STAT 383C: Statistical Modeling I **Fall 2015**

Lecture 14 — October 13

Lecturer: Purnamrita Sarkar *Scribe: Mingzhang Yin, Can Gokalp*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

Note: The latex template was borrowed from EECS, U.C. Berkeley.

14.1 Conditional Independence

Definition 14.1. *For data point i; $X_{i1}, X_{i2}, ..., X_{ip}$ conditionally independent given $Y_i = y_i$*

$$\begin{aligned} P(Y = y|X_{i1} = x_1, X_{i2} = x_2, ..., X_{ip} = x_p) \propto & P(X_{i1} = x_1, X_{i2} = x_2, ..., X_{ip}|Y = y)P(Y = y) \\ \propto & \prod_{i=1}^{p} P(X_{ij} = x_j|Y = y)P(Y = y) \end{aligned}$$

Now each parameter

$$\begin{aligned} \theta_{jy} = P(X_{ij} = 1|Y = y) = & \frac{\sum_{i=1}^{n} \mathbb{1}_{(X_{ij}=1, y_i=y)}}{\sum_{i=1}^{n} \mathbb{1}_{(y_i=y)}} \\ P(X_{i1}, X_{i2}|Y = y) = & \theta_{1y}\theta_{2y} \end{aligned}$$

The problem we have here is that if we don't have that observation for any of the θ_{jy}'s, that is to say if $\sum_{i=1}^{n} \mathbb{1}_{(X_{ij}=1, y_i=y)} = 0$ for a θ_{jy} then we would have $P(X_{i1}, X_{i2}|Y = y) = \theta_{1y}\theta_{2y} = 0$. Therefore we add a smoothing term to θ_{jy}. Now each parameter would be;

$$\theta_{jy} = P(X_{ij} = 1|Y = y) = \frac{\sum_{i=1}^{n} \mathbb{1}_{(X_{ij}=1, y_i=y)} + 1}{\sum_{i=1}^{n} \mathbb{1}_{(y_i=y)} + 2}$$

14.2 Logistic Regression vs. Naive Bayes

Both models perform classification into two classes. Both models have features, **X**, and classes Y=1 or Y=0. Naive Bayes (generative model) estimates a joint a probability for p(x,y) equal to $p(y) \times p(x|y)$ from the training data, and uses Bayes Rule to predict $p(y|x)$

for new test cases. Gaussian Naive Bayes (GNB) uses a multivariate Gaussian for the probability of the features. In contrast, logistic regression models logit$p(y|x)$ directly as a linear function (discriminative). Let us compare these two methods using some pictures first. But before that we will need some definitions.

Gaussian Naive Bayes We will consider two types of GNB. Note that in GNB we model:

$$X|Y = 1 \sim N(\mu_1, \Sigma_1)$$
$$X|Y = 0 \sim N(\mu_0, \Sigma_0)$$

Note that by assumption, Σ_1 and Σ_0 are diagonal.

Definition 14.2. *We will denote by GNB_1 a Naive Bayes model with $\Sigma_0 \neq \Sigma_1$. Recall from the last lecture that this leads to a quadratic decision boundary.*

are diagonal and the decision boundary is quadratic.

Definition 14.3. *We denote by GNB2, the case with $\Sigma_0 = \Sigma_1$. Here the decision boundary is linear.*

Several questions are considered when comparing Gaussian Naive Bayes (GNB) or Logistic Regression (LR) for classification. Questions consider the comparison of GNB_2 models and LR. For example, which is a preferred model to use– logistic regression or naive bayes? For any GNB_2 model, can there be a logistic regression decision boundary? When does the LR fit everything but the GNB_2 does not? When comparing GNB_1 and LR, is one a model a strict subset of the other?

GNB vs LR GNB-1 can learn quadratic decision boundaries, but LR cannot. On the other hand, LR can learn the following which GNB-1 or GNB-2 cannot.

On the other hand whatever decision boundary GNB-2 learns is linear and hence can be learned by Logistic Regression. So Logistic Regression is strictly a better classifier than GNB-2, since any classifier learned by GNB-2 can be learned by LR as well, but not the opposite.

In real life, we would not have the parameters at hand and so we would use the decision boundary after plugging in the MLE of each of the models parameters $\pi, \mu_1, \Sigma_1, \mu_0$ and Σ_0. Below are some examples of linear and quadratic decision boundaries.

If $\Sigma_0 \neq \Sigma_1$ then we have a quadratic decision boundary as in figure **??**:

$$\hat{\mu_0} = \frac{\sum_1^n X_i \mathbb{1}(Y_i = 0)}{\sum_1^n \mathbb{1}(Y_i = 0)} \tag{14.1}$$

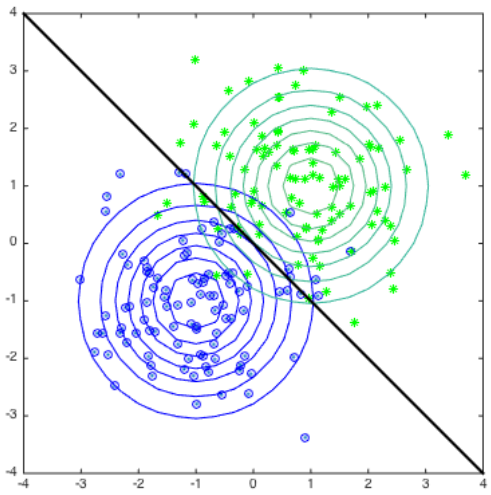

Figure 14.1: Linear decision boundary with equal class proportions

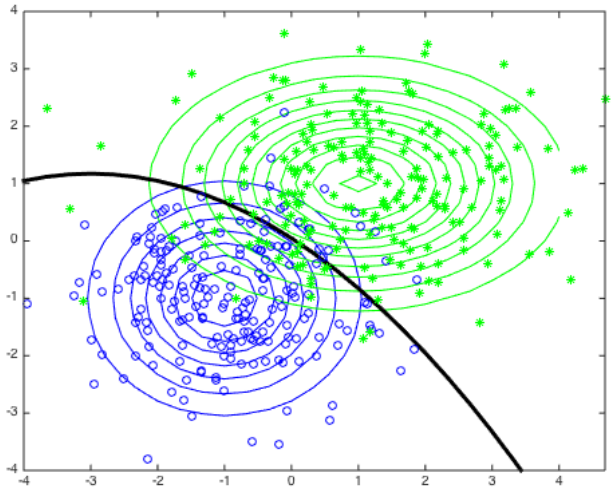

Figure 14.2: Quadratic decision boundary with equal class proportions

Also, for $j = p$,

$$\hat{\sigma_{0j}}^2 = \frac{\sum_1^n (X_{ij} - \hat{\mu_{0j}})^2 \mathbb{1}(Y_i = 0)}{\sum_1^n \mathbb{1}(Y_i = 0)} \tag{14.2}$$

For an unbiased estimate we would normalize by $\sum_1^n \mathbb{1}(Y_i = 0) - 1$. If we get the posterior distribution of Y, the question is how to make predictions with a given X? Actually, if $P(Y = 1|X) > P(Y = 0|X)$, we would label Y as 1. This classification criterion can be

expressed as

$$\begin{aligned} log\frac{P(Y=1|x)}{P(Y=0|x)} &= log\frac{f(X|Y=1)p(Y=1)}{f(X|Y=1)p(Y=1)} \\ &= -\frac{(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}{2} + \frac{(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)}{2} + log\frac{\pi}{1-\pi} \\ &> 0 \end{aligned} \tag{14.3}$$

Here $\pi = P(Y=1)$. If we further assume $\Sigma_1 = \Sigma_0 = \Sigma$, then we would get a linear decision boundary

$$(\mu_1-\mu_0)^T\Sigma^{-1}(x - \frac{\mu_0+\mu_1}{2}) + log\frac{\pi}{1-\pi} > 0 \tag{14.4}$$

Whenever we get data x, we can plug in the left side of the inequality and label y as 0 if it is bigger than 0.

Asymptotic behavior For linear classifiers, it is generally believed that LR is preferable over GNB-2, since it makes no assumption about conditional independence and hence can estimate a richer class of linear models for classification. However Ng and Jordan showed in their very nice paper [1] that this is only part of the story. Since GNB-2 makes simplifying assumptions about the model it needs far fewer data-points to reach its (possibly worse) asymptotic error rate than LR. More concretely they showed that if the data is in k dimensions, then LR needs $O(k)$ data-points to converge to its asymptotic error rate (which is lower than that of GNB-2), whereas GNB-2 needs only $O(\log k)$ samples.

14.3 Generalized linear models

14.3.1 Logistic regression of categorical data

In previous lecture we introduce logistic regression and IRLS (iteratively reweighed least squares), but the data we are dealing with is listed by "subject number", which is corresponding to each observation. A simple example is shown in Table 14.1a. However, as we can see, several observations may share the same covariate vector, like in the Table 14.1a, X_i indexed with subject No. 1, 3, 5 share the same covariate vector $< 1, 1 >$, we say they are in the same covariate class. It would be more reasonable and more efficient if we list the data by the covariate class, a simple example is in Table 14.1b. $\boldsymbol{\beta}$An easy example is:

Now consider implementing logistic regression with this kind of categorical data. Denote m_i as the number of observations in class X_i, $y_i \sim \text{Bin}(m_i, \pi_i)$. Then link function is

$$g(\pi_i) = \log\frac{\pi_i}{1-\pi_i} = X_i^T\boldsymbol{\beta}.$$

Now let us do the maximum likelihood estimation for logistic regression, the conditional likelihood and the conditional log likelihood are

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{n}\binom{m_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{m_i-y_i} = \prod_{i=1}^{n}\binom{m_i}{y_i}(\frac{\pi_i}{1-\pi_i})^{y_i}(1-\pi_i)^{m_i},\\ \ell &= \sum_{i=1}^{n}\left(y_i X_i^T\boldsymbol{\beta} - m_i\log\left(1+\exp(X_i^T\boldsymbol{\beta})\right)\right)+\text{const},\end{aligned}$$

respectively.

Take the derivative of conditional log likelihood ℓ and set to 0, we can get:

$$\begin{aligned}\frac{\partial\ell}{\partial\beta_r} &= \sum_{i=1}^{n}y_i X_{ir} - \sum_{i=1}^{n}m_i\underbrace{\frac{\exp(X_i^T\boldsymbol{\beta})}{1+\exp(X_i^T\boldsymbol{\beta})}}_{\pi_i}X_{ir}\\ &= \sum_{i=1}^{n}(y_i - m_i\pi_i)X_{ir} = 0,\\ \frac{\partial\ell}{\partial\boldsymbol{\beta}} &= \boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{\mu}) = 0,\end{aligned}$$

where $\mu_i = m_i\pi_i$.

Note that there is not closed form solution as π_i is a non-linear function of $\boldsymbol{\beta}$. So we need to compute $\boldsymbol{\beta}$ iteratively using Newton-Raphson method, with update rule:

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}_t - H^{-1}\frac{\partial\ell}{\partial\boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^t},$$

which require us to compute the Hessian. An important intermediate step to get Hessian is:

$$\begin{aligned}\frac{\mathrm{d}\pi_i}{\mathrm{d}\beta_s} &= \frac{\mathrm{d}}{\mathrm{d}\beta_s}\left(\frac{\exp(X_i^T\boldsymbol{\beta})}{1+\exp(X_i^T\boldsymbol{\beta})}\right)\\ &= \frac{\exp(X_i^T\boldsymbol{\beta})}{\left(1+\exp(-X_i^T\boldsymbol{\beta})\right)^2}X_{is}\\ &= X_{is}\pi_i(1-\pi_i).\end{aligned}$$

No.	X_i	y_i
1	$<1,1>$	1
2	$<1,2>$	0
3	$<1,1>$	1
4	$<2,1>$	1
5	$<1,1>$	0

(a) Data listed by subject No.

X_i	m_i	y_i
$<1,1>$	3	2
$<1,2>$	1	0
$<2,1>$	1	1

(b) Data listed by covariate class

Table 14.1: Alternative ways of presenting the same data.

Then Hessian is

$$\begin{aligned} H_{rs} &= \frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} = -\sum_{i=1}^{n} m_i X_{ir} \left(\frac{\mathrm{d}}{\mathrm{d}\beta_s} \pi_i \right) = -\sum_{i=1}^{n} m_i X_{ir} X_{is} \pi_i (1-\pi_i), \\ H &= -\boldsymbol{X}^T W \boldsymbol{X}, \end{aligned}$$

where W is a diagonal matrix with $W_{ii} = m_i \pi_i (1-\pi_i)$, which can be seen as the variance of y_i (note that y_i is from binomial distribution).

Plugging the Hessian to the update rule of $\boldsymbol{\beta}$, we get the IRLS for this problem:

$$\begin{aligned} \boldsymbol{\beta}^{t+1} &= \boldsymbol{\beta}^t + (\boldsymbol{X}^T W \boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{\mu})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^t} \\ &= (\boldsymbol{X}^T W \boldsymbol{X})^{-1} \boldsymbol{X}^T W \underbrace{\left(\boldsymbol{X}\boldsymbol{\beta}^t + W^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \right)}_{z_t}, \end{aligned}$$

Note that W here is actually W^t, which changes every iteration, and that is where "reweighed" comes from. The stop criterion should be something like

$$\frac{|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t|}{|\boldsymbol{\beta}^t|} < 10^{-c},$$

where $c > 0$ can be chosen by the user. This is in case that $|\boldsymbol{\beta}^t|$ is very small and may be in same order as 10^{-c}.

Analogy 14.1. *(Estimation of $\boldsymbol{y}$: LS and IRLS.)*

1. *Least squares is to minimize squared error*

   $$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

   where the model is

   $$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I});$$

2. *IRLS is to minimize weighed squared error*

   $$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

   , which arises from MLE estimation for the ***heteroskedastic*** *model*

   $$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \Sigma).$$

14.3.2 Logistic regression of count data

Now we have applied logistic regression to categorical data, what about count data? The procedures are very similar, we use Poisson distribution, which is widely used to model count data, as an example here.

Suppose $y_i \sim \text{Pois}(\lambda_i)$, where $\lambda_i = \exp(X_i^T \boldsymbol{\beta})$.

First write out the conditional likelihood $\mathcal{L}$ and conditional log likelihood ℓ, get the derivative of ℓ and set to 0:

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \\
\ell &= \sum_{i=1}^{n} \left(-\lambda_i + y_i X_i^T \boldsymbol{\beta}\right) + \text{const}, \\
\frac{\partial \ell}{\partial \beta_r} &= \sum_{i=1}^{n} \left(y_i X_{ir} - \lambda_i X_{ir}\right) = \sum_{i=1}^{n} \left(y_i - \lambda_i\right) X_{ir} = 0, \\
\frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}),
\end{aligned}$$

where $\mu_i = \lambda_i$.

There is also no closed form solution as λ_i is not invariant to $\boldsymbol{\beta}$. And use Newton-Raphson method too:

$$\begin{aligned}
\boldsymbol{\beta}^{t+1} &= \boldsymbol{\beta}_t - H^{-1} \frac{\partial \ell}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^t}, \\
H_{rs} &= \frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} = -\frac{\mathrm{d}}{\mathrm{d}\beta_s} \sum_{i=1}^{n} \lambda_i X_{ir} = -\sum_{i=1}^{n} \lambda_i X_{is} X_{ir}, \\
H &= -\boldsymbol{X}^T W \boldsymbol{X},
\end{aligned}$$

where W is a diagonal matrix with $W_{ii} = \lambda_i$, which can be seen as the variance of y_i (note that y_i is from Poisson distribution).

The IRLS steps are exactly the same as what we have when discussing logistic regression for categorical data.

Note that a fan fact is that diagonal of W are variance of y_i for both categorical data and count data, this leads to an analogy with least squares.

Analogy 14.2. *(Estimation of $\boldsymbol{\beta}$: LS and LR)*

1. *Least squares estimates $\boldsymbol{\beta}$ as*

$$\hat{\boldsymbol{\beta}}_{ls} \sim \mathcal{N}\left(\boldsymbol{\beta}, (\boldsymbol{X}^T \boldsymbol{X})^{-1} \sigma^2\right);$$

2. Logistic regression estimates $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}_{logistic} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\beta}, (\boldsymbol{X}^T W \boldsymbol{X})^{-1}\right).$$

So we get

$$\begin{aligned} \mathrm{E}\left[\hat{\boldsymbol{\beta}}\right] &\rightarrow \boldsymbol{\beta}, \\ \mathrm{cov}(\hat{\boldsymbol{\beta}}) &\rightarrow (\boldsymbol{X}^T W \boldsymbol{X})^{-1}. \end{aligned}$$

References

[1] McCullagh, Peter, and John A. Nelder. Generalized linear models. CRC press., 1989.