CSE 788.04: Topics in Machine Learning	**Lecture Date: April 11th, 2012**

Lecture 6: Bayesian Logistic Regression

Lecturer: Brian Kulis *Scribe: Ziqi Huang*

1 Logistic Regression

Logistic Regression is an approach to learning functions of the form $f : X \rightarrow Y$ or $P(Y|X)$, in the case where Y is discrete-valued, and $X =< X_1 ... X_n >$ is any vector containing discrete or contiuous variables. For a two-class classification problem, the posterior probability of Y can be written as follows:

$$\begin{aligned} P(Y=1|X) &= \frac{1}{1+exp(-\omega-\sum_{i=1}^{n}\omega_i X_i)} \\ &= \sigma(\omega^T X_i) \end{aligned} \tag{1}$$

and

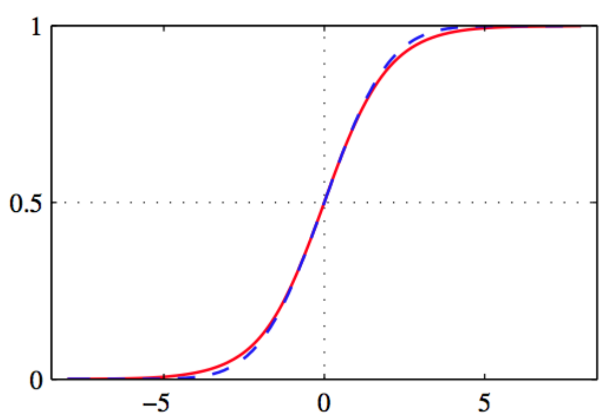
$$\begin{aligned} P(Y=0|X) &= \frac{exp(-\omega-\sum_{i=1}^{n}\omega_i X_i)}{1+exp(\omega+\sum_{i=1}^{n}\omega_i X_i)} \\ &= 1-\sigma(\omega^T X_i) \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ is the *logistic sigmoid* function defined by

$$\sigma(a) = \frac{1}{1+exp(-a).} \tag{3}$$

which is plotted in Figure1. (Note we are implicitly redefining the data X_i to add an extra dimension with

Figure 1: Plot of the logistic sigmoid function.



a 1, as in linear regression, and then redefining ω appropriately.) The term sigmoid means S-shaped. This type of function is sometimes also called a squashing function because it maps the whole real axis into a finite interval.It satisfies the following symmetry property

$$\sigma(-a) \quad = \quad 1 - \sigma(a) \tag{4}$$

Interestingly, the parametric form of $P(Y|X)$ used by Logistic Regression is precisely the form implied by the assumption of a Gaussian Naive Bayes classifier.

1.1 Form of $P(Y|X)$ for Gaussian Naive Bayes Classifier

We derive the form of $P(Y|X)$ entailed by the assumption of a Gaussian Naive Bayes (GNB) classifier. Consider a GNB based on the following modeling assumption:

- Y is Boolean, governed by a Bernoulli distribution, with parameter $\pi = P(Y = 1)$.
- $X =< X_1 ... X_n >$, where each X_i is a continuous random variable.
- For each X_i, $P(X_i|Y = y_k)$ is a Guassian distribution of the form $N(\mu_{ik}, \sigma_i)$ (in many cases, this will simply be $N(\mu_k, \sigma)$).
- For all i and $j \neq i$, X_i and X_j are conditionally independent given Y.

Note here we are assuming the standard deviations σ_i vary from point to point, but do not depend on Y.

We now derive the parametric form of $P(Y|X)$ that follows from this set of GNB assumptions. In general, Bayes rule allows us to write

$$P(Y = 1|X) \quad = \quad \frac{P(Y = 1)P(X|Y = 1)}{P(Y = 1)P(X|Y = 1) + P(Y = 0)P(X|Y = 0)} \tag{5}$$

Dividing both the numerator and denominator by the numerator yields:

$$\begin{aligned}
P(Y = 1|X) &= \frac{1}{1 + \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}} && (6)\\
&= \frac{1}{1 + exp(ln\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)})} && (7)\\
&= \frac{1}{1 + exp(ln\frac{P(Y=0)}{P(Y=1)} + \sum_i ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)})} && (8)\\
&= \frac{1}{1 + exp(ln\frac{1-\pi}{\pi} + \sum_i ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)})} && (9)
\end{aligned}$$

Note the final step expresses P(Y=0) and P(Y=1) in terms of the binomial parameter π.

Now consider just the summation in the denominator of equation (9). Given our assumption that $P(X_i|Y = y_k)$ is Gaussian, we can expand this term as follows:

$$\sum_i \ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)} = \sum_i \ln \frac{\frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{-(X_i-\mu_{i0}^2)}{2\sigma_i^2})}{\frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{-(X_i-\mu_{i1}^2)}{2\sigma_i^2})} \tag{10}$$

$$= \sum_i \ln exp(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}) \tag{11}$$

$$= \sum_i (\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}) \tag{12}$$

$$= \sum_i (\frac{(X_i^2-2X_i\mu_{i1}+\mu_{i1}^2)-(X_i^2-2X_i\mu i0+\mu_{i0}^2)}{2\sigma_i^2}) \tag{13}$$

$$= \sum_i (\frac{2X_i(\mu_{i0}-\mu_{i1})+\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}) \tag{14}$$

$$= \sum_i (\frac{\mu_{i0}-\mu_{i1}}{sigma^2} X_i + \frac{\mu_{i1}^2}{2\sigma_i^2}) \tag{15}$$

Note this expression is a linear weighted sum of the $X_i's$. Substituting expression (15) back into equation (9), we have

$$P(Y=1|X) = \frac{1}{1+exp(\ln\frac{1-\pi}{\pi}+\sum_i(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i+\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}))} \tag{16}$$

Or equivalently,

$$P(Y=1|X) = \frac{1}{1+exp(\omega_0+\sum_{i=1}^n \omega_i X_i)} \tag{17}$$

where the weights $\omega_1 ... \omega_n$ are given by

$$\omega_i = \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} \tag{18}$$

and

$$\omega_0 = \ln\frac{1-\pi}{\pi} + \sum_i \frac{\mu_{i1}^2-\mu i0^2}{2\sigma_i^2} \tag{19}$$

Then we can derive

$$P(Y=1|X) = \sigma(\omega^T X_i) \tag{20}$$

And also we have

$$P(Y=0|X) = 1-\sigma(\omega^T X_i) \tag{21}$$

To summarize, the logistic form arises naturally from a generative model. However, since the number of parameters in a generative model is often more than the number of parameters in the logistic regression model, one often prefers working directly with the logistic regression model to find the parameters W. This is a *discriminative* approach to classification, as we directly model the probabilities over the class labels.

2 Estimating Parameters for Logistic Regression

One reasonable approach to traning Logistic Regression is to choose parameter values that maximize the conditional data likelihood. We choose parameters

$$W \leftarrow \arg\max_{W} \prod_{i} P(Y_i|X_i, W)$$

where $W =< \omega_0, \omega_1 ... \omega_n >$ is the vector of parameters to be estimated, Y^l denotes the observed value of Y in the l^{th} training example, and X^l denotes the observed value in the l^{th} training example. Equivalently, we can work with the log of the conditional likelihood:

$$W \leftarrow \arg\max_{W} \prod_{i} \ln P(Y_i|X_i, W)$$

And

$$\begin{aligned}
\ln P(Y_i|X_i, W) &= \sum_{i=1}^{n} Y_i \ln P(Y_i = 1|X_i, W) + (1 - Y_i) \ln P(Y_i = 0|X_i, W) && (22) \\
&= \sum_{i=1}^{n} Y_i \ln \sigma(\omega^T X_i) + (1 - Y_i) \ln (1 - \sigma(\omega^T X_i)) && (23) \\
&= \sum_{i=1}^{n} Y_i \ln \frac{\sigma(\omega^T X_i)}{(1 - \sigma(\omega^T X_i)} + \ln (1 - \sigma(\omega^T X_i)) && (24)
\end{aligned}$$

As usual, we can define an error function by taking the negative logarithm of the likelihood, which gives the crossentropy error function in the form

$$\begin{aligned}
E(W) &= -\ln p(Y'|W) && (25) \\
&= -\sum_{n=1}^{N} y'_n \ln y_n + (1 - y'_n) \ln (1 - y_n) && (26)
\end{aligned}$$

Unfortunately, there is no closed form solution to maximizing the likelihood with respect to W. The singularity can be avoided by inclusion of a prior and finding a MAP solution for w, or equivalently by adding a regularization term to the error function.

2.1 Iterative reweighted least squares

In the case of the linear regression models, the maximum likelihood solution, on the assumption of a Gaussian noise model, leads to a closed-form solution. However, for logistic regression, there is no longer a closed-form solution, due to the nonlinearity of the logistic sigmoid function. However, the departure from a quadratic form is not substantial. To be precise, the error function is concave, as we shall see shortly, and hence has a unique minimum. Furthermore, the error function can be minimized by an efficient iterative technique based on the $Newton - Raphson$ iterative optimization scheme, which uses a local quadratic approximation to the log likelihood function.

$$W^{(new)} = W^{(old)} - [\nabla^2 \sigma(W^{(old)})]^{-1} \nabla f(W^{(old)}) \qquad (27)$$

Then we can derive

$$\begin{aligned}
\nabla E(W) &= \sum_{n=1}^{N} (W^T X_n - Y') X_n && (28) \\
&= X^T X W - X^T Y' && (29)
\end{aligned}$$

$$\begin{aligned} \nabla^2 E(W) &= \sum_{n=1}^{N} X_n X_n^T & (30)\\ &= X^T X & (31) \end{aligned}$$

Plug in equation (27), we can derive

$$\begin{aligned} W^{(new)} &= W^{(old)} - (X^T X)^{-1}\{X^T X W^{(old)} - X^T Y\} & (32)\\ &= (X^T X)^{-1} X^T Y & (33) \end{aligned}$$

Now let us apply the Newton-Raphson update to the cross-entropy error function (26) for the logistic regression model.

$$\begin{aligned} \nabla E(W) &= \sum_{n=1}^{N} (y_n - y_n')x_n & (34)\\ &= X^T(Y - Y') & (35) \end{aligned}$$

$$\begin{aligned} H = \nabla\nabla E(W) &= \sum_{n=1}^{N} y_n(1-y_n)x_n x_n^T & (36)\\ &= X^T R X & (37) \end{aligned}$$

where $R = y_n(1 - y_n)$, then we can derive

$$\begin{aligned} W^{(new)} &= W^{(old)} - (X^T R X)^{-1} X^T (Y - Y') & (38)\\ &= (X^T R X)^{-1}\{X^T R X W^{(old)} - X^T (Y - Y')\} & (39)\\ &= (X^T R X)^{-1} X^T R z & (40) \end{aligned}$$

where $z = XW^{(old)} - R^{-1}(Y - Y')$.

2.2 Regularization in Logistic Regression

Overfitting the training data is a problem that can arise in Logistic Regression, especially when data is very high dimensional and training data is sparse. One approach to reducing overfitting is regularization, in which we create a modified penalized log likelihood function, which penelizes large value of W. One approach is to use the penalized log likelihood function

$$W \leftarrow \arg\max_W \sum \ln P(Y_i|X_i, w) - \frac{\lambda}{2}\|W\|^2 \quad (41)$$

Which adds a penalty proportional to the squared magnitude of W. Here λ is a constant that determines the strength of this penalty term.

3 The Bayesian Setting

Now we have some assmputions:

- $P(W) \propto N(\mu_0, \sigma_0^2)$

- $P(Y|X,W) \propto \sigma(W^T X)$
- $P(W|Y,X) \propto P(Y|X,W)P(W) \propto \sigma(W^T X)P(W)$

Then for a new data X_{new}, we can derive the predictive distribution:

$$P(Y_{new}|X,\bar{Y},X_{new}) = \int P(Y_{new}|\bar{Y},X_{new})P(W|\bar{Y},X)d_W \tag{42}$$

Since $P(Y_{new}|\bar{Y},X_{new})$ is proportional to logistic sigmoid distribution and $P(W|\bar{Y},X)$ is proportional to Normal distribution, there is no closed form for $P(Y_{new}|X,\bar{Y},X_{new})$. There are several approaches to approximating the predictive distribution: the Laplace approximation, variational methods, and Monte Carlo sampling are three of the main ones. Below we focus on the Laplace approximation.

4 The Laplace Approximation

In this section, we introduce a framework called the Laplace approximation, that aims to find a Gaussian approximation to a probability density defined over a set of continuous variables.

We assume that there is a $f(x)$ where $\int exp(Nf(x))dx$ has no closed form. In the Laplace method the goal is to find a Gaussian approximation $g(z)$ which is centred on a mode of the distribution $f(x)$. The first step is to find a mode for $f(x)$, in other words a point x_0 such that $f'(x_0) = 0$. A Gaussian distribution has the property that its logarithm is a quadratic function of the variables. We therefore consider a Taylor expansion of $f(x)$

$$f(x) \approx f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 \tag{43}$$

Since $f'(x_0) = 0$

$$f(x) \approx f(x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 \tag{44}$$

Therefore,

$$\begin{aligned}
\int exp(-Nf(x))dx &= \int exp\Big(-N\Big(f(x_0) + \frac{|f''(x_0)|}{2!}(x-x_0)^2\Big)\Big)dx && (45)\\
&= exp(-Nf(x_0)\int exp\Big(-N\frac{|f''(x_0)|}{2!}(x-x_0)^2\Big)dx && (46)\\
&= \sqrt{\frac{2\pi}{N|f''(x_0)|}}exp(-Nf(x_0)) && (47)
\end{aligned}$$

So we can get a approximate closed form solution for $f(x)$. Note that the approximate integral is accurate to order $O(1/N)$. Finally, given a distribution $p(x)$, using the Laplace approximation we form a Gaussian approximation with mean x_0 and precision $|p''(x_0)|$, where x_0 is a mode of p.

4.1 Example: $P(W|Y) \propto P(Y|W,X)P(W)$

As we introduced before, for the predictive distribution

$$\begin{aligned}
P(Y_{new}|X,\bar{Y},X_{new}) &= \int P(Y_{new}|\bar{W},X_{new})P(W|\bar{Y},X)d_W && (48)\\
&= \int \sigma(W^T X)P(W|\bar{Y},X)d_w && (49)
\end{aligned}$$

we cannot get a closed form solution. We will use the Laplace approximation to approximate the posterior $P(W|\bar{Y}, X)$ as a Gaussian. We further approximate $P(Y_{new}|\bar{Y}, X_{new})$ as a probit function $\phi(\lambda a)$ and get an approximate solution for the predictive distribution. Because $P(W|\bar{Y}, X)d_W$ is approximated as Gaussian, we know that the marginal distribution will also be Gaussian. Since

$$\sigma(W^T X) = \int \delta(a - W^T X)\sigma(a)d_a \tag{50}$$

Then we derive

$$\int \sigma(W^T X)P(W|\bar{Y}, X)d_w = \int \sigma(a)P(a)d_a \tag{51}$$

where $P(a) = \int \delta(a - W^T X)P(W|\bar{Y}, X)d_w$ then we can derive

$$\int \phi(\lambda a)N(a|\mu, \sigma)d_a \approx \sigma(k(\sigma^2)\mu) \tag{52}$$

where $k(\sigma^2) = (1 + \frac{\pi\sigma^2 i}{8})^{-\frac{1}{2}}$ and

$$\begin{aligned} \mu_a &= \int p(a)ad_a && (53)\\ &= \int P(W|\bar{Y}, X)W^T X d_W && (54)\\ &= W_{MAP}^T X && (55) \end{aligned}$$

and also we can derive

$$\begin{aligned} \sigma_a^2 = \int p(a)(a^2 - \mu^2)d_a & && (56)\\ &= \int P(W|\bar{Y}, X)(W^T X^2 - m_N^T X^2)d_w && (57)\\ &= X^T S_N X && (58) \end{aligned}$$