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

Lecture 10 — September 29

Lecturer: Purnamrita Sarkar *Scribe: Mingzhang Yin*

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

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

10.1 Naive Bayes

10.1.1 Naive Bayes for discrete random variable

Suppose $X_1, X_2, \ldots, X_k$ and Y are Boolean variables, i.e. $X_i \in \{0,1\}, Y \in \{0,1\}$. For example, consider the spam email classification problem, in which $Y = 1$ means the email is spam and each X_i represents a certain feature. For instance, X_i can be the key word "Buy now" or "Weight loss", and $X_i = 1$ means such key word appears in the email X. Suppose there are k features and according to the Bayes Rule:

$$P(Y = y|X_1 = x_1, \ldots, X_k = x_k) = \frac{P(X_1 = x_1, \ldots, X_k = x_k|Y = y)P(Y = y)}{P(X_1 = x_1, \ldots, X_k = x_k)} \tag{10.1}$$

If we denote a set of parameters as

$$\theta_{(x_1,\ldots,x_n),j} \equiv P(X_1 = x_1, \ldots, X_k = x_k|Y = y_j) \tag{10.2}$$

Then in order to estimate the posterior distribution, we have $(2^k - 1) * 2 + 1$ parameters needed to estimate. This can be unrealistic in practical learning domains.

Definition 1. (Conditional Independence)
Given random variables X,Y,and Z, we say X is **conditionally independent** *of Y given Z, if and only if the probability distribution of X is independent of Y given Z:*

$$P(X = x|Y = y, Z = z) = P(X = x|Z = z)(\forall x, y, z) \tag{10.3}$$

Assume X_i are conditionally independent given Y, then we get

$$P(y = 1|X_1 = x_1, \ldots, X_k = x_k) \propto \prod_i^k P(X_i = x_i|Y = y) * P(Y = y) \tag{10.4}$$

Here there are k parameters $P(X_j = 1|Y = y)$, so the total number of parameters is $2k + 1$. Denote the parameters by

$$\theta_{jy} = P(X_j = 1|Y = y) \tag{10.5}$$

Then $\theta_{j1} = P(X_j = 1|Y = 1)$. Our estimator for θ_{j1} is

$$\hat{\theta}_{j1} = \frac{\sum_{i=1}^{n} \mathbb{1}\{X_{ij} = 1, Y_i = 1\}}{\sum_{i}^{n} \mathbb{1}\{Y_i = 1\}} \tag{10.6}$$

For one data point (X_i, Y_i):

$$P(X_i|Y_i)P(Y_i) = \prod_{j=1}^{k} P(X_{ij} = x_{ij}|Y_i = y_i)P(Y_i = y_i) \tag{10.7}$$

Note we can write $P(X_{ij} = x_{ij}|Y_i = y_i)$ as

$$\theta_{jy_i}^{x_{ij}\mathbb{1}(Y_i=y_i)}(1 - \theta_{jy_i})^{(1-x_{ij})\mathbb{1}(Y_i=y_i)} \tag{10.8}$$

so

$$P(X|Y)P(Y) = \prod_{i=1}^{n}\prod_{j=1}^{k} \theta_{jy_i}^{x_{ij}\mathbb{1}(Y_i=y_i)}(1 - \theta_{jy_i})^{(1-x_{ij})\mathbb{1}(Y_i=y_i)} P(Y_i = y_i) \tag{10.9}$$

Compare the expression above with the likelihood function of binomial distribution, for a fixed j, we proved that $\hat{\theta}_{jy_i}$ is the MLE of θ_{jy_i}, where

$$\hat{\theta}_{jy_i} = \frac{\sum_{i=1}^{n} \mathbb{1}\{X_{ij} = 1, Y_i = y_i\}}{\sum_{i}^{n} \mathbb{1}\{Y_i = y_i\}} \tag{10.10}$$

One shortcoming of this maximum likelihood estimate is that it can result in θ estimates of zero. To avoid this, it is common to smooth the estimate by adding some hallucinated example which are spread evenly on the possible values of X_i.

$$\hat{\theta}_{jy_i} = \frac{\sum_{i=1}^{n} \mathbb{1}\{X_{ij} = 1, Y_i = y_i\} + 1}{\sum_{i}^{n} \mathbb{1}\{Y_i = y_i\} + 2} \tag{10.11}$$

And this approach is called Laplace smoothing.

10.1.2 Naive Bayes for continuous random variable

When $X_1, \ldots, X_n$ are continuous variables, we assume that $(X_1, \ldots, X_n)^T|y = 0 \sim N(\mu_0, \sigma_0^2 I_{k\times k})$ and $(X_1, \ldots, X_n)^T|y = 1 \sim N(\mu_1, \sigma_1^2 I_{k\times k})$. Similar to the discrete case, the parameters can be estimated by their maximum likelihood estimate:

$$\hat{\mu_0} = \frac{\sum_{1}^{n} X_i \mathbb{1}(Y_i = 0)}{\sum_{1}^{n} \mathbb{1}(Y_i = 0)} \tag{10.12}$$

$$\hat{\sigma_0}^2 = \frac{\sum_{1}^{n} (X_i - \hat{\mu_0})^2 \mathbb{1}(Y_i = 0)}{\sum_{1}^{n} \mathbb{1}(Y_i = 0)} \tag{10.13}$$

For an unbiased estimate we would use $\hat{\sigma_0}^2 = \frac{\sum_1^n (X_i - \hat{\mu_0})^2 \mathbb{1}(Y_i=0)}{\sum_1^n \mathbb{1}(Y_i=0)-1}$. 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

$$\begin{aligned} log\frac{P(Y=1|x)}{P(Y=0|x)} &= log\frac{f(X|Y=1)p(Y=1)}{f(X|Y=1)p(Y=1)} \\ &= -\frac{(x-\mu_1)^T(x-\mu_1)}{2\sigma_1^2} + \frac{(x-\mu_0)^T(x-\mu_0)}{2\sigma_0^2} + log\frac{\pi}{1-\pi} \\ &> 0 \end{aligned} \tag{10.14}$$

Here $\pi = P(Y = 1)$. If we further assume $\sigma_1 = \sigma_0 = \sigma$, then we would get a linear decision boundary

$$(\mu_1 - \mu_0)^T(X - \frac{\mu_0+\mu_1}{2}) + \sigma^2 log\frac{\pi}{1-\pi} > 0 \tag{10.15}$$

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.