Universität Stuttgart
IPVS – Institute for Parallel and Distributed Systems

Analytic Computing

Advanced Topics in Machine Learning
8 Learning - Part 2

Prof. Dr. Steffen Staab
Dr. Rafika Boutalbi
Zihao Wang
https://www.ipvs.uni-stuttgart.de/departments/ac/

Learning Objectives

Bayesian Parameter Estimation and Prediction

Bayesian Prediction

Estimation in Bayesian Network

Priors in Markov Random Fields

Disclaimer

Figures and examples not marked otherwise are taken from the book by Koller & Friedman

1 Bayesian Parameter Estimation

(cf Chapter 11 Bayesian Updates of Beliefs in ML Lecture)

Without / with background knowledge

$\mathrm{val}(\mathrm{Thumbtack}) = \{\mathrm{H}, \mathrm{S}\}$,
$\mathrm{H} = \mathrm{Head},\ \mathrm{S} = \mathrm{Sideways}$

<H, S, S, S, S,H, S, S, S, H>

MLE: $\hat{\theta}_H = 0.3$

$\mathrm{val}(\mathrm{Coin}) = \{\mathrm{H}, T\}$,
$\mathrm{H} = \mathrm{Head},\ \mathrm{T} = \mathrm{Tail}$

<H, T, H, H, T ,H, H, T, H, H>

MLE: $\hat{\theta}_H = 0.7$???

Remember: Chapter 2 ML Lecture

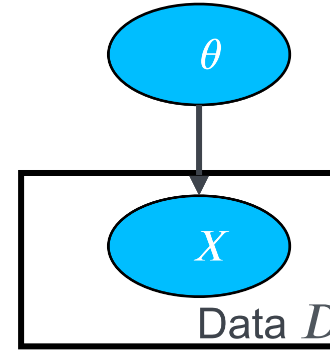
Limitations of MLE

- Tossing a coin 10 times and seeing 3 heads up
- Tossing a coin 100 times and seeing 30 heads up
- Tossing a coin 1000 times and seeing 300 heads up

All the same for MLE!

Not all the same for our beliefs!

Bayesian inference about priors



$$P\left(x[1],\ldots,x[M],\theta\right) =$$

$$= P\left(x[1],\ldots,x[M] \mid \theta\right)P(\theta) =$$

$$= P(\theta)\prod_{m=1}^{M} P(x[m] \mid \theta) =$$

$$= P(\theta)\theta^{k(\mathrm{Head})}(1-\theta)^{k(\mathrm{Tail})}$$

Posterior:

$$P\left(\theta \mid x[1],\ldots,x[M]\right) = \frac{\overbrace{P\left(x[1],\ldots,x[M] \mid \theta\right)P(\theta)}^{L(\mathscr{D};\theta)}}{\underbrace{P\left(x[1],\ldots,x[M]\right)}_{\text{constant relative to } \theta}}$$

Dirichlet Distribution

- Its Probability Density Function f is defined by

$$f\left(x_1, \ldots, x_K; \alpha_1, \ldots \alpha_K\right) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \text{ with}$$

$$\sum_{i=1}^{K} x_i = 1, \text{ and } \forall i : x_i \geq 0$$

- Dirichlet Distribution is a generalization of the Beta distribution
  - $B(\boldsymbol{\alpha})$ being the multi-variate Beta function
  - α_i are the **hyperparameters** / shape parameters / concentration parameters / **pseudocounts**
- Beta distribution means K=2
- Dirichlet distribution is the conjugate prior to the multinomial distribution (Beta distribution is the conjugate prior to the binomial distribution)

Dirichlet Priors and Posteriors

Posterior Likelihood Prior

$$P(\theta \mid \mathscr{D}) \propto P(\mathscr{D} \mid \theta)\ P(\theta)$$

$$P(\mathscr{D} \mid \theta) = \prod_{i=1}^{l} \theta_i^{k_i} \qquad P(\theta) \propto \prod_{i=1}^{l} \theta_i^{\alpha_i - 1}$$

If $P(\theta)$ is Dirichlet and the likelihood is multinomial, then the posterior is also Dirichlet

- Prior is $\mathrm{Dir}(\alpha_1, \ldots, \alpha_l)$
- Data counts are $k_1, \ldots, k_l$
- Posterior is $\mathrm{Dir}(\alpha_1 + k_1, \ldots, \alpha_l + k_l)$

Dirichlet is a conjugate prior for the multinomial

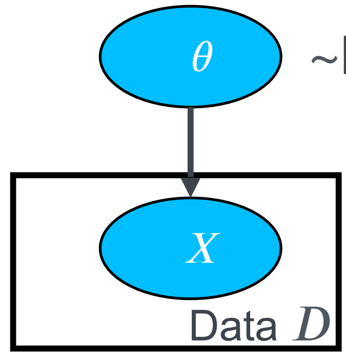
If there is no strong reason to choose otherwise, pick a conjugate prior for your distribution (often) allowing for simple (often: closed form) solution

Bayesian Parameter Estimation

- Bayesian parameter estimation less prone to overfitting than maximum likelihood estimation
- determining the Bayes estimate (Learning!) boils down to inference

2 Bayesian Prediction

Bayesian prediction

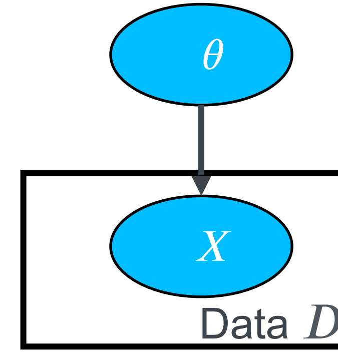


$\sim$Dirichlet$(\alpha_1, \ldots, \alpha_l)$

$$P(X) = \int P(X \mid \theta) P(\theta) d\theta$$

$$P\left(x_i \mid \theta\right) = \frac{1}{Z} \int \theta_i \prod \theta^{\alpha_i - 1} d\theta = \frac{\alpha_i}{\sum_i \alpha_i}$$

Bayesian Prediction



$$P\Big(x[M+1] \;\Big|\; x[1], \ldots, x[M]\Big) =$$

$$= \int P\Big(x[M+1] \;\Big|\; \theta,\; x[1], \ldots, x[M]\Big) \underbrace{P\Big(\theta \mid x[1], \ldots, x[M]\Big)}_{\text{Posterior: } Dir(\alpha_1 + k_1, \ldots, \alpha_l + k_l)} d\theta =$$

$$= \int P\Big(x[M+1] \;\Big|\; \theta\Big)\, P\Big(\theta \mid x[1], \ldots, x[M]\Big) d\theta$$

$$P\Big(x[M+1] = x^i \;\Big|\; x[1], \ldots, x[M]\Big) = \frac{\alpha_i + k_i}{\sum (\alpha_j + k_j)}$$

Example

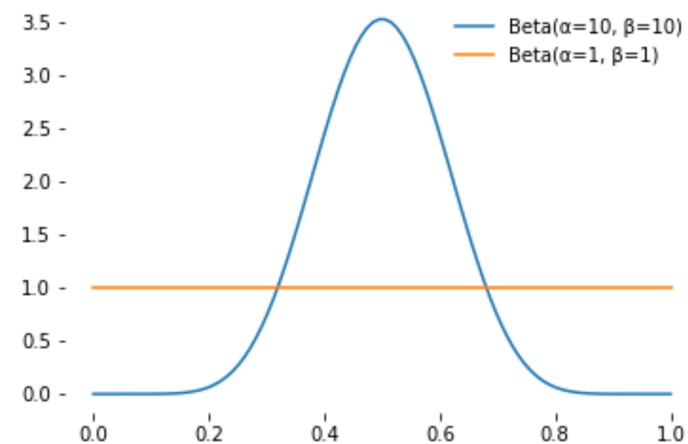
Maximum likelihood: $P(\text{tail}) = 0.8$

Bayesian estimate with uniform prior $\alpha_{\text{tail}} = \alpha_{\text{head}} = 1$ (Laplace smoothing):

$$P(\text{tail}) = \frac{4+1}{5+2} = \frac{5}{7} \approx 0.71$$

Bayesian estimate with $\alpha_{\text{tail}} = \alpha_{\text{head}} = 10$:

$$P(\text{tail}) = \frac{4+10}{5+20} = \frac{14}{25} \approx 0.56$$



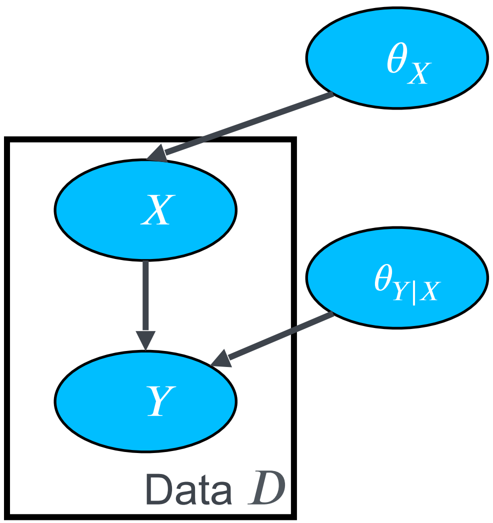
Summary

- Bayesian prediction combines sufficient statistics from imaginary Dirichlet samples and real data samples
- Asymptotically the same as MLE
- Dirichlet hyperparameters determine the prior beliefs and their strengths

Question 2: Bayesian Prediction

3 Bayesian Parameter Estimation for Bayesian Networks

Bayesian Estimation in BNs

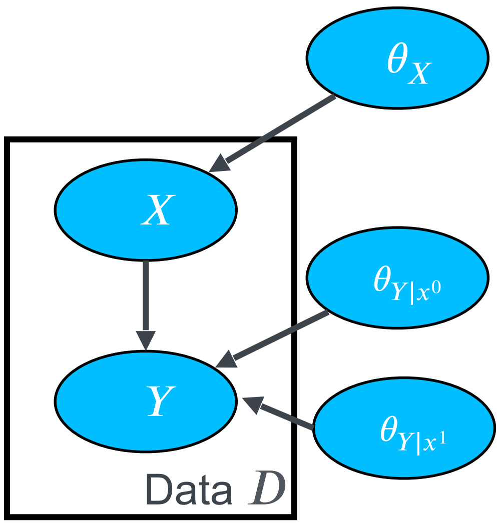


- Instance pairs $(X[m], Y[m])$ are independent from other instance pairs $(X[i], Y[i])$ given $\theta_X, \theta_{Y|X}$
- A priori, parameters $\theta_X, \theta_{Y|X}$ are independent, thus

$$P(\theta) = \prod_m P(\theta_{X_i|\text{Parents}(X_i)})$$

- Given complete data, parameters $\theta_X, \theta_{Y|X}$ are independent

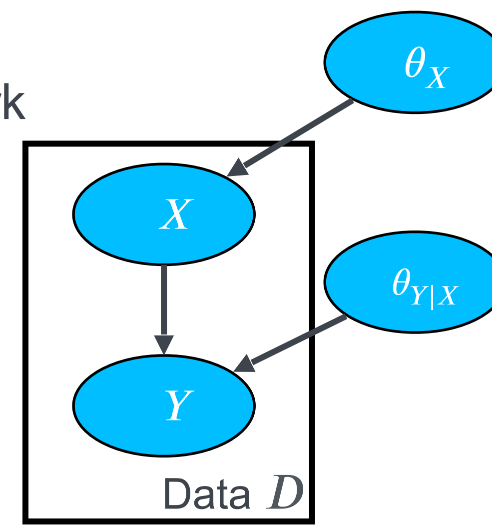
Bayesian Estimation in BNs



- A priori, parameters $\theta_X, \theta_{Y|X}$ are independent
- Given complete data $\mathscr{D}$, parameters $\theta_X, \theta_{Y|X}$ are independent
  - also $\theta_{Y|x^0}$ and $\theta_{Y|x^1}$ are context specific independent when given the data $\mathscr{D}$
  - $P(\theta \mid \mathscr{D}) = P(\theta_X \mid \mathscr{D}) P\left(\theta_{Y|x^0} \mid \mathscr{D}\right) P(\theta_{Y|x^1} \mid \mathscr{D})$
- Posteriors of θ can be computed independently
  - for multinomial $\theta_{X_i|\mathrm{Pa}(X_i)}$ if prior is $\mathrm{Dir}(\alpha_{x^1|\boldsymbol{u}}, \ldots, \alpha_{x^l|\boldsymbol{u}})$
  - then: posterior is $\mathrm{Dir}\left(\alpha_{x^1|\boldsymbol{u}} + k\left(x^1 \middle| \boldsymbol{u}\right), \ldots, \alpha_{x^l|\boldsymbol{u}} + k\left(x^l \middle| \boldsymbol{u}\right)\right)$

Consistent assignment of priors for Bayesian Network

- We need hyperparameter $\alpha_{x|\boldsymbol{u}}$ for each Node X, value x, and parent assignment $\boldsymbol{u}$
  - Define equivalent sample size parameter α (strength of initial belief)
  - Define initial distribution θ_0
    - typically uniform $(\mathrm{Dir}(1,\ldots,1))$
  - Define prior that balances consistently over the network
    - $\alpha_{x|\boldsymbol{u}} := \alpha \bullet P(x, u \,|\, \theta_0)$
  - Example:
    - $\alpha_X = (\frac{1}{2}, \frac{1}{2})$
    - $\alpha_{Y|x^0} = (\frac{1}{4}, \frac{1}{4}), \alpha_{Y|x^1} = (\frac{1}{4}, \frac{1}{4})$



Summary

- In Bayesian networks, if parameters are independent a priori, then they are also independent in the posterior
- For multinomial Bayesian networks, estimation uses sufficient statistics $k(x, \boldsymbol{u})$

MLE

$$\tilde{\theta}_{x|\boldsymbol{u}} = \frac{k(x, \boldsymbol{u})}{k(\boldsymbol{u})}$$

Bayesian (Dirichlet)

$$\tilde{\theta}_{x|\boldsymbol{u}} = \frac{\alpha_{x|\boldsymbol{u}} + k(x, \boldsymbol{u})}{\alpha_{\boldsymbol{u}} + k(\boldsymbol{u})}$$

- Bayesian methods require choice of prior
  - can be assigned as prior network and equivalent sample size
- Bayesian methods tend to converge much, much better than MLE

4 MAP Estimation for MRFs

Setting hyperparameters in MRFs, CRFs

- How to include priors into MRFs, CRFs
  - given that probabilities, likelihood, posterior are not straightforwardly implemented
- MAP inference to acquire the parameters

$$\operatorname{argmax}_\theta P(D,\theta) =$$

$$= \operatorname{argmax}_\theta \Big(P\big(D \mid \theta\big) P(\theta) \Big) =$$

$$= \operatorname{argmax}_\theta \log \Big(P\big(D \mid \theta\big) P(\theta) \Big) =$$

$$= \operatorname{argmax}_\theta \Big(\ell\big(P\big(D \mid \theta\big)\big) + \log P(\theta) \Big)$$

Regularization

Gaussian Prior

$$P\left(\theta;\sigma^2\right) = \prod_{i=1}^{l} \frac{1}{\sqrt{2\pi\sigma}} e^{\left(-\frac{\theta_i^2}{2\sigma^2}\right)}$$

Laplacian Prior

$$P\left(\theta;\beta\right) = \prod_{i=1}^{l} \frac{1}{2\beta} e^{\left(-\frac{\left|\theta_i\right|}{\beta}\right)}$$

Mean 0

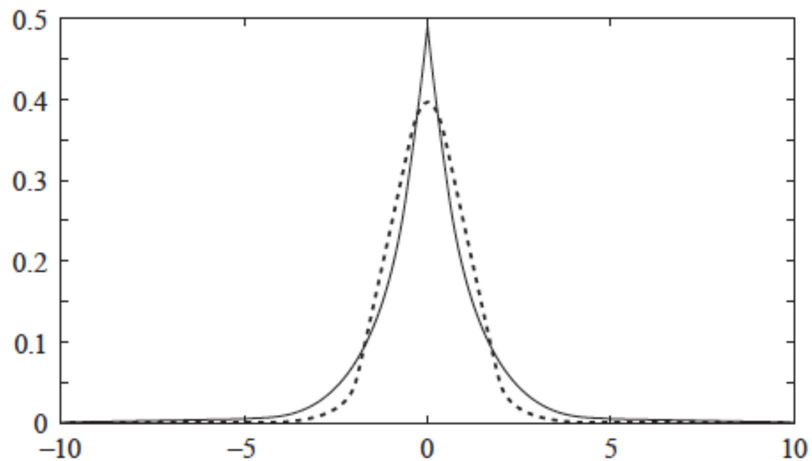


Figure 20.3 **Laplacian distribution ($\beta = 1$) and Gaussian distribution ($\sigma^2 = 1$**

Regularization

Gaussian Prior

$$P(\theta;\sigma^2) = \prod_{i=1}^{l} \frac{1}{\sqrt{2\pi\sigma}} e^{\left(-\frac{\theta_i^2}{2\sigma^2}\right)}$$

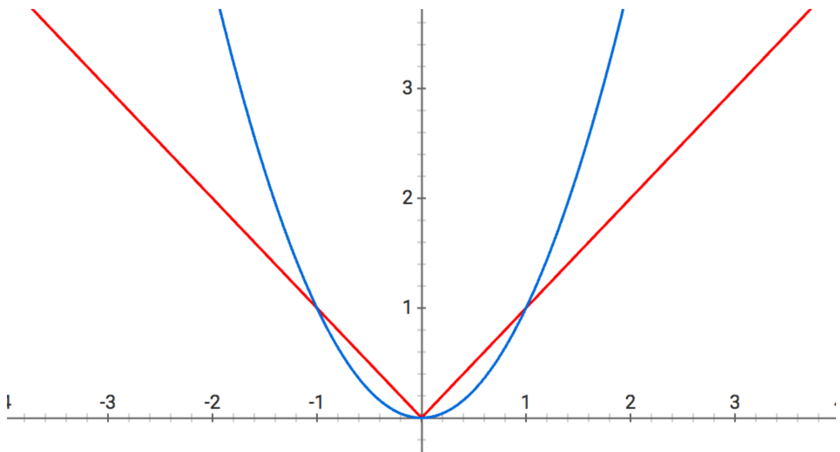
$$\log P(\theta;\sigma^2) = \text{const} - \frac{1}{2\sigma^2}\sum_{i=1}^{l}\theta_i^2$$

Laplacian Prior

$$P(\theta;\beta) = \prod_{i=1}^{l} \frac{1}{2\beta} e^{\left(-\frac{|\theta_i|}{\beta}\right)}$$

$$\log P(\theta;\beta) = \text{const} - \frac{1}{\beta}\sum_{i=1}^{l}|\theta_i|$$

L2 regularization



L1 regularization

Summary

- In undirected models, parameter coupling prevents efficient Bayesian estimation
- However, one may still use parameter priors to avoid overfitting of MLE
- L1 induces sparse solutions
  - feature selection / step towards structure learning

Question 4:
MAP estimation for MRFs

Universität Stuttgart
IPVS

Thank you!

Steffen Staab

E-Mail Steffen.staab@ipvs.uni-stuttgart.de
Telefon +49 (0) 711 685- To be defined
www. ipvs.uni-stuttgart.de/departments/ac/

Universität Stuttgart
Analytic Computing, IPVS
Universitätsstraße 32, 50569 Stuttgart