Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 20
March 17, 2023

Sampling methods

Some more observations from the last model through an example.

Sampling methods

Some more observations from the last model through an example.

We call a node y, a descendant of a node x if there is a path from x to y in which each step of the path follows the directions

Sampling methods

Some more observations from the last model through an example.

We call a node y, a descendant of a node x if there is a path from x to y in which each step of the path follows the directions

D-separation

$\rightarrow$ Consider a directed graph in which A, B, C are arbitrary nonintersecting sets of nodes, whose union may be smaller than the total set of nodes in the graph

Sampling methods

Some more observations from the last model through an example.

We call a node y, a descendant of a node x if there is a path from x to y in which each step of the path follows the directions

D-separation

- $\rightarrow$ Consider a directed graph in which A, B, C are arbitrary nonintersecting sets of nodes, whose union may be smaller than the total set of nodes in the graph
- $\rightarrow$ We wish to ascertain whether a particular conditional independence statement $A \perp\!\!\!\perp B | C$ is implied by a given directed acyclic graph!

Sampling methods

Consider all possible paths from any node in A to any node in B. Any such path is called blocked if it includes a node such that either

(a) the arrows on the path meet either head-to-tail or tail-to-tail at the node which is in C, or

(b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C

Sampling methods

Consider all possible paths from any node in A to any node in B. Any such path is called blocked if it includes a node such that either

(a) the arrows on the path meet either head-to-tail or tail-to-tail at the node which is in C, or

(b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C

If all the paths are blocked, then A is said to be d-separated from B by C, and the joint distribution over all the variables in the graph will satisfy $A \perp\!\!\!\perp B | C$

Sampling methods

Consider all possible paths from any node in A to any node in B. Any such path is called blocked if it includes a node such that either

(a) the arrows on the path meet either head-to-tail or tail-to-tail at the node which is in C, or

(b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C

If all the paths are blocked, then A is said to be d-separated from B by C, and the joint distribution over all the variables in the graph will satisfy $A \perp\!\!\!\perp B | C$

Example Consider $p(a, b, c, d, e) = p(a)p(e)p(d|a, e)p(b|e)p(c|d)$. Then

$\rightarrow$ Justify: $a \not\!\perp\!\!\!\perp b | c$

Sampling methods

Consider all possible paths from any node in A to any node in B. Any such path is called blocked if it includes a node such that either

(a) the arrows on the path meet either head-to-tail or tail-to-tail at the node which is in C, or

(b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C

If all the paths are blocked, then A is said to be d-separated from B by C, and the joint distribution over all the variables in the graph will satisfy $A \perp\!\!\!\perp B | C$

Example Consider $p(a, b, c, d, e) = p(a)p(e)p(d|a, e)p(b|e)p(c|d)$. Then

- $\rightarrow$ Justify: $a \not\perp\!\!\!\perp b | c$
- $\rightarrow$ Justify: $a \perp\!\!\!\perp b | e$

Sampling methods

Consider all possible paths from any node in A to any node in B. Any such path is called blocked if it includes a node such that either

(a) the arrows on the path meet either head-to-tail or tail-to-tail at the node which is in C, or

(b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C

If all the paths are blocked, then A is said to be d-separated from B by C, and the joint distribution over all the variables in the graph will satisfy $A \perp\!\!\!\perp B | C$

Example Consider $p(a, b, c, d, e) = p(a)p(e)p(d|a, e)p(b|e)p(c|d)$. Then

- $\rightarrow$ Justify: $a \not\perp\!\!\!\perp b | c$
- $\rightarrow$ Justify: $a \perp\!\!\!\perp b | e$

Question Develop an algorithm for D-separation for DAGs.

Sampling methods

Exponential family of distributions - The exponential family of distributions over $\mathbf{x}$, given parameters $\boldsymbol{\eta}$ is said to be distributions of the form

$$p(\mathbf{x};\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\},$$

where $\mathbf{x}$ may be scalar or vector, and may be continuous and discrete, $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_K]^T$ is called the vector of natural parameters of the distribution, and $\mathbf{u}(\mathbf{x}) = [u_1(\mathbf{x}), \ldots, u_K(x)]^T$ is the vector of sufficient statistics, each sufficient statistic $u_k(\mathbf{x})$ being a function of $\mathbf{x}$, $h(\mathbf{x})$ is the base measure which is a function of $\mathbf{x}$ independent of $\boldsymbol{\eta}$, and $g(\boldsymbol{\eta})$ is the partition function such that

$$\frac{1}{g(\boldsymbol{\eta})} = \int \exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))h(\mathbf{x})d\mathbf{x}$$

for continuous rvs and $\frac{1}{g(\boldsymbol{\eta})} = \sum_{\mathbf{x}} \exp(\boldsymbol{\eta}^T\mathbf{u}(x))h(\mathbf{x})$ for discrete rv

Sampling methods

Sufficient statistic Let $p(\mathbf{x}, \theta)$ be the distribution of a rv $\mathbf{X}$ that depends on θ. A function $f(\mathbf{x})$ is a sufficient statistic for the estimate of θ if the likelihood $p(\mathbf{x}, \theta)$ of the parameters θ depends on $\mathbf{x}$ only through the function $f(\mathbf{x})$.

Sampling methods

Sufficient statistic Let $p(\mathbf{x}, \theta)$ be the distribution of a rv $\mathbf{X}$ that depends on θ. A function $f(\mathbf{x})$ is a sufficient statistic for the estimate of θ if the likelihood $p(\mathbf{x}, \theta)$ of the parameters θ depends on $\mathbf{x}$ only through the function $f(\mathbf{x})$.
For example, $X \sim \mathcal{N}(0, \sigma^2)$, the function $f(x) = x^2$ can be easily seen to be sufficient for the estimate of the variance σ^2

Sampling methods

Sufficient statistic Let $p(\mathbf{x}, \theta)$ be the distribution of a rv $\mathbf{X}$ that depends on θ. A function $f(\mathbf{x})$ is a sufficient statistic for the estimate of θ if the likelihood $p(\mathbf{x}, \theta)$ of the parameters θ depends on $\mathbf{x}$ only through the function $f(\mathbf{x})$.
For example, $X \sim \mathcal{N}(0, \sigma^2)$, the function $f(x) = x^2$ can be easily seen to be sufficient for the estimate of the variance σ^2

Bernoulli -

$$
\begin{aligned}
p(x, \mu) &= \mu^x (1-\mu)^{1-x} = \exp\{x \ln \mu + (1-x)\ln(1-\mu)\} \\
&= (1-\mu) \exp\left\{ \ln\left(\frac{\mu}{1-\mu}\right) x \right\}
\end{aligned}
$$

Sampling methods

Sufficient statistic Let $p(\mathbf{x}, \theta)$ be the distribution of a rv $\mathbf{X}$ that depends on θ. A function $f(\mathbf{x})$ is a sufficient statistic for the estimate of θ if the likelihood $p(\mathbf{x}, \theta)$ of the parameters θ depends on $\mathbf{x}$ only through the function $f(\mathbf{x})$.
For example, $X \sim \mathcal{N}(0, \sigma^2)$, the function $f(x) = x^2$ can be easily seen to be sufficient for the estimate of the variance σ^2

Bernoulli -

$$
\begin{aligned}
p(x, \mu) &= \mu^x (1-\mu)^{1-x} = \exp\{x \ln \mu + (1-x)\ln(1-\mu)\} \\
&= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\}
\end{aligned}
$$

Set $\eta = \ln\left(\frac{\mu}{1-\mu}\right)$ and $g(\eta) = \frac{1}{1+\exp(-\eta)} \Rightarrow p(x, \mu) = g(-\eta)\exp(\eta x)$, g is called the logistic sigmoid function

Sampling methods

Gaussian -

$$
\begin{aligned}
p(x; \mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^1/2} \exp\left\{-\frac{1}{-2\sigma^2}(x-\mu)^2\right\} \\
&= \frac{1}{(2\pi\sigma^2)^1/2} \exp\left\{-\frac{1}{-2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \\
&= h(\mathbf{x})g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T u(\mathbf{x})\}
\end{aligned}
$$

$\boldsymbol{\eta} = \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix}$, $u(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$, $h(x) = (2\pi)^{-1/2}$,
$g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right)$

Sampling methods

Examples of Exponential Families

Bernoulli: distribution on $(0, 1)$

Categorical: distribution on $\{1, 2, \ldots, k\}$

Gaussian: distribution on $\mathbb{R}^d$

Beta: distribution on $[0, 1]$ (including uniform)

Dirichlet: distribution on discrete probabilities

Wishart: distribution on positive-definite matrices

Poisson: distribution on non-negative integers.

Gamma: distribution on positive real numbers

many more....

Sampling methods

Maximum likelihood and sufficient statistics How to estimate the values of the parameters from a data which supposedly follows a distribution from the exponential family?

Recall that

→ The gradient of a differentiable function $f(\mathbf{x})$ with $\mathbf{x} = [x_1, \ldots, x_d] \in \mathbb{R}^d$ is defined as

$$\nabla f(\mathbf{x}) = [\partial f(\mathbf{x})/\partial x_1, \ldots, \partial f(\mathbf{x})/\partial x_d]^T$$

Sampling methods

Maximum likelihood and sufficient statistics How to estimate the values of the parameters from a data which supposedly follows a distribution from the exponential family?
Recall that

→ The gradient of a differentiable function $f(\mathbf{x})$ with $\mathbf{x} = [x_1, \ldots, x_d] \in \mathbb{R}^d$ is defined as

$$\nabla f(\mathbf{x}) = [\partial f(\mathbf{x})/\partial x_1, \ldots, \partial f(\mathbf{x})/\partial x_d]^T$$

→ The Hessian of $f(\mathbf{x})$ is a $d \times d$ matrix with ij-th entry $\partial^2 f(\mathbf{x})/\partial x_i \partial x_j$

Sampling methods

Taking the gradient both sides of $g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} d\mathbf{x} = 1$ we obtain

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} d\mathbf{x} +$$

Sampling methods

Taking the gradient both sides of $g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} d\mathbf{x} = 1$ we obtain

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} d\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) d\mathbf{x} = 0$$

Sampling methods

Taking the gradient both sides of $g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} d\mathbf{x} = 1$ we obtain

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} d\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) d\mathbf{x} = 0$$

Which implies

$$-\frac{1}{g(\boldsymbol{\eta})} \nabla g(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Sampling methods

Taking the gradient both sides of $g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} d\mathbf{x} = 1$ we obtain

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} d\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) d\mathbf{x} = 0$$

Which implies

$$-\frac{1}{g(\boldsymbol{\eta})} \nabla g(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Thus

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Sampling methods

Taking the gradient both sides of $g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} d\mathbf{x} = 1$ we obtain

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} d\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) d\mathbf{x} = 0$$

Which implies

$$-\frac{1}{g(\boldsymbol{\eta})} \nabla g(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Thus

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Similarly the covariance matrix of $\mathbf{u}(\mathbf{x})$ can be expressed in terms of the second derivative of $g(\boldsymbol{\eta})$, and the higher order moments. The covariance matrix is also equal to the Fisher information matrix for natural parameters

Sampling methods

Estimation of η_{ML} Consider a set of iid data denoted by $\mathbf{X} = \{X_1, \ldots, X_N\}$. Then the likelihood function is

Sampling methods

Estimation of $\boldsymbol{\eta}_{ML}$ Consider a set of iid data denoted by $\mathbf{X} = \{X_1, \ldots, X_N\}$. Then the likelihood function is

$$p(\mathbf{X}, \boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n) \right) g(\boldsymbol{\eta})^N \exp\left\{ \boldsymbol{\eta}^T \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) \right\}$$

Sampling methods

Estimation of $\boldsymbol{\eta}_{ML}$ Consider a set of iid data denoted by $\mathbf{X} = \{X_1, \ldots, X_N\}$. Then the likelihood function is

$$p(\mathbf{X}, \boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^T \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\}$$

Setting the gradient of $\ln p(\mathbf{X}, \boldsymbol{\eta})$ wrt $\boldsymbol{\eta}$ to zero, we obtain the following condition to be satisfied by the maximum likelihood estimator $\boldsymbol{\eta}_{ML}$

$$-\nabla \ln g(\boldsymbol{\eta}_{ML}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n),$$

which can in principle be solved to obtain $\boldsymbol{\eta}_{ML}$

Sampling methods

Estimation of $\boldsymbol{\eta}_{ML}$ Consider a set of iid data denoted by $\mathbf{X} = \{X_1, \ldots, X_N\}$. Then the likelihood function is

$$p(\mathbf{X}, \boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^T \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\}$$

Setting the gradient of $\ln p(\mathbf{X}, \boldsymbol{\eta})$ wrt $\boldsymbol{\eta}$ to zero, we obtain the following condition to be satisfied by the maximum likelihood estimator $\boldsymbol{\eta}_{ML}$

$$-\nabla \ln g(\boldsymbol{\eta}_{ML}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n),$$

which can in principle be solved to obtain $\boldsymbol{\eta}_{ML}$
Observation The MLE depends on the data through $\sum_n \mathbf{u}(\mathbf{x}_n)$, which is therefore called the sufficient statistic of the distribution

Sampling methods

Question Verify the above MLE for multivariate Gaussian distribution

Sampling methods

Question Verify the above MLE for multivariate Gaussian distribution

The sufficient statistics are:

$$\sum_{n=1}^{N} \mathbf{x}_n \text{ and } \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T$$

Sampling methods

Question Verify the above MLE for multivariate Gaussian distribution

The sufficient statistics are:

$$\sum_{n=1}^{N} \mathbf{x}_n \text{ and } \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T$$

The ML estimates are:

$$\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n, \text{ and } \boldsymbol{\Sigma}_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T$$

Sampling methods

Observation Note that the partition function $g(\boldsymbol{\eta})$ for the exponential family

$$p(\mathbf{x};\boldsymbol{\eta}) = g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\}h(\mathbf{x}),$$

normalizes the distribution.

Sampling methods

Observation Note that the partition function $g(\boldsymbol{\eta})$ for the exponential family

$$p(\mathbf{x};\boldsymbol{\eta}) = g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\}h(\mathbf{x}),$$

normalizes the distribution.
Then the unnormalized distribution

$$\widetilde{p}(\mathbf{x},\boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right)h(\mathbf{x})$$

Sampling methods

Observation Note that the partition function $g(\boldsymbol{\eta})$ for the exponential family

$$p(\mathbf{x};\boldsymbol{\eta}) = g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\}h(\mathbf{x}),$$

normalizes the distribution.
Then the unnormalized distribution

$$\widetilde{p}(\mathbf{x},\boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\right)h(\mathbf{x})$$

Then

$$\ln\widetilde{p}(\mathbf{x},\boldsymbol{\eta}) = \boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}) + \ln h(\mathbf{x})$$

is known as energy function, which is linear in $\boldsymbol{\eta}$.

Sampling methods

Observation Note that the partition function $g(\boldsymbol{\eta})$ for the exponential family

$$p(\mathbf{x}; \boldsymbol{\eta}) = g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\} h(\mathbf{x}),$$

normalizes the distribution.
Then the unnormalized distribution

$$\widetilde{p}(\mathbf{x}, \boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right) h(\mathbf{x})$$

Then

$$\ln \widetilde{p}(\mathbf{x}, \boldsymbol{\eta}) = \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) + \ln h(\mathbf{x})$$

is known as energy function, which is linear in $\boldsymbol{\eta}$.
Thus the $p(\mathbf{x}; \boldsymbol{\eta})$ is referred as log-linear