25

GENERATIVE MODELS

25.1 Naive Bayes classifier (NBC)

25.1.1 Overview

Given observations $(x^{(i)}, y^{(i)}), i = 1, ..., N, x = (x_1, ..., x_K)$ is a multi-dimensional feature, the goal of **Naive Bayes classifier (NBC)** is to estimate $P(Y|X)$. In the previous section of linear classification models, we directly model conditional distribution $P(Y|X)$ by proposing linear function forms for decision functions (e.g., logistic regressions, linear Gaussian discriminant analysis). This type of modes are known as **discriminative models** [1]. On the other hand, $P(Y|X)$ can also be arrived via Bayes rule which says

$$P(Y = y|x) = \frac{P(X = x|Y = y)P(Y = y)}{\sum_{y \in \mathcal{Y}} P(X = x|Y = y)P(Y = y)} \propto P(X = x|Y = y)P(Y = y).$$

Models employing the Bayes rule are known as **generative model**, where the probabilities, $P(X|Y)$ and $P(Y)$ are modeled and estimated. As a comparison, in discriminant models, only $P(Y|X)$ is modeled.

In our following sections, we will go through different types of NBCs, which differ from each other by their assumptions with regards to $P(X|Y)$. They have the name 'naive' because their simplified **conditional independence** assumption, where different components of X are independent conditioning on Y. The conditional independent assumption significantly simplify the estimation of $P(X|Y)$ and enables the algorithm to scale to problems with enormous number of features.

25.1.2 Binomial NBC

In Binomial NBC, the multi-dimensional feature vector $X_1, ..., X_K$ are all binary $\{0, 1\}$ random variables, whose conditional distribution $X_i|Y$ is assumed to be Bernoulli distribution [Definition 12.1.1] with parameters $\theta_{ic}, i = 1, ..., K, c \in \{1, ..., C\}$.

Specifically, we assume

$$P(X_i = x_i|Y = c) = \theta_{ic}^{x_i}(1 - \theta_{ic})^{1-x_i}, x_i \in \{0, 1\},$$

and conditional independence gives

$$P(X_1 = x_1, ..., X_K = x_K|Y = c) = \prod_{i=1}^{K} P(X_i = x_i|Y = c)$$

Via Bayes rule, the probability of $Y = c$ at given observation $X_1 = x_1, ..., X_K = x_K$ is

$$P(Y = c|X_1 = x_1, ..., X_K = x_K) \propto P(Y = c) \prod_{i=1}^{K} P(X_i = x_i|Y = c).$$

To fully evaluate such classification probability, we need to estimate $P(Y = c)$ and $P(X_i = x_i|Y = c)$ respectively. We first start with

Lemma 25.1.1 (MLE for Binomial NBC). *Suppose we have N observations, $\mathcal{D} = \{x^{(i)}, y^{(i)}\}, i = 1, 2, ..., N$. Let $\pi_c \triangleq P(Y = c)$. The log-likelihood function respect to parameter π and θ is given by*

$$L(\pi, \theta) = \sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{K} \sum_{c=1}^{C} \sum_{i:y^{(i)}=c} x_j^{(i)} \log \theta_{jc}^{(1)} + (1 - x_j^{(i)}) \log \theta_{jc}^{(0)}.$$

By maximizing L, we have

- *The estimation of prior probability is*

$$\hat{\pi}_c \triangleq P(Y = c) = \frac{N_c}{N},$$

  where N is the total count and N_c is the count of class c.
- *The estimation of θ_{ic} is*

$$\hat{\theta}_{jc}^{(k)} = \frac{N_{jc}^{(k)}}{N_c}, k = 0, 1$$

Proof. (1) To maximize $\sum_{c=1}^{C} N_c \log \pi_c$ with the constraint $\sum_c \pi_c = 1$, we have

$$\hat{\pi}_c = \frac{N_c}{N}.$$

(2) For each θ_{jc}, we maximize

$$\begin{aligned} &\sum_{i:y^{(i)}=c} x_j^{(i)} \log \theta_{jc}^{(0)} + (1 - x_j^{(i)}) \log \theta_{jc}^{(1)} \\ =&N_{jc} \log \theta_{jc} + (N_c - N_{jc}) \log(1 - \theta_{jc}) \end{aligned}$$

where $N_{jc} = \sum_{i:y^{(i)}=c} x_j^{(i)}, N_c = \sum_{i:y^{(i)}=c} 1$. Clearly, this term will be maximized when

$$\hat{\theta}_{jc}^{(k)} = \frac{N_{jc}^{(k)}}{N_c}, k = 0, 1$$

□

In the testing stage, using estimation from **MLE can give ill-defined results**, as we explain here in a more details. Suppose during the training stage, one θ_{jc} is estimated to be zero. Then for a test example that contains feature j will be predicted to have zero probability belonging to class c.

There are different strategies to remedy this issue. The simplest one is called **add-one-smoothing**, also known as **Laplace Smoothing**, where we modify the The estimation of θ_{ic} is

$$\hat{\theta}_{jc}^{(k)} = \frac{N_{jc}^{(k)} + 1}{N_c + 2}, k = 0, 1.$$

The add-one-smoothing technique is equivalent to introduce a single observation into each dimensionality of x, with a total of $2KC$.

The straight-forward generation is add-λ-smoothing, which gives the estimation of θ_{ic} as

$$\hat{\theta}_{jc}^{(k)} = \frac{N_{jc}^{(k)} + \lambda}{N_c + 2\lambda}$$

Furthermore, we can assume $\theta_{jc} \sim Beta(\alpha, \beta)$, then the MAP estimator [Theorem 25.3.1] for θ_{jc} is

$$\hat{\theta}_{jc} = \frac{N_{jc}^{(k)} + (\alpha - 1)}{N_c + (\alpha - 1) + (\beta - 1)}.$$

25.1.3 Multinomial NBC

In Multinomial NBC, the feature vectors $X_1, ..., X_K$ are all discrete random variables taking values in $\{1, 2, ..., S_i\}, i = 1, .., K$, whose conditional distribution $X_i|Y$ is assumed to be Multivariate Bernoulli distribution with parameters $\theta_{ic} \in \mathbb{R}^{S_i}.i = 1, ..., K, c \in \{1, ..., C\}$.

Specifically, we assume

$$P(X_i = x_i|Y = c) = \theta_{ic}^{(x_i)}, x_i \in \{1, ..., S_i\},$$

and conditional independence giving

$$P(X_1 = x_1, ..., X_K = x_K|Y = c) = \prod_{i=1}^{K} P(X_i = x_i|Y = c)$$

Via Bayes rule, the probability of $Y = c$ at given observation $X_1 = x_1, ..., X_K = x_K$ is

$$P(Y = c|X_1 = x_1, ..., X_K = x_K) \propto P(Y = c) \prod_{i=1}^{K} P(X_i = x_i|Y = c).$$

To fully evaluate such classification probability, we need to estimate $P(Y = c)$ and $P(X_i = x_i|Y = c)$ respectively. We first start with

Lemma 25.1.2 (MLE for Multinomial NBC). *Suppose we have N observations, $\mathcal{D} = \{x^{(i)}, y^{(i)}\}, i = 1, 2, ..., N$. Let $\pi_c \triangleq P(Y = c)$. The log-likelihood function respect to parameter π and θ is given by*

$$L(\pi, \theta) = \sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{K} \sum_{c=1}^{C} \sum_{i:y^{(i)}=c} x_j^{(i)} \log \theta_{jc}^{(x_j^{(i)})}$$

By maximizing L, we have

- *The estimation of prior probability*

$$\hat{\pi}_c \triangleq P(Y = c) = \frac{N_c}{N}$$

, where N is the total count and N_c is the count of class c.

- *The estimation of vector θ_{ic} is*

$$\hat{\theta}_{jc} = (\frac{N_{jc}^{(1)}}{N_c}, \frac{N_{jc}^{(2)}}{N_c}, \cdots, \frac{N_{jc}^{(S_j)}}{N_c})$$

where $N_{jc}^{(k)} \triangleq \sum_{i=1}^{N} \mathbb{I}(x_j^{(i)} = k, y^{(i)} = c)$ is the number that feature $x_j = k$ occurs in class c.

Proof. (1) To maximize $\sum_{c=1}^{C} N_c \log \pi_c$ with the constraint $\sum_c \pi_c = 1$, we have

$$\hat{\pi}_c = \frac{N_c}{N}.$$

(2) For each θ_{jc}^k, we maximize

$$\sum_{i:y^{(i)}=c, x_j^{(k)}=c} x_j^{(i)} \log \theta_{jc}^{(k)},$$

with the constrict $\sum_k \theta_{jc}^{(k)} = 1$. We have

$$\hat{\theta}_{jc} = (\frac{N_{jc}^{(1)}}{N_c}, \frac{N_{jc}^{(2)}}{N_c}, \cdots, \frac{N_{jc}^{(S_j)}}{N_c})$$

where $N_{jc}^{(k)} \triangleq \sum_{i=1}^{N} \mathbb{I}(x_j^{(i)} = k, y^{(i)} = c)$ is the number that feature $x_j = k$ occurs in class c. □

To ensure the stability of MLE, we have the add-one-smoothing technique, which gives the estimation of θ_{ic} is

$$\hat{\theta}_{jc}^{(k)} = \frac{N_{jc}^{(k)} + 1}{N_c + S_j},$$

and the add-λ-smoothing, which gives the estimation of θ_{ic} as

$$\hat{\theta}_{jc}^{(k)} = \frac{N_{jc}^{(k)} + \lambda}{N_c + S_j \lambda}$$

Similarly, if we assume $\theta_{jc} \sim Dir(\alpha_1, ..., \alpha_S)$, then the MAP estimator [Theorem 25.3.2] for θ_{jc} is

$$\hat{\theta}_{jc}^{(k)} = \frac{N_{jc}^{(k)} + (\alpha_k - 1)}{N_c + \sum_{k=1}^{S} (\alpha_k - 1)}$$

Finally, the algorithm is showed in algorithm 36.

Algorithm 36: Multinomial Naive Bayes classification

Input: Training data set $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$, where each sample $x_i = x_i^{(1)}, ..., x_i^{(m)})$ has m feature. Assume feature $x^{(j)}$ can take values $x^{(j)} \in \{a_{j1}, ..., a_{jS_j}\}$. Label $y_i \in \{c_1, ..., c_K\}$. Test data x.

1 Calculate prior and conditional probabilities:

$$P(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k)}{N}, k = 1, 2, \cdots, K$$

$$P\left(X^{(j)} = a_{jl} | Y = c_k\right) = \frac{\sum_{i=1}^{N} I\left(x_i^{(j)} = a_{jl}, y_i = c_k\right)}{\sum_{i=1}^{N} I(y_i = c_k)}$$

$$j = 1, 2, \cdots, n; \quad l = 1, 2, \cdots, S_j; \quad k = 1, 2, \cdots, K$$

2 For the given test data $x = \left(x^{(1)}, x^{(2)}, \cdots, x^{(n)}\right)^T$, compute

$$P(Y = c_k) \prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)} | Y = c_k\right), \quad k = 1, 2, \cdots, K.$$

3 The label for test data x is given by

$$y = \arg\max_{\varepsilon_k} P(Y = c_k) \prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)} | Y = c_k\right).$$

Output: the label for test data x.

25.1.4 Gaussian NBC

In Gaussian NBC, the feature vectors $X_1, ..., X_K$ are random variables with support in $\mathbb{R}$, whose conditional distribution $X_i|Y$ is assumed to be Gaussian distribution with parameters $(\mu_{ic}, \sigma_{ic}).i = 1, ..., K, c \in \{1, ..., C\}$.

Specifically, we assume

$$P(X_i = x_i | Y = c) = \frac{1}{\sqrt{2\pi\sigma_{ic}^2}} \exp(-\frac{(x_i - \mu_{ic})^2}{2\sigma_{ic}^2}),$$

and conditional independence giving

$$P(X_1 = x_1, ..., X_K = x_K | Y = c) = \prod_{i=1}^{K} P(X_i = x_i | Y = c)$$

Via Bayes rule, the probability of $Y = c$ given observation $X_1 = x_1, ..., X_K = x_K$ is

$$P(Y = c | X_1 = x_1, ..., X_K = x_K) \propto P(Y = c) \prod_{i=1}^{K} P(X_i = x_i | Y = c).$$

To fully evaluate such classification probability, we need to estimate $P(Y = c)$ and $P(X_i = x_i | Y = c)$ respectively. We first start with

Lemma 25.1.3 (MLE for Gaussian NBC). *Suppose we have N observations, $\mathcal{D} = \{x^{(i)}, y^{(i)}\}, i = 1, 2, ..., N$. Let $\pi_c \triangleq P(Y = c)$. The log-likelihood function respect to parameter π and (μ, σ) are given by*

$$L(\pi, \theta) = \sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{K} \sum_{c=1}^{C} \sum_{i:y^{(i)}=c} (-\frac{1}{2} \log(2\pi) - \frac{\sigma_{jc}}{2} - \frac{(x_i - \mu_{ic})^2}{2\sigma_{ic}^2})$$

By maximizing L, we have

- *The estimation of prior probability*

$$\hat{\pi}_c \triangleq P(Y = c) = \frac{N_c}{N}$$

  , where N is the total count and N_c is the count of class c.
- *The estimation of μ_{ic} is*

$$\hat{\mu}_{jc} = \frac{\sum_{i:y^{(i)}=c} (x_j^{(i)}}{N_c}.$$

- *The estimation of σ_{ic} is*

$$\hat{\sigma}_{jc} = \frac{\sum_{i:y^{(i)}=c} (x_j^{(i)} - \hat{\mu}_{jc})}{N_c}.$$

Proof. See MLE for Gaussian distributions. □

Remark 25.1.1 (connections to linear and quadratic Gaussian discriminant models). Gaussian NBC employs similar Gaussian distribution conditioning on the label as the linear and quadratic Gaussian discriminant model in subsection 24.2.1. Some of key difference and implications are

- Consider two classes, if $\sigma_{jc} = \sigma$, then the decision boundary is linear; otherwise, it is curved.
- Gaussian NBC can almost be seen as linear and Quadratic Gaussian discriminator without correlation among features.

25.1.5 Discussion

Thanks to the conditional independence assumption, NBC becomes a simple and easy use classifier in practice. It can be successfully trained on small data set, and it is good for text classification. However, the relationship among the features will not be modeled due to the conditional independence assumption.

The NBC can be extended to different directions:

- By specifying other conditional distributions other than multinomial or Gaussian, we can introduce different types of NBCs.
- By specifying some features to follow discrete distributions some to follow continuous distributions, we have a mixed NBC.

25.2 Application

25.2.1 Classifying documents using bag of words

Document classification is the problem of classifying text documents into different categories. For example, classifying news articles into different categories like politics, sports, etc.

One simple approach is to represent each document as a binary vector, which records whether each word is present or not, so $x_{ij} = 1$ if only if word j occurs in document i, otherwise $x_{ij} = 0$. We can then use the following class conditional density:

$$\begin{aligned} p(x_i|y_i = c, \theta) &= \prod_{j=1}^{D} \text{Ber}(x_{ij}|\theta_{jc}) \\ &= \prod_{j=1}^{D} \theta_{jc}^{x_{ij}} (1 - \theta_{jc})^{1-x_{ij}} \end{aligned}$$

This is called the **Bernoulli product model**, or the **binary independence model**.

25.2.2 Credit card fraud prediction

Now we apply the Gaussian Naive Bayes classifier to fraud detection problem. We have transaction data that contains a number of features and are labeld as 'genuine' or "fraud'. The feature s includes transition amount, time, and other confidential features. In the first stage feature screening, we plot the histogram of each continuous feature with respect to label.

It is clear that there are some features' distribution/histogram conditioned on the label are similar; thus they will highly not contribute to classification. What is more, including these feature will make the model are complicated and likely lead to overfitting.

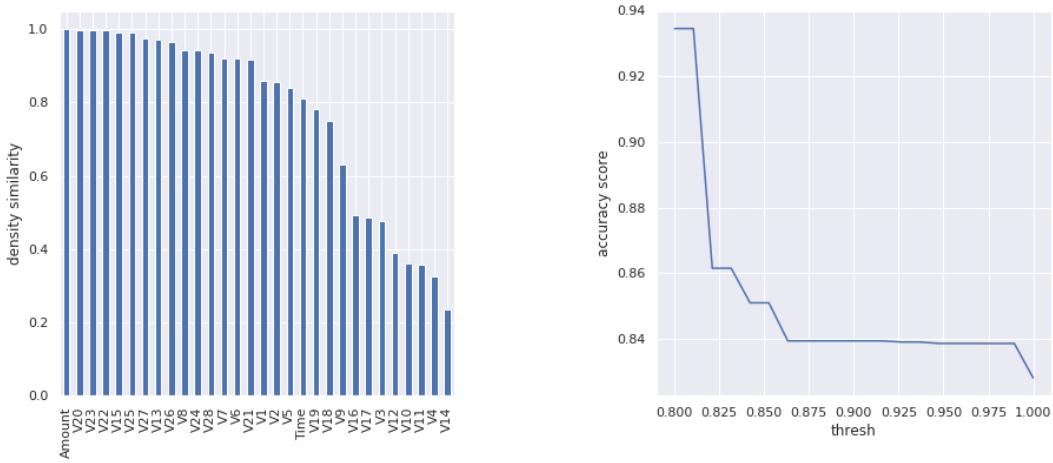
We characterize the similarity in the conditional distribution via a quantity, density similarity, defined as

$$s(d_i) = d_i^{fraud} \cdot d_j^{genuine},$$

where $d_i^{fraud}, d_i^{genuine}$ is the conditional probability vector for feature i with label **fraud** and **genuine**, respectively. The results are showed in Figure 25.2.1 (a). We can see that some features can have density similarity as large as 1.

We use different thresholds to filter out features with high density similarity in order to build simple models that can generalize well. Figure 25.2.1 (b) shows that a threshold of 0.8 yields the best prediction performance on the testing set.

Figure 25.2.1: Histogram of the features group by class label (fraud vs. genuine).



(a) Decision boundary of a two-class classification problem. The conditional distributions are represented by the contours of the Gaussian distribution.

(b) Predictive performance of model vs. density similarity threshold.

Figure 25.2.2: Feature density similarity and predictive performance of model

25.3 Supporting mathematical results

25.3.1 Beta-binomial model

25.3.1.1 *The model*

Definition 25.3.1 (Beta-binomial model). *A Beta-binomial model with pre-specified parameter n has parameter a and b, and can be represented by the following Hierachical model*

$$Y|\theta \sim Binom(n,\theta), \theta \sim Beta(a,b).$$

In particular, the density functions for $Y|\theta$ and θ given by

$$Pr_{Y|\theta}(Y=y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{(n-y)}, y=0,1,...,n;$$

$$f(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}, a>0, b>0;$$

where

$$B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta.$$

Lemma 25.3.1 (density functions for Beta-binomial model). *Let (Y,θ) follow a Beta-binomial model with parameter (a,b), we have*

- *the joint density function is given by*

$$Pr_{Y,\theta}(Y=y,\theta) = \binom{n}{y}\theta^y(1-\theta)^{(n-y)}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}, y=0,1,...,n;$$

- *the marginal density function for y is given by*

$$Pr_Y(Y=y) = \binom{n}{y}\frac{B(y+a,n-y+b)}{B(a,b)}.$$

where

$$B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta.$$

Proof. (1) the joint density function is given by

$$Pr_{Y,\theta}(Y=y,\theta) = \binom{n}{y}\theta^y(1-\theta)^{(n-y)}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}, y=0,1,...,n;$$

(2)the marginal density function for y is given by

$$Pr_Y(Y=y) = \binom{n}{y} \frac{B(y+a, n-y+b)}{B(a,b)}.$$

□

Lemma 25.3.2 (basic statistical properties of Beta-binomial model). *Let (Y, θ) follow a Beta-binomial model with parameter (a, b), we have*

- $$E[Y|\theta] = n\theta, E[\theta] = \frac{a}{a+b}, E[Y] = n\frac{a}{a+b}.$$
- $$Var[Y|\theta] = n\theta(1-\theta), Var[Y] = n\mu_\theta(1-\mu_\theta)(1+(n-1)\frac{1}{a+b+1}),$$

  where $\mu_\theta = a/(a+b)$.

Proof. (1)

$$E[Y] = E[E[Y|\theta]] = E[n\theta] = nE[\theta] = n\frac{a}{a+b}.$$

(2) Note that

$$\begin{aligned} Var[Y] &= E[Var[Y|\theta]] + Var[E[Y|\theta]] \\ &= E[n\theta(1-\theta)] + Var[n\theta] \\ &= E[n\theta] - nE[\theta^2] + n^2 Var[\theta] \\ &= n(E[\theta] - E[\theta]^2 + nVar[\theta])] \end{aligned}$$

where we use the property of conditional variance [Theorem 11.5.1). then we use the following facts [Lemma 12.1.28),

- $$E[\theta] = \frac{a}{a+b}.$$
- $$E[\theta^2] = \frac{a(a+1)}{(a+b)(a+b+1)}$$
- $$Var[\theta] = \frac{ab}{(a+b)^2(a+b+1)}.$$

□

25.3.1.2 *Parameter inference*

Theorem 25.3.1 (posterior distribution and estimation of θ). *[2, p. 75] Given an iid random experiment data set $\mathcal{D}$ consisting of N_1 results of labeling 1 and N_0 results of the labeling 0. Consider a Beta-binomial model with parameter (a, b). It follows that*

- *The posterior distribution of parameter θ given the data $\mathcal{D}$ is*

$$\theta|\mathcal{D} \sim Beta(N_1 + a, N_0 + b),$$

that is

$$f_{\theta|\mathcal{D}} = \frac{1}{B(a + N_1, b + N_0)} \theta^{a+N_1-1}(1-\theta)^{b+N_0-1}.$$

- *The mean and variance of the posterior θ is given by*

$$E[\theta|\mathcal{D}] = \frac{a + N_1}{a + b + N_1 + N_0},$$

$$Var[\theta|\mathcal{D}] = \frac{(a + N_1)(b + N_0)}{(a + N_1 + b + N_0)^2(a + b + 1 + N_1 + N_0)},$$

- *The maximum likelihood function is given by*

$$L(\mathcal{D}|\theta) \propto \theta^{N_1}(1-\theta)^{N_0}$$

and the maximum likelihood estimation of θ is given by

$$\theta_{MLE} = \frac{N_1}{N_1 + N_0}.$$

- *The maximum a posterior estimation of θ is given by*

$$\theta_{MAP} = \frac{a + N_1 - 1}{a + b + N_1 + N_0 - 2},$$

when $a = b = 1$, $\theta_{MAP} = \theta_{MLE}$.

Proof. (1) Note that

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta) \\ &\propto \theta^{N_1}(1-\theta)^{N_0}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \end{aligned}$$

After normalization, we can show $\theta|\mathcal{D}$ follows $Beta(N_1 + a, N_0 + b)$. (2)Lemma 12.1.28. (3)(4)Lemma 12.1.28. □

Lemma 25.3.3 (sequential learning property). *Suppose we have two data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ with sufficient statistics N_1', N_0' and N_1'', N_0''. And denote $N_1 = N_1' + N_1''$ and $N_0 = N_0' + N_0''$. In batch mode, we have*

$$p(\theta|\mathcal{D}', \mathcal{D}'') \propto Beta(\theta|N_1 + a, N_0 + b).$$

and in sequential mode, we also have

$$p(\theta|\mathcal{D}', \mathcal{D}'') \propto Beta(\theta|N_1 + a, N_0 + b).$$

Proof. Use Theorem 25.3.1, we have

$$\begin{aligned} p(\theta|\mathcal{D}', \mathcal{D}'') &\propto p(\mathcal{D}''|\theta)p(\theta|\mathcal{D}') \\ &\propto Bin(N_1''|\theta, N_1'' + N_0'')Beta(\theta|N_1' + a, N_0' + b) \\ &\propto Bin(\theta|N_1' + N_2'' + a, N_0' + N_0'' + b) \\ = Beta(N_1 + a, N_0 + b) \end{aligned}$$

where we used the conditional Bayesian law [Theorem 11.4.1). □

25.3.2 Dirichlet-multinomial model

25.3.2.1 *The model*

Definition 25.3.2 (Dirichlet-multinomial model). *A Dirichlet-multinomial model with pre-specified parameter N(total counts) and K(number of classes) has parameters $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K)$, and can be represented by the following Hierachical model*

$$Y|\theta \sim Mulnom(n, \theta), \theta \sim Dir(\theta).$$

In particular, the density functions for $Y|\theta$ and θ given by

$$Pr_{Y|\theta}(Y = (n_1, n_2, ..., n_K)|\theta) = \binom{N}{n_1 \cdots n_K} \theta^y (1-\theta)^{(n-y)}, y = 0, 1, ..., n;$$

$$f(\theta|\alpha) = Dir(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

with support $\theta \in \{\theta_k : 0 \leq \theta_k \leq 1, \sum_k x_\theta = 1, \forall k = 1, 2..., K\}$, *and* $B(a)$ *is a normalization constant given as*

$$B(a) = \frac{\prod_{k=1}^{K} \Gamma(a_k)}{\Gamma(\sum_k a_k)},$$

where $\Gamma(\cdot)$ *is the Gamma function.*

Example 25.3.1. Consider tossing a tie with 6 faces. The count on the each distinct faces from a total number count N can be modeled by the Dirichlet-multinomial model.

Lemma 25.3.4 (density functions for Dirichlet-Multinomial model). *Let* (Y, θ) *follow a Beta-binomial model with parameter* (a, b)*, we have*

- *the joint density function for* (Y, θ) *with parameter* α *is given by*

$$\begin{aligned} Pr_{Y,\theta}(Y = (y_1, y_2, ..., y_K), \theta) &= \binom{N}{N_1 \cdots N_K} \prod_{k=1}^{K} \theta_k^{\alpha_k + y_k - 1} \frac{1}{B(\alpha)} \\ &= \prod_{k=1}^{K} \theta_k^{\alpha_k + y_k - 1} \frac{1}{B(\alpha')} \end{aligned}$$

  where $\alpha' = \alpha + (y_1, y_2, ..., y_K)$.
- *the marginal density function for* Y_i *with parameter* α *is given by*

$$Pr_Y(Y_i = y) = \binom{N}{y_i} \frac{B(y + \theta_i, N - y + (\alpha_0 - \alpha_i))}{B(\alpha_i, \alpha_0 - \alpha_i)}.$$

where

$$B(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1} dx.$$

Proof. (1) the joint density function is given by

$$Pr_{Y,\theta}(Y, \theta) = Pr(Y|\theta) f(\theta).$$

(2)the marginal density function for y is given by

$$Pr_Y(Y = y) = \binom{n}{y} \frac{B(y + a, n - y + b)}{B(a, b)}.$$

□

Remark 25.3.1 (connections between Dirichlet-multinomial model and Beta-binomial model). In the Dirichlet-multinomial model, we have multiple classes. However, if we are only interested in the counts in one particular class(i.e., the marginal distribution of one types of result), we can convert it to an equivalent Beta-binomial model.

Lemma 25.3.5 (basic statistical properties of Dirichlet-Multinomial model). *Let* $(Y, \alpha), Y = (Y_1, Y_2, ..., Y_K), \alpha = (\alpha_1, ..., \alpha_K)$ *follow a Dirichlet-Multinomial model with parameter* α*. Further denote* $\alpha_0 = \sum_{i=1}^{K} \alpha_i$*. Then we have*

- $$E[Y_i|\alpha] = N\alpha_i, E[\alpha_i] = \frac{\alpha_i}{\alpha_0}, E[Y] = n\frac{a}{a+b}.$$
- $$Var[Y|\theta] = n\theta(1-\theta), Var[Y] = n\mu_\theta(1-\mu_\theta)(1+(n-1)\frac{1}{a+b+1}),$$

  where $\mu_\theta = a/(a+b)$.

Proof. (1)

$$E[Y_i] = E[E[Y_i|\theta_i]] = E[N\theta_i] = NE[\theta_i] = N\frac{\alpha_i}{\alpha_0}.$$

(2) Note that

$$\begin{aligned} Var[Y] &= E[Var[Y|\theta]] + Var[E[Y|\theta]] \\ &= E[n\theta(1-\theta)] + Var[n\theta] \\ &= E[n\theta] - nE[\theta^2] + n^2 Var[\theta] \\ &= n(E[\theta] - E[\theta]^2 + nVar[\theta])] \end{aligned}$$

where we use the property of conditional variance [Theorem 11.5.1). then we use the following facts [Lemma 12.1.28),

- $$E[\theta] = \frac{a}{a+b}.$$
- $$E[\theta^2] = \frac{a(a+1)}{(a+b)(a+b+1)}$$
- $$Var[\theta] = \frac{ab}{(a+b)^2(a+b+1)}.$$

□

Remark 25.3.2. Note that once θ_i is known, we can have the marginal distribution of Y_i, even without knowing other θ_{-i}.

25.3.2.2 *Parameter inference*

Theorem 25.3.2 (parameter inference).

- *(likelihood)Suppose we observe N dice rolls, $\mathcal{D} = \{x_1, x_2, \cdots, x_N\}$, where $x_i \in \{1, 2, \cdots, K\}$. The likelihood has the form*

$$p(\mathcal{D}|\theta) = \binom{N}{N_1 \cdots N_k} \prod_{k=1}^{K} \theta_k^{N_k} \quad \text{where } N_k = \sum_{i=1}^{N} \mathbb{I}(y_i = k)$$

- *(Prior)*

$$Dir(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \mathbb{I}(\theta \in S_K)$$

- *(Posterior)*

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta) \\ &\propto \prod_{k=1}^{K} \theta_k^{N_k} \theta_k^{\alpha_k - 1} = \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \\ &= Dir(\theta|\alpha_1 + N_1, \cdots, \alpha_K + N_K) \end{aligned}$$

From Equation Lemma 25.3.5, the MAP estimate is given by

$$\hat{\theta}_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}$$

If we use a uniform prior, $\alpha_k = 1$, we recover the MLE:

$$\hat{\theta}_k = \frac{N_k}{N}$$

BIBLIOGRAPHY

1. Jordan, A. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems* **14,** 841 (2002).
2. Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).