical approximation. To this end, given samples $x_1, \ldots, x_n$ from p, we define the *empirical cross entropy* as

$$\hat{H}(p,q) = -\frac{1}{n}\sum_{i=1}^{n} \log\big(q(x_i)\big)$$

In the Appendix (A.1), I show that the empirical cross entropy converges to the regular cross entropy in the limit as $n \to \infty$, justifying its role as an empirical approximation.

You might have noticed that the expression for empirical cross entropy is exactly the negative of the average log-likelihood. In fact, we find that the average log-likelihood maximization problem is equivalent to the empirical cross entropy minimization problem.

$$\begin{aligned}\max_{\theta_1,\ldots,\theta_k} \frac{1}{n}\log \mathcal{L}(\theta_1,\ldots,\theta_k \mid \mathcal{D}) &= \max_{\theta_1,\ldots,\theta_k} \frac{1}{n}\sum_{i=1}^{n} \log\Big(\hat{\Pr}_\theta(Y = y_i \mid \mathbf{X} = \mathbf{x}_i)\Big) \\ &= \min_{\theta_1,\ldots,\theta_k} -\frac{1}{n}\sum_{i=1}^{n} \log\Big(\hat{\Pr}_\theta(Y = y_i \mid \mathbf{X} = \mathbf{x}_i)\Big) \\ &= \min_{\theta_1,\ldots,\theta_k} \hat{H}\big(\Pr(Y \mid \mathbf{X}), \hat{\Pr}_\theta(Y \mid \mathbf{X})\big)\end{aligned} \tag{5}$$

When determining the optimal parameters to maximize the likelihood of our dataset, we are simultaneously determining the parameters that yield a distribution with minimum cross entropy to the data's distribution. In fact, it holds that the Principle of Maximum Likelihood is equivalent to the Principle of Minimum Cross Entropy for *any* generative probability model, not just for logistic regression (at no point in our proof did we assume any particular form of $\hat{\Pr}_\theta(Y \mid \mathbf{X})$). This is a powerful concept and a main factor for using the cross entropy loss function for neural networks and other machine learning models.

3 The Balance Equations

We now turn to the problem of solving for the optimal θ_i, transitioning from information theory to the realm of mathematical optimization. The log-likelihood maximization/cross entropy minimization problem (5) happens to be a *convex* optimization problem. Convex problems represent an important class of optimization problems because they do not suffer from local minima and thus admit efficient solution methods. For convex problems, a necessary and sufficient condition for optimality is that the gradient of the objective function equals zero at optimum.

$$\theta_1^*, \ldots, \theta_k^* \text{ optimal} \quad \longleftrightarrow \quad \nabla_{\theta_i} \frac{1}{n}\log \mathcal{L}(\theta_1^*,\ldots,\theta_k^* \mid \mathcal{D}) = 0, \quad \text{for } i = 1,\ldots,k$$

Note that logistic regression does not admit an analytic solution, so we will not be able to derive a formula to solve for θ_i^* directly. Instead, numerical optimization procedures are required to determine the optimal θ_i^* (see Minka, 2003). Nonetheless, the optimality conditions will allow us to characterize important properties of the optimal distribution without actually having to solve for θ_i^*—this is the power of convexity.

To apply the first-order optimality conditions to our problem, we first take the gradient of log-likelihood objective with respect to θ_t $(1 \leq t \leq k)$.

$$\begin{aligned}
\nabla_{\theta_t} \ \log \mathcal{L}(\theta_1, \ldots, \theta_k \mid \mathcal{D}) &= \nabla_{\theta_t} \ \sum_{i=1}^{n} \log \Big(\hat{\Pr}_\theta(Y = y_i \mid \mathbf{X} = \mathbf{x}_i) \Big) \\
&= \nabla_{\theta_t} \ \sum_{i=1}^{n} \log \left(\frac{e^{\theta_{y_i}^\top \mathbf{x}_i}}{\sum_{j=1}^{k} e^{\theta_j^\top \mathbf{x}_i}} \right) \\
&= \nabla_{\theta_t} \ \sum_{i=1}^{n} \theta_{y_i}^\top \mathbf{x}_i - \log \left(\sum_{j=1}^{k} e^{\theta_j^\top \mathbf{x}_i} \right) \\
&= \sum_{i=1}^{n} \mathbb{1}[y_i = t] \, \mathbf{x}_i - \frac{e^{\theta_t^\top \mathbf{x}_i}}{\sum_{j=1}^{k} e^{\theta_j^\top \mathbf{x}_i}} \mathbf{x}_i
\end{aligned}$$

Next, we enforce that the gradient must equal zero at optimum.

$$\begin{aligned}
\nabla_{\theta_t} \ \log \mathcal{L}(\theta_1, \ldots, \theta_k \mid \mathcal{D}) &= 0 \\
\sum_{i=1}^{n} \mathbb{1}[y_i = t] \, \mathbf{x}_i &= \sum_{i=1}^{n} \frac{e^{\theta_t^\top \mathbf{x}_i}}{\sum_{j=1}^{k} e^{\theta_j^\top \mathbf{x}_i}} \mathbf{x}_i
\end{aligned}$$

We recognize the expression on the left as a probability from (4), which yields the following *balance equations*[4].

$$\sum_{i=1}^{n} \mathbb{1}[y_i = t] \, \mathbf{x}_i = \sum_{i=1}^{n} \hat{\Pr}_\theta\big(Y = y_t \mid \mathbf{X} = \mathbf{x}_i\big) \, \mathbf{x}_i \quad \text{for } t = 1, \ldots, k \tag{6}$$

Factoring out $\mathbf{x}_i$ on both sides yields a simplified form, which we will reason with for the rest of this section.

$$\sum_{i=1}^{n} \mathbb{1}[y_i = t] = \sum_{i=1}^{n} \hat{\Pr}_\theta\big(Y = t \mid \mathbf{X} = \mathbf{x}_i\big) \quad \text{for } t = 1, \ldots, k$$

The balance equations state that at optimum, the amount of probability mass allocated to class i in the training set equals the amount of probability mass allocated to class i by the logistic regression model. Further, we can interpret both quantities as an expectation over the training set $\mathcal{D}$,

$$\mathbb{E}_{\mathcal{D}}[Y \mid \text{class}(\mathbf{X}) = t] = \hat{\mathbb{E}}_{\mathcal{D}}[Y \mid \text{class}(\mathbf{X}) = t] \quad \text{for } t = 1, \ldots, k$$

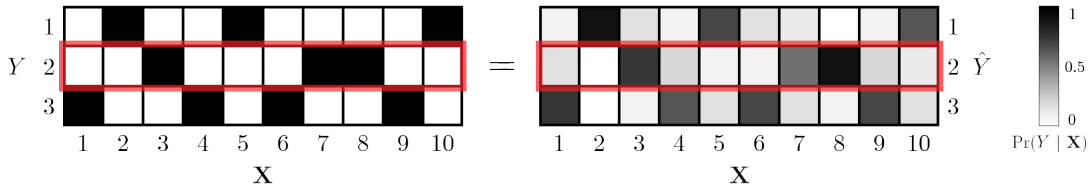
where $\mathbb{E}_{\mathcal{D}}$ is the expectation over the true training data distribution and $\hat{\mathbb{E}}_{\mathcal{D}}$ is the expectation over our model's predicted distribution. By the law of total expectation, this also implies that $\mathbb{E}_{\mathcal{D}}[Y] = \hat{\mathbb{E}}_{\mathcal{D}}[Y]$.

Visualizing the Balance Equations

[4] I am borrowing this terminology (inspired by Markov chain theory) from John Mount's excellent resource The Equivalence of Logistic Regression and Maximum Entropy Models

Consider a simple example with $k = 3$ classes and $n = 10$ training points, shown in Figure 2 below. The training set distribution, an indicator function for the correct class, is pictured on the left. Suppose we train a logistic regression model on this dataset and have found the optimal weights. The model distribution, i.e. the distribution yielded by evaluating the model on the training set, is pictured on the right. For each $\mathbf{x}_i$, the model outputs a probability distribution over $Y \in \{1, 2, 3\}$.

Figure 2: Visual Representation of the Balance Equations



The balance equations state that the sum over the probabilities in each row on the left equals the sum over the corresponding row on the right. This is outlined in red for class $t = 2$. Seen in this light, the balance equations seem like very reasonable conditions for a classifier—we would like to allocate the same amount of probability mass to each class in the model distribution as in training distribution. In some sense, the balance equations ensure that the training distribution's probability mass is 'preserved' for each class in our model.

4 Maximum Entropy Models

In 1957, the physicist E.T. James noted striking similarities between formulas from statistical mechanics and information theory. In the paper Information Theory and Statistical Mechanics, James argued that thermodynamic entropy and information entropy are actually the same concept. Moreover, he demonstrated that statistical mechanics can be re-interpreted as a form of statistical inference that is not bound to the theory of physics, but relies on information theory instead. With this insight, James formalized the *Principle of Maximum Entropy*:

> "in making inferences on the basis of partial information, we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have."
>
> E.T. James, *Information Theory and Statistical Mechanics* (pg. 623)

When conducting inference to determine a probability distribution p, we would like to make as few assumptions about p as possible—a form of Occam's razor, one might argue. The Principle of Maximum Entropy states that among the infinite set of distributions available, one should choose the distribution that maximizes the entropy of p given your prior assumptions

The solution the general maximum entropy problem (below) is the uniform distribution, $p_i^* = 1/n$.

$$\max_p \; -\sum_{i=1}^{n} p_i \; \log(p_i) \quad \text{s.t.} \sum_{i=1}^{n} p_i = 1, \quad p \geq 0 \tag{7}$$

Intuitively, the uniform distribution is perfectly random and thus perfectly unpredictable; more predictable distributions should carry less information. With this in mind, the Principle of Maximum Entropy can further be interpreted as choosing a distribution that is as *uniform* as possible subject to the given constraints. In (7), we only require that p is a valid probability distribution, but impose no further constraints. However, in practice we

must model some prior assumptions about the distribution, since the uniform distribution is useless for prediction.

If we would like to apply the Principle of Maximum Entropy to our problem, how should we proceed? Since we are not allowed to make any apriori assumptions about the model, pretend that we have forgotten about the logit function, likelihood maximization, or cross entropy minimization. We are simply given the dataset $\mathcal{D}$ and tasked with solving for an optimal distribution given some constraint: a constraint that is too general would yield an overly uniform distribution, but a constraint that is too data-specific is likely to overfit.
Suppose that we choose to impose the balance equations (6) as a constraint. We maximize over the $k \times n$ values the distribution $\hat{\Pr}_\theta(Y \mid \mathbf{X})$ can take on in the training set, where $p_{ij} = \hat{\Pr}_\theta\big(Y = i \mid \mathbf{X} = \mathbf{x}_j\big)$. This yields the optimization problem

$$
\begin{aligned}
H(p^*) \;= \max_{p} \quad & -\sum_{i=1}^{k}\sum_{j=1}^{n} p_{ij}\log(p_{ij}) && && (8)\\
\text{s.t.} \quad & \sum_{j=1}^{n} \mathbb{1}\left[y_j = i\right]\, \mathbf{x}_j = \sum_{j=1}^{n} p_{ij}\, \mathbf{x}_j && \text{for } i = 1,\dots,k\\
& \sum_{i=1}^{k} p_{ij} = 1, && \text{for } j = 1,\dots,n\\
& p \geq 0
\end{aligned}
$$

Optimization theory states that the *primal* problem over the primal variable $p \in \mathbf{R}^{k\times n}$ is mathematically equivalent to another *dual* problem over the dual variable $\Theta \in \mathbf{R}^{k\times n}$, where $\Theta = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_k \end{bmatrix}$. If you are familiar with optimization theory, I derive the dual problem in the Appendix (A.2). Otherwise, you can take for granted that the dual is given by

$$
d^* = \max_{\theta_1,\dots,\theta_k} \sum_{j=1}^{n} \log\left(\frac{\exp\left(\theta_{y_i}^\top \mathbf{x}_j\right)}{\sum_{i=1}^{k}\exp\left(\theta_i^\top \mathbf{x}_j\right)}\right)
$$

which astonishingly happens to be our original likelihood maximization problem! We know from Section 3 that the dual does not have an analytic solution and must be solved using numerical methods. However, given optimal dual variables $\theta_1^*, \dots, \theta_k^*$, duality allows us to recover an optimal solution to the primal. The key result of this section is that the optimal primal solution is precisely the same formula as multinomial logistic regression.

$$
p_{ij}^* = \frac{\exp\left(\theta_i^{*\top} \mathbf{x}_j\right)}{\sum_{t=1}^{k}\exp\left(\theta_t^{*\top} \mathbf{x}_j\right)}
$$

What makes these results so surprising is that we only employed the Principle of Maximum Entropy and the balance equations to derive logistic regression, arriving at our initial formulas without any mention of the logit function. Granted that the balance equations are derived from the gradients of our likelihood maximization problem, they are relatively mild constraints to impose, only requiring that expectation is conserved across each class.

5 Conclusion

In this note, I have presented a number of equivalent perspectives from which we can study logistic regression, drawing from classical statistics, information theory, and optimization theory. We have shown that logistic regression is

- a generalized linear model with a logit link function

- the distribution that maximizes the likelihood of the training data $\mathcal{D}$
- the distribution that minimizes the cross-entropy to the training data's distribution
- the distribution that maximizes entropy over the training data subject to the balance equations

Logistic regression is usually taught using the first two interpretations, which unfortunately yield the least insight into what logistic regression is actually learning. Using *information theory*, we were able to demonstrate that maximizing the log-likelihood is equivalent to minimizing the information gain between the true data distribution and our model's distribution. There are many ways to determine the optimal weights, but only log-likelihood maximization allows for this information theoretical interpretation. Using *optimization theory*, we were able to show that the dual of log-likelihood maximization is actually an entropy maximization problem, to which logistic regression is the optimal solution. This explains why modelling the log-odds as a linear function was a not just a sensible decision, but the optimal choice to make.

A Appendix

A.1 Limit of the Empirical Cross Entropy

Consider the empirical cross entropy from $\hat{\Pr}_\theta(Y \mid \mathbf{X})$ to $\Pr(Y \mid \mathbf{X})$.

$$\begin{aligned}\hat{H}\left(\Pr(Y \mid \mathbf{X}), \hat{\Pr}_\theta(Y \mid \mathbf{X})\right) &= -\frac{1}{n}\sum_{i=1}^{n} \log\left(\hat{\Pr}_\theta(Y = y_i \mid \mathbf{X} = \mathbf{x}_i)\right) \\ &= -\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[Y = y_i \mid \mathbf{X} = \mathbf{x}_i\right] \log\left(\hat{\Pr}_\theta(Y = y_i \mid \mathbf{X} = \mathbf{x}_i)\right)\end{aligned}$$

By the weak law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[Y = y \mid \mathbf{X} = \mathbf{x}\right]$ converges to $\Pr(Y = y, \mathbf{X} = \mathbf{x})$ as $n \to \infty$. Moreover, we assume that with infinite samples, all $(\mathbf{x}, y)$ pairs are exhausted[5], yielding the following expression in the limit.

$$\begin{aligned}\lim_{n\to\infty} \hat{H}\left(\Pr(Y \mid \mathbf{X}), \hat{\Pr}_\theta(Y \mid \mathbf{X})\right) &= -\sum_{\mathbf{x}\in\mathbf{X}}\sum_{y\in Y} \Pr(Y = y, \mathbf{X} = \mathbf{x}) \log\left(\hat{\Pr}_\theta(Y = y \mid \mathbf{X} = \mathbf{x})\right) \\ &= -\sum_{\mathbf{x}\in\mathbf{X}} \Pr(\mathbf{X} = \mathbf{x}) \sum_{y\in Y} \Pr(Y = y \mid \mathbf{X} = \mathbf{x}) \log\left(\hat{\Pr}_\theta(Y = y \mid \mathbf{X} = \mathbf{x})\right) \\ &= H\left(\Pr(Y \mid \mathbf{X}),\ \hat{\Pr}_\theta(Y \mid \mathbf{X})\right)\end{aligned}$$

Note that we have used a different expression for the cross entropy here, since we are dealing with conditional distributions. We cannot use the standard definition since we are dealing with distributions over 2 random variables. With a slight change in notation, we note that

$$H\left(\Pr(Y \mid \mathbf{X}),\ \hat{\Pr}_\theta(Y \mid \mathbf{X})\right) = H\left(Y, \hat{Y}_\theta \mid \mathbf{X}\right) = \mathbb{E}_{\mathbf{X}}\left[H\left(Y, \hat{Y}_\theta \mid \mathbf{X} = \mathbf{x}\right)\right]$$

We might call this quantity the *conditional cross entropy*. The definition is analogous to the more standard *conditional entropy*, which is defined as

$$\begin{aligned}H(Y \mid X) &= -\sum_{x}\sum_{y} p(y, x) \log\left(p(y|x)\right) \\ &= \sum_{x} p(x)\Big[-\sum_{y} p(y|x) \log\left(p(y|x)\right)\Big] \\ &= \mathbb{E}_X\left[H(Y \mid X = x)\right]\end{aligned}$$

The conditional entropy measures the average information information content of Y when the value of X is assumed to be known. The entropy of Y might change when conditioned on different values of X, so we take the expectation over X. With the conditional entropy in our toolkit, we can further decompose the conditional cross entropy.

$$\begin{aligned}H\left(Y, \hat{Y}_\theta \mid \mathbf{X}\right) &= -\sum_{\mathbf{x}\in\mathbf{X}} \Pr(\mathbf{X} = \mathbf{x}) \sum_{y\in Y} \Pr(Y = y \mid \mathbf{X} = \mathbf{x}) \log\left(\hat{\Pr}_\theta(Y = y \mid \mathbf{X} = \mathbf{x})\right) \\ &= -\sum_{\mathbf{x}\in\mathbf{X}} \Pr(\mathbf{X} = \mathbf{x}) \sum_{y\in Y} \Pr(Y = y \mid \mathbf{X} = \mathbf{x}) \left(\frac{\log\left(\hat{\Pr}_\theta(Y = y \mid \mathbf{X} = \mathbf{x})\right)}{\log\left(\Pr(Y = y \mid \mathbf{X} = \mathbf{x})\right)} + \log\left(\Pr(Y = y \mid \mathbf{X} = \mathbf{x})\right)\right) \\ &= D_{KL}\left(Y \,\|\, \hat{Y}_\theta \mid \mathbf{X}\right) + H(Y \mid \mathbf{X})\end{aligned}$$

[5] I am cheating slightly here, since $X \in \mathbf{R}^n$ is not necessarily countable. If X is a binary/categorical variable (as is often the case with one-hot encoded models in practice) there is no issue. Even so, by the weak law of large numbers,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H\left(Y, \hat{Y}_\theta \mid \mathbf{X} = \mathbf{x}_i\right) = \mathbb{E}_{\mathbf{X}}\left[H\left(Y, \hat{Y}_\theta \mid \mathbf{X} = \mathbf{x}\right)\right]$$

Here, just as with the conditional entropy and cross entropy, the conditional KL-divergence is given by

$$D_{KL}\left(Y \parallel \hat{Y}_\theta \mid \mathbf{X}\right) = \mathbb{E}_{\mathbf{X}}\left[D_{KL}\left(\Pr(Y \mid \mathbf{X}=\mathbf{x}) \parallel \hat{\Pr}_\theta(Y \mid \mathbf{X}=\mathbf{x})\right)\right]$$

After some algebra, we have shown that the empirical cross entropy converges to the standard cross entropy in the limit. Given enough samples, this justifies using the empirical cross entropy as a Monte Carlo estimate for the true cross entropy. With some more algebra, we proved that the conditional cross entropy decomposes into the conditional KL-divergence and conditional entropy, just as the regular cross entropy.

A.2 Dual of the Maximum Entropy Problem

In this section, we wish to solve for the dual of the optimization problem

$$\begin{aligned}
-H(p^*) \;= \min_{p} \quad & \sum_{i=1}^{k}\sum_{j=1}^{n} p_{ij}\log(p_{ij}) \\
\text{s.t.} \quad & \sum_{j=1}^{n} \mathbb{1}\left[y_j = i\right]\,\mathbf{x}_j = \sum_{j=1}^{n} p_{ij}\,\mathbf{x}_j \quad \text{for } i = 1,\ldots,k \\
& \sum_{i=1}^{k} p_{ij} = 1 \qquad\qquad\qquad\qquad\quad \text{for } j = 1,\ldots,n \\
& p \geq 0
\end{aligned}$$

Note that we have converted (8) into standard form by negating the objective, transforming the problem into a minimization. This is a *convex* optimization problem in p with strictly feasible constraints, so strong duality must hold (by Slater's condition). The Lagrangian of this problem is given by

$$\mathcal{L}(p,\theta,\mu) = \sum_{i=1}^{k}\sum_{j=1}^{n} p_{ij}\log(p_{ij}) + \sum_{i=1}^{k}\theta_i^\top\left(\sum_{j=1}^{n}\mathbb{1}[y_j=i]\,\mathbf{x}_j - \sum_{j=1}^{n} p_{ij}\,\mathbf{x}_j\right) + \sum_{j=1}^{n}\mu_j\left(\sum_{i=1}^{k} p_{ij} - 1\right)$$

where $p \in \mathbf{R}^{k\times n}$ is the primal variable, $\theta_1, ..., \theta_k \in \mathbf{R}^n$ are Lagrange multipliers (i.e. dual variables) for the k balance equation constraints, and $\mu \in \mathbf{R}^n$ is the multiplier for the sum-to-one constraint. We omit a multiplier for the non-negativity constraint, which is common practice and not necessary for this problem.

Strong duality states that the primal minimization problem (8) is equivalent to another dual maximization problem over the Lagrange multipliers (θ, μ).

$$-H(p^*) = \min_{p}\max_{\theta,\mu}\,\mathcal{L}(p,\theta,\mu) = \max_{\theta,\mu}\min_{p}\,\mathcal{L}(p,\theta,\mu)$$

Since $\mathcal{L}(p,\theta,\mu)$ is convex in p, we can solve for $\min_p\,\mathcal{L}(p,\theta,\mu)$ directly by setting partial derivates to zero.

$$\frac{\partial}{\partial p_{ij}}\mathcal{L}(p,\theta,\mu) = 1 + \log(p_{ij}) + \theta_i^\top\mathbf{x}_j + \mu_j = 0$$

$$p_{ij} = \exp\left(\theta_i^\top\mathbf{x}_j + \mu_j + 1\right)$$

We notice that μ can be eliminated directly by enforcing the sum-to-one constraint, $\sum_{i=1}^{k} p_{ij} = 1$

$$\sum_{i=1}^{k} \exp\left(\theta_i^\top \mathbf{x}_j + \mu_j + 1\right) = 1$$

$$e^{\mu_j} \sum_{i=1}^{k} \exp\left(\theta_i^\top \mathbf{x}_j + 1\right) = 1$$

$$e^{\mu_j} = \frac{1}{\sum_{i=1}^{k} \exp\left(\theta_i^\top \mathbf{x}_j + 1\right)}$$

Plugging this back into our expression for p_{ij}, we find

$$p_{ij} = \exp\left(\theta_i^\top \mathbf{x}_j + 1\right) \exp\left(\mu_j\right) = \frac{\exp\left(\theta_i^\top \mathbf{x}_j + 1\right)}{\sum_{t=1}^{k} \exp\left(\theta_t^\top \mathbf{x}_j + 1\right)} = \frac{\exp\left(\theta_i^\top \mathbf{x}_j\right)}{\sum_{t=1}^{k} \exp\left(\theta_t^\top \mathbf{x}_j\right)}$$

Note that this expression for p_{ij} is identical to the multinomial logistic regression formula (4). In fact, when we insert our expression for p_{ij} back into the Lagrangian, we recover a dual problem equivalent to our initial maximum likelihood problem (5).

$$\begin{aligned}
\mathcal{L}(p, \theta) &= \sum_{i=1}^{k} \sum_{j=1}^{n} p_{ij} \log(p_{ij}) + \sum_{i=1}^{k} \theta_i^\top \left(\sum_{j=1}^{n} \mathbb{1}[y_j = i]\, \mathbf{x}_j - \sum_{j=1}^{n} p_{ij}\, \mathbf{x}_j \right) \\
&= \sum_{i=1}^{k} \sum_{j=1}^{n} p_{ij} \log(p_{ij}) + \mathbb{1}[y_j = i]\, \theta_i^\top \mathbf{x}_j - p_{ij} \theta_i^\top \mathbf{x}_j \\
&= \sum_{j=1}^{n} \sum_{i=1}^{k} \frac{\exp\left(\theta_i^\top \mathbf{x}_j\right)}{\sum_{t=1}^{k} \exp\left(\theta_t^\top \mathbf{x}_j\right)} \left[\theta_i^\top \mathbf{x}_j - \log\left(\sum_{t=1}^{k} \exp\left(\theta_t^\top \mathbf{x}_j\right) \right) \right] \\
&\qquad + \mathbb{1}[y_j = i]\, \theta_i^\top \mathbf{x}_j - \frac{\exp\left(\theta_i^\top \mathbf{x}_j\right)}{\sum_{t=1}^{k} \exp\left(\theta_t^\top \mathbf{x}_j\right)} \theta_i^\top \mathbf{x}_j \\
&= \sum_{j=1}^{n} \theta_{y_i}^\top \mathbf{x}_j + \log\left(\sum_{t=1}^{k} \exp\left(\theta_t^\top \mathbf{x}_j\right) \right) \sum_{i=1}^{k} \frac{\exp\left(\theta_i^\top \mathbf{x}_j\right)}{\sum_{t=1}^{k} \exp\left(\theta_t^\top \mathbf{x}_j\right)} \\
&= \sum_{j=1}^{n} \theta_{y_i}^\top \mathbf{x}_j + \log\left(\sum_{i=1}^{k} \exp\left(\theta_i^\top \mathbf{x}_j\right) \right) \\
&= \sum_{j=1}^{n} \log\left(\frac{\exp\left(\theta_{y_i}^\top \mathbf{x}_j\right)}{\sum_{i=1}^{k} \exp\left(\theta_i^\top \mathbf{x}_j\right)} \right)
\end{aligned}$$

The dual problem is thus given by

$$-H(p^*) = \max_{\theta_1, \ldots, \theta_k} \min_{p} \mathcal{L}(\theta, p) = \max_{\theta_1, \ldots, \theta_k} \sum_{j=1}^{n} \log\left(\frac{\exp\left(\theta_{y_i}^\top \mathbf{x}_j\right)}{\sum_{i=1}^{k} \exp\left(\theta_i^\top \mathbf{x}_j\right)} \right)$$

As mentioned in Section 3, this problem has no analytic solution. However, given the optimal dual variable θ^*, the *Karush–Kuhn–Tucker (KKT) conditions* guarantee that any p^* that satisfies $\nabla_p \mathcal{L}(\theta^*, p^*) = 0$ must be primal optimal as well (we have already solved for this condition above). Thus, a primal optimal solution can be expressed in terms of optimal dual variables.

$$p_{ij}^* = \frac{\exp\left(\theta_i^{*\top} \mathbf{x}_j\right)}{\sum_{t=1}^{k} \exp\left(\theta_t^{*\top} \mathbf{x}_j\right)}$$