Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 18
March 15, 2023

Monte Carlo methods

Observation from LDA The posterior distribution is a primary concept to make predictions. Finding $\mathbb{E}[L(\mathbf{x})] = \mathbb{E}[\mathbf{b}^T\mathbf{x}]$ and $Var[L(\mathbf{x})]$ is crucial

Monte Carlo methods

Observation from LDA The posterior distribution is a primary concept to make predictions. Finding $\mathbb{E}[L(\mathbf{x})] = \mathbb{E}[\mathbf{b}^T\mathbf{x}]$ and $Var[L(\mathbf{x})]$ is crucial

Problem Find the expected value of a function $f(\mathbf{x})$ with respect to a probability distribution $p(\mathbf{x})$ using a numerical sampling!

Monte Carlo methods

Observation from LDA The posterior distribution is a primary concept to make predictions. Finding $\mathbb{E}[L(\mathbf{x})] = \mathbb{E}[\mathbf{b}^T\mathbf{x}]$ and $Var[L(\mathbf{x})]$ is crucial

Problem Find the expected value of a function $f(\mathbf{x})$ with respect to a probability distribution $p(\mathbf{x})$ using a numerical sampling!

$\rightarrow$ If f is continuous then $\mathbb{E}[f] = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ and replace the integral by summation for discrete random variables

Monte Carlo methods

Observation from LDA The posterior distribution is a primary concept to make predictions. Finding $\mathbb{E}[L(\mathbf{x})] = \mathbb{E}[\mathbf{b}^T\mathbf{x}]$ and $Var[L(\mathbf{x})]$ is crucial

Problem Find the expected value of a function $f(\mathbf{x})$ with respect to a probability distribution $p(\mathbf{x})$ using a numerical sampling!

- $\rightarrow$ If f is continuous then $\mathbb{E}[f] = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ and replace the integral by summation for discrete random variables
- $\rightarrow$ The idea of sampling methods is to obtain a set of samples from $\{\mathbf{x}_l : 1 \leq l \leq L\}$ drawn independently from the distribution $p(\mathbf{x})$

Monte Carlo methods

Observation from LDA The posterior distribution is a primary concept to make predictions. Finding $\mathbb{E}[L(\mathbf{x})] = \mathbb{E}[\mathbf{b}^T\mathbf{x}]$ and $Var[L(\mathbf{x})]$ is crucial

Problem Find the expected value of a function $f(\mathbf{x})$ with respect to a probability distribution $p(\mathbf{x})$ using a numerical sampling!

- → If f is continuous then $\mathbb{E}[f] = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ and replace the integral by summation for discrete random variables
- → The idea of sampling methods is to obtain a set of samples from $\{\mathbf{x}_l : 1 \leq l \leq L\}$ drawn independently from the distribution $p(\mathbf{x})$
- → Then the expectation can be approximated by a finite sum

$$\widehat{f} = \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{x}_l)$$

Monte Carlo methods

Observation from LDA The posterior distribution is a primary concept to make predictions. Finding $\mathbb{E}[L(\mathbf{x})] = \mathbb{E}[\mathbf{b}^T\mathbf{x}]$ and $Var[L(\mathbf{x})]$ is crucial

Problem Find the expected value of a function $f(\mathbf{x})$ with respect to a probability distribution $p(\mathbf{x})$ using a numerical sampling!

→ If f is continuous then $\mathbb{E}[f] = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ and replace the integral by summation for discrete random variables

→ The idea of sampling methods is to obtain a set of samples from $\{\mathbf{x}_l : 1 \leq l \leq L\}$ drawn independently from the distribution $p(\mathbf{x})$

→ Then the expectation can be approximated by a finite sum

$$\widehat{f} = \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{x}_l)$$

→ The variance of this estimator is

$$var[\widehat{f}] = \frac{1}{L}\mathbb{E}[(f - \mathbb{E}[f])^2]$$

Sampling methods

Note

$\rightarrow$ Accuracy of the estimator does not depend on the dimension of $\mathbf{x}$

Sampling methods

Note

- $\rightarrow$ Accuracy of the estimator does not depend on the dimension of $\mathbf{x}$
- $\rightarrow$ In principle, high accuracy may be achievable with a relatively small number of samples $\mathbf{x}_l$

Sampling methods

Note

- $\rightarrow$ Accuracy of the estimator does not depend on the dimension of $\mathbf{x}$
- $\rightarrow$ In principle, high accuracy may be achievable with a relatively small number of samples $\mathbf{x}_l$

Problem Are the samples independent?

Sampling methods

Graphical model A convenient representation of joint distribution $p(\mathbf{x})$

Sampling methods

Graphical model A convenient representation of joint distribution $p(\mathbf{x})$

$\rightarrow$ Graph: $G = (V, E)$, where V is called the vertex set, $E \subset V \times V$ is called the edge set

Sampling methods

Graphical model A convenient representation of joint distribution $p(\mathbf{x})$

- $\rightarrow$ Graph: $G = (V, E)$, where V is called the vertex set, $E \subset V \times V$ is called the edge set
- $\rightarrow$ The nodes represent random variables, and edges represent probabilistic relationship between those random variables

Sampling methods

Graphical model A convenient representation of joint distribution $p(\mathbf{x})$

- $\rightarrow$ Graph: $G = (V, E)$, where V is called the vertex set, $E \subset V \times V$ is called the edge set
- $\rightarrow$ The nodes represent random variables, and edges represent probabilistic relationship between those random variables
- $\rightarrow$ The graph captures the way in which a joint distribution over all the random variables can be decomposed into a product of factors each depending only on a subset of random variables

Sampling methods

Graphical model A convenient representation of joint distribution $p(\mathbf{x})$

- $\rightarrow$ Graph: $G = (V, E)$, where V is called the vertex set, $E \subset V \times V$ is called the edge set
- $\rightarrow$ The nodes represent random variables, and edges represent probabilistic relationship between those random variables
- $\rightarrow$ The graph captures the way in which a joint distribution over all the random variables can be decomposed into a product of factors each depending only on a subset of random variables

Belief networks/Bayesian networks/directed graphical models - the edges in these graphs are directed with other certain structure. The edges express causal relationships between the variables

Sampling methods

Graphical model A convenient representation of joint distribution $p(\mathbf{x})$

$\rightarrow$ Graph: $G = (V, E)$, where V is called the vertex set, $E \subset V \times V$ is called the edge set

$\rightarrow$ The nodes represent random variables, and edges represent probabilistic relationship between those random variables

$\rightarrow$ The graph captures the way in which a joint distribution over all the random variables can be decomposed into a product of factors each depending only on a subset of random variables

Belief networks/Bayesian networks/directed graphical models - the edges in these graphs are directed with other certain structure. The edges express causal relationships between the variables

Markov random fields/undirected graphical models - the edges express constraints between random variables

Sampling methods

Bayesian networks Consider a joint random variables $\mathbf{X} = (X_1, X_2, X_3)$. Then by product rule:

$$p(a, b, c) \quad = \quad p(X_1 = a, X_2 = b, X_3 = c) = p(c|a, b)\, p(a, b)$$

Sampling methods

Bayesian networks Consider a joint random variables $\mathbf{X} = (X_1, X_2, X_3)$. Then by product rule:

$$\begin{aligned} p(a, b, c) &= p(X_1 = a, X_2 = b, X_3 = c) = p(c|a, b)\, p(a, b) \\ &= p(c|a, b)\, p(b|a)\, p(a) \end{aligned}$$

Sampling methods

Bayesian networks Consider a joint random variables $\mathbf{X} = (X_1, X_2, X_3)$. Then by product rule:

$$\begin{aligned} p(a,b,c) &= p(X_1 = a, X_2 = b, X_3 = c) = p(c|a,b)\, p(a,b) \\ &= p(c|a,b)\, p(b|a)\, p(a) \end{aligned}$$

This is valid for any choice of the distribution. The graphical model representation of this joint distribution is given by:
For each conditional distribution, we add directed edge to the graph from the nodes corresponding to the random variables on which the distribution is conditional

Sampling methods

Thus a link going from a node a to a node b, we say that node a is the parent of node b, and we say that the node b is the child of a

Sampling methods

Thus a link going from a node a to a node b, we say that node a is the parent of node b, and we say that the node b is the child of a

From $p(a, b, c) = p(c|a, b)\, p(b|a)\, p(a)$, observe that the left-hand side is symmetrical wrt the variables a, b, c but the right-hand side is not.

Sampling methods

Thus a link going from a node a to a node b, we say that node a is the parent of node b, and we say that the node b is the child of a

From $p(a, b, c) = p(c|a, b)\, p(b|a)\, p(a)$, observe that the left-hand side is symmetrical wrt the variables a, b, c but the right-hand side is not. In general, for K random variables,

$$p(X_1, \ldots, X_K) = p(X_K|X_1, \ldots, X_{K-1}) \ldots p(X_2|X_1)\, p(X_1)$$

Sampling methods

Thus a link going from a node a to a node b, we say that node a is the parent of node b, and we say that the node b is the child of a

From $p(a, b, c) = p(c|a, b)\, p(b|a)\, p(a)$, observe that the left-hand side is symmetrical wrt the variables a, b, c but the right-hand side is not. In general, for K random variables,

$$p(X_1, \ldots, X_K) = p(X_K|X_1, \ldots, X_{K-1}) \ldots p(X_2|X_1)\, p(X_1)$$

Then observe that, for any K, the corresponding Graphical model is fully connected, i.e. any two pair of nodes there is an link connecting them.

Sampling methods

Consider the graph

$$p(\mathbf{X}) = p(X_1)p(X_2)p(X_3)p(X_4|X_1, X_2, X_3)p(X_5|X_1, X_3)p(X_6|X_4)p(X_7|X_4, X_5)$$

Sampling methods

Consider the graph

$$p(\mathbf{X}) = p(X_1)p(X_2)p(X_3)p(X_4|X_1, X_2, X_3)p(X_5|X_1, X_3)p(X_6|X_4)p(X_7|X_4, X_5)$$

For any graph with K nodes, the joint distribution is given by

$$p(\mathbf{X}) = \prod_{k=1}^{K} p(X_k|pa_k)$$

where pa_k denotes the set of parents of X_k, and $\mathbf{X} = (X_1, \ldots, X_K)$

Sampling methods

Directed acyclic graphs (DAGs) - a directed graph without cycles i.e. there are no closed paths within the graph

Sampling methods

Directed acyclic graphs (DAGs) - a directed graph without cycles i.e. there are no closed paths within the graph

Ancestral sampling Suppose that the random variables have been ordered such that there are no links from any node to any lower numbered node i.e. each node has a higher number than any of its parents.

Sampling methods

Directed acyclic graphs (DAGs) - a directed graph without cycles i.e. there are no closed paths within the graph

Ancestral sampling Suppose that the random variables have been ordered such that there are no links from any node to any lower numbered node i.e. each node has a higher number than any of its parents.

Goal: Draw a sample $\widehat{X}_1, \ldots, \widehat{X}_K$ from the joint distribution, starting from the lowest-numbered node

Sampling methods

Directed acyclic graphs (DAGs) - a directed graph without cycles i.e. there are no closed paths within the graph

Ancestral sampling Suppose that the random variables have been ordered such that there are no links from any node to any lower numbered node i.e. each node has a higher number than any of its parents.

Goal: Draw a sample $\widehat{X}_1, \ldots, \widehat{X}_K$ from the joint distribution, starting from the lowest-numbered node

In practical applications, the higher numbered nodes correspond to the observed values and the lower numbered nodes correspond to the latent variables. For example, if an image is considered as an observation then object, position, and orientation will be the latent variables. (Draw a graphical model!)

Sampling methods

$\rightarrow$ Draw a sample from the distribution $p(X_1)$, call it $\widehat{x}_1$

Sampling methods

→ Draw a sample from the distribution $p(X_1)$, call it $\widehat{x}_1$

→ Do the same for each conditional distribution $p(X_k|pa_k)$ attached with all the nodes sequentially in order

Sampling methods

- $\rightarrow$ Draw a sample from the distribution $p(X_1)$, call it $\widehat{x}_1$
- $\rightarrow$ Do the same for each conditional distribution $p(X_k|pa_k)$ attached with all the nodes sequentially in order
- $\rightarrow$ Once we have sampled from the final random variable x_K, we will have achieved our objective of obtaining a sample from the joint distribution

Sampling methods

- → Draw a sample from the distribution $p(X_1)$, call it $\widehat{x}_1$
- → Do the same for each conditional distribution $p(X_k|pa_k)$ attached with all the nodes sequentially in order
- → Once we have sampled from the final random variable x_K, we will have achieved our objective of obtaining a sample from the joint distribution
- → To obtain a sample from some marginal distribution corresponding to a subset of variables, the sampled values for the required nodes are taken out ignoring the samples values for the remaining nodes

  For example, to draw a sample from a distribution $p(x_2, x_4)$, we sample from the full joint distribution and then retain the values $\widehat{x}_2, \widehat{x}_4$ and discard the remaining values $\{\widehat{x}_{j\neq 2,4}\}$

Monte Carlo methods

Linear-Gaussian models - a multivariate Gaussian can be expressed as a directed graph corresponding to a linear-Gaussian model

Monte Carlo methods

Linear-Gaussian models - a multivariate Gaussian can be expressed as a directed graph corresponding to a linear-Gaussian model

Consider an arbitrary directed acyclic graph with n nodes, which is attached with a random variable X_i having Gaussian distribution. The mean of this distribution is taken to be a linear combination of the states of the parent nodes pa_i of node i:

$$p(X_i|pa_i) = \mathcal{N}\left(X_i \,\middle|\, \sum_{j \in pa_i} w_{ij}x_j + b_i, v_i\right),$$

where w_{ij} and b_i are parameters governing the mean, and v_i is the variance of the conditional distribution for X_i

Sampling methods

Then setting $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Then

$$\begin{aligned}\ln p(\mathbf{x}) &= \sum_{i=1}^{n} \ln p(x_i | pa_i) \\ &= -\sum_{i=1}^{n} \frac{1}{2v_i}\left(x_i - \sum_{j\in pa_i} w_{ij}x_j - b_i\right)^2 + \text{constant}\end{aligned}$$

Sampling methods

Then setting $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Then

$$\begin{aligned}\ln p(\mathbf{x}) &= \sum_{i=1}^{n} \ln p(x_i|pa_i) \\ &= -\sum_{i=1}^{n} \frac{1}{2v_i}\left(x_i - \sum_{j\in pa_i} w_{ij}x_j - b_i\right)^2 + \text{constant}\end{aligned}$$

Finding mean and covariance of the joint distribution recursively:

$\rightarrow$ From the expression of $p(X_i|pa_i)$,

$$X_i = \sum_{j\in pa_i} w_{ij}X_j + b_i + \sqrt{v_i}\epsilon_i,$$

where ϵ_i is a zero mean, unit variance Gaussian rv satisfying $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i\epsilon_j] = I_{ij}$, the ij-th entry of the identity matrix

Sampling methods

Then setting $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Then

$$\begin{aligned}\ln p(\mathbf{x}) &= \sum_{i=1}^{n} \ln p(x_i|pa_i) \\ &= -\sum_{i=1}^{n} \frac{1}{2v_i}\left(x_i - \sum_{j\in pa_i} w_{ij}x_j - b_i\right)^2 + \text{constant}\end{aligned}$$

Finding mean and covariance of the joint distribution recursively:

→ From the expression of $p(X_i|pa_i)$,

$$X_i = \sum_{j\in pa_i} w_{ij}X_j + b_i + \sqrt{v_i}\epsilon_i,$$

where ϵ_i is a zero mean, unit variance Gaussian rv satisfying $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i\epsilon_j] = I_{ij}$, the ij-th entry of the identity matrix

→ Then

$$\mathbb{E}[X_i] = \sum_{j\in pa_i} w_{ij}\mathbb{E}[X_j] + b_i$$

Sampling methods

Further

$$
\begin{aligned}
Cov[X_i, X_j] &= \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] \\
&= \mathbb{E}\left[(X_i - \mathbb{E}[X_i])\left\{\sum_{k\in pa_j} w_{jk}(X_k - \mathbb{E}(X_k)) + \sqrt{v_j}\epsilon_j\right\}\right] \\
&= \sum_{k\in pa_j} w_{jk}\, Cov[X_i, X_k] + I_{ij} v_j
\end{aligned}
$$

Thus covariance can similarly be evaluated recursively starting from the lowest numbered node

Sampling methods

Example Suppose the graph corresponding to a graphical model has no links and only with n isolated nodes i.e. random variables.

$\rightarrow$ There are no w_{ij} and hence only n parameters v_i

Sampling methods

Example Suppose the graph corresponding to a graphical model has no links and only with n isolated nodes i.e. random variables.

- $\rightarrow$ There are no w_{ij} and hence only n parameters v_i
- $\rightarrow$ From the recursive relations of $\mathbb{E}[X_i]$ and $Cov[X_i, X_j]$, we see that the mean of $\mathbf{X}$ is $(b_1, b_2, \ldots, b_n)$

Sampling methods

Example Suppose the graph corresponding to a graphical model has no links and only with n isolated nodes i.e. random variables.

- $\rightarrow$ There are no w_{ij} and hence only n parameters v_i
- $\rightarrow$ From the recursive relations of $\mathbb{E}[X_i]$ and $Cov[X_i, X_j]$, we see that the mean of $\mathbf{X}$ is $(b_1, b_2, \ldots, b_n)$
- $\rightarrow$ The covariance matrix is diagonal of the form $diag(v_1, \ldots, v_n)$

Sampling methods

Example Suppose the graph corresponding to a graphical model has no links and only with n isolated nodes i.e. random variables.

- $\rightarrow$ There are no w_{ij} and hence only n parameters v_i
- $\rightarrow$ From the recursive relations of $\mathbb{E}[X_i]$ and $Cov[X_i, X_j]$, we see that the mean of $\mathbf{X}$ is $(b_1, b_2, \ldots, b_n)$
- $\rightarrow$ The covariance matrix is diagonal of the form $diag(v_1, \ldots, v_n)$
- $\rightarrow$ The joint distribution has $2n$ parameters and represents a set of n independent univariate Gaussian distributions

Sampling methods

Example Suppose the graph corresponding to a Graphical model is fully connected i.e. each node has all lower numbered nodes as parents

- $\rightarrow$ The matrix defined by w_{ij} has $i-1$ nonzero entries in each row with the diagonal entry 0

Sampling methods

Example Suppose the graph corresponding to a Graphical model is fully connected i.e. each node has all lower numbered nodes as parents

- $\rightarrow$ The matrix defined by w_{ij} has $i - 1$ nonzero entries in each row with the diagonal entry 0
- $\rightarrow$ The total number of parameters w_{ij} is $n(n-1)/2$

Sampling methods

Example Suppose the graph corresponding to a Graphical model is fully connected i.e. each node has all lower numbered nodes as parents

- $\rightarrow$ The matrix defined by w_{ij} has $i-1$ nonzero entries in each row with the diagonal entry 0
- $\rightarrow$ The total number of parameters w_{ij} is $n(n-1)/2$
- $\rightarrow$ The total number of independent parameters w_{ij} and v_i is then $n(n+1)/2$ corresponding to a general symmetric matrix

Sampling methods

Example Suppose the graph corresponding to a Graphical model is fully connected i.e. each node has all lower numbered nodes as parents

- $\rightarrow$ The matrix defined by w_{ij} has $i-1$ nonzero entries in each row with the diagonal entry 0
- $\rightarrow$ The total number of parameters w_{ij} is $n(n-1)/2$
- $\rightarrow$ The total number of independent parameters w_{ij} and v_i is then $n(n+1)/2$ corresponding to a general symmetric matrix

Example

$$\begin{aligned}
\boldsymbol{\mu} &= (b_1, b_2 + w_{21}b_1, b_3 + w_{32}b_2 + w_{32}w_{21}b_1) \\
\boldsymbol{\Sigma} &= \begin{bmatrix} v_1 & w_{21}v_1 & w_{32}w_{21}v_1 \\ w_{21}v_1 & v_2 + w_{21}^2 v_1 & w_{32}(v_2 + w_{21}^2 v_1) \\ w_{32}w_{21}v_1 & w_{32}(v_2 + w_{21}^2 v_1) & v_3 + w_{32}^2(v_2 + w_{21}^2 v_1) \end{bmatrix}
\end{aligned}$$