Pattern Recognition And Machine Learning
8.1 Bayesian Networks

Hyewon Park

University of Seoul, Statistics

2019-01-22

Outline

- Probabilistic graphical models
  - Polynomial regression
- Generative models
  - Ancestral sampling
- Discrete variables
- Linear-Gaussian models

Probabilistic Graphical Models

A Bayesian polynomial regression model

- How do we use directed graphs to describe probability distributions?

A Bayesian polynomial regression model

- How do we use directed graphs to describe probability distributions?
- We consider a Bayesian polynomial regression model

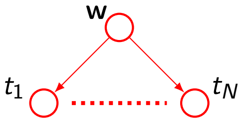
$$p(\mathbf{w} \mid \alpha) = N(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}).$$

  - $\mathbf{w}$: A vector of polynomial coefficients.
  - α : Precision of the Gaussian prior over $\mathbf{w}$.
  - $\mathbf{t} = (t_1, \dots t_N)^T$: Observed data.

A Bayesian polynomial regression model

- Focussing just on the random variables.

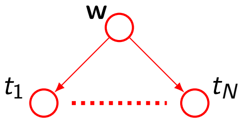
$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n \mid \mathbf{w}).$$



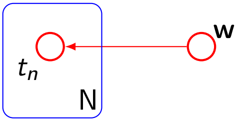
A Bayesian polynomial regression model

- Focussing just on the random variables.

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n \mid \mathbf{w}).$$



- We draw a single representative node t_n and then surround this with a box, called a plate.



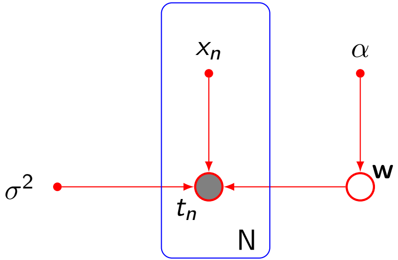
  - N indicating that there are N nodes of this kind.

A Bayesian polynomial regression model

- Consider the parameters and stochastic variables of a model.

$$p(\mathbf{t}, \mathbf{w} \mid \mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w} \mid \alpha) \prod_{n=1}^{N} p(t_n \mid \mathbf{w}, x_n, \sigma^2).$$

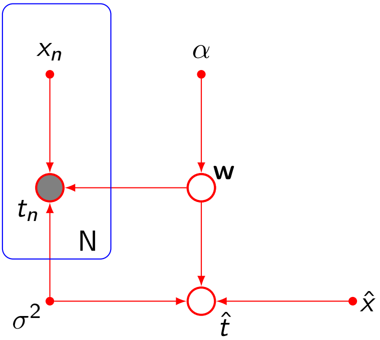
  - $\mathbf{x} = (x_1, \ldots, x_N)^T$: Input data.
  - $\mathbf{w}$: Not observed, latent variables.
  - σ^2 : Noise variance.
- We will denote observed variables by shading the corresponding nodes.



Prediction for a new input value

- $\hat{x}$: A new input value.

$$p(\hat{t}, \mathbf{t}, \mathbf{w} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n \mid x_n, \mathbf{w}, \sigma^2)\right] p(\mathbf{w} \mid \alpha) p(\hat{t} \mid \hat{x}, \mathbf{w}, \sigma^2)$$



Generative Models

Ancestral sampling

- Consider a joint distribution $p(x_1, \dots, x_K)$.
- The joint distribution factorizes,

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k \mid \mathrm{pa}_k).$$

- Each node has a higher number than any of its parents.
  - Example)

$$x_1 \rightarrow x_2 \rightarrow x_3$$

- Our goal : To draw a sample $\hat{x}_1, \dots, \hat{x}_K$ form the joint distribution.

Ancestral sampling from the joint distribution

1. We draw a sample from the distribution $p(x_1)$, which we call $\hat{x}_1$.

Ancestral sampling from the joint distribution

1. We draw a sample from the distribution $p(x_1)$, which we call $\hat{x}_1$.
2. We then work through each of the nodes in order.

   For node n, we draw a sample from the conditional distribution $p(x_n \mid \mathrm{pa}_n)$.

Ancestral sampling from the joint distribution

1. We draw a sample from the distribution $p(x_1)$, which we call $\hat{x}_1$.
2. We then work through each of the nodes in order.

   For node n, we draw a sample from the conditional distribution $p(x_n \mid \mathrm{pa}_n)$.
3. We have sample from the final variable x_K, we will have achieved our objective of obtaining a sample from the joint distribution!

Ancestral sampling from the marginal distribution

- We simply take the sampled values for the required nodes and ignore the sampled values for the remaining nodes.
- Example) Draw a sample from the distribution $p(x_2, x_4)$.
  1. Sample from the full joint distribution.
  2. Retain the values $\hat{x}_2, \hat{x}_4$.
  3. Discard the remaining values $\{\hat{x}_{j \neq 2,4}\}$.

Practical applications of probabilistic models

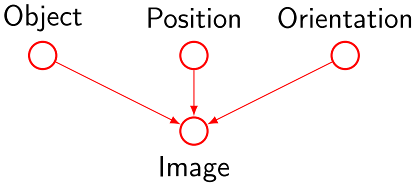
- Higher numbered variables : Terminal nodes of graph, observations.
- Lower numbered variables : Latent variables.
- The primary role of the latent variables.
  - To allow a complicated distribution over the observed variables to be represented in terms of a model constructed from simpler conditional distributions.
  - Typically exponential family.

Causal model

- We can interpret such models as expressing the processes by which the observed data arose.
- Example) An object recognition task.
  - Observed data point : An image of one of the objects.
  - Latent variables : The position and orientation of the object.
  - Our goal : To find the posterior distribution over objects.
    - We integrate over all possible positions and orentations.

Generative model

- We can represent this problem using graphical model.



- The graphical model captures the causal process by which the observed data was generated.
- For this reason, such models are often called generative models.
  - It is possible to generate synthetic data points from this model.

Discrete Variables

A single discrete variable

- The probability distribution $p(\mathbf{x} \mid \boldsymbol{\mu})$ for a single discrete variable $\mathbf{x}$ having K possible states (One-hot encoding).
  - Example) $K = 6,\ x_3 = 1 \rightarrow \mathbf{x} = (0, 0, 1, 0, 0, 0)^T$.

$$p(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}, \ \ \sum_{k=1}^{K} \mu_k = 1.$$

- $K - 1$ values for μ_k need to be specified in order to define the distribution.

Two discrete variables

- We have two discrete variables, x_1 and x_2, each of which has K states.
- We wish to model their joint distribution.

$$p(\mathbf{x}_1, \mathbf{x}_2 \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \prod_{l=1}^{K} \mu_{kl}^{x_{1k} x_{2l}}$$

  - μ_{kl} : The probability of ovserving both $x_{1k} = 1$ and $x_{2l} = 1$.
  - x_{1k} : The k^{th} component of $\mathbf{x}_1$.
  - x_{2l} : The l^{th} component of $\mathbf{x}_2$.
  - $\sum_k \sum_l \mu_{kl} = 1$.
- This distribution is governed by $K^2 - 1$ parameters.

Two discrete variables

- We have two discrete variables, x_1 and x_2, each of which has K states.
- We wish to model their joint distribution.

$$p(\mathbf{x}_1, \mathbf{x}_2 \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \prod_{l=1}^{K} \mu_{kl}^{x_{1k} x_{2l}}$$

  - μ_{kl} : The probability of ovserving both $x_{1k} = 1$ and $x_{2l} = 1$.
  - x_{1k} : The k^{th} component of $\mathbf{x}_1$.
  - x_{2l} : The l^{th} component of $\mathbf{x}_2$.
  - $\sum_k \sum_l \mu_{kl} = 1$.
- This distribution is governed by $K^2 - 1$ parameters.
- The total number of parameters that must be specified for an arbitrary joint distribution over M variables is $K^M - 1$.

Two discrete variables

- Using the product rule, $p(\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x}_2 \mid \mathbf{x}_1)p(\mathbf{x}_1)$.
  - The marginal distribution $p(\mathbf{x}_1)$ is governed by $K - 1$ parameters.
  - the conditional distribution $p(\mathbf{x}_2 \mid \mathbf{x}_1)$ requires the specification of $K - 1$ parameters for each of the K possible values of $\mathbf{x}_1$.
  - The total number of parameters : $(K - 1) + K(K - 1) = K^2 - 1$.

Two discrete variables

- Using the product rule, $p(\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x}_2 \mid \mathbf{x}_1)p(\mathbf{x}_1)$.
  - The marginal distribution $p(\mathbf{x}_1)$ is governed by $K - 1$ parameters.
  - the conditional distribution $p(\mathbf{x}_2 \mid \mathbf{x}_1)$ requires the specification of $K - 1$ parameters for each of the K possible values of $\mathbf{x}_1$.
  - The total number of parameters : $(K - 1) + K(K - 1) = K^2 - 1$.

- Suppose that the variables $\mathbf{x}_1$, and $\mathbf{x}_2$ are independent.
  - Each variable is then described by a separate multinomial distribution.
  - The total number of parameters : $2(K - 1)$.

General case

- We have M discrete variables $\mathbf{x}_1, \ldots \mathbf{x}_M$.
- Total number of parameters
  - Fully connected graph : $K^M - 1$
  - No links in the graph : $M(K-1)$

General case

- We have M discrete variables $\mathbf{x}_1, \dots \mathbf{x}_M$.
- Total number of parameters
  - Fully connected graph : $K^M - 1$
  - No links in the graph : $M(K-1)$
- Chain of nodes : $(K-1) + (M-1)K(K-1)$

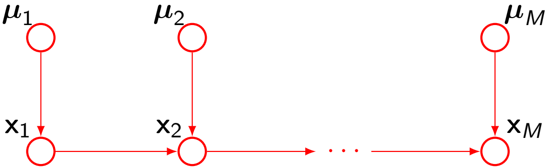
  - $p(\mathbf{x}_1)$ requires $K-1$ parameters.
  - Each of the $M-1$ conditional distributions $p(\mathbf{x}_i|\mathbf{x}_{i-1})$, for $i = 2, \dots, M$, requires $K(K-1)$ parameters.

A Bayesian model

- Using Dirichlet priors for parameters.
  - prior distribution of the multinomial distribution.

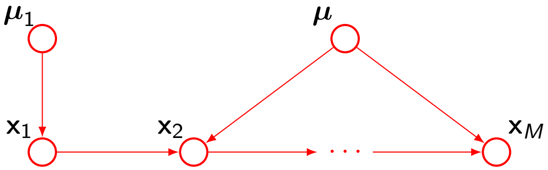
$$\frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

  - $\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)},\ \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$

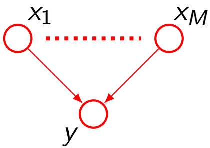


Reduce the number of independent parameters - Sharing

- Also known as tying of parameters.
- A single set of parameters $\boldsymbol{\mu}$ shared amongst all of the conditional distributions $p(\mathbf{x}_i \mid \mathbf{x}_{i-1})$.
  - $p(\mathbf{x}_1)$ requires $K-1$ parameters.
  - All of the conditional distributions $p(\mathbf{x}_i|\mathbf{x}_{i-1})$, for $i = 2, \ldots, M$, are governed by same set of $K(K-1)$ parameters.
  - The total number of parameters : $K^2 - 1$



Reduce the number of independent parameters - Parameterized models for the conditional distributions



- All of the nodes represent binary variables.
- Each of the parent variables x_i is goverened by a single parameter μ_i representing the probability $p(x_i = 1)$.
- The conditional distribution $p(y \mid x_i, \ldots x_M)$ requires 2^M parameters.
- The number of parameters grows exponentially with M.

Reduce the number of independent parameters - Parameterized models for the conditional distributions

- Using a logistic sigmoid function acting on a linear combination of the parent variables.

$$p(y = 1 \mid x_1, \dots, x_M) = \sigma\left(w_0 + \sum_{i=1}^{M} w_i x_i\right) = \sigma(\mathbf{w}^T \mathbf{x})$$

  - $\sigma(a) = (1 + \exp(-a))^{-1}$, the losigtic sigmoid.
  - $\mathbf{x} = (x_0, x_1, \dots, x_M)^T$, $x_0 = 1$.
  - $\mathbf{w} = (w_0, w_1, \dots, w_M)^T$, a vector of $M + 1$ parameters.
- The number of parameters grows linearly with M.

Linear-Gaussian Models

Linear-Gaussian models

- Consider an arbitrary directed acyclic graph over D variables.
- The node i represents a single continuous random variable x_i having a Gaussian distribution.

$$p(x_i \mid \mathrm{pa}_i) = N\left(x_i \,\middle|\, \sum_{j \in \mathrm{pa}_i} w_{ij} x_j + b_i, v_i\right).$$

  - The mean of this distribution is taken to be a linear combination of the states of its parent nodes pa_i of node i.
  - w_{ij}, b_j : The parameters governing the mean.
  - v_i : The variance of the conditional distribution for x_i

Linear-Gaussian models

- The log of the joint distribution,

$$\ln p(\mathbf{x}) = \sum_{i=1}^{D} \ln p(x_i \mid \mathrm{pa}_i)$$

$$= -\sum_{i=1}^{D} \frac{1}{2v_i} \left(x_i - \sum_{j \in \mathrm{pa}_i} w_{ij} x_j - b_i \right)^2 + \text{const.}$$

  - $\mathbf{x} = (x_1, \ldots, x_D)^T$.
  - const : Terms independent of x.

- $p(\mathbf{x})$ is a multivariate Gaussian.

Linear-Gaussian models : Mean

- Each variable x_i (conditional on the states of its parents) has a Gaussian distribution.

$$x_i = \sum_{j \in \mathrm{pa}_i} w_{ij} x_j + b_i + \sqrt{v_i} \epsilon_i .$$

  - ϵ_i : A zero mean, unit variance Gaussian random variable.
  - $\mathbb{E}[\epsilon_i] = 0, \ \mathbb{E}[\epsilon_i \epsilon_j] = I_{ij}.$
- $\mathbb{E}[x_i] = \sum_{j \in \mathrm{pa}_i} w_{ij} \mathbb{E}[x_j] + b_i.$
- Thus we can find the $\mathbb{E}[\mathbf{x}] = (\mathbb{E}[x_1], \ldots, \mathbb{E}[x_D])^T.$

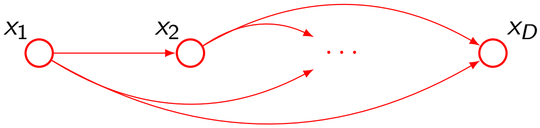
Linear-Gaussian models : Covariance

$$
\begin{aligned}
\mathrm{cov}[x_i, x_j] &= \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])] \\
&= \mathbb{E}\left[(x_i - \mathbb{E}[x_i])\left\{\sum_{k \in \mathrm{pa}_i} w_{jk}(x_k - \mathbb{E}[x_k]) + \sqrt{v_j}\epsilon_j\right\}\right] \\
&= \sum_{k \in \mathrm{pa}_i} w_{jk}\mathrm{cov}[x_i, x_k] + I_{ij}v_j.
\end{aligned}
$$

No links in the graph

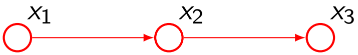
- There are no parameters w_{ij}.
- There are D parameters b_i and D parameters v_i
- $\mathbb{E}[p(\mathbf{x})] = (b_1, \ldots b_D)^T, \ \Sigma = diag(v_1, \ldots v_D)$.
- The total number of parameters : $2D$.

Fully connected graph



- Each node has all lower numbered nodes as parents.
- w_{ij} : A lower triangular matrix ($i-1$ entries on the i^{th} row).
- The total number of parameters w_{ij} : $\dfrac{D(D-1)}{2}$.
- The total number of parameters $\{w_{ij}\}$ and $\{v_i\}$: $\dfrac{D(D+1)}{2}$.
- The total number of parameters : $\dfrac{D(D+1)}{2} + D$.

Chain of nodes



$$\boldsymbol{\mu} = (b_1,\ b_2 + w_{21} b_1,\ b_3 + w_{32} b_2 + w_{32} w_{21} b_1)^T,$$

$$\Sigma = \begin{pmatrix} v_1 & w_{21} v_1 & w_{32} w_{21} v_1 \\ w_{21} v_1 & v_2 + w_{21}^2 v_1 & w_{32}(v_2 + w_{21}^2 v_1) \\ w_{32} w_{21} v_1 & w_{32}(v_2 + w_{21}^2 v_1) & v_3 + w_{32}^2 (v_2 + w_{21}^2 v_1) \end{pmatrix}.$$

- The total number of parameters w_{ij} : $D - 1$.
- The total number of parameters $\{w_{ij}\}$ and $\{v_i\}$: $2D - 1$.
- The total number of parameters : $3D - 1$.

Summary

- Probabilistic graphical models.
  - Plate : Multiple nodes are expressed more compactly.
  - Shading : Ovserved variables.
- Generative models.
  - The graphical model captures the causal process by which the observed data was generated.
  - Ancestral sampling.
- Discrete variables and linear-Gaussian models.
  - The total number of parameters.
  - Alternative way to reduce the number of independent parameters. Sharing, parameterized models.

Thank you!