COS 424: Interacting with Data

Lecturer: Professor David Blei
Scribe: Pete DiMaggio
Lecture #17
4/10/2008

Logistic Regression

We are going to use the same type of machinery from linear regression to do *classification*, as illustrated by the graphical model in Figure 1. Let us reconsider *binary classification*

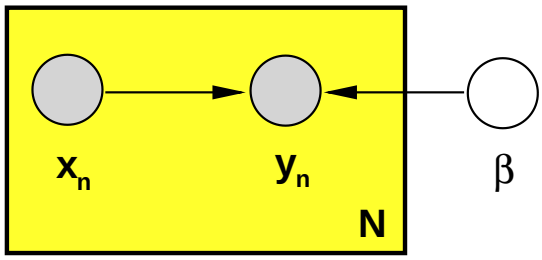

Figure 1: Graphical model for linear and logical regression

(Note that the classes are now $\{0, 1\}$ instead of $\{-1, 1\}$). Recall in linear regression that $y_n \sim N(\beta^T x, \sigma^2)$, which is not appropriate for binary classification because it will predict non $\{0, 1\}$ values, as well as values less than 0 and greater than 1 (as shown by the blue line Figure 2). Furthermore, the addition of an outlier point will skew the fit of the line (as illustrated by the red point and line in Figure 2). In classification $y_n \in \{0, 1\}$, but we still want y_n to be a linear function of x_n. So the question becomes: how do we combine x_n and β in a linear fashion to obtain y_n?

Bernoulli:

Given the constraint $y_n \in \{0, 1\}$, a natural choice is to model y_n as a Bernoulli distribution, as shown in Equation 1.

$$p(y|x) = \mu(x)^y \ (1 - \mu(x))^{1-y} \tag{1}$$

Where the parameter to the Bernoulli, μ, is a function of the input x. So what is this function?
Q: Is $\mu(x) = \beta^T x$?
A: NO! We require $0 \leq \mu(x) \leq 1$, and as shown in Figure 2, these values will exceed this interval. However, we can "squash" $\beta^T x$ to be confined to the interval $[0, 1]$ by using what is known as a *logistic function*: logistic($\beta^T x \ : \ \Re \rightarrow \{0, 1\}$)

Logistic Function:

$$\mu(x) = \frac{1}{1 + e^{-\eta(x)}} \tag{2}$$

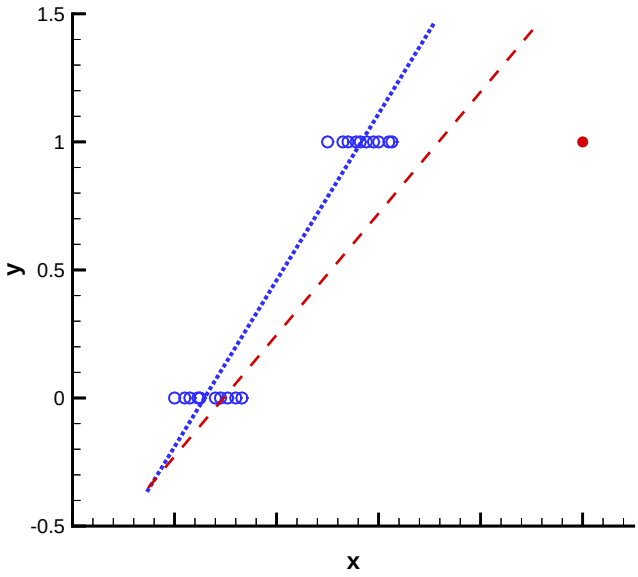

Figure 2: Illustration of linear regression fit for points that have a value of 1 or 0. We see that the fit, represented by the blue line, results in predictions that extend beyond 0 and 1 and thus is not suitable for the purpose of binary classification. Furthermore, the existence of outliers, represented by the single red point, can bias the regression towards this point, as illustrated by the red line.

$$\eta(x) = \beta^T x \tag{3}$$

This specifies the model:

$$y_n \sim Bernoulli(\mu(x)) \tag{4}$$

Where $\mu(x)$ is defined in Equations 2 and 3. A 1-D illustration of this function is provided in Figure 3 for several different values of β, where we see that larger values of beta result in steeper curve near $\beta x = 0$. In Figure 3, we also see that as βx approaches ∞, $\mu(x)$ goes to 1 and as βx approaches $-\infty$, $\mu(x)$ goes to 0.
Let us also examine the logistic function as a 2-D plot, as shown in Figure 4. We can see from this figure that the logistic regression model implicity places a *separating hyperplane*, $\beta^T x = 0$, in the input space. The classifications are now measured in a probabilistic sense, where $p(y = 1|x, \beta) = \mu(\beta^T x)$. As illustrated in Figure 4, points far away from $\beta^T x = 0$ all have a probability of 1, which implies that only the closest points matter when training the model (as was the case with support vector machines). What's also interesting is that while outliers can skew the fit of linear regression models (as shown in Figure 1), they do not have such an effect on logistic regression models.

MLE of β:

Maximum likelihood estimation in logistic regression is very similar to maximizing the margin of separation in support vector machines (that is, the emphasis is placed on points near the boundary). What remains is to determine the appropriate parameter β for our

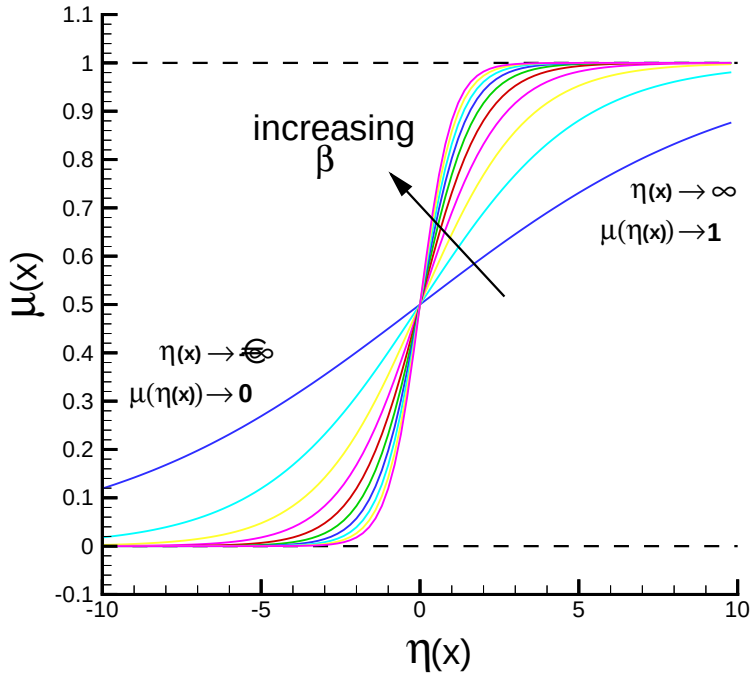

Figure 3: Plot of the logistic function (as defined in Equation 2) in a single dimension. Notice that increasing values of β result in a steeper curve near $\beta x = 0$.

model, which is given by maximizing the log likelihood:

$$\hat{\beta} = \max_{\beta} \log p(y_{1:N}|x_{1:N}, \beta) \tag{5}$$

Where our data is $\{(x_n, y_n)\}_{n=1}^N$. The log likelihood is given by the expression:

$$\log p(y_{1:N}|x_{1:N}, \beta) = \sum_{n=1}^{N} \log p(y_n|x_n, \beta) \tag{6}$$

$$= \sum_{n=1}^{N} \log \left(\mu(x_n)^{y_n} \left(1 - \mu(x_n)\right)^{1-y_n}\right) \tag{7}$$

Note in the above equation that we condition on x (as compared to naive Bayes) and we have suppressed the dependence on β. Taking the logarithm of the terms yields:

$$L = \sum_{n=1}^{N} y_n \log \mu(x_n) \; + \; (1 - y_n) \log \left(1 - \mu(x_n)\right) = \sum_{n=1}^{N} L_n \tag{8}$$

Our main objective is to find the optimal β, so we must derive an expression for $\frac{dL}{d\beta_i}$, which using the chain rule is:

$$\frac{dL}{d\beta_i} = \sum_{n=1}^{N} \frac{dL}{d\mu(x_n)} \frac{d\mu(x_n)}{d\beta_i} \tag{9}$$

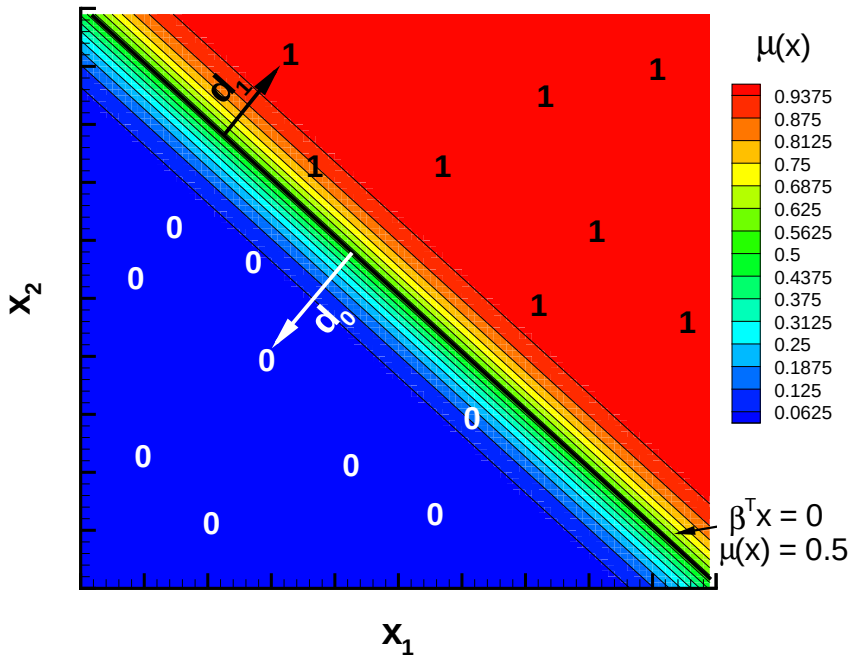

Figure 4: Two-dimensional plot of the logistic function to illustrate classification. We notice that the line corresponding to $\beta^T x = 0$ results in a hyperplane that separates the points labeled as 0 or 1. Also note that points far from this hyperplane either have a value of $\mu(x) = 0$ (towards the lower-left of the figure) or $\mu(x) = 1$ (towards the upper-right of the figure), implying that outliers do not have a significant impact on the model. In this figure, $d_1 = \frac{\beta^T x_1}{||\beta||} > 0$ and $d_0 = \frac{\beta^T x_0}{||\beta||} < 0$.

To keep a clean notation, we will define $\mu(x_n) = \mu_n$. Using the calculus of logarithms, we can easily find the first term of Equation 9 by differentiating Equation 8 with respect to μ_n:

$$\frac{dL}{d\mu_n} = \frac{y_n}{\mu_n} - \frac{(1 - y_n)}{(1 - \mu_n)} \tag{10}$$

However, the second term of Equation 9 requires some additional calculus. Now we can express $\frac{d\mu_n}{d\beta_i}$ using the chain rule as:

$$\frac{d\mu_n}{d\beta_i} = \frac{d\mu_n}{d\eta_n} \frac{d\eta_n}{d\beta_i} \tag{11}$$

The second term in Equation 11 is the easiest to compute, which from Equation 3 we see is simply $\frac{d(\beta^T x_n)}{d\beta_i} = x_{ni}$. Computing $\frac{d\mu_n}{d\eta_n}$ is a little trickier as $\mu_n = \frac{1}{1+e^{-\eta_n}}$. However, from Equation 2 we can define the *inverse function* is defined as:

$$\eta_n \equiv \log \frac{\mu_n}{1 - \mu_n} \tag{12}$$

This is also referred to as the *log odds*. Now if we differentiate Equation 12 with respect to μ_n using the chain rule and calculus of logarithms, we get:

$$\frac{d\eta_n}{d\mu_n} = \frac{1-\mu_n}{\mu_n}\left(\frac{(1-\mu_n)-\mu_n}{(1-\mu_n)^2}\right) = \frac{1}{\mu_n(1-\mu_n)} \tag{13}$$

Simply inverting the above derivative gives us the derivative of μ_n with respect to η_n:

$$\frac{d\mu_n}{d\eta_n} = \mu_n\,(1-\mu_n) \tag{14}$$

Now combining the first and second terms for Equation 11, we get the following expression:

$$\frac{d\mu_n}{d\beta_i} = \mu_n\,(1-\mu_n)\,x_{ni} \tag{15}$$

Plugging this in for the second term for $\frac{dL}{d\beta_i}$ in Equation 9, and combining the expression in Equation 10, yields:

$$\begin{aligned}
\frac{dL}{d\beta_i} &= \sum_{n=1}^{N} \frac{((1-\mu_n)y_n - (1-y_n)\mu_n)}{(\mu_n(1-\mu_n))} \cdot (\mu_n\,(1-\mu_n)\,x_{ni}) \\
&= \sum_{n=1}^{N} (1-\mu_n)y_n\,x_{ni} - (1-y_n)\mu_n\,x_{ni} \\
&= \sum_{n=1}^{N} y_n\,x_{ni} - \mu_n\,x_{ni} = \sum_{n=1}^{N}(y_n-\mu_n)x_{ni}
\end{aligned} \tag{16}$$

Where again, our dependency on β enters through the definition of μ_n (see Equation 2). To close this derivation of the MLE, note that:

$$E[y_n|x_n,\beta] = p(y_n|x_n,\beta) = \mu_n \tag{17}$$

And therefore:

$$\frac{dL}{d\beta_i} = \sum_{n=1}^{N}(y_n - E[y_n|x_n,\beta])x_{ni} \tag{18}$$

Notice that this is analogous to the expression we had from linear regression:

$$\frac{dL^{\text{LinReg}}}{d\beta_i} = \sum_{n=1}^{N}(y_n - \beta^T x_n)x_{ni} \tag{19}$$

Final Comments

We can apply *regularization* techniques to the logistic model as we did with linear regression (see notes on Linear Regression for details).

$$L^{\text{Reg}} = \sum_{n=1}^{N} y_n\mu_n + (1-y_n)\mu_n + ||\beta||_q \tag{20}$$

Finally, let us examine the connection to the naive Bayes model. The logistic regression model is what is known as a *discriminative model*, whereas the naive Bayes model is what is known as a *generative model*. The differences between these models are highlighted in Figure 5. It has been empirically observed that discriminative models outperform generative models for larger amounts of data. The reasoning for this is still debated. Is it because large amounts of data can potentially contain more outliers? Is the original underlying model incorrect?

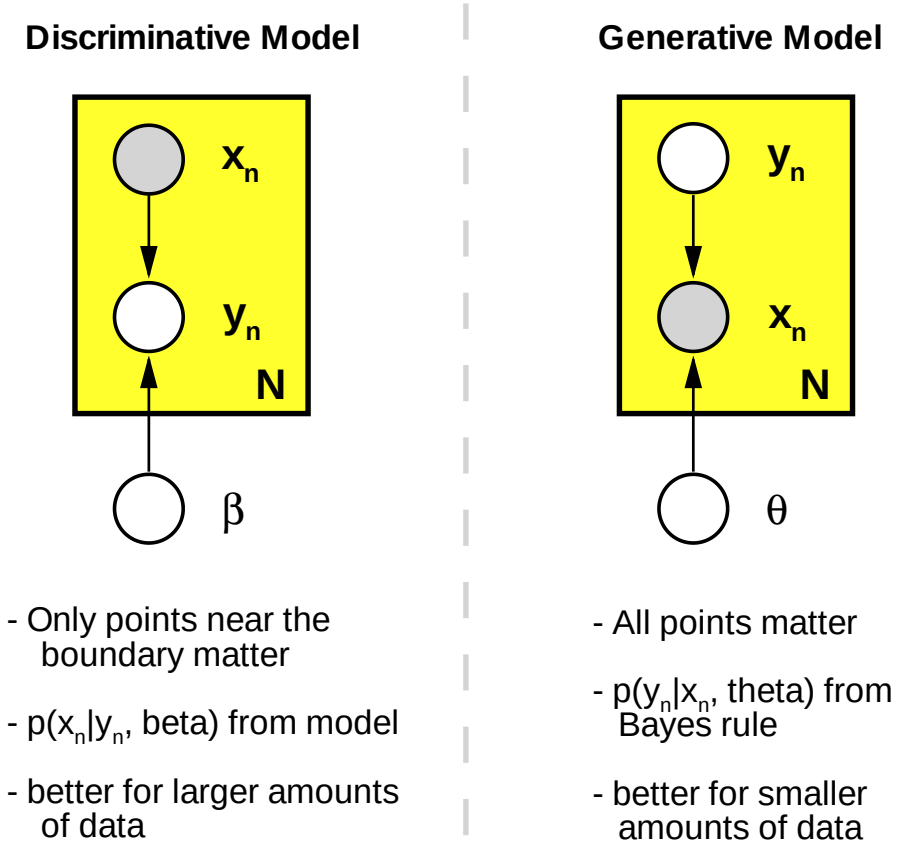

Figure 5: Comparison of discriminative and generative models. Note that the above probabilities should read $p(y_n|x_n, \beta)$ and $p(y_n|x_n, \theta)$, respectively.