Lecture Notes 3: Multinomial Distribution and Basic Kernel

Professor: Zhihua Zhang *Scribe:Ruotian Luo, Yuxi Zhang, Cheng Chen*

2 The Multivariate Normal Distributions

Lemma 2.2. ***(homework)****Prove* $\mathbf{\Sigma}_{11.2} = \mathbf{\Theta}_{11}^{-1}$ *and* $\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} = \mathbf{\Theta}_{12}\mathbf{\Theta}_{22}^{-1}$.

Theorem 2.4. *Assume* $X = [X_1, \ldots X_m]^T \sim \mathcal{N}_m(0, \mathbf{\Sigma})$, $\mathbf{\Sigma} = (\sigma_{ij})$ *and* $\mathbf{\Theta} = \mathbf{\Sigma}^{-1} = (\theta_{ij})$. *Then, we have*

1. $X_i \perp\!\!\!\perp X_j$ *iff* $\sigma_{ij} = 0$
2. $X_i \perp\!\!\!\perp X_j | X_{\{1\ldots m\}\setminus\{i,j\}}$ *iff* $\theta_{ij} = 0$
3. $X_i | X_{\{1\ldots m\}\setminus\{i\}} \sim \mathcal{N}\left(\sum_{j\neq i} \frac{\theta_{ij}}{\theta_{ii}} X_j, \theta_{ii}^{-1}\right)$

Proof. Without loss of generality, we assume that $i = 1$ and $j = 2$.

1. Let $\mathbf{y}_1 = [X_1, X_2]^T$ and $\mathbf{y}_2 = [X_3, \ldots, X_m]^T$. We have
$$\begin{aligned}\mathbf{y}_1 &\sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_{11}) \\ \mathbf{y}_1|\mathbf{y}_2 &\sim \mathcal{N}(\boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2), \mathbf{\Sigma}_{11.2}).\end{aligned}$$
where the subscript $1, 2$ and 11.2 are with respect to $\mathbf{y}_1$ and $\mathbf{y}_2$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$. Then we have $X_1 \perp\!\!\!\perp X_2$ iff $\sigma_{12} = 0$.

2. According to lemma 2.2, we have
$$\mathbf{\Sigma}_{11.2} = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix}^{-1} = \frac{1}{\theta_{11}\theta_{22} - \theta_{12}\theta_{21}} \begin{bmatrix} \theta_{22} & -\theta_{21} \\ -\theta_{12} & \theta_{11} \end{bmatrix}.$$
Hence we have $X_1 \perp\!\!\!\perp X_2 | X_{\{1\ldots m\}\setminus\{1,2\}}$ iff $\theta_{12} = \theta_{21} = 0$

3. Let $\mathbf{z}_1 = X_1$ and $\mathbf{z}_2 = [X_2, \ldots, X_m]^T$. We have
$$\mathbf{z}_1|\mathbf{z}_2 \sim \mathcal{N}(\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{z}_2, \mathbf{\Sigma}_{11.2}).$$
where the subscript $1, 2$ and 11.2 are with respect to $\mathbf{z}_1$ and $\mathbf{z}_2$. According to lemma 2.2,
$$\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{z}_2 = \mathbf{\Theta}_{11}^{-1}\mathbf{\Theta}_{21}\mathbf{z}_2 = \sum_{j\neq 1} \frac{\theta_{1j}}{\theta_{11}} X_j$$
Then we have $X_i | X_{\{1\ldots m\}\setminus\{i\}} \sim \mathcal{N}\left(\sum_{j\neq i} \frac{\theta_{ij}}{\theta_{ii}} X_j, \theta_{ii}^{-1}\right)$

□

3 Multinomial Distribution

Theorem 3.1 (Multinomial Theorem)**.** *Let k and n be positive integers. Let $\mathscr{A}$ be the set of vector $\mathbf{x} = (x_1, \cdots, x_k)^T$,such that each x_i is a nonnegative integer and $\sum_{i=1}^{k} x_i = n$. Then, for any real number $p_1, \cdots, p_k$*

$$(p_1 + \cdots + p_k)^n = \sum_{\mathbf{x}\in\mathscr{A}} \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

Definition 3.1 (Multinomial Distribution)**.** *We say $X = (X_1, \cdots, X_k)^T$ has a multinomial distribution of dimension k-1 with parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^T$ and n if p.m.f. is*

$$\mathcal{M}_{k-1}(X = \mathbf{x}|\boldsymbol{\theta}, n) = \frac{n!}{\prod\limits_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i}$$

in which, $0 \le \theta_j \le 1$, $\sum\limits_i \theta_i = 1$, $n = 1, 2, \cdots$, $\mathbf{x} = (x_1, \ldots, x_k), x_i = 0, 1, 2, \ldots, n, \sum\limits_{i=1}^{k} x_i = n$.

Proposition 3.1. *Basic properties of Multinomial Distribution*

- $\mathbb{E}(x_i) = n\theta_i$
- $\mathrm{Var}(x_i) = n\theta_i(1 - \theta_i)$
- $\mathrm{Cov}(x_i, x_j) = -n\theta_i\theta_j$
- $\boldsymbol{\Sigma} = Var(\mathbf{x}) = \mathrm{Cov}(\mathbf{x}) = n(\mathrm{diag}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}^T)$
- $\boldsymbol{\Sigma}\mathbf{1}_k = 0$

We can prove that Σ is a **PSD** matrix.

$$\frac{1}{n}\mathbf{a}^T\Sigma\mathbf{a} = \frac{1}{n}[\sum_{i=1}^{k} a_i^2\theta_i - (\sum_{i=1}^{k} a_i\theta_i)^2]$$

Because quadratic function is convex function, thus we have

$$(\sum_{i=1}^{k} a_i\theta_i)^2 \le \sum_{i=1}^{k} a_i^2\theta_i$$

So $\mathbf{a}^T\Sigma\mathbf{a} \succeq= 0$, i.e. $\Sigma \succeq 0$

Theorem 3.2. *The marginal distribution of $X^{(m)} = (X_1, \cdots, X_m)^T$,$m < k$, is multinomial*

$$\mathcal{M}_{m-1}(\mathbf{x}^{(m)}|(\theta_1 \cdots \theta_m)^T, n)$$

also can be written as

$$\mathcal{M}_m((\mathbf{x}^{(m)}, n - \sum_{i=1}^{m} x_i)|(\theta_1 \cdots \theta_m, 1 - \sum_{i=1}^{m} \theta_i)^T, n)$$

Proof. Assume that $\mathcal{B} = \left\{ (x_{m+1}, \ldots, x_k) : \sum\limits_{i=m+1}^{k} x_i = n - \sum\limits_{i=1}^{m} x_i \right\}$

$$\begin{aligned}
& f(x_1, \ldots, x_m) \\
= & \sum_{(x_{m+1},\cdots,x_k)\in\mathcal{B}} \frac{n!}{x_1! \cdots x_k!} \theta_1^{x_1} \cdots \theta_k^{x_k} \\
= & \frac{n!\theta_1^{x_1} \cdots \theta_m^{x_m}}{x_1! \cdots x_m!} \sum_{(x_{m+1},\cdots,x_k)\in\mathcal{B}} \frac{\theta_{m+1}^{x_{m+1}} \cdots \theta_k^{x_k}}{x_{m+1}! \cdots x_k!} \\
= & \frac{n!\theta_1^{x_1} \cdots \theta_m^{x_m}}{x_1! \cdots x_m!} \frac{\left(1 - \sum\limits_{i=1}^{m} \theta_i\right)^{n - \sum\limits_{i=1}^{m} x_i}}{\left(n - \sum\limits_{i=1}^{m} x_i\right)!} \\
= & \frac{n!}{x_1! \cdots x_m! \left(n - \sum\limits_{i=1}^{m} x_i\right)!} \theta_1^{x_1} \cdots \theta_m^{x_m} \left(1 - \sum_{i=1}^{m} \theta_i\right)^{n - \sum\limits_{i=1}^{m} x_i}
\end{aligned}$$

□

Example 3.1. *The one dimension marginal distribution of the multinomial distribution is binomial distribution:*

$$f_{X_k}(x_k) = \frac{n!}{x_k!(n - x_k)!} \theta_k^{x_k} (1 - \theta_k)^{n - x_k}$$

Theorem 3.3. *The conditional distribution of $X^{(m)}$ given the remaining x_i's is also multinomial:*

$$f(x_1, \ldots, x_m | x_{m+1}, \ldots, x_k) = \mathcal{M}_{m-1} \left(\mathbf{x}^{(m)} | \left(\frac{\theta_1}{\sum\limits_{i=1}^{m} \theta_i}, \ldots, \frac{\theta_m}{\sum\limits_{i=1}^{m} \theta_i} \right)^T, \sum_{i=1}^{m} x_i \right)$$

.

Proof.

$$\begin{aligned}
& f(x_1,\cdots,x_m|x_{m+1},\ldots,x_k) \\
= & \frac{f(x_1,\cdots,x_k)}{f(x_{m+1},\cdots,x_k)} \\
= & \frac{\frac{n!\theta_1^{x_1}\cdots\theta_k^{x_k}}{x_1!\ldots x_k!}}{\frac{n!\theta_{m+1}^{x_{m+1}}\cdots\theta_k^{x_k}\left(1-\sum\limits_{i=m+1}^{k}\theta_i\right)^{n-\sum\limits_{i=m+1}^{k}x_i}}{x_{m+1}!\cdots x_k!\left(n-\sum\limits_{i=m+1}^{k}x_i\right)!}} \\
= & \frac{\theta_1^{x_1}\cdots\theta_m^{x_m}\left(\sum\limits_{i=1}^{m}x_i\right)!}{x_1!\cdots x_m!\left(\sum\limits_{i=1}^{m}\theta_i\right)^{\sum\limits_{i=1}^{m}x_i}} \\
= & \frac{\left(\sum\limits_{i=1}^{m}x_i\right)!}{x_1!\ldots x_m!}\left(\frac{\theta_1}{\sum\limits_{i=1}^{m}\theta_i}\right)^{x_1}\cdots\left(\frac{\theta_m}{\sum\limits_{i=1}^{m}\theta_i}\right)^{x_m}
\end{aligned}$$

□

Example 3.2. $f(x_1,\cdots,x_{k-1}|x_k) = \frac{(n-x_k)!}{x_1!\ldots x_{k-1}!}\left(\frac{\theta_1}{1-\theta_k}\right)^{x_1}\ldots\left(\frac{\theta_{k-1}}{1-\theta_k}\right)^{x_{k-1}}$

Theorem 3.4. *Assume* $X = (X_1,\ldots,X_k)^T \sim \mathcal{M}_{k-1}(\mathbf{x}|\boldsymbol{\theta},n)$*,* $\hat{Y} = (Y_1,\cdots,Y_t)^T$*,* $t < k$*,* $\boldsymbol{\phi} = (\phi_1,\cdots,\phi_t)$*, and* $Y_1 = X_1+\cdots+X_{i_1}, Y_2 = X_{i_1+1}+\cdots+X_{i_2},\cdots$*,* $\phi_1 = \theta_1+\cdots+\theta_{i_1}, \phi_2 = \theta_{i_1+1}+\cdots+\theta_{i_2},\cdots$*. Then*

$$\hat{\mathbf{Y}} \sim \mathcal{M}_{t-1}(\mathbf{y}|\boldsymbol{\phi},n)$$

Theorem 3.5. *If Z is the sum of m independent random vector having multinomial density parameters* $(\boldsymbol{\theta},n_i)$*,*$i = 1,\cdots,m$*, then*

$$Z \sim \mathcal{M}(\hat{\mathbf{z}}|\boldsymbol{\theta},n_1+\cdots+n_m)$$

We could express a multi-label classification problem as a multinomial distribution $\mathcal{M}(\mathbf{x}|\boldsymbol{\theta},1)$(because we express a class as a vector which there is only one item larger than zero and others equals to zero).

We could express a multi-classfication problem as a multinomial distribution. We set all the m random vector to $(\boldsymbol{\theta},1)$(because we express a class as a vector which there is only one item larger than 0 and others equals to zero), thus $\mathbf{Z} \sim \mathcal{M}(\hat{\mathbf{z}}|\boldsymbol{\theta},m)$.
To estimate distribution, we can minimize $-\log\sum\mathcal{M}(\mathbf{y}_k|h_\theta(x_k),1)$.
Commonly, we can not do least squares estimate on multi-label classification problem. However, when the number of class is large, we can use normal distribution approximate the multinomial distribution.

4 Reproducing Kernels

Theorem 4.1 (Cover's Theorem)**.** *A complex pattern-classification problem cast in a high-dimensional space nonlinearly is more likely to be linearly separable than in a low-dimensional space.*

Before we introduce kernel, we first find a way to reflect a vector to a higher dimension. For $\mathbf{x} \in \mathbb{R}^p$, and $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^r, r > p$, we called $\mathbf{x}$ the **input**, and the corresponding $\boldsymbol{\phi}(\mathbf{x})$ **feature**.
Actually we do not need to calculate $\boldsymbol{\phi}(\mathbf{x})$ explicitly, we can instead calculate the inner product of $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\phi}(\mathbf{x}_j)$ and we introduce kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle$.

Definition 4.1. *Let $\mathcal{X} \subset \mathbb{R}^p$ be a nonempty set. A function $\mathbf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel.*

Here we give definition to some specific kinds of kernel function.

Definition 4.2. *A function $\mathbf{K}$ is called symmetric kernel if $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{K}(\mathbf{x}_j, \mathbf{x}_i)$ for any $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$*

Definition 4.3. *A kernel $\mathbf{K}$ is positive definite if and only if*

$$\sum_{j,k=1}^{n} \alpha_j \alpha_k \mathbf{K}(\mathbf{x}_j, \mathbf{x}_k) \geq 0$$

for all $n \in \mathbb{N}, \{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subseteq \mathcal{X}$ and $\{\alpha_1, \cdots, \alpha_n\} \subseteq \mathbb{R}$

Definition 4.4. *We call the symmetric kernel $\mathbf{K}$ is conditionally positive definite if and only if*

$$\sum_{j,k=1}^{n} \alpha_j \alpha_k \mathbf{K}(\mathbf{x}_j, \mathbf{x}_k) \geq 0$$

for all $n \geq 2$, $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subseteq \mathcal{X}$ and $\{\alpha_1, \cdots, \alpha_n\} \subseteq \mathbb{R}$ such that $\sum_{i=1}^{n} \alpha_i = 0$.

Definition 4.5. *If $\mathbf{K}$ is conditionally positive definite, then we call that $-\mathbf{K}$ is negative definite.*

Proposition 4.1. *For any kernel $\mathbf{K}$, $\mathbf{K}(\mathbf{x}, \mathbf{x}) \geq 0$*

Example 4.1. *Kernel $\mathbf{K}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$ is negative definite.*

Proof. If $\sum_{i=1}^{n} \alpha_i = 0$, then

$$\begin{aligned}
&\sum_{j,k=1}^{n} \alpha_j \alpha_k \|\mathbf{x}_j - \mathbf{x}_k\|_2^2 \\
=&\sum_{j,k=1}^{n} \alpha_j \alpha_k (\mathbf{x}_j^T \mathbf{x}_j - 2\mathbf{x}_j^T \mathbf{x}_k + \mathbf{x}_k^T \mathbf{x}_k) \\
=&-2\|\sum_{j=1}^{n} \alpha_j \mathbf{x}_j\|_2^2 \leq 0
\end{aligned}$$

□