Learning

Wilker Aziz

w.aziz@uva.nl

https://probabll.github.io

Outline and goals

Module 5 introduces *Learning Algorithms* for BNs and MNs (Chapters 17, ...).

ILOs After this module the student

- can estimate parameters for a BN via MLE;
- can estimate parameters for an MN via (approximations to) MLE;

Textbook for this course: Koller and Friedman [1].

Overview of Module 1

HC6: Parameter estimation for BNs.

LC1: Parameter estimation in code.

HC2: Parameter estimation for MNs.

WC1: exercises.

Table of contents

Parameter Estimation

Prescribing a Categorical Distribution

Let X be a categorical rv with outcomes in $\mathsf{Val}(X) = \{x^1, \ldots, x^K\}$.

Prescribing a Categorical Distribution

Let X be a categorical rv with outcomes in $\mathsf{Val}(X) = \{x^1, \ldots, x^K\}$.

If we model with **tabular CPDs**, we can say that there is a 'table' (rather a 'row vector') $\boldsymbol{\theta} = (\theta_{x^1}, \ldots, \theta_{x^K})$ which stores the K **parameters** needed to prescribe the distribution of X:

- for any $x \in \mathsf{Val}(X)$, $0 \leq \theta_x \leq 1$
- $\sum_{x \in \mathsf{Val}(X)} \theta_x = 1$

Then, under this choice, $P(X = x) = \theta_x$.

Prescribing a Categorical Distribution

Let X be a categorical rv with outcomes in $\mathsf{Val}(X) = \{x^1, \ldots, x^K\}$.

If we model with **tabular CPDs**, we can say that there is a 'table' (rather a 'row vector') $\boldsymbol{\theta} = (\theta_{x^1}, \ldots, \theta_{x^K})$ which stores the K **parameters** needed to prescribe the distribution of X:

- for any $x \in \mathsf{Val}(X)$, $0 \leq \theta_x \leq 1$
- $\sum_{x \in \mathsf{Val}(X)} \theta_x = 1$

Then, under this choice, $P(X = x) = \theta_x$.

Suppose we have a dataset $\mathcal{D} = \{x[1], \ldots, x[M]\}$ of M observations of realisations of X.

How can we use **data** to inform our choice of numerical values for the parameters $\boldsymbol{\theta}$?

Prescribing a Categorical Distribution

Let X be a categorical rv with outcomes in $\mathsf{Val}(X) = \{x^1, \dots, x^K\}$.

If we model with **tabular CPDs**, we can say that there is a 'table' (rather a 'row vector') $\boldsymbol{\theta} = (\theta_{x^1}, \dots, \theta_{x^K})$ which stores the K **parameters** needed to prescribe the distribution of X:

- for any $x \in \mathsf{Val}(X)$, $0 \leq \theta_x \leq 1$
- $\sum_{x \in \mathsf{Val}(X)} \theta_x = 1$

Then, under this choice, $P(X = x) = \theta_x$.

Suppose we have a dataset $\mathcal{D} = \{x[1], \dots, x[M]\}$ of M observations of realisations of X.

How can we use **data** to inform our choice of numerical values for the parameters $\boldsymbol{\theta}$? This is a **statistical inference** problem.

Frequentist Inference

First, characterise the model's likelihood function given the available data $\mathcal{D}$:

$$L(\boldsymbol{\theta}; \mathcal{D}) = \prod_{m=1}^{M} P(X = x[m]) = \prod_{m=1}^{M} \theta_{x[m]}$$

The likelihood function $L(\boldsymbol{\theta}; \mathcal{D})$ assigns 'worth' or 'utility' to a choice of parameter $\boldsymbol{\theta}$. This utility is defined to be the probability mass that our model assigns to $\mathcal{D}$ under the assumption that the data samples were drawn IID from our model using the current choice of $\boldsymbol{\theta}$.

Frequentist Inference

First, characterise the model's likelihood function given the available data $\mathcal{D}$:

$$L(\boldsymbol{\theta}; \mathcal{D}) = \prod_{m=1}^{M} P(X = x[m]) = \prod_{m=1}^{M} \theta_{x[m]}$$

The likelihood function $L(\boldsymbol{\theta}; \mathcal{D})$ assigns 'worth' or 'utility' to a choice of parameter $\boldsymbol{\theta}$. This utility is defined to be the probability mass that our model assigns to $\mathcal{D}$ under the assumption that the data samples were drawn IID from our model using the current choice of $\boldsymbol{\theta}$.

Second, pick the parameter value that yields maximum likelihood:

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta} \in \Delta_{K-1}}{\operatorname{argmax}} \; L(\boldsymbol{\theta}; \mathcal{D}) = \underset{\boldsymbol{\theta} \in \Delta_{K-1}}{\operatorname{argmax}} \; \underbrace{\log L(\boldsymbol{\theta}; \mathcal{D})}_{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})}$$

Frequentist Inference

First, characterise the model's likelihood function given the available data $\mathcal{D}$:

$$L(\boldsymbol{\theta}; \mathcal{D}) = \prod_{m=1}^{M} P(X = x[m]) = \prod_{m=1}^{M} \theta_{x[m]}$$

The likelihood function $L(\boldsymbol{\theta}; \mathcal{D})$ assigns 'worth' or 'utility' to a choice of parameter $\boldsymbol{\theta}$. This utility is defined to be the probability mass that our model assigns to $\mathcal{D}$ under the assumption that the data samples were drawn IID from our model using the current choice of $\boldsymbol{\theta}$.

Second, pick the parameter value that yields maximum likelihood:

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta} \in \Delta_{K-1}}{\operatorname{argmax}}\ L(\boldsymbol{\theta}; \mathcal{D}) = \underset{\boldsymbol{\theta} \in \Delta_{K-1}}{\operatorname{argmax}}\ \underbrace{\log L(\boldsymbol{\theta}; \mathcal{D})}_{\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})} = \sum_{m=1}^{M} \log \theta_{x[m]}$$

This is known as **maximum likelihood estimation (MLE).**

Bayesian Inference

Bayesian inference treats parameters as rvs on their own right.

It also uses the likelihood function, but in a very different way, as a means to update beliefs about parameters:

$$\underbrace{p(\boldsymbol{\theta}|\mathcal{D})}_{\text{posterior}} \propto \underbrace{p(\boldsymbol{\theta})}_{\text{prior}} \underbrace{L(\boldsymbol{\theta};\mathcal{D})}_{\text{likelihood}}$$

Rather than looking for a parameter value judged to be 'right' (or 'optimum')—a point estimate—Bayesian inference attempts to estimate parameters that are *probable* in light of the available data and model assumptions as captured by the likelihood function.

Bayesian Inference

Bayesian inference treats parameters as rvs on their own right.

It also uses the likelihood function, but in a very different way, as a means to update beliefs about parameters:

$$\underbrace{p(\boldsymbol{\theta}|\mathcal{D})}_{\text{posterior}} \propto \underbrace{p(\boldsymbol{\theta})}_{\text{prior}} \underbrace{L(\boldsymbol{\theta};\mathcal{D})}_{\text{likelihood}}$$

Rather than looking for a parameter value judged to be 'right' (or 'optimum')—a point estimate—Bayesian inference attempts to estimate parameters that are *probable* in light of the available data and model assumptions as captured by the likelihood function.

In this course, we will focus on Frequentist inference (not because it's 'right' or to be preferred in general, simply because Bayesian inference requires a longer course).

MLE for the Parameters of a Categorical Distribution

X is a categorical rv with outcomes in $\mathsf{Val}(X) = \{x^1, \ldots, x^K\}$. We model its distribution in tabular form, that is, we introduce a parameter vector $\boldsymbol{\theta}_X = (\theta_{x^1}, \ldots, \theta_{x^K})$ such that $P(X = x) = \theta_x$.

We have a dataset $\mathcal{D} = \{x[1], \ldots, x[M]\}$ of M observations of X.

Define the helper 'counting' function $M[o] = \sum_{m=1}^{M} [x[m] = o]$.

The Iverson bracket $[\alpha]$ is 1 if the logical predicate α is True and 0 otherwise. MLE for Categorical distribution: if you're curious I derived it here.

MLE for the Parameters of a Categorical Distribution

X is a categorical rv with outcomes in $\text{Val}(X) = \{x^1, \ldots, x^K\}$. We model its distribution in tabular form, that is, we introduce a parameter vector $\boldsymbol{\theta}_X = (\theta_{x^1}, \ldots, \theta_{x^K})$ such that $P(X = x) = \theta_x$.

We have a dataset $\mathcal{D} = \{x[1], \ldots, x[M]\}$ of M observations of X.

Define the helper 'counting' function $M[o] = \sum_{m=1}^{M} [x[m] = o]$.

The maximum likelihood estimate of θ_x for any $x \in \text{Val}(X)$ is given by

$$\theta_x = \frac{M[x]}{\sum_{x' \in \text{Val}(X)} M[x']} = \frac{M[x]}{M} \tag{1}$$

The Iverson bracket $[\alpha]$ is 1 if the logical predicate α is True and 0 otherwise. MLE for Categorical distribution: if you're curious I derived it here.

MLE for a Categorical Distribution – Example

The coherence C of a course may be high c^1 or low c^0. Let's do MLE for a tabular representation of $P(C)$.

Course	Votes
CS0001	$c^1, c^1, c^1, c^0, c^0, c^0, c^0, c^1, c^1, c^0$
CS0002	$c^1, c^1, c^1, c^0, c^0, c^0, c^1, c^1, c^1, c^1$

Observations from course evaluation surveys.

MLE for $P(C)$ using the data above: $\boldsymbol{\theta} = (\theta_{c^0}, \theta_{c^1}) = (\quad , \quad)$.

MLE for a Categorical Distribution – Example

The coherence C of a course may be high c^1 or low c^0. Let's do MLE for a tabular representation of $P(C)$.

Course	Votes
CS0001	$c^1, c^1, c^1, c^0, c^0, c^0, c^0, c^1, c^1, c^0$
CS0002	$c^1, c^1, c^1, c^0, c^0, c^0, c^1, c^1, c^1, c^1$

Observations from course evaluation surveys.

MLE for $P(C)$ using the data above: $\boldsymbol{\theta} = (\theta_{c^0}, \theta_{c^1}) = (8/20, 12/20)$.

MLE for the Parameters of a Tabular CPD

Now, we also have an rv Y taking on values in $\mathrm{Val}(Y) = \{y^1, \ldots, y^V\}$.

We are modelling $P(Y|X)$ in tabular representation. Hence, we introduce a collection of parameter vectors $\boldsymbol{\theta}_{Y|X} = (\boldsymbol{\theta}_{Y|x^1}, \ldots, \boldsymbol{\theta}_{Y|x^K})$.

- Each $\boldsymbol{\theta}_{Y|x}$ is a vector $(\theta_{y^1|x}, \ldots, \theta_{y^V|x})$ for some $x \in \mathrm{Val}(X)$
- such that $P(Y = y|X = x) = \theta_{y|x}$.

Now our observations are M pairs $\mathcal{D} = \{(x[1], y[1]), \ldots, (x[M], y[M])\}$.

Define a new 'counting' function $M[c, o] = \sum_{m=1}^{M} [x[m] = c][y[m] = o]$.

MLE for the Parameters of a Tabular CPD

Now, we also have an rv Y taking on values in $\mathrm{Val}(Y) = \{y^1, \ldots, y^V\}$.

We are modelling $P(Y|X)$ in tabular representation. Hence, we introduce a collection of parameter vectors $\boldsymbol{\theta}_{Y|X} = (\boldsymbol{\theta}_{Y|x^1}, \ldots, \boldsymbol{\theta}_{Y|x^K})$.

- Each $\boldsymbol{\theta}_{Y|x}$ is a vector $(\theta_{y^1|x}, \ldots, \theta_{y^V|x})$ for some $x \in \mathrm{Val}(X)$
- such that $P(Y = y|X = x) = \theta_{y|x}$.

Now our observations are M pairs $\mathcal{D} = \{(x[1], y[1]), \ldots, (x[M], y[M])\}$.

Define a new 'counting' function $M[c, o] = \sum_{m=1}^{M} [x[m] = c][y[m] = o]$.

The MLE of $\theta_{y|x}$ for any $x \in \mathrm{Val}(X)$ and $y \in \mathrm{Val}(Y)$ is given by

$$\theta_{y|x} = \frac{M[x, y]}{\sum_{y' \in \mathrm{Val}(Y)} M[x, y']} = \frac{M[x, y]}{M[x]} \tag{2}$$

MLE for a Categorical CPD – Example

The course may be difficult d^1 or easy d^0. Let's do MLE for a tabular representation of $P(D)$.

Course	Votes
CS0001	$(c^1, d^0), (c^1, d^0), (c^1, d^0), (c^0, d^1), (c^0, d^1)$
	$(c^0, d^0), (c^0, d^1), (c^1, d^0), (c^1, d^0), (c^0, d^1)$
CS0002	$(c^1, d^1), (c^1, d^1), (c^1, d^1), (c^0, d^1), (c^0, d^1)$
	$(c^0, d^1), (c^1, d^1), (c^1, d^1), (c^1, d^1), (c^1, d^0)$

Observations from course evaluation surveys.

MLE for $P(D|C)$ using the data above:

$$\boldsymbol{\theta}_{D|c^0} = (\theta_{d^0|c^0}, \theta_{d^1|c^0}) = (\quad , \quad)$$
$$\boldsymbol{\theta}_{D|c^1} = (\theta_{d^0|c^1}, \theta_{d^1|c^1}) = (\quad , \quad)$$

MLE for a Categorical CPD – Example

The course may be difficult d^1 or easy d^0. Let's do MLE for a tabular representation of $P(D)$.

Course	Votes
CS0001	$(c^1, d^0), (c^1, d^0), (c^1, d^0), (c^0, d^1), (c^0, d^1)$
	$(c^0, d^0), (c^0, d^1), (c^1, d^0), (c^1, d^0), (c^0, d^1)$
CS0002	$(c^1, d^1), (c^1, d^1), (c^1, d^1), (c^0, d^1), (c^0, d^1)$
	$(c^0, d^1), (c^1, d^1), (c^1, d^1), (c^1, d^1), (c^1, d^0)$

Observations from course evaluation surveys.

MLE for $P(D|C)$ using the data above:

$\boldsymbol{\theta}_{D|c^0} = (\theta_{d^0|c^0}, \theta_{d^1|c^0}) = (1/8, 7/8)$

$\boldsymbol{\theta}_{D|c^1} = (\theta_{d^0|c^1}, \theta_{d^1|c^1}) = (\quad , \quad)$

MLE for a Categorical CPD – Example

The course may be difficult d^1 or easy d^0. Let's do MLE for a tabular representation of $P(D)$.

Course	Votes
CS0001	$(c^1, d^0), (c^1, d^0), (c^1, d^0), (c^0, d^1), (c^0, d^1)$
	$(c^0, d^0), (c^0, d^1), (c^1, d^0), (c^1, d^0), (c^0, d^1)$
CS0002	$(c^1, d^1), (c^1, d^1), (c^1, d^1), (c^0, d^1), (c^0, d^1)$
	$(c^0, d^1), (c^1, d^1), (c^1, d^1), (c^1, d^1), (c^1, d^0)$

Observations from course evaluation surveys.

MLE for $P(D|C)$ using the data above:

$$\boldsymbol{\theta}_{D|c^0} = (\theta_{d^0|c^0}, \theta_{d^1|c^0}) = (1/8, 7/8)$$
$$\boldsymbol{\theta}_{D|c^1} = (\theta_{d^0|c^1}, \theta_{d^1|c^1}) = (6/12, 6/12)$$

MLE for Bayesian Networks

MLE for Bayesian Networks

In a BN we have a CPD for each node X_i given the node's parents $\mathrm{Pa}(X_i)$. For CPDs in tabular representation, this means we have a collection of parameters $\boldsymbol{\theta}_{X_i|\mathrm{Pa}(X_i)}$ for each node X_i.

MLE for Bayesian Networks

In a BN we have a CPD for each node X_i given the node's parents $\mathrm{Pa}(X_i)$. For CPDs in tabular representation, this means we have a collection of parameters $\boldsymbol{\theta}_{X_i|\mathrm{Pa}(X_i)}$ for each node X_i.

The likelihood $L(\boldsymbol{\theta}; \mathcal{D})$ of a BN is $L(\boldsymbol{\theta}; \mathcal{D}) = \prod_{m=1}^{M} P(\mathbf{X} = \mathbf{x}[m])$

$$= \prod_{m=1}^{M} \prod_{i} P(X_i = x[m]_i | \mathrm{Pa}(X_i) = \mathrm{pa}(x[m]_i))$$

MLE for Bayesian Networks

In a BN we have a CPD for each node X_i given the node's parents $\mathrm{Pa}(X_i)$. For CPDs in tabular representation, this means we have a collection of parameters $\boldsymbol{\theta}_{X_i|\mathrm{Pa}(X_i)}$ for each node X_i.

The likelihood $L(\boldsymbol{\theta};\mathcal{D})$ of a BN is $L(\boldsymbol{\theta};\mathcal{D}) = \prod_{m=1}^{M} P(\boldsymbol{X} = \boldsymbol{x}[m])$

$$= \prod_{m=1}^{M} \prod_{i} P(X_i = x[m]_i | \mathrm{Pa}(X_i) = \mathrm{pa}(x[m]_i))$$

$$= \prod_{i} \underbrace{\prod_{m=1}^{M} P(X_i = x[m]_i | \mathrm{Pa}(X_i) = \mathrm{pa}(x[m]_i))}_{L(\boldsymbol{\theta}_{X_i|\mathrm{Pa}(X_i)};\mathcal{D})}$$

decomposes in terms of **local likelihood functions**, one per X_i, where

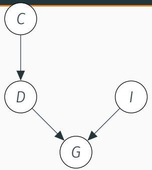
$$L(\boldsymbol{\theta}_{X_i|\mathrm{Pa}(X_i)};\mathcal{D}) = \prod_{m=1}^{M} \theta_{x[m]_i|\mathrm{pa}(x[m]_i)}$$

MLE for Bayesian Networks - Algorithm

Because the BN likelihood decomposes, so long as we have complete observations (that is, we learn by observing joint assignments of *all* rvs), we can solve independent MLE problems, one per rv.

For each rv X_i, the MLE for the parameters of the tabular CPD $P(X_i|\mathrm{Pa}(X_i))$ is

- $\theta_{o|\boldsymbol{u}} = \frac{M[\boldsymbol{u},o]}{M[\boldsymbol{u}]}$ for any $o \in \mathrm{Val}(X_i)$ and any $\boldsymbol{u} \in \mathrm{Val}(\mathrm{Pa}(X_i))$



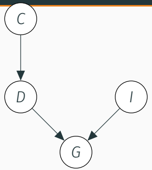
Student	Record
s1000	(c^0, d^1, i^1, g^1)
s1001	(c^0, d^0, i^1, g^1)
s1002	(c^0, d^1, i^0, g^1)
s1003	(c^1, d^0, i^1, g^2)
s1004	(c^1, d^0, i^1, g^2)
s1005	(c^1, d^0, i^0, g^1)
s1006	(c^1, d^0, i^0, g^3)
s1007	(c^0, d^1, i^1, g^3)
s1008	(c^0, d^1, i^1, g^2)
s1009	(c^1, d^0, i^1, g^1)

Observations from CS0001.

θ_{i^0}	θ_{i^1}

pa	$\theta_{g^1\|d,i}$	$\theta_{g^2\|d,i}$	$\theta_{g^3\|d,i}$
d^0, i^0			
d^0, i^1			
d^1, i^0			
d^1, i^1			

MLE for $P(I)$ and $P(G|I)$



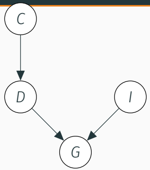
Student	Record
s1000	(c^0, d^1, i^1, g^1)
s1001	(c^0, d^0, i^1, g^1)
s1002	(c^0, d^1, i^0, g^1)
s1003	(c^1, d^0, i^1, g^2)
s1004	(c^1, d^0, i^1, g^2)
s1005	(c^1, d^0, i^0, g^1)
s1006	(c^1, d^0, i^0, g^3)
s1007	(c^0, d^1, i^1, g^3)
s1008	(c^0, d^1, i^1, g^2)
s1009	(c^1, d^0, i^1, g^1)

Observations from CS0001.

θ_{i^0}	θ_{i^1}
$^3/_{10}$	$^7/_{10}$

pa	$\theta_{g^1\|d,i}$	$\theta_{g^2\|d,i}$	$\theta_{g^3\|d,i}$
d^0, i^0			
d^0, i^1			
d^1, i^0			
d^1, i^1			

MLE for $P(I)$ and $P(G|I)$



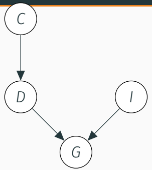
Student	Record
s1000	(c^0, d^1, i^1, g^1)
s1001	(c^0, d^0, i^1, g^1)
s1002	(c^0, d^1, i^0, g^1)
s1003	(c^1, d^0, i^1, g^2)
s1004	(c^1, d^0, i^1, g^2)
s1005	(c^1, d^0, i^0, g^1)
s1006	(c^1, d^0, i^0, g^3)
s1007	(c^0, d^1, i^1, g^3)
s1008	(c^0, d^1, i^1, g^2)
s1009	(c^1, d^0, i^1, g^1)

Observations from CS0001.

θ_{i^0}	θ_{i^1}
$^3/_{10}$	$^7/_{10}$

pa	$\theta_{g^1\|d,i}$	$\theta_{g^2\|d,i}$	$\theta_{g^3\|d,i}$
d^0, i^0	$^1/_2$	$^0/_2$	$^1/_2$
d^0, i^1			
d^1, i^0			
d^1, i^1			

MLE for $P(I)$ and $P(G|I)$



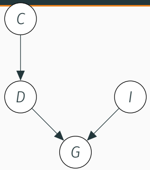
Student	Record
s1000	(c^0, d^1, i^1, g^1)
s1001	(c^0, d^0, i^1, g^1)
s1002	(c^0, d^1, i^0, g^1)
s1003	(c^1, d^0, i^1, g^2)
s1004	(c^1, d^0, i^1, g^2)
s1005	(c^1, d^0, i^0, g^1)
s1006	(c^1, d^0, i^0, g^3)
s1007	(c^0, d^1, i^1, g^3)
s1008	(c^0, d^1, i^1, g^2)
s1009	(c^1, d^0, i^1, g^1)

Observations from CS0001.

θ_{i^0}	θ_{i^1}
$3/10$	$7/10$

pa	$\theta_{g^1\|d,i}$	$\theta_{g^2\|d,i}$	$\theta_{g^3\|d,i}$
d^0, i^0	$1/2$	$0/2$	$1/2$
d^0, i^1	$2/3$	$1/3$	$0/3$
d^1, i^0			
d^1, i^1			

MLE for $P(I)$ and $P(G|I)$



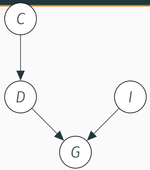
Student	Record
s1000	(c^0, d^1, i^1, g^1)
s1001	(c^0, d^0, i^1, g^1)
s1002	(c^0, d^1, i^0, g^1)
s1003	(c^1, d^0, i^1, g^2)
s1004	(c^1, d^0, i^1, g^2)
s1005	(c^1, d^0, i^0, g^1)
s1006	(c^1, d^0, i^0, g^3)
s1007	(c^0, d^1, i^1, g^3)
s1008	(c^0, d^1, i^1, g^2)
s1009	(c^1, d^0, i^1, g^1)

Observations from CS0001.

θ_{i^0}	θ_{i^1}
$3/10$	$7/10$

pa	$\theta_{g^1\|d,i}$	$\theta_{g^2\|d,i}$	$\theta_{g^3\|d,i}$
d^0, i^0	$1/2$	$0/2$	$1/2$
d^0, i^1	$2/3$	$1/3$	$0/3$
d^1, i^0	$1/1$	$0/1$	$0/1$
d^1, i^1			

MLE for $P(I)$ and $P(G|I)$



Student	Record
s1000	(c^0, d^1, i^1, g^1)
s1001	(c^0, d^0, i^1, g^1)
s1002	(c^0, d^1, i^0, g^1)
s1003	(c^1, d^0, i^1, g^2)
s1004	(c^1, d^0, i^1, g^2)
s1005	(c^1, d^0, i^0, g^1)
s1006	(c^1, d^0, i^0, g^3)
s1007	(c^0, d^1, i^1, g^3)
s1008	(c^0, d^1, i^1, g^2)
s1009	(c^1, d^0, i^1, g^1)

Observations from CS0001.

θ_{i^0}	θ_{i^1}
$3/10$	$7/10$

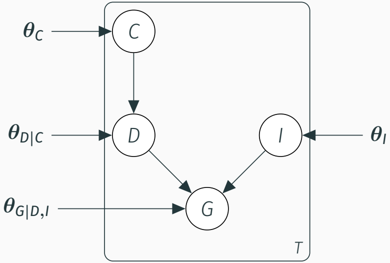
pa	$\theta_{g^1\|d,i}$	$\theta_{g^2\|d,i}$	$\theta_{g^3\|d,i}$
d^0, i^0	$1/2$	$0/2$	$1/2$
d^0, i^1	$2/3$	$1/3$	$0/3$
d^1, i^0	$1/1$	$0/1$	$0/1$
d^1, i^1	$1/3$	$1/3$	$1/3$

MLE for $P(I)$ and $P(G|I)$

Parameter Sharing with Template Models

It is very common to think of a PGM as a *template* to be instantiated given certain meta-data.

For example, we can imagine that our simplified student BN can be used to model T students who take a certain course:

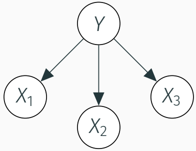


- Students are independent of one another.
- There are 4 tabular CPDs that are reused across all students.

Think of the plate as a ‘for loop’, the variables inside are ‘instantiated’ for each ‘data record’ out of T such data records. The assignments in iteration t_1 are independent of the assignments in iteration t_2.

Suppose we have 3 'predictors' X_1, X_2, X_3 for a condition Y.

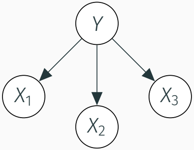
A condition might be something like COVID19 y^1 (True) or y^0 (False) and predictors might be things like: cough x_1^1 or x_1^0, high temperature x_2^1 or x_2^0, headache x_3^1 or x_3^0).



This is known as a **naive Bayes** model, commonly used for **classification** via the decision rule $\mathrm{argmax}_y \; P(Y = y | X_1, X_2, X_3)$

Suppose we have 3 ‘predictors’ X_1, X_2, X_3 for a condition Y.

A condition might be something like COVID19 y^1 (True) or y^0 (False) and predictors might be things like: cough x_1^1 or x_1^0, high temperature x_2^1 or x_2^0, headache x_3^1 or x_3^0).

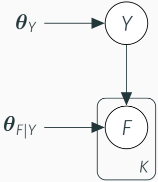


This is known as a **naive Bayes** model, commonly used for **classification** via the decision rule $\text{argmax}_y \; P(Y = y|X_1, X_2, X_3)$

In a naive Bayes model, we assume that the distribution of $X_1|Y = y$ is the same as the distribution of $X_2|Y = y$ and of $X_3|Y = y$.

That is, there is one CPD $P(F|Y)$ for ‘feature value’ given ‘class’ and we take X_1, X_2, X_3 to be 3 feature values drawn from that CPD.

With K ‘predictors’ $X_1, \ldots, X_K$ for a condition Y. The NB model is a template BN:



- $P(Y = y) = \theta_y$
- $P(F = f|Y = y) = \theta_{f|y}$

Evaluating the MLE Solution

‘Intrinsically’

- Assess $L(\boldsymbol{\theta}; \mathcal{H})$ for a dataset ‘heldout’ from training;

‘Extrinsically’ (task-driven)

- use the model in a predictive task (e.g., classification) and measure its task performance for some heldout dataset.

MLE for Markov Networks

References

[1] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.