STA414: Statistical Methods for Machine Learning II Lecture Notes

Yiliu Cao

May 3, 2024

Contents

1 Probabilistic models

1.1 An overview of probabilistic models

We have the random vector $X = (X_1, X_2, \cdots X_n)$, and we want to compute the relationship between each random variable. The joint distribution is $p(x) = p(x_1, \cdots, x_d)$. Denote the input data $\mathbf{x}$ (high-dimensional), and output y (discrete or continuous). In general, we have two models:

Regression:

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x,y)}{\int p(x,y)\, dy}$$

Classification/Clustering:

$$p(c|x) = \frac{p(x,c)}{\sum_c p(x,c)}$$

Observed vs Unobserved random variables

Supervised classification(learning):

- We **KNOW** what to predict
- **Supervised Dataset:** $\{x^{(i)}, c^{(i)}\}_{i=1}^{N} \sim p(x,c)$
- The class labels are observed.

Unsupervised classification(learning):

- We do **NOT KNOW** what to predict
- **Unsupervised Dataset:** $\{x^{(i)}\}_{i=1}^{N} \sim p(x) = \sum_c p(x,c)$
- We only observe the inputs $\mathbf{x}$

In order to estimate the unknown distribution $p(x)$, we have few assumptions:

1. **IID Data**: we assume the samples $x^{(i)}$ are independent and identically distributed.
2. **Parametrized distribution**: $p(x|\theta)$ comes from a parametrized family $\mathcal{P} = \{p(x|\theta) : \theta \in \Theta\}$

Maximum Likelihood Estimation(MLE)

MLE is the method to estimate the parameters of an assume probability distribution, given some observed data. Technically, we can use MLE to estimate any parameters we want. More specifically:

- Let $x^{(i)} \sim p_* = p(x|\theta_*)$ for $i = 1, \ldots, N$ be i.i.d. random variables.
- The joint of $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ is $p(\mathcal{D}|\theta_*) = \prod_i p(x^{(i)}|\theta_*)$.

- Assume we observe data $\mathcal{D}$ and θ_* is unknown. The likelihood function is:

$$\mathcal{L}(\theta; \mathcal{D}) = p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta)$$

- The log-likelihood function:

$$\ell(\theta; \mathcal{D}) = \log \mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^{N} \log p(x^{(i)}|\theta)$$

Here is an example of MLE

Sufficient Statistics and Exponential Families

A **sufficient statistics** is a function of the data that conveys exactly the same information about the parameter as the entire data.
In addition, we can writing any exponential family member in the form:

$$p(x|\eta) = h(x) \exp\{\eta^\top T(x) - A(\eta)\}$$

where

$$\begin{aligned} T(x) &: \text{sufficient statistics} \\ \eta &: \text{natural parameter} \\ A(\eta) &: \text{log-partition function} \\ h(x) &: \text{carrying measure} \end{aligned}$$

Moreover, let $X \sim p(x|\eta)$, then we have $E[T(X)] = A'(\eta)$
One example of exponential family

1.2 Statistical decision theory

Suppose we have an input vector x and the corresponding target, we want to predict the label given a new input. Notice that here we assume the output is the label/class which is discrete. However, the output can also be continuous (regression).

Intuitively, for a given new input x, we have:

$$p(\mathcal{C}_k|x) = \frac{p(x|\mathcal{C}_k)p(\mathcal{C}_k)}{p(x)}$$

We then pick the $\mathcal{C}_k$ with the highest probability.
Decision Rule: Divide the input space to $\mathcal{R}_1$ & $\mathcal{R}_2$ such that all points in $\mathcal{R}_k$ are assigned to class $\mathcal{C}_k$. We want to make mistakes as less as possible; equivalently, we want to minimize the **misclassification rate**.
For $k \in \{1,\, 2\}$:

$$p(\text{mistake}) = p(x \in \mathcal{R}_2,\, \mathcal{C}_1) + p(x \in \mathcal{R}_2,\, \mathcal{C}_1) = \int_{\mathcal{R}_1} p\left(x, \mathcal{C}_2\right) dx + \int_{\mathcal{R}_2} p\left(x, \mathcal{C}_1\right) dx \quad (1.1)$$

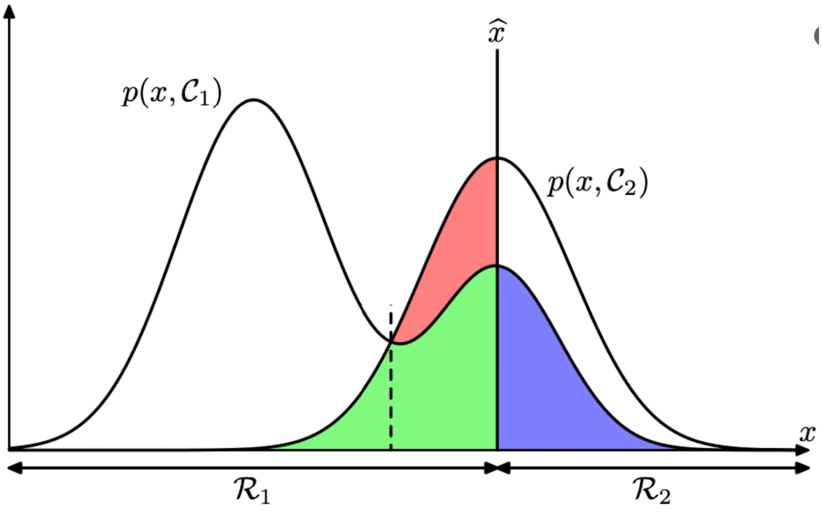

Figure 1: Misclassification Rate

1. **RedGreen regions**: inputs that belong to $\mathcal{C}_2$ but assigns to $\mathcal{R}_1$ as they are under $p\left(x, \mathcal{C}_2\right)$.
2. **Blue regions**: inputs that belong to $\mathcal{C}_1$ but assigns to $\mathcal{R}_2$ as they are under $p\left(x, \mathcal{C}_1\right)$
3. Therefore, for any data **x**, if $p\left(x, \mathcal{C}_1\right) > p\left(x, \mathcal{C}_2\right)$, then we assign this point to $\mathcal{C}_1$, vice-versa. Therefore, $\mathcal{R} = \{x : \ p\left(x, \mathcal{C}_1\right) > p\left(x, \mathcal{C}_2\right)\}$

We want to minimize the misclassification error rate ⇒ minimize the **loss**

Loss Function

Loss function measures the loss incurred by taking of any available decisions.

For discrete case

we denote L_{ij} as the $(k,\, j)$ element of the loss matrix. We want to minimize the expected loss
Therefore:

$$\begin{aligned}\mathbb{E}[L] &= \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p\left(x, \mathcal{C}_k\right) dx \\ &= \sum_j \int_{\mathcal{R}_j} \sum_k L_{kj} p\left(x, \mathcal{C}_k\right) dx\end{aligned}$$

Define $g_j(x) = \sum_k L_{kj} p\left(x, \mathcal{C}_k\right)$. Notice that $g_j(x) \geq 0$ and

$$\mathbb{E}[L] = \sum_j \int_{\mathcal{R}_j} g_j(x) dx$$

Thus, minimizing $\mathbb{E}[L]$ is equivalent to choosing

$$\mathcal{R}_j = \{x : g_j(x) < g_i(x) \text{ for all } i \neq j\} \tag{1.2}$$

$$\Rightarrow \mathcal{R}_j = \left\{x : \sum_k L_{kj} p\left(\mathcal{C}_k \mid x\right) < \sum_k L_{ki} p\left(\mathcal{C}_k \mid x\right) \text{ for all } i \neq j\right\} \tag{1.3}$$

For regression

- Consider the input/target $(x,\ t)$, where t is continuous and the joint density is $p(x,\ t)$
- The regression function is $y(t)$
- The loss function is $L(y(x),\ t) = (y(x) - t)^2$

Therefore the expected loss will be:

$$\begin{aligned}\mathbb{E}[L] &= \iint L(y(x), t) p(x, t) dx dt \\ &= \iint (y(x) - \mathbb{E}[t \mid x])^2 p(x, t) dx dt + \iint (\mathbb{E}[t \mid x] - t)^2 p(x, t) dx dt\end{aligned}$$

Full derivations here
The second term is the conditional variance of $t|x$ and does not depend on $y(x)$ and hence the expected loss is minimized when $y(x) = \mathbb{E}[t|x]$. Therefore, we can see that the loss function will change the decision rule significantly; however, we can always reject the option or not making a decision.

2 Graphical Models

2.1 Introduction to graphical models

Remember our goal is to specify the joint distribution N random variables $p(x_1, \cdots, x_N) = p(x)$. If we assume each x_i is binary such that $x_i \in \{0,\ 1\}$, then we need $\mathbf{2^N - 1}$ parameters to specify $p(x)$. For example, $p(x_1 = 0,\ x_2 = 0,\ \cdots, x_N = 0)$ or $p(x_1 = 1,\ x_2 = 0,\ \cdots, x_N = 0)$.
Equivalently, we can specify the joint distribution $p(x)$ as:

$$\begin{aligned}p\left(x_1, x_2, \ldots, x_N\right) &= \prod_{j=1}^{N} p\left(x_j \mid x_1, x_2, \ldots, x_{j-1}\right) \\ &= p(x_1|x_0) p(x_2|x_1,\ x_0) \cdots\end{aligned}$$

Thus total number of parameters is $1 + 2 + 4 + \cdots 2^{N-1} = 2^N - 1$
We can see that it requires a huge number of parameters to specify the joint distribution. We want to draw relationships between variables.

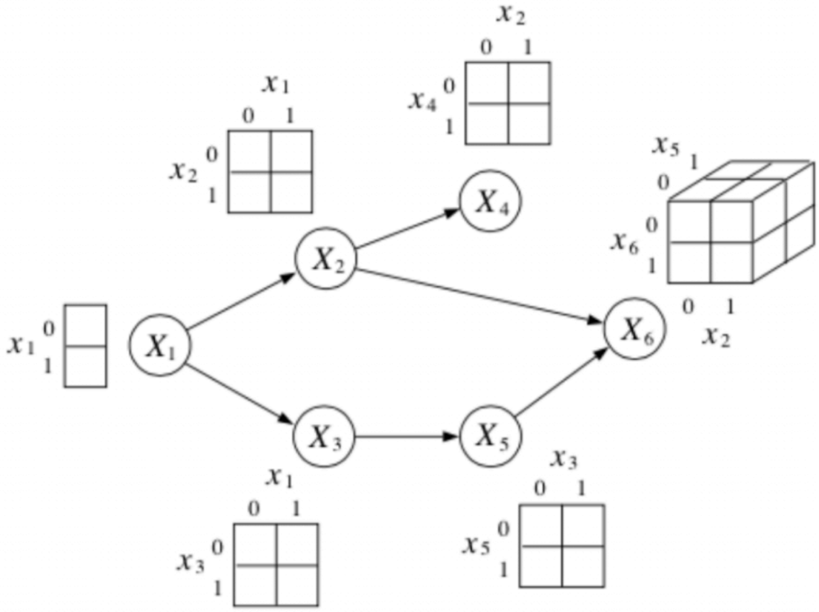

Figure 2: An example of conditional probability tables(CPT)

Condition independence

For three random variables x_A, x_B x_C, if x_A, x_B are conditionally independent given x_C, then we write $x_A \perp x_B \mid x_C$. The following conditions are equivalent:

- $x_A \perp x_B \mid x_C$
- $p(x_A,\, x_B|x_C) = p(x_A|x_C)p(x_B|x_C)$
- $p(x_A|x_B,\, x_C) = p(x_A|x_C)$
- $p(x_B|x_A,\, x_C) = p(x_B|x_C)$

2.2 Directed Acyclic Graphical Models

A directed cyclic graphical model encode a particular form of factorization of the joint distribution. The form of factorization is various.

Figure 2 shows an example of conditional probability. From the graph, we only need $2^1 * 4 + 2^0 + 2^2 = 13 < 2^6 - 1$ parameters.
D-separation: If C d-separates A and B, then $x_A \perp x_B \mid x_C \; \forall a \in A,\, b \in B$

Bayes ball algorithm

Bayes ball determines the conditional independence/dependence in a DAG (I personally found this part most ambiguous). There are three fundamental Bayes ball algorithms which are causal chain, common cause and explaining away. For each one, we will under it intuitively by drawing a story.

1. **Causal chain**

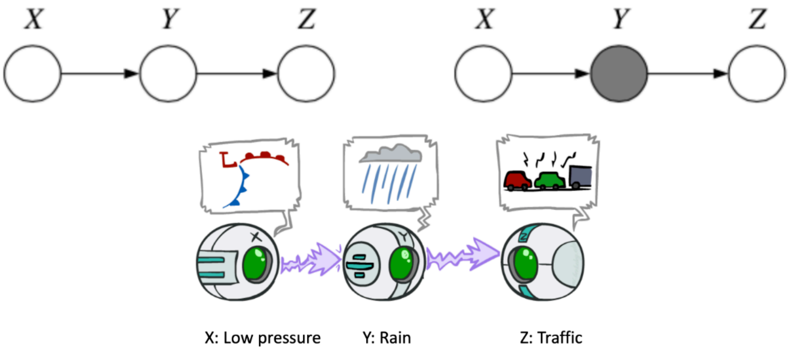

Figure 3: An illustration of causal chain

$$
\begin{aligned}
p(z \mid x, y) &= \frac{p(x, y, z)}{p(x, y)} \\
&= \frac{p(x)p(y \mid x)p(z \mid y)}{p(x)p(y \mid x)} \\
&= p(z \mid y) \\
\Rightarrow X &\text{ and } Z \text{ d-separated given } Y
\end{aligned}
$$

2. **Common cause**

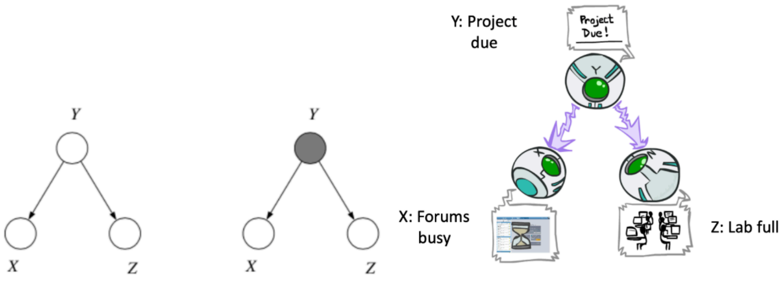

Figure 4: An illustration of common cause

$$
\begin{aligned}
p(x, z \mid y) &= \frac{p(x, y, z)}{p(y)} \\
&= \frac{p(y)p(x \mid y)p(z \mid y)}{p(y)} \\
&= p(x \mid y)p(z \mid y) \\
\Rightarrow X &\text{ and } Z \text{ d-separated given } Y
\end{aligned}
$$

3. **Explaining away**

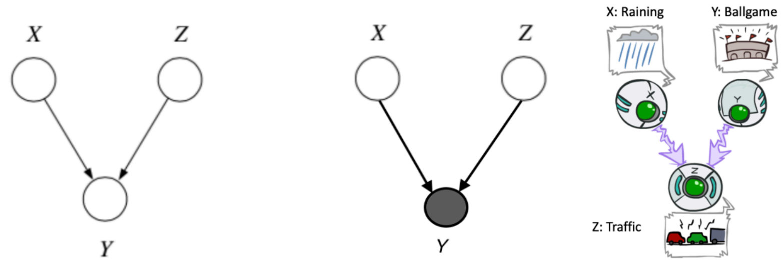

Figure 5: An illustration of explaining away

$$\begin{aligned} p(z \mid x, y) &= \frac{p(x)p(z)p(y \mid x, z)}{p(x)p(y \mid x)} \\ &= \frac{p(z)p(y \mid x, z)}{p(y \mid x)} \neq p(z \mid y) \\ &\Rightarrow X \text{ and } Z \text{ are NOT d-separated given } Y \end{aligned}$$

In general, the Bayes ball works as follows:

1. Shade all nodes x_C (these are observed)
2. Place ”balls” at each node in x_A (or x_B)
3. Let the ”balls” ”bounce” around according to some rules. If any of the balls reach any of the nodes in x_B from x_A then $x_A \not\perp x_B \mid x_C$. Otherwise $x_A \perp x_B \mid x_C$

Example
Question: Is $x_2 \perp x_3 \mid \{x_1, x_6\}$

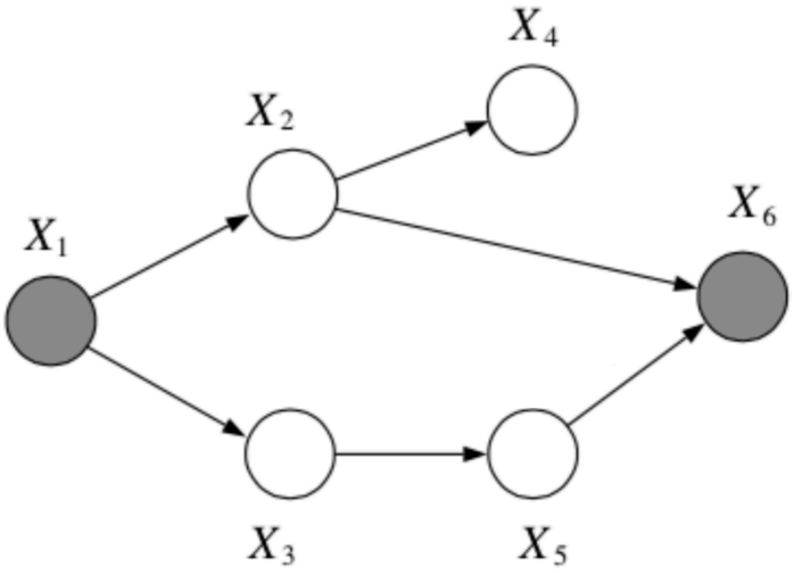

Figure 6: An illustration of explaining away

Answer: No. By Bayes ball algorithm, x_2 can travel to x_5, and hence can travel to x_3.

Moralization

Like I said, I personally do not like Bayes algorithm. Instead, another one called ”moralization” is more straightforward and easier to use. We can follow the procedure:

1. **Draw the ancestral graph**
   We only keep the ancestor of the mentioned nodes. That said, in the previous example, we only keep the ancestors of $\{x_2,\ x_3,\ x_1\ x_6\}$. Hence we have the entire graph except the node x_4. Note the ancestors includes **their parents, parents' parents etc**.

2. **”Moralize” the ancestral graph by ”marrying” the parents**
   If two nodes have the same children, such as x_2 and x_5, then we draw a line between these two nodes.

3. **”Disorient” the graph**
   Ignore the directions by replacing the arrows to edges.

4. **Delete the givens and their edges**
   In the previous example, the givens are the x_1 and x_4.

5. **Find the answer**
   After we finished the step 1 to 4, we then justify whether the two nodes are connected or disconnected. If **connected**, then the two nodes are conditionally **dependent**. Otherwise **disconnected**, the two nodes are conditionally **independent**. In the previous example, we can easily find that x_2 and x_3 are d-separated by x_1 and x_6.

Question: What about the marginal independence, such as $x_2 \perp x_3$?
Answer: We use the same way as above without step 4.

2.3 Undirected Graphical Models

The undirected graphical models are also called the **Markov random fields (MRFs)**. Compare to graphical models, we have no more directed edges; instead, the dependencies are now described as undirected graphs. Moreover, **Markov blanket** is the set of nodes that makes X_i conditionally independent of all other nodes. **Clique** is a subset of nodes that every two nodes are connected by an edge. **Maximal clique** a clique that can not be extended by including one more adjacent vertex.

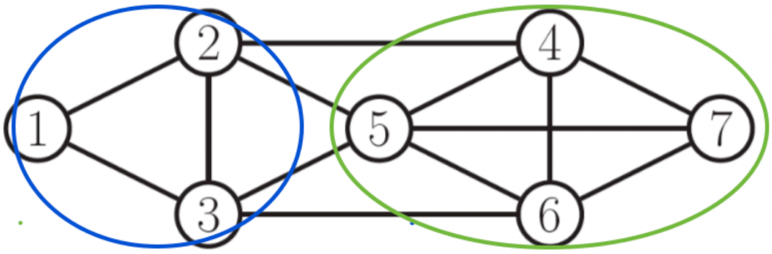

Figure 7: An example of Markov random fields: {1, 2, 3} is a clique and {4, 5, 6, 7} is a maximal clique

Distribution induced by MRFs

- Let $X = (X_1, \ldots, X_m)$ be the set of all random variables in our graph G.
- Let $\mathcal{C}$ be the set of all maximal cliques of G.
- The distribution p of X factorizes with respect to G if

  $$p(x) \propto \prod_{C \in \mathcal{C}} \psi_{\mathcal{C}}\left(x_C\right)$$

  for some nonnegative potential functions ψ_C, where $x_C = \left(x_i\right)_{i \in C}$.

The density can be factorized to cliques is also called the **Hammersley-Clifford Theorem**.
Global markov properties: $X_A \perp X_B \mid X_S$ if the sets A and B are separated by S in G (every path from A to B has to pass S).
hhhhhsss

3 Appendix

3.1 Example 1

The Bernoulli Naïve Bayes model parameterized by θ and π defines the following joint probability of x and c,

$$p(x, c|\theta, \pi) = p(c|\pi)p(x|c, \theta) = p(c|\pi) \prod_{j=1}^{D} p(x_j|c, \theta),$$

where $x_j|c, \theta \sim \text{Bernoulli}(\theta_{jc})$, i.e. $p(x_j|c, \theta) = \theta_{jc}^{x_j}(1-\theta_{jc})^{1-x_j}$, and $c|\pi$ follows a simple categorical distribution, i.e. $p(c|\pi) = \pi_c$.

Solution:

For $\mathbf{x^1}$, its likelihood function is:

$$\begin{aligned} L(\theta, \pi; x^1, c) &= \prod_{c=1}^{10} \left[p(x^1, c \mid \theta, \pi)\right]^{1\{c^1=c\}} && (3.1)\\ &= \prod_{c=1}^{10} \left[p(c \mid \pi) \prod_{j=1}^{784} p(x_j^1 \mid c, \theta)\right]^{1\{c^1=c\}} && (3.2)\\ &= \prod_{c=1}^{10} \left[\pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j^1} (1-\theta_{jc})^{1-x_j^1}\right]^{1\{c^1=c\}} && (3.3) \end{aligned}$$

Therefore, the joint likelihood function for $\mathbf{x^1}, \cdots \mathbf{x^n}$ is:

$$\begin{aligned} L(\theta, \pi) &= \prod_{i=1}^{n} \prod_{c=1}^{10} \left[\pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j^i} (1-\theta_{jc})^{1-x_j^i}\right]^{1\{c^i=c\}} && (3.4)\\ \Rightarrow l(\theta, \pi) &= \log\left(\prod_{i=1}^{n} \prod_{c=1}^{10} \left[\pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j^i} (1-\theta_{jc})^{1-x_j^i}\right]^{1\{c^i=c\}}\right) && (3.5)\\ &= \sum_{i=1}^{n} \sum_{c=1}^{10} 1\{c^i = c\} \left\{\log(\pi_c) + \sum_{j=1}^{784} \left[x_j^i \log(\theta_{jc}) + (1-x_j^i) \log(1-\theta_{jc})\right]\right\} && (3.6)\\ &= \sum_{i=1}^{n} \sum_{c=1}^{9} 1\{c^i = c\} \left\{\log(\pi_c) + \sum_{j=1}^{784} \left[x_j^i \log(\theta_{jc}) + (1-x_j^i) \log(1-\theta_{jc})\right]\right\} && (3.7)\\ &+ \sum_{i=1}^{n} 1\{c^i = 10\} \left\{\log(1 - \sum_{c=1}^{9} \pi_c) + \sum_{j=1}^{784} \left[x_j^i \log(\theta_{j,10}) + (1-x_j^i) \log(1-\theta_{j,10})\right]\right\} && (3.8) \end{aligned}$$

If we pick any $c \in [C]$ and $j \in [D]$:

$$\frac{\partial l(\theta, \pi)}{\partial \theta_{jc}} = \sum_{i=1}^{n} 1\{c^i = c\} \left(\frac{x_j^i}{\theta_{jc}} - \frac{1-x_j^i}{1-\theta_{jc}}\right) \qquad (3.9)$$

Letting it equal to zero, we have:

$$\hat{\theta_{jc}} = \frac{\sum_{i=1}^n 1\{c^i = c\} x_j^i}{\sum_{i=1}^n 1\{c^i = c\}} \tag{3.10}$$

For π_c:

$$\frac{\partial l(\theta, \pi)}{\partial \pi_c} = \sum_{i=1}^n 1\{c^i = c\} \frac{1}{\pi_c} - \sum_{i=1}^n 1\{c^i = 10\} \frac{1}{1 - \sum_{c=1}^9 \pi_c} \tag{3.11}$$

Letting $n_c = \sum_{i=1}^n 1\{c^i = c\}$, when it equals to zero, we have:

$$n_c(1 - \sum_{c=1}^9 \hat{\pi_c}) = \hat{\pi_c} n_{10}, \quad \text{where } 1 \leq a \leq 9 \tag{3.12}$$

Summation on both sides over $1 \leq c \leq 9$, we have:

$$\sum_{c=1}^9 n_c(1 - \sum_{c=1}^9 \hat{\pi_c}) = \sum_{c=1}^9 \hat{\pi_c} n_{10} \tag{3.13}$$

$$\Rightarrow (n - n_{10})(1 - \sum_{c=1}^9 \hat{\pi_c}) = n_{10} \sum_{c=1}^9 \hat{\pi_c} \tag{3.14}$$

$$\sum_{c=1}^9 \hat{\pi_c} = \frac{n - n_{10}}{n} \tag{3.15}$$

Substituting back, we will have:

$$\hat{\pi_c} = \frac{n_c}{n}, \quad 1 \leq c \leq 9 \tag{3.16}$$

3.2 Example 2

We can write this distribution as an exponential family

$$\begin{aligned} p(x \mid \theta) &= \theta^x (1-\theta)^{1-x} && (3.17) \\ &= \exp\{x \log(\theta) + (1-x)\log(1-\theta)\} && (3.18) \\ &= \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right\} && (3.19) \end{aligned}$$

Here,

$$\begin{aligned} T(x) &= x \\ \eta &= \log\left(\frac{\theta}{1-\theta}\right) \\ A(\eta) &= \log(1 + e^\eta) \\ h(x) &= 1 \end{aligned}$$

Notice that $A'(\eta) = \frac{e^\eta}{1+e^\eta} = \theta$ is the mean of $T(X) = X$ and $A''(\eta) = \frac{e^\eta}{(1+e^\eta)^2} = \theta(1-\theta)$ is the variance of X.

3.3 Derivations 1

We add and subtract $\mathbb{E}[t \mid x]$ and write

$$\begin{aligned}
\mathbb{E}[L] &= \iint (y(x) - t)^2 p(x,t) dx dt \\
&= \iint (y(x) - \mathbb{E}[t \mid x] + \mathbb{E}[t \mid x] - t)^2 p(x,t) dx dt \\
&= \iint (y(x) - \mathbb{E}[t \mid x])^2 p(x,t) dx dt + \iint (\mathbb{E}[t \mid x] - t)^2 p(x,t) dx dt \\
&\quad + 2 \iint (y(x) - \mathbb{E}[t \mid x])(\mathbb{E}[t \mid x] - t) p(x,t) dx dt
\end{aligned}$$

The last term is zero since

$$\begin{aligned}
&\iint (y(x) - \mathbb{E}[t \mid x])(\mathbb{E}[t \mid x] - t) p(x,t) dx dt \\
&= \iint (y(x) - \mathbb{E}[t \mid x])(\mathbb{E}[t \mid x] - t) p(t \mid x) p(x) dx dt \\
&= \int (y(x) - \mathbb{E}[t \mid x]) \{ \underbrace{\int (\mathbb{E}[t \mid x] - t) p(t \mid x) dt}_{=0} \} p(x) dx = 0
\end{aligned}$$