Logistic Regression

Ed Scerbo

Note: In the following, it shall be convenient to write X, Y when we wish to indicate random quantities and x, y when we wish to indicate particular but unspecified values those random quantities assume. This same convention shall be in force for subscripted variables, as well.

Suppose pairs (x, y) with $x \in \mathbb{R}^p, y \in \{0, 1\}$ are drawn from some probability distribution D. Given that $X = x$, we would like to predict the probability that $Y = 1$. In other words, we would like to estimate the function $f(x) = \mathbb{P}(Y = 1|X = x)$. (Note that, since the only possible values for Y are 0 and 1, we have $1 - f(x) = \mathbb{P}(Y = 0|X = x)$.)

To do so, we use the method of **maximum likelihood estimation**: Choose some hypothesis class $\mathcal{H}$, and select a sample $((x_i, y_i))_{i=1}^n$ by making independent random draws from D. For any $h \in \mathcal{H}$, its **likelihood** $\mathcal{L}(h)$ is the probability that each $Y_i = y_i$ given that $X_i = x_i$ for $i = 1, ..., n$, under the additional assumption that $f = h$. Since our sample was constructed via independent draws, this is

$$\begin{aligned}
\mathcal{L}(h) &= \mathbb{P}(Y_1 = y_1, ..., Y_n = y_n|X_1 = x_1, ..., X_n = x_n) \\
&= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i|X_i = x_i) \\
&= \prod_{y_i=1} h(x_i) \prod_{y_i=0} (1 - h(x_i)).
\end{aligned}$$

We would like to find $h \in \mathcal{H}$ that maximizes $\mathcal{L}(h)$. It will be easier to maximize $\ln \mathcal{L}(h)$; in this note, however, to make the connection with a general model-fitting program a bit clearer, we shall *minimize* the *negative* average log-likelihood:

$$\begin{aligned}-\frac{1}{n}\ln\mathcal{L}(h) &= -\frac{1}{n}\sum_{y_i=1}\ln h(x_i) - \frac{1}{n}\sum_{y_i=0}\ln(1-h(x_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n}[-y_i\ln h(x_i) - (1-y_i)\ln(1-h(x_i))].\end{aligned}$$

Digression: Define

$$l(y,\hat{y}) = -y\ln\hat{y} - (1-y)\ln(1-\hat{y}).$$

l is the pointwise **cross-entropy** loss function, which arises in information theory as a way of gauging the dissimilarity between two probability distributions. Thus we see that maximizing likelihood is the same as minimizing the average cross-entropy loss

$$L(h) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, h(x_i)).$$

From this point of view, the above is an instance of the general model-fitting program wherein we select a model class and a loss function and minimize the loss over all possible models in the class relative to our training data.

For logistic regression, we take $\mathcal{H}$ to be the set of all functions of the form $h_w(x) = \sigma(w\cdot x)$, where $w \in \mathbb{R}^{p+1}$ and $\sigma : \mathbb{R}\to\mathbb{R}$ denotes the sigmoid function

$$\sigma(s) = \frac{e^s}{e^s+1} = \frac{1}{1+e^{-s}}.$$

(Here we tacitly append to $x \in \mathbb{R}^p$ a "bias term" of 1 so that the equation $w\cdot x = 0$ can denote *any* hyperplane in $\mathbb{R}^p$, not just one through the origin. Alternately we could have written $h_{w,b}(x) = \sigma(w\cdot x + b)$ where $w \in \mathbb{R}^p$ and $b \in \mathbb{R}$, but this seems clunkier.) There are many reasons for working with this class, principal among them that it captures the following often reasonable assumption: Even if the population is not linearly separable, there may well be a hyperplane in $\mathbb{R}^p$ that does a decent job splitting the data, with the property that the further to one side or the other of this hyperplane a given point is, the more likely it is to have one label or the other.

Remark: We here record two useful properties of σ: For any $s \in \mathbb{R}$, we have

- $\sigma(s) + \sigma(-s) = 1$, and
- $\sigma'(s) = \sigma(s)\sigma(-s) = \sigma(s)(1 - \sigma(s))$.

Replacing h by h_w in the above expression and using the first bullet point, we see that we need to find $w \in \mathbb{R}^{p+1}$ that minimizes

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}[-y_i \ln \sigma(w \cdot x_i) - (1 - y_i)\ln \sigma(-w \cdot x_i)].$$

Note that L is convex: A trivial calculation shows that $-\ln \sigma(s)$ is convex, which implies that for any vector $x \in \mathbb{R}^{p+1}$, the function $w \mapsto -\ln \sigma(w \cdot x)$ is convex. Since L is a linear combination of such functions with nonnegative weights, L is convex.

There is no way in general to minimize this function analytically. However, since L is convex, it is amenable to numerical methods. A simple calculation shows that for any $x \in \mathbb{R}^{p+1}$,

$$\nabla_w \ln \sigma(w \cdot x) = \sigma(-w \cdot x)x,$$

and so

$$\begin{aligned}\nabla L(w) &= \frac{1}{n}\sum_{i=1}^{n}[-y_i\sigma(-w \cdot x_i)x_i + (1 - y_i)\sigma(w \cdot x_i)x_i]\\ &= \frac{1}{n}\sum_{i=1}^{n}[-y_i(1 - \sigma(w \cdot x_i)) + (1 - y_i)\sigma(w \cdot x_i)]x_i\\ &= \frac{1}{n}\sum_{i=1}^{n}[\sigma(w \cdot x_i) - y_i]x_i\\ &= \frac{1}{n}X^T(\sigma(Xw) - y).\end{aligned}$$

Here we are abusing notation by letting X denote the $n \times (p + 1)$ matrix whose i^{th} row is x_i. (Note that this matrix includes a “bias column” of all 1’s because we are appending a “bias term” of 1 to each x_i.) We are also treating w and y as column vectors. Finally, by $\sigma(Xw)$ we mean the vector obtained by applying σ to each entry of Xw.

In practice, L is often minimized via Newton's method, wherefore we should compute the Hessian of L. A trivial application of Chain Rule yields

$$\nabla^2 L(w) = \frac{1}{n} X^T \Sigma(w) X,$$

where $\Sigma(w)$ denotes the $n \times n$ diagonal matrix with diagonal entries $\sigma'(w \cdot x_i), i = 1, ..., n$. Note that for any $w \in \mathbb{R}^{p+1}$, $\Sigma(w)$ is positive-definite, whence $\nabla^2 L(w)$ is positive-semidefinite, yielding another proof of the convexity of L.

Note: In practice, we run into a problem trying to minimize L in the case where our training set is linearly separable: Suppose $((x_i, y_i))_{i=1}^n$ is linearly separable, and let w_0 denote a vector of weights such that for each i with $y_i = 1$ we have $w_0 \cdot x_i > 0$ and for each i with $y_i = 0$ we have $w_0 \cdot x_i < 0$. We may write

$$L(w_0) = -\frac{1}{n}\Big[\sum_{y_i=1} \ln \sigma(w_0 \cdot x_i) + \sum_{y_i=0} \ln \sigma(-w_0 \cdot x_i)\Big].$$

The above assumptions imply that for each summand in this expression, the argument to σ is a *positive* quantity. It follows immediately that, if we let t denote some positive number, then $L(tw_0) \to 0$ as $t \to \infty$. Since L is always positive and is convex, it follows that attempting to minimize $L(w)$ in this case forces $|w| \to \infty$. To avoid this, logistic regression algorithms are always implemented with some form of regularization, which restricts how big $|w|$ can be.

As an alternative to the above treatment, we can also discuss logistic regression in terms of Kullback-Leibler (KL) divergence: Let Ω be an at most countable set, and let P, Q be probability measures on Ω such that P is absolutely continuous with respect to Q, i.e., $Q(\omega) = 0 \implies P(\omega) = 0$. The **KL divergence** between P and Q is defined by

$$D_{KL}(P\|Q) = \sum_{\omega\in\Omega} P(\omega) \ln \frac{P(\omega)}{Q(\omega)},$$

with the convention that whenever $P(\omega) = 0$, the corresponding term is defined to be 0. Note that $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$, so D_{KL} is not a metric on the space of probability measures on Ω. However, one can show via Jensen's inequality that $D_{KL}(P\|Q)$ is always nonnegative and that $D_{KL}(P\|Q) = 0 \iff P = Q$. Note also that intuitively, if $\omega \in \Omega$ is such that $P(\omega)$ is small but $Q(\omega)$ is big, then the corresponding term $P(\omega) \ln \frac{P(\omega)}{Q(\omega)}$ is small. On the other hand, if $P(\omega)$ is big but $Q(\omega)$ is small, then the corresponding term is big. That is, intuitively, KL divergence is a measure of how much $Q(\omega)$ differs from $P(\omega)$ mostly for $\omega \in \Omega$ that are P-likely.

Return to the general logistic regression setting from the beginning of this note: Draw some point (x, y) from the underlying distribution D, and let h be some element of the hypothesis class $\mathcal{H}$. Once we know the value of y, the probability that it equals 1 is, in fact, y. Let P denote the distribution on $\Omega = \{0, 1\}$ with $P(1) = y$, and let Q denote the distribution predicted by h, i.e., $Q(1) = h(x)$. Then the KL divergence between P and Q is

$$\begin{aligned} D_{KL}(P\|Q) &= y \ln \frac{y}{h(x)} + (1-y) \ln \frac{1-y}{1-h(x)} \\ &= l(y, h(x)) + y \ln y + (1-y)\ln(1-y) \\ &= l(y, h(x)), \end{aligned}$$

where l is the pointwise cross-entropy loss function defined above and the third equality holds because $y \in \{0, 1\}$.

Now, given a training set $((x_i, y_i))_{i=1}^n$ and some $h \in \mathcal{H}$, let P_i denote the distribution on $\Omega = \{0, 1\}$ with $P_i(1) = y_i$, and let Q_i denote the distribution predicted by h, i.e., $Q_i(1) = h(x_i)$. Then the average KL divergence between these distributions is

$$\frac{1}{n} \sum_{i=1}^{n} D_{KL}(P_i\|Q_i) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, h(x_i)).$$

Thus we see that, in addition to representing maximum likelihood, minimizing average cross-entropy loss is equivalent to minimizing the average KL divergence between the distributions P_i and Q_i determined by the training set and elements of the hypothesis class.