Parametric Learning

Probabilistic Graphical Models

Jerónimo Hernández-González

Summary

- Probabilistic learning for graphical models
- Maximum likelihood learning
- Bayesian approach

A typical scenario

Sources of information:

- Data (which is assumed to be generated by repeatedly sampling a probability distribution P^* – i.i.d. sample)
- Domain experts (from which we can elicit valuable information about P^*)

Objective: Combine sources to infer a model M which approximates P^*

Restrict the set of possible probability distributions, $\mathcal{M}$, for our approximation $M \in \mathcal{M}$:

PGMs

Learning **always** implies to make assumptions

Model learning

Learning algorithm

Approximate P^* by learning a PGM, M, from data
$D = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N\}$ which is assumed to be i.i.d. sampled from P^*

$$A : D \rightarrow M \equiv (G, \boldsymbol{\Theta})$$

We want to extract information from D about P^* to encode in:

- The structure of the PGM, G (Structural learning):
  - NP-complete combinatorial optimization problem
  - Efficient heuristics: Local search, genetic algorithms, ...
- The parameters of the PGM, $\boldsymbol{\Theta}$ (Parametric learning)

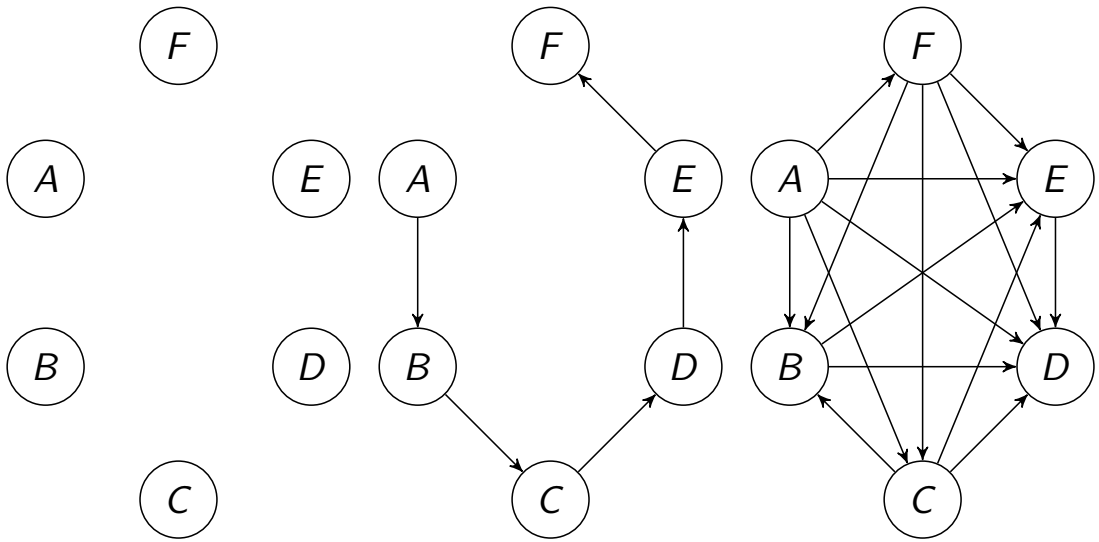
** It might be complemented with domain expert information

Trade-off to avoid **overfitting**!

no. parameters vs. no. training samples

Model learning

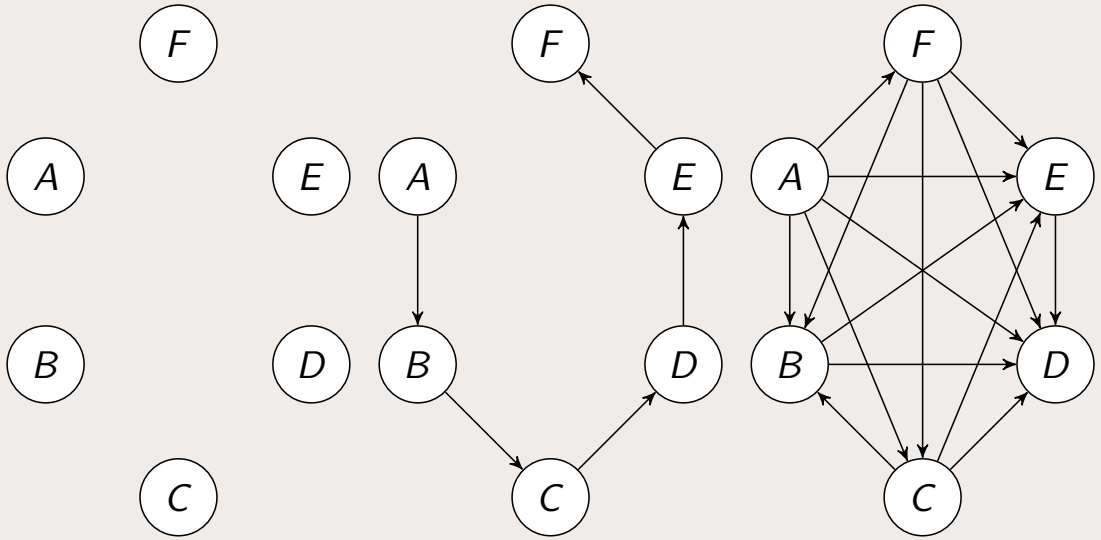
Fitting and generalization



Exercise

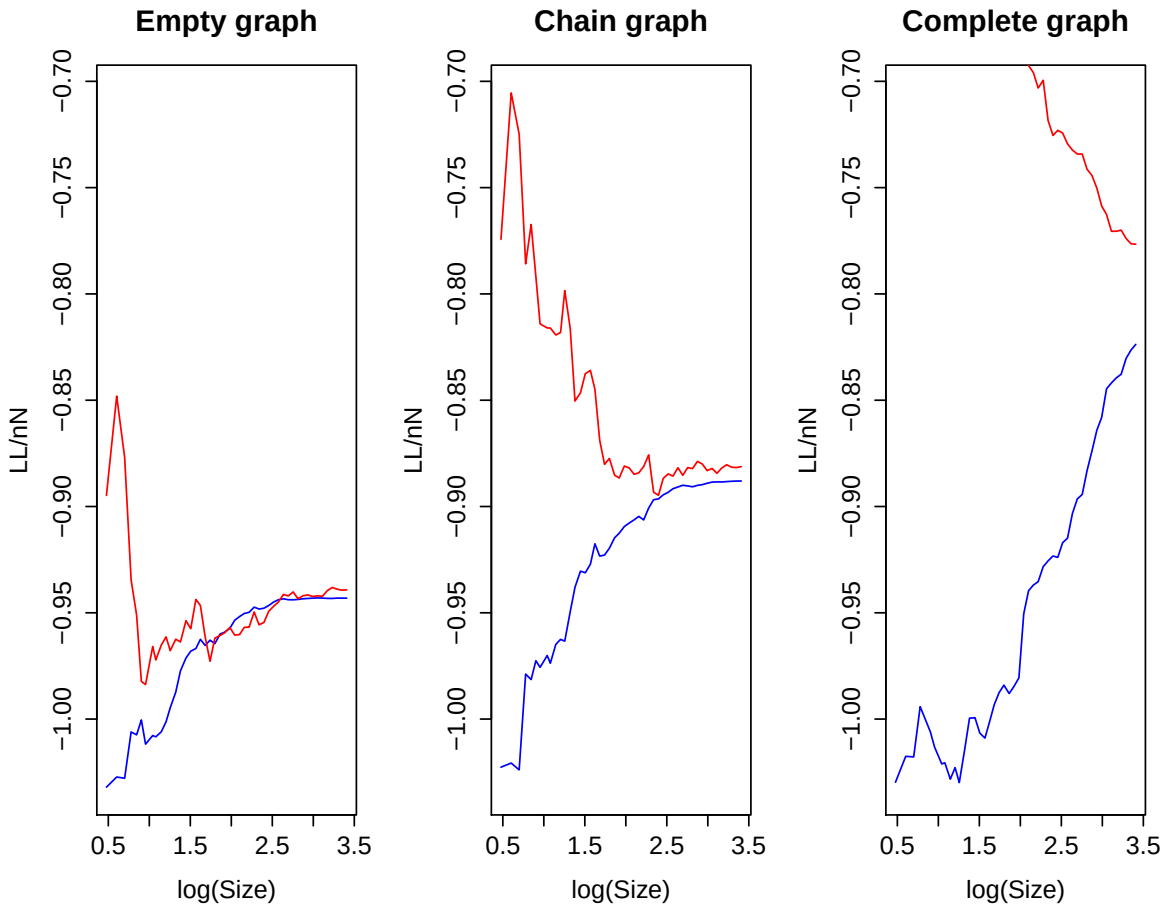
Number of parameters

Which is the number of parameters in each of these models? Let us assume that all the variables have cardinality 3.



Model learning

Fitting and generalization



Training vs. validation Log-Likelihood

Model learning

Learn a PGM, $M \equiv (G, \Theta)$, from data

- Structural learning, G
- Parametric learning, Θ

Learning scenarios when restricting to PGMs:

- Both the structure G and the CPDs Θ are known $\rightarrow$ No learning
- The structure G of the PGM is known $\rightarrow$ Learn the parameters Θ
- Only the variables $\boldsymbol{X}$ of the PGM are known $\rightarrow$ Learn both the structure G and its parameters Θ
- Nothing $\rightarrow$ Decide about the variables $\boldsymbol{X}$, and learn both the structure G and its parameters Θ

**These tasks are harder if some values in D are missing

Facing practical problems (rev.)

What to do?

1. Be smart about how to represent joint distributions
   - Bayesian networks, probabilistic graphical models
2. **Be smart about how to estimate probabilities (parameters) from sparse data**
   - Maximum likelihood estimates: θ that maximize $P(data|\theta)$
   - Maximum a posteriori estimates: θ that maximize $P(\theta|data)$

Parametric Learning

Probabilistic Graphical Models

Jerónimo Hernández-González

Facing practical problems (rev.)

What to do?

1. Be smart about how to represent joint distributions
   - Bayesian networks, probabilistic graphical models
2. **Be smart about how to estimate probabilities (parameters) from sparse data**
   - Maximum likelihood estimates: θ that maximize $P(data|\theta)$

$$\hat{\theta}^{MLE}_{\boldsymbol{x}_S} = \frac{N(\boldsymbol{x}_S)}{N}$$

Possible overfitting!

   - Maximum a posteriori estimates: θ that maximize $P(\theta|data)$

$$\hat{\theta}^{MAP}_{\boldsymbol{x}_S} = \frac{N(\boldsymbol{x}_S) + \alpha_{\boldsymbol{x}_S}}{N + \alpha_0}$$

Priors: Expert knowledge (Non-informative priors: $\alpha_{\boldsymbol{x}_S} = \frac{\alpha_0}{r_S}$)

Given a coin X

Estimate the probability that it will turn up heads ($X = 1$) or tails ($X = 0$)

You flip the coin repeatedly, observing:

- heads, N_1 times
- tails, N_0 times

Your estimate for $P(X = 1)$ –heads– is...?

X=1 X=0

Exercise: Parametric learning

Estimating Probability of Heads

$$\theta = P(X = \text{ heads })$$

Test A:

- 3 flips
- 2 heads
- 1 tails

X=1 X=0

Exercise: Parametric learning

Estimating Probability of Heads

$$\theta = P(X = \text{ heads })$$

Test A:

- 3 flips
- 2 heads
- 1 tails

Test B:

- 100 flips
- 54 heads
- 46 tails

X=1 X=0

$$P(X = \textit{heads}) = \theta \qquad P(X = \textit{tails}) = 1 - \theta$$

Flips produce data D with N_1 heads and N_0 tails

- flips are i.i.d. heads and tails (Bernoulli)
- N_1 and N_0 are counts that sum the outcomes (Binomial)
- Likelihood

$$P(D|\theta) = P(N_1, N_0|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

Maximum likelihood parameters

Let $D = \{x^1, ..., x^N\}$ be the available data where the random variable X takes r different values. The ML parameters are

$$\hat{p}(x) = \theta_x = \frac{N(x)}{N}$$

for $x \in \Omega_x$, and where $\hat{p}(x)$ is called the empirical distribution

- The likelihood of the parameters θ_x is given by

$$\mathcal{L}(\boldsymbol{\theta}|D) = p(D|\boldsymbol{\theta}) = \prod_{x=1}^{r} \theta_x^{N(x)}$$

- Tends to the true distribution as N increases
- As r is comparatively larger, the risk of overfitting increases

Exercise: Maximum Likelihood Estimate for θ

$$\hat{\theta} = \arg\max_\theta P(D|\theta)$$
$$= \arg\max_\theta \log P(D|\theta)$$
$$= \arg\max_\theta \log \theta^{N_1}(1-\theta)^{N_0}$$

Set derivative to zero:
$\frac{d}{d\theta} \log P(D|\theta) = 0$

Remember that $\frac{d \log \theta}{d\theta} = \frac{1}{\theta}$

Parametric learning

Maximum likelihood estimation

$$
\begin{aligned}
P(D|\theta) &= \prod_{i=1}^{N} P(x^i|\theta) \\
&= \prod_{i=1}^{N} \Big(\delta[x_i = 1]\theta + \delta[x_i = 0](1-\theta)\Big) = \\
&= \prod_{x^i \in D: x^i = 1} \theta \cdot \prod_{x^i \in D: x^i = 0} (1-\theta) = \\
&= \theta^{\#[x=1 \text{ in } D]} \cdot (1-\theta)^{\#[x=0 \text{ in } D]} = \\
&= \theta^{N_1} \cdot (1-\theta)^{N_0}
\end{aligned}
$$

Parametric learning

Maximum likelihood estimation

$$
\begin{aligned}
P(D|\theta) &= \prod_{i=1}^{N} P(x^i|\theta) \\
&= \prod_{i=1}^{N} \Big(\delta[x_i = 1]\theta + \delta[x_i = 0](1-\theta)\Big) = \\
&= \prod_{x^i \in D: x^i = 1} \theta \cdot \prod_{x^i \in D: x^i = 0} (1-\theta) = \\
&= \theta^{\#[x=1 \text{ in } D]} \cdot (1-\theta)^{\#[x=0 \text{ in } D]} = \\
&= \theta^{N_1} \cdot (1-\theta)^{N_0}
\end{aligned}
$$

$$
\begin{aligned}
\log P(D|\theta) &= \log \theta^{N_1} \cdot (1-\theta)^{N_0} \\
&= N_1 \log \theta + N_0 \log(1-\theta)
\end{aligned}
$$

$$\frac{d}{d\theta} \log P(D|\theta) = 0$$

$$\frac{d}{d\theta} \Big[N_1 \log \theta + N_0 \log(1-\theta) \Big] = 0$$

$$N_1 \frac{1}{\theta} - N_0 \frac{1}{1-\theta} = 0$$

$$\frac{1-\theta}{\theta} = \frac{N_0}{N_1} \quad ; \quad \frac{1}{\theta} = \frac{N_0 + N_1}{N_1} \quad ; \quad \theta = \frac{N_1}{N_0 + N_1}$$

$$\theta_{MLE} = \arg \max_{\theta} Pr(D|\theta) = \frac{N_1}{N_0 + N_1}$$

Parametric learning

Maximum likelihood estimation

$$\frac{d}{d\theta} \log P(D|\theta) = 0$$

$$\frac{d}{d\theta} \Big[N_1 \log \theta + N_0 \log(1-\theta) \Big] = 0$$

$$N_1 \frac{1}{\theta} - N_0 \frac{1}{1-\theta} = 0$$

$$\frac{1-\theta}{\theta} = \frac{N_0}{N_1} \quad ; \quad \frac{1}{\theta} = \frac{N_0 + N_1}{N_1} \quad ; \quad \theta = \frac{N_1}{N_0 + N_1}$$

$$\theta_{MLE} = \arg \max_{\theta} Pr(D|\theta) = \frac{N_1}{N_0 + N_1} \approx \frac{N_1 + 1}{N_0 + N_1 + 2}$$

(Laplace) smoothing

Sufficient statistics

- To find θ_{MLE}, we only need to calculate N_0 and N_1

$$\theta_{MLE} = \frac{N_1}{N_0 + N_1}$$

- A statistic is a function $s(D)$ from samples D to a vector in $\mathbb{R}^k$
- A statistic is sufficient if no other statistic (calculated from the same sample) provides additional information about the value of the parameter
- N_0 and N_1 are sufficient statistics

Exercise

Sufficient statistics

Suppose that you are playing with a 4-sided die and you want to check if it is biased.
You rolled it 60 times and get: 20 times it came up 1, 15 times 2, 15 times 3, and 10 times 4.
Let θ_1 be the true probability of the die landing on 1, and similarly for θ_2, θ_3, and θ_4. Which is a sufficient statistic for this data to estimate the parameters of this simple multinomial model?

a) The sum of all die rolls.

b) The total number of times that the dice was rolled.

c) A vector with four components, where each component is the number of times a specific side was observed.

d) None of these are sufficient statistics.

MLE is a simple principle for parameter learning given D

- A flip yields a value $X \sim Bernoulli(\theta)$

$$P(X) = \theta^X (1-\theta)^{(1-X)}$$

- Likelihood is uniquely determined by sufficient statistics that summarize D

  Given a dataset D of i.i.d. flips with n_1 heads and n_0 tails, the likelihood is:

$$P(D|\theta) = \theta^{n_1}(1-\theta)^{n_0}$$

  and the MLE estimate:

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta) = \frac{n_1}{n_1 + n_0}$$

- Closed form for many distributions

X=1 X=0

Bernoulli:

$$P(X = \text{heads}) = \theta$$

$$P(X = \text{tails}) = 1-\theta$$

Exercise

Maximum likelihood estimation

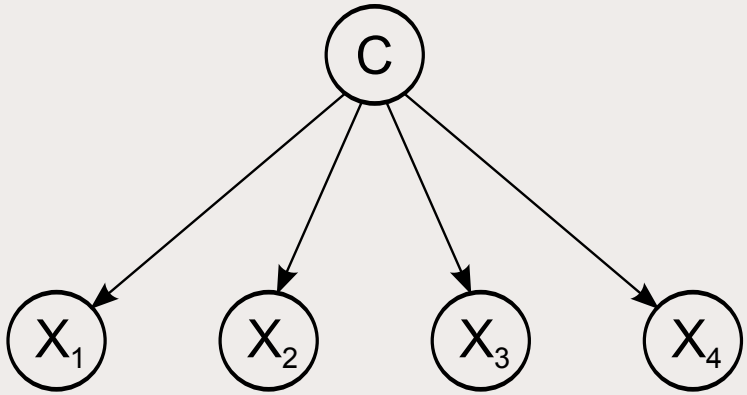
Suppose that you are playing with a 4-sided die and you suspect that it is biased. You rolled it 60 times and 20 times it came up 1, 15 times it came up 2, 15 times 3, and 10 times 4. Let θ_1 be the true probability of the die landing on 1, and similarly for θ_2, θ_3, and θ_4.

Which is the unique Maximum Likelihood Estimate (MLE) of the parameter θ_1?

Exercise

Estimating the parameters of NB from data

X_1	X_2	X_3	X_4	C
1	0	1	0	0
0	1	1	0	1
1	0	0	1	0
0	1	0	1	0
1	0	1	0	1
0	1	0	0	1
1	1	1	0	0
1	1	1	1	1
1	0	0	1	0
1	0	1	1	0



$$p(C)$$

$$p(X_1|C=0),\ p(X_1|C=1)$$

$$p(X_2|C=0),\ p(X_2|C=1)$$

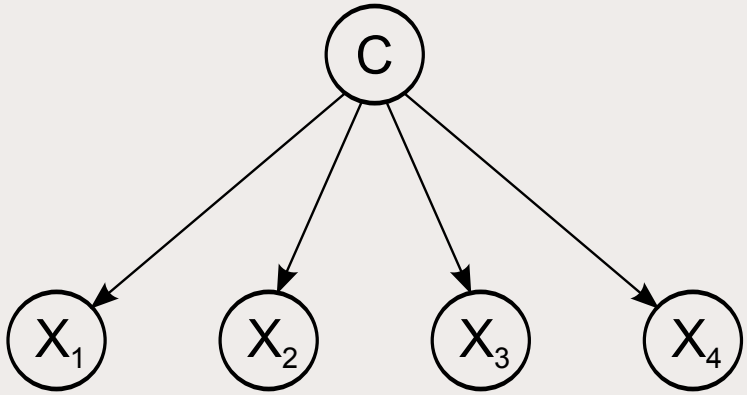
$$p(X_3|C=0),\ p(X_3|C=1)$$

$$p(X_4|C=0),\ p(X_4|C=1)$$

Exercise

Estimating the parameters of NB from data

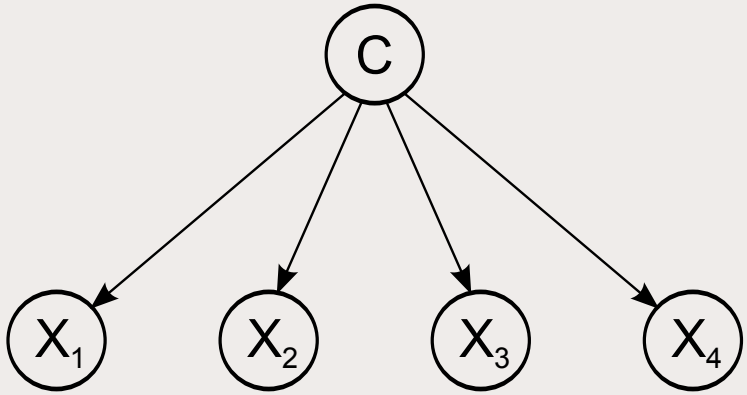
X_1	X_2	X_3	X_4	C
1	0	1	0	0
0	1	1	0	1
1	0	0	1	0
0	1	0	1	0
1	0	1	0	1
0	1	0	0	1
1	1	1	0	0
1	1	1	1	1
1	0	0	1	0
1	0	1	1	0



C	$p(C)$
0	
1	

Estimating the parameters of NB from data

X_1	X_2	X_3	X_4	C
1	0	1	0	0
0	1	1	0	1
1	0	0	1	0
0	1	0	1	0
1	0	1	0	1
0	1	0	0	1
1	1	1	0	0
1	1	1	1	1
1	0	0	1	0
1	0	1	1	0

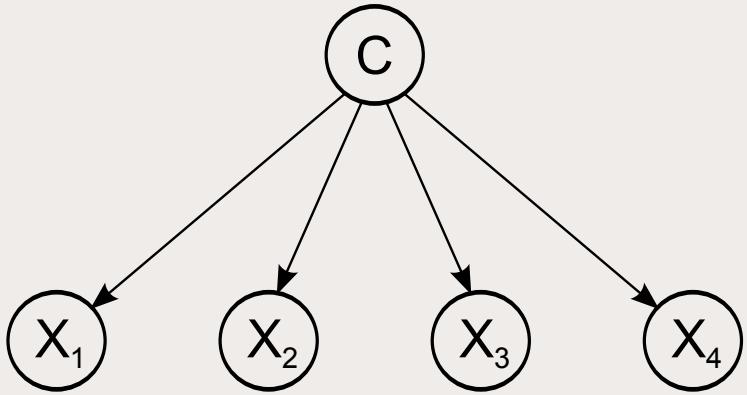


C	$p(C)$
0	$6/10 = 0{,}6$
1	$4/10 = 0{,}4$

Exercise

Estimating the parameters of NB from data

X_1	X_2	X_3	X_4	C
1	0	1	0	0
0	1	1	0	1
1	0	0	1	0
0	1	0	1	0
1	0	1	0	1
0	1	0	0	1
1	1	1	0	0
1	1	1	1	1
1	0	0	1	0
1	0	1	1	0

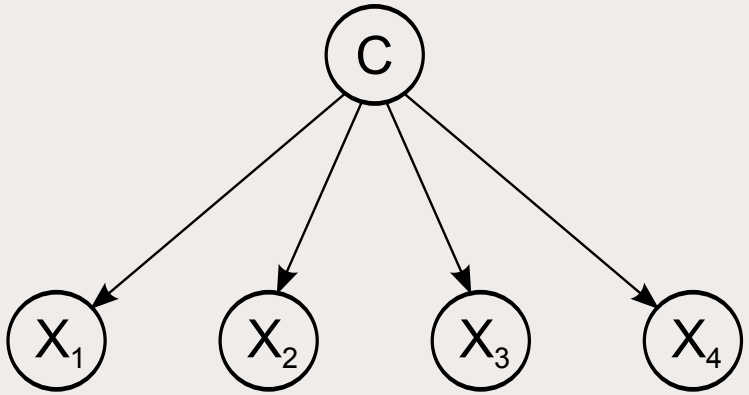


X_1	C	$p(X_1 \mid C)$
0	0	
1	0	
0	1	
1	1	

Exercise

Estimating the parameters of NB from data

X_1	X_2	X_3	X_4	C
1	0	1	0	0
0	1	1	0	1
1	0	0	1	0
0	1	0	1	0
1	0	1	0	1
0	1	0	0	1
1	1	1	0	0
1	1	1	1	1
1	0	0	1	0
1	0	1	1	0

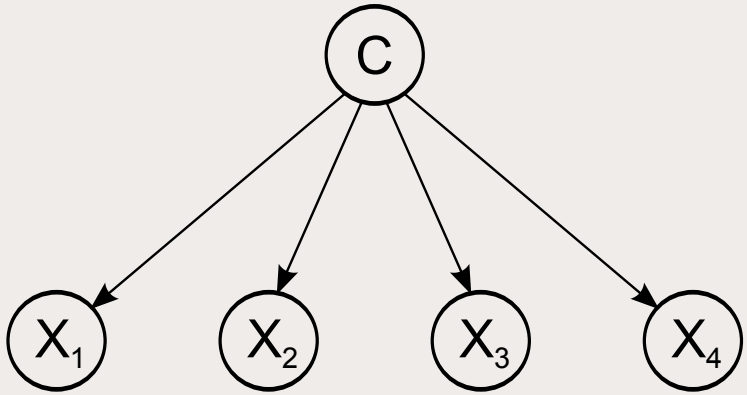


X_1	C	$p(X_1\|C)$
0	0	$1/6 = 0{,}17$
1	0	$5/6 = 0{,}83$
0	1	$2/4 = 0{,}50$
1	1	$2/4 = 0{,}50$

Exercise

Estimating the parameters of NB from data

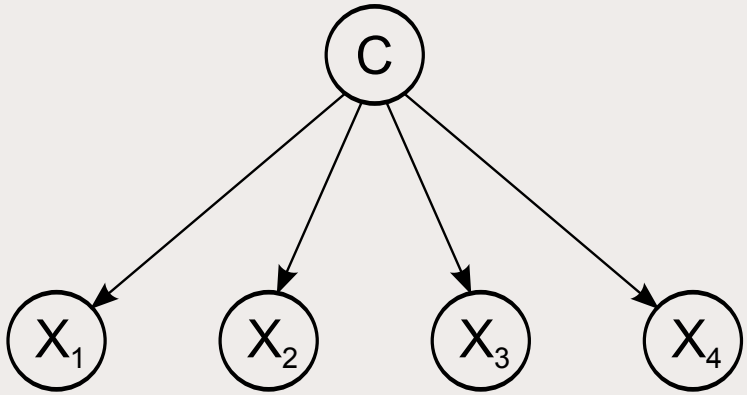
X_1	X_2	X_3	X_4	C
1	0	1	0	0
0	1	1	0	1
1	0	0	1	0
0	1	0	1	0
1	0	1	0	1
0	1	0	0	1
1	1	1	0	0
1	1	1	1	1
1	0	0	1	0
1	0	1	1	0



X_2	C	$p(X_2 \mid C)$
0	0	
1	0	
0	1	
1	1	

Estimating the parameters of NB from data

X_1	X_2	X_3	X_4	C
1	0	1	0	0
0	1	1	0	1
1	0	0	1	0
0	1	0	1	0
1	0	1	0	1
0	1	0	0	1
1	1	1	0	0
1	1	1	1	1
1	0	0	1	0
1	0	1	1	0



X_2	C	$p(X_2\|C)$
0	0	$4/6 = 0{,}33$
1	0	$2/6 = 0{,}67$
0	1	$1/4 = 0{,}25$
1	1	$3/4 = 0{,}75$

Summary

MLE problems and possible solutions

Selecting the model M that maximizes the likelihood overfits to statistical noise:

- Parametric overfitting: Parameters fit random noise in training data
  Solution: Use regularization or parameter priors
- Structural overfitting: Likelihood increases with structural complexity
  Solution: Bound/penalize model complexity

Exercise

Parametric Learning in MNs and BNs

Compared to learning parameters in Bayesian networks, learning in Markov networks is generally...

a) more difficult because we cannot push in sums to decouple the likelihood function, allowing independent parallel optimizations, as we can in Bayesian networks with CPDs.

b) equally difficult, as both require an inference step at each iteration.

c) equally difficult, though MN inference will be better by a constant factor difference in the computation time as we do not need to worry about directionality.

d) less difficult because we must separately optimize decoupled portions of the likelihood function in a Bayesian network, while we can optimize portions together in a Markov network.

Parametric Learning

Probabilistic Graphical Models

Jerónimo Hernández-González

Principle 1 (maximum likelihood):

Choose parameters θ that maximize

$$P(data|\theta)$$

Principle 2 (maximum a posteriori):

Choose parameters θ that maximize

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{P(data)}$$

Using domain expert knowledge

To build a model, we can...

- estimated it from data (we just saw it)
- elicit from expert knowledge

Are both approaches incompatible?

Using domain expert knowledge

To build a model, we can...

- estimated it from data (we just saw it)
- elicit from expert knowledge

Are both approaches incompatible? No!
Bayesian estimation

Combine a priori (expert?) knowledge about the model with the evidence gathered from data

Any new piece of data, used to update our knowledge

Idea

- Posterior:

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{P(data)} \propto P(data|\theta)P(\theta)$$

- Prior:

** Prob. distr. over a parameter: not easy to see concept

$$P(\theta)$$

- Likelihood:

$$P(data|\theta)$$

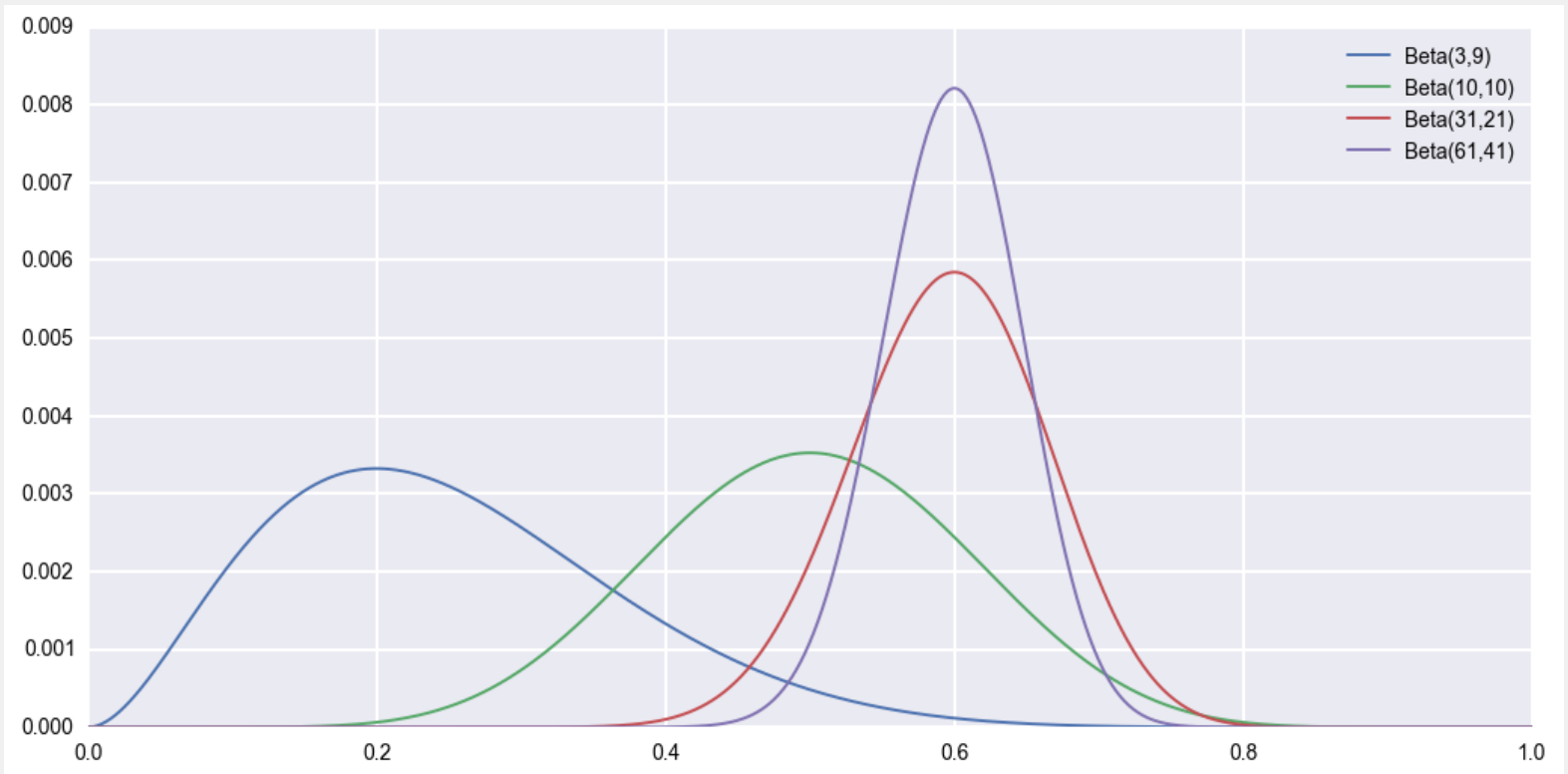
MAP estimator: the parameter that maximizes the posterior distr.

$$\arg\max_{\theta} p(\theta|data) = \arg\max_{\theta} \frac{p(data|\theta)p(\theta)}{p(data)} = \arg\max_{\theta} p(data|\theta)p(\theta)$$

$$\arg\max_{\theta} \text{Likelihood} \times \text{Prior}$$

Beta prior distribution

$$P(\theta) = \frac{\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}}{B(\beta_1, \beta_0)} \sim Beta(\beta_1, \beta_0)$$



Parametric learning

Maximum a Posteriori

- Likelihood is a Binomial:

$$P(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

- Prior is a Beta:

$$P(\theta) = \frac{\theta^{\beta_1-1}(1-\theta)^{\beta_0-1}}{B(\beta_1,\beta_0)} \sim Beta(\beta_1,\beta_0)$$

- Then, Posterior is also a Beta:

$$P(\theta|D) \sim Beta(N_1+\beta_1, N_0+\beta_0)$$

Example # 1

Coin flip problem

X=1 X=0

MAP estimate

$$\hat{\theta}^{MAP} = \frac{N_1+\beta_1-1}{(N_1+\beta_1-1)+(N_0+\beta_0-1)}$$

** Mode of the Beta posterior distribution

Parametric learning

Maximum a Posteriori

- Likelihood is a Multinomial:

$$P(D|\theta) = \theta_1^{N_1}\theta_2^{N_2}\dots\theta_r^{N_r}$$

Example # 2

Dice roll problem

- Prior is a Dirichlet (conjugate prior):

$$P(\theta) = \frac{\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\dots\theta_r^{\alpha_r-1}}{B(\alpha_1,\dots,\alpha_r)} \sim Dirichlet(\alpha_1,\dots,\alpha_r)$$

- Then, Posterior is also a Dirichlet:

$$P(\theta|D) \sim Dirichlet(N_1+\alpha_1,\dots,N_r+\alpha_r)$$

MAP estimate

$$\hat{\theta}_i^{MAP} = \frac{N_i+\alpha_i-1}{\sum_{j=1}^{r}(N_j+\alpha_j-1)}$$

** Mode of the Dirichlet posterior distribution

Bayesian approach

- The parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_r)$ are distributed according to a Dirichlet with (hyper)parameters:

$$p(\boldsymbol{\theta}) \sim Dirichlet(\alpha_1, \ldots, \alpha_r)$$

$$p(\boldsymbol{\theta}|D) \sim Dirichlet(N_1 + \alpha_1, \ldots, N_r + \alpha_r)$$

- $\alpha_0 = \sum_i \alpha_i$, is known as the equivalent sample size
- We always can use non-informative priors: $\alpha_i = \frac{\alpha_0}{r}$

$\sim$ Smoothed version of the maximum likelihood estimate

Equivalent sample size

Posterior mean of a Beta distribution, $B(\beta_1, \beta_0)$:

$$\frac{\beta_1+N_1}{\beta_1+\beta_0+N} = \frac{\beta_1}{\beta_1+\beta_0+N} + \frac{N_1}{\beta_1+\beta_0+N}$$

$$= \frac{\beta_1+\beta_0}{\beta_1+\beta_0+N} \cdot \frac{\beta_1}{\beta_1+\beta_0} + \frac{N}{\beta_1+\beta_0+N} \cdot \frac{N_1}{N}$$

$$= \text{weight\_prior} \times \text{mean\_prior} + \text{weight\_data} \times \text{mean\_data}$$

where $\text{weight\_prior} + \text{weight\_data} = 1$

The update is a weighted sum of mean_prior and mean_data!

** Note the relevance of $\beta_0 + \beta_1$ to balance the weights!

Summary

Alternatives

- Maximum likelihood: Given some data and a family of distributions, select the distribution that fits the data best.

  * Possible overfitting

- Maximum a posteriori: Given some data, a family of distributions and a prior belief for each of them, and refine your belief according to the data. The MAP estimate is the mode of the refined belief (posterior).

  * It might be more complex

Parametric Learning

Probabilistic Graphical Models

Jerónimo Hernández-González