

# STA414: Statistical Methods for Machine Learning II

## Lecture Notes

Yiliu Cao

May 3, 2024

### Contents

<b>1</b>	<b>Probabilistic models</b>	<b>2</b>
1.1	An overview of probabilistic models	2
1.2	Statistical decision theory	3
<b>2</b>	<b>Graphical Models</b>	<b>5</b>
2.1	Introduction to graphical models	5
2.2	Directed Acyclic Graphical Models	6
2.3	Undirected Graphical Models	9
<b>3</b>	<b>Appendix</b>	<b>13</b>
3.1	Example 1	13
3.2	Example 2	14
3.3	Derivations 1	15

# 1 Probabilistic models

## 1.1 An overview of probabilistic models

We have the random vector  $X = (X_1, X_2, \dots, X_n)$ , and we want to compute the relationship between each random variable. The joint distribution is  $p(x) = p(x_1, \dots, x_d)$ . 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  $i = 1, \dots, N$  be i.i.d. random variables.
- The joint of  $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$  is  $p(\mathcal{D}|\theta_*) = \prod_i p(x^{(i)}|\theta_*)$ .

- Assume we observe data  $\mathcal{D}$  and  $\theta_*$  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

$T(x)$  : sufficient statistics

$\eta$  : natural parameter

$A(\eta)$  : log-partition function

$h(x)$  : carrying measure

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  $k \in \{1, 2\}$ :

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

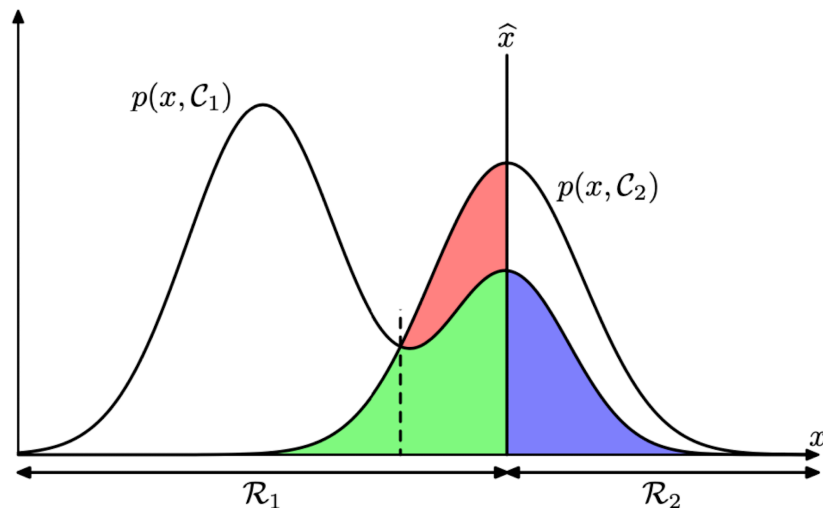


Figure 1.1: Misclassification Rate

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

We want to minimize the misclassification error rate  $\Rightarrow$  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  $(i, j)$  element of the loss matrix. We want to minimize the expected loss

Therefore:

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

Define  $g_j(x) = \sum_k L_{kj} p(x, \mathcal{C}_k)$ . 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

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

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

### For regression

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

Therefore the expected loss will be:

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

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  $p(x_1, \dots, x_N) = p(x)$ . If we assume each  $x_i$  is binary such that  $x_i \in \{0, 1\}$ , then we need  $2^N - 1$  parameters to specify  $p(x)$ . For example,  $p(x_1 = 0, x_2 = 0, \dots, x_N = 0)$  or  $p(x_1 = 1, x_2 = 0, \dots, x_N = 0)$ .

Equivalently, we can specify the joint distribution  $p(x)$  as:

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

Thus total number of parameters is  $1 + 2 + 4 + \dots + 2^{N-1} = 2^N - 1$

We can see that it requires a huge number of parameters to specify the joint distribution. We want to draw relationships between variables.

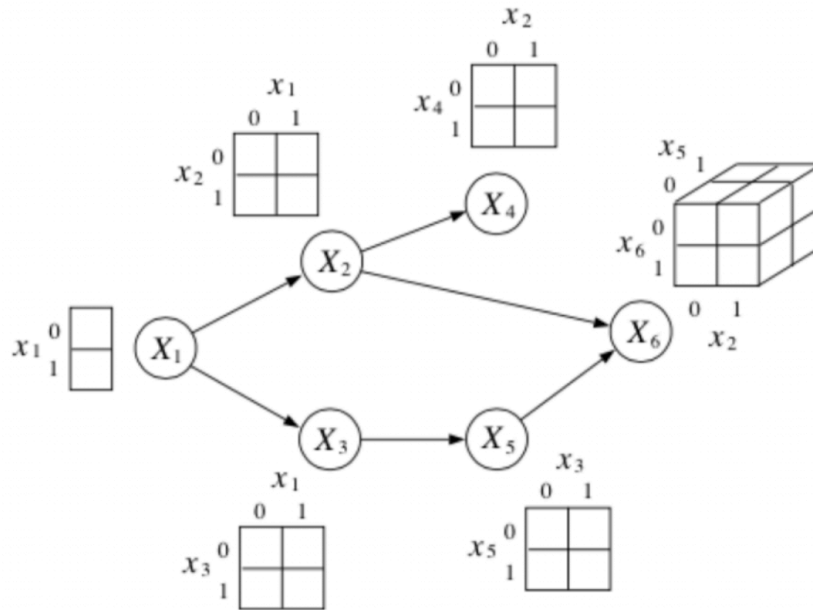


Figure 2.1: An example of conditional probability tables(CPT)

### 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:

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

## 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

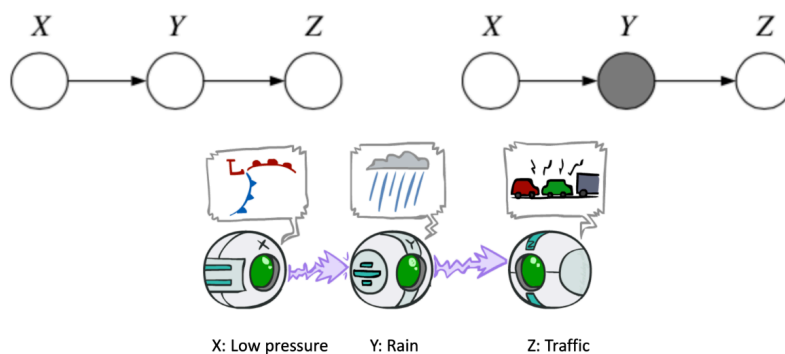


Figure 2.2: An illustration of causal chain

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

## 2. Common cause

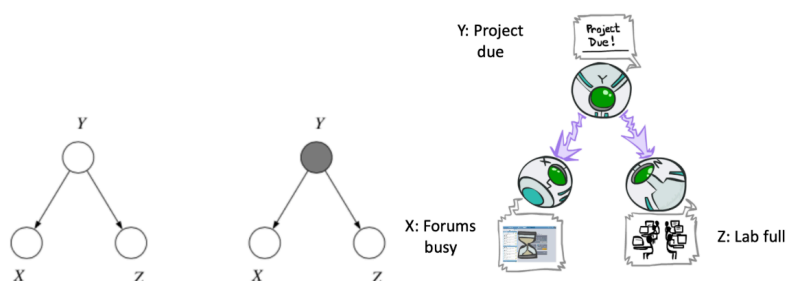


Figure 2.3: An illustration of common cause

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

## 3. Explaining away

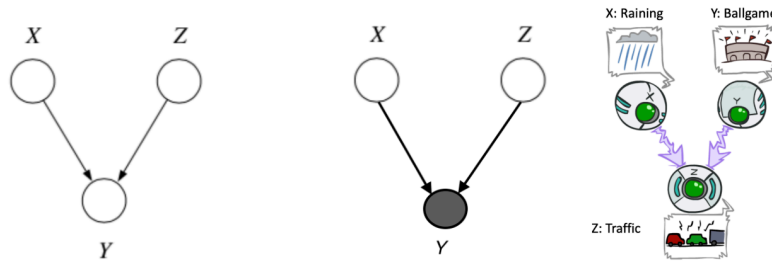


Figure 2.4: An illustration of explaining away

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

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  $x_2 \perp x_3 \mid \{x_1, x_6\}$

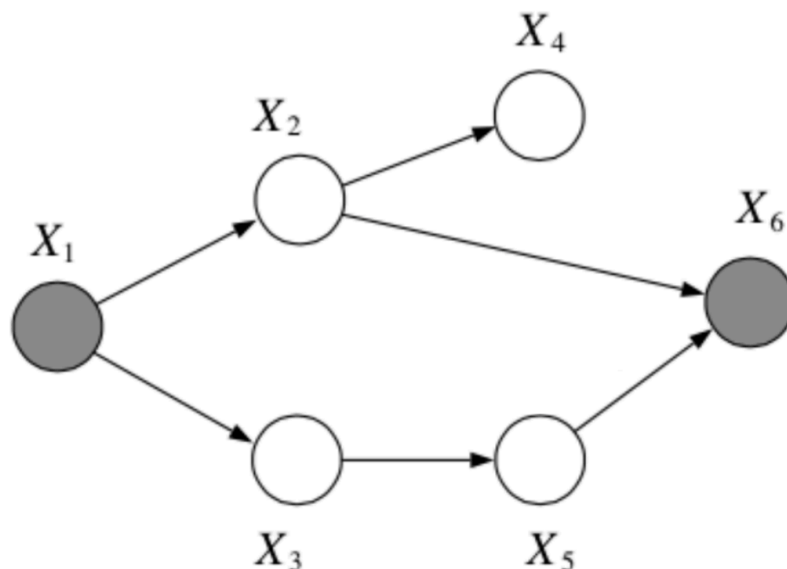


Figure 2.5: An illustration of explaining away

*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.

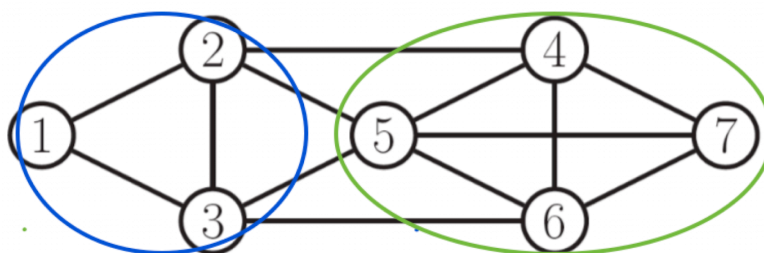


Figure 2.6: An example of Markov random fields:  $\{1, 2, 3\}$  is a clique and  $\{4, 5, 6, 7\}$  is a maximal clique

### Distribution induced by MRFs

- Let  $X = (X_1, \dots, X_m)$  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

$$p(x) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

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

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  $S$ ). Therefore, the joint distribution described on above is:

$$p(x) \propto \psi_{1,2,3}(x_1, x_2, x_3) \psi_{2,3,5}(x_2, x_3, x_5) \psi_{2,4,5}(x_2, x_4, x_5) \\ \times \psi_{3,5,6}(x_3, x_5, x_6) \psi_{4,5,6,7}(x_4, x_5, x_6, x_7)$$

Not all DAGMs can be represented as MRFs. .

### Relations between exponential families and MRFs

- Consider a parametric family of factorized distributions

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \psi_C(x_C \mid \theta_C), \quad \theta = (\theta_C)_{C \in \mathcal{C}}.$$

- We can write this in an exponential form:

$$p(x \mid \theta) = \exp\left\{\sum_{C \in \mathcal{C}} \log \psi_C(x_C \mid \theta_C) - \underbrace{\log Z(\theta)}_{=A(\theta)}\right\}$$

- Suppose the potentials have a log-linear form

$$\log \psi_C(x_C \mid \theta_C) = \theta_C^\top \phi_C(x_C)$$

we then get the exponential family

$$p(x \mid \theta) = \exp\left\{\sum_{C \in \mathcal{C}} \theta_C^\top \phi_C(x_C) - \underbrace{\log Z(\theta)}_{=A(\theta)}\right\}$$

*Question:* When the potentials have a log-linear form?

*Solution:* Suppose we have the random vector  $x_1$  and  $x_2$ , and they are all binary. Then

there are four possible values of  $(x_1, x_2)$ :  $(0, 0)$ ,  $(1, 0)$ ,  $(0, 1)$ ,  $(1, 1)$ . We take

$$\theta_{1,2} := \begin{bmatrix} \log \psi_{1,2}(0, 0) \\ \log \psi_{1,2}(0, 1) \\ \log \psi_{1,2}(1, 0) \\ \log \psi_{1,2}(1, 1) \end{bmatrix} \in \mathbb{R}^4$$

and let  $\psi_{1,2}(x_1, x_2)$  be the function that satisfies

$$\phi_{1,2}(0, 0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \phi_{1,2}(0, 1) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \phi_{1,2}(1, 0) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \phi_{1,2}(1, 1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

$$\Rightarrow \phi_{1,2}(x_1, x_2) = \begin{bmatrix} (1-x_1)(1-x_2) \\ (1-x_1)x_2 \\ x_1(1-x_2) \\ x_1x_2 \end{bmatrix}$$

We then have a linear form:

$$\log \psi_C(x_{12} \mid \theta_{12}) = \theta_{12}^\top \phi_{12}(x_{12})$$

### Ising model

The Ising model is an example of MRFs, which is used to model magnets. It has a form of potential function:

$$\psi_{st}(x_s, x_t) = e^{J_{st}x_sx_t}$$

Equivalently:

$$\begin{aligned} \psi_{st}(-1, -1) &= \psi_{st}(1, 1) = e^{J_{st}} \\ \psi_{st}(1, -1) &= \psi_{st}(-1, 1) = e^{-J_{st}} \end{aligned}$$

$$\psi_{st}(x_s, x_t) = 0 \text{ if the two nodes are not connected}$$

In terms of the distribution in Ising model, we also include the node potential  $\psi_s(x_s) = e^{b_sx_s}$ :

$$p(x) \propto \prod_{s \sim t} \psi_{st}(x_s, x_t) \prod_s \psi_s(x_s) = \exp \left\{ \sum_{s \sim t} J_{st}x_sx_t + \sum_s b_sx_s \right\}$$

### Recall: Multivariate normal distribution

$X = (X_1, \dots, X_m) : \mu \in \mathbb{R}^m$  and  $\Sigma$  symmetric positive definite  $m \times m$  matrix. Write  $X \sim N_m(\mu, \Sigma)$  if the density of the vector  $X$  is

$$f(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{m/2}} (\det \Sigma)^{-1/2} \exp \left( -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right).$$

Denote  $K = \Sigma^{-1}$  then

$$f(\mathbf{x}; \mu, \Sigma) \propto \prod_s e^{-\frac{1}{2} K_{ss}(x_s - \mu_s)^2} \prod_{s < t} e^{-K_{st}(x_s - \mu_s)(x_t - \mu_t)}$$

Intuitively, we can also visualize the conditional independencies between each variables (similar to Bayes ball). Just like the Ising model, here we can use concentration matrix  $K$  to represent the relationship between variables.

Conditional independence:

- $X_i \perp X_j$  if and only if  $\Sigma_{ij} = 0$ .
- $X_i \perp X_j \mid X_C$  if and only if  $\Sigma_{ij} - \Sigma_{i,C}\Sigma_{C,C}^{-1}\Sigma_{C,j} = 0$
- Let  $R = V \setminus \{i, j\}$ . The following are equivalent:
  - $X_i \perp X_j \mid X_R$
  - $\Sigma_{ij} - \Sigma_{i,R}\Sigma_{R,R}^{-1}\Sigma_{R,j} = 0$
  - $(\Sigma^{-1})_{ij} = 0$

Hence we have the **Gaussian Graphical models**

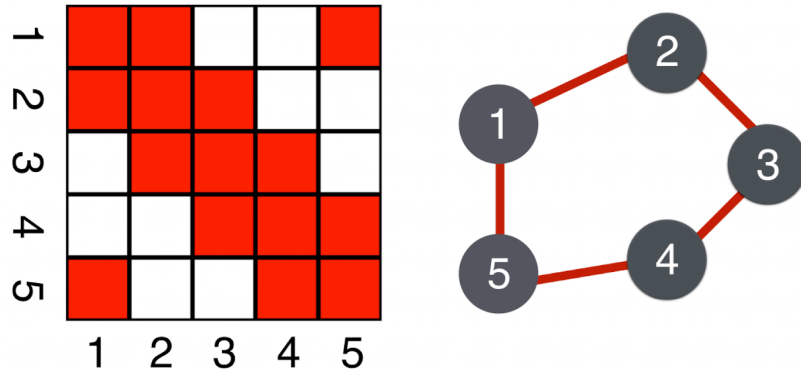


Figure 2.7: An example of Gaussian Graphical model

$K_{ij} = 0$  if and only if  $X_i \perp X_j \mid X_{rest}$ . For example,  $X_1 \perp X_4 \mid X_{rest}$

### 3 Appendix

#### 3.1 Example 1

The Bernoulli Naïve Bayes model parameterized by  $\theta$  and  $\pi$  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:

$$L(\theta, \pi; x^1, c) = \prod_{c=1}^{10} [p(x^1, c | \theta, \pi)]^{1\{c^1=c\}} \quad (3.1)$$

$$= \prod_{c=1}^{10} \left[ p(c | \pi) \prod_{j=1}^{784} p(x_j^1 | c, \theta) \right]^{1\{c^1=c\}} \quad (3.2)$$

$$= \prod_{c=1}^{10} \left[ \pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j^1} (1 - \theta_{jc})^{1-x_j^1} \right]^{1\{c^1=c\}} \quad (3.3)$$

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

$$L(\theta, \pi) = \prod_{i=1}^n \prod_{c=1}^{10} \left[ \pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j^i} (1 - \theta_{jc})^{1-x_j^i} \right]^{1\{c^i=c\}} \quad (3.4)$$

$$\Rightarrow l(\theta, \pi) = \log \left( \prod_{i=1}^n \prod_{c=1}^{10} \left[ \pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j^i} (1 - \theta_{jc})^{1-x_j^i} \right]^{1\{c^i=c\}} \right) \quad (3.5)$$

$$= \sum_{i=1}^n \sum_{c=1}^{10} 1\{c^i = c\} \left\{ \log(\pi_c) + \sum_{j=1}^{784} [x_j^i \log(\theta_{jc}) + (1 - x_j^i) \log(1 - \theta_{jc})] \right\} \quad (3.6)$$

$$= \sum_{i=1}^n \sum_{c=1}^9 1\{c^i = c\} \left\{ \log(\pi_c) + \sum_{j=1}^{784} [x_j^i \log(\theta_{jc}) + (1 - x_j^i) \log(1 - \theta_{jc})] \right\} \quad (3.7)$$

$$+ \sum_{i=1}^n 1\{c^i = 10\} \left\{ \log(1 - \sum_{c=1}^9 \pi_c) + \sum_{j=1}^{784} [x_j^i \log(\theta_{j,10}) + (1 - x_j^i) \log(1 - \theta_{j,10})] \right\} \quad (3.8)$$

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

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

Letting it equal to zero, we have:

$$\hat{\theta}_{jc} = \frac{\sum_{i=1}^n 1\{c^i = c\} x_j^i}{\sum_{i=1}^n 1\{c^i = c\}} \quad (3.10)$$

For  $\pi_c$ :

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

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

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

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

$$\sum_{c=1}^9 n_c(1 - \sum_{c=1}^9 \hat{\pi}_c) = \sum_{c=1}^9 \hat{\pi}_c n_{10} \quad (3.13)$$

$$\Rightarrow (n - n_{10})(1 - \sum_{c=1}^9 \hat{\pi}_c) = n_{10} \sum_{c=1}^9 \hat{\pi}_c \quad (3.14)$$

$$\sum_{c=1}^9 \hat{\pi}_c = \frac{n - n_{10}}{n} \quad (3.15)$$

Substituting back, we will have:

$$\hat{\pi}_c = \frac{n_c}{n}, \quad 1 \leq c \leq 9 \quad (3.16)$$

## 3.2 Example 2

We can write this distribution as an exponential family

$$p(x | \theta) = \theta^x (1 - \theta)^{1-x} \quad (3.17)$$

$$= \exp \{x \log(\theta) + (1 - x) \log(1 - \theta)\} \quad (3.18)$$

$$= \exp \left\{ x \log \left( \frac{\theta}{1 - \theta} \right) + \log(1 - \theta) \right\} \quad (3.19)$$

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

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