Logistic Regression
MIT 15.097 Course Notes

Cynthia Rudin

Thanks to Ashia Wilson
Credit: J.S. Cramer's "The Origin of Logistic Regression"

Origins: 19th Century.

- Studying growth of populations and the course of chemical reactions using

$$\frac{d}{dt}W(t) = \beta W(t) \quad \Rightarrow \quad W(t) = Ae^{\beta t}$$

  which is a good model for unopposed growth, like the US population's growth at the time.

- Adolphe Quetelet (1796 - 1874), Belgian astronomer turned statistician, knew it produced impossible values and asked his pupil Pierre-François Verhulst (1804-1849) to help him work on a more realistic model. They chose

$$\frac{d}{dt}W(t) = \beta W(t) - \Phi(W(t))$$

  to resist further growth, and with the choice of Φ to be a quadratic function, they got:

$$\frac{d}{dt}W(t) = \beta W(t)(\Omega - W(t)),$$

where Ω is the saturation limit of W. Writing $P(t) = \frac{W(t)}{\Omega}$ as the proportion of saturation limit:

$$\frac{d}{dt}P(t) = \beta P(t)(1 - P(t))$$

$$P(t) = \frac{e^{(\alpha+\beta t)}}{1 + e^{(\alpha+\beta t)}}$$

where α is the constant from the integration. Verhulst's "logistic" function has values between 0 and 1.

Draw the logistic function

He published in 3 papers between 1838 and 1847. The first paper demonstrated that the curve agrees very well with the actual course of the population in France, Belgium, Essex, and Russia for periods up to 1833. He did not say how he fitted the curves. In the second paper he tried to fit the logistic to 3 points using 20-30 years of data (which in general is a not a great way to get a good model). His estimates of the limiting population Ω of 6.6 million for Belgium and 40 million for France were a little off - these populations are now 11 million for Belgium and 65 million for France. In another paper, the estimate was corrected, and they estimated 9.5 million for Belgium (pretty close!).

Verhulst died young in poor heath, and his work was forgotten about, but reinvented in 1920 by Raymond Pearl and Lowell Reed who were studying population growth of the US. They also tried to fit the logistic function to population growth, and estimated Ω for the US to be 197 million (the current population is 312 million). Actually, Pearl and collaborators spent 20 years applying the logistic growth curve to almost any living population (fruit flies, humans in North Africa, cantaloupes).

Verhulst's work was rediscovered just after Pearl and Reed's first paper in 1920, but they didn't acknowledge him in their second paper, and only in a footnote in a third paper (by Pearl) in 1922. They cited him in 1923, but didn't use his terminology and called his papers "long since forgotten." The name logistic was revived by Yule, in a presidential address to the Royal Statistical Society in 1925.

There was a lot of debate over whether the logistic function could replace the cdf of the normal distribution. The person who really showed this could happen was Joseph Berkson (1899-1982), who was the chief statistician at the Mayo Clinic, and who was a collaborator of Reed's. But because of his personality and attacks on the method of maximum likelihood, controversy ensued... it took until around the 1960's to resolve it!

Let's start with a probabilistic model. Assume $\mathbf{x} \in \mathbf{R}^n$, and that $\mathbf{x}$ is deterministic (chosen, not random):

$$Y \sim \text{Bernoulli}(P(Y = 1|\mathbf{x})).$$

Here Y takes either 0 or 1, but we need a ± 1 for classification. I'll write $\tilde{0}$ for now to stand for either 0 or -1. The model is:

$$\underbrace{\ln\left(\frac{P(Y=1|\mathbf{x},\boldsymbol{\lambda})}{P(Y=\tilde{0}|\mathbf{x},\boldsymbol{\lambda})}\right)}_{\text{"odds ratio"}} = \boldsymbol{\lambda}^T\mathbf{x}$$

Why does this model make any sense at all? We want to make a linear combination of feature values, like in regular regression, which explains the right hand side. But that can take any real values. We need to turn it into a model for $P(Y=1|\mathbf{x},\boldsymbol{\lambda})$, which takes values only between 0 and 1. The odds ratio turns those probabilities into positive real numbers, then the log turns it into any real numbers.

$$\frac{P(Y=1|\mathbf{x},\boldsymbol{\lambda})}{P(Y=\tilde{0}|\mathbf{x},\boldsymbol{\lambda})} = e^{\boldsymbol{\lambda}^T\mathbf{x}}$$

$$P(Y=1|\mathbf{x},\boldsymbol{\lambda}) = e^{\boldsymbol{\lambda}^T\mathbf{x}}P(Y=\tilde{0}|\mathbf{x},\boldsymbol{\lambda}) = e^{\boldsymbol{\lambda}^T\mathbf{x}}(1-P(Y=1|\mathbf{x},\boldsymbol{\lambda}))$$

$$P(Y=1|\mathbf{x},\boldsymbol{\lambda})\left[1+e^{\boldsymbol{\lambda}^T\mathbf{x}}\right] = e^{\boldsymbol{\lambda}^T\mathbf{x}}$$

$$P(Y=1|\mathbf{x},\boldsymbol{\lambda}) = \frac{e^{\boldsymbol{\lambda}^T\mathbf{x}}}{1+e^{\boldsymbol{\lambda}^T\mathbf{x}}} \leftarrow \text{logistic function}$$

We'll use **maximum likelihood estimation**. We choose parameters $\boldsymbol{\lambda}$ to maximize the likelihood of the data given the model.

$$L(\boldsymbol{\lambda}) := P(Y_1=y_1,\ldots,Y_m=y_m|\boldsymbol{\lambda},\mathbf{x}_1,\ldots,\mathbf{x}_m) \overset{\text{iid}}{=} \prod_{i=1}^{m} P(Y_i=y_i|\boldsymbol{\lambda},\mathbf{x}_i).$$

Choose

$$\boldsymbol{\lambda}^* \in \operatorname*{argmax}_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) = \operatorname*{argmax}_{\boldsymbol{\lambda}} \underbrace{\log L(\boldsymbol{\lambda})}_{\text{log-likelihood}}$$

where we take the log for convenience since it doesn't effect the argmax.

But first, we'll simplify. We need $P(Y=y_i|\boldsymbol{\lambda},\mathbf{x}_i)$ for both $y_i=1$ and $y_i=-1$.

$$\text{If } \begin{cases} y_i = 1, & \text{need } P(Y=1|\boldsymbol{\lambda},\mathbf{x}_i) := p_i = \frac{e^{\boldsymbol{\lambda}^T\mathbf{x}_i}}{1+e^{\boldsymbol{\lambda}^T\mathbf{x}_i}} = \frac{1}{1+e^{-\boldsymbol{\lambda}^T\mathbf{x}_i}} = \frac{1}{1+e^{-y_i\boldsymbol{\lambda}^T\mathbf{x}_i}} \\ y_i = -1 \text{ (i.e., } \tilde{0}), & \text{need } P(Y=\tilde{0}|\boldsymbol{\lambda},\mathbf{x}_i) = 1-p_i = \frac{1+e^{\boldsymbol{\lambda}^T\mathbf{x}_i}}{1+e^{\boldsymbol{\lambda}^T\mathbf{x}_i}} - \frac{e^{\boldsymbol{\lambda}^T\mathbf{x}_i}}{1+e^{\boldsymbol{\lambda}^T\mathbf{x}_i}} = \frac{1}{1+e^{\boldsymbol{\lambda}^T\mathbf{x}_i}} \\ & \qquad = \frac{1}{1+e^{-y_i\boldsymbol{\lambda}^T\mathbf{x}_i}}. \end{cases}$$

So, we can just write

$$P(Y = y_i|\boldsymbol{\lambda}, \mathbf{x}_i) = \frac{1}{1 + e^{-y_i \boldsymbol{\lambda}^T \mathbf{x}_i}}.$$

Then,

$$\begin{aligned} \boldsymbol{\lambda}^* &\in \operatorname*{argmax}_{\boldsymbol{\lambda}} \log L(\boldsymbol{\lambda}) \\ &= \operatorname*{argmax}_{\boldsymbol{\lambda}} \sum_{i=1}^{m} \log \frac{1}{1 + e^{-y_i \boldsymbol{\lambda}^T \mathbf{x}_i}} \\ &= \operatorname*{argmin}_{\boldsymbol{\lambda}} \sum_{i=1}^{m} \log(1 + e^{-y_i \boldsymbol{\lambda}^T \mathbf{x}_i}). \end{aligned}$$

This agrees with the "frequentist" derivation we had before.

The loss is convex in $\boldsymbol{\lambda}$ so we can minimize by gradient descent.

MIT OpenCourseWare
http://ocw.mit.edu

15.097 Prediction: Machine Learning and Statistics
Spring 2012

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.