Graphical Models in Computer Vision

Andreas Geiger

Max Planck Institute for Intelligent Systems
Perceiving Systems

May 9, 2016

Syllabus

11.04.2016	Introduction
18.04.2016	Graphical Models 1
25.04.2016	Graphical Models 2 (Sand 6/7)
02.05.2016	Graphical Models 3
09.05.2016	Graphical Models 4
23.05.2016	Body Models 1
30.05.2016	Body Models 2
06.06.2016	Body Models 3
13.06.2016	Body Models 4
20.06.2016	Stereo
27.06.2016	Optical Flow
04.07.2016	Segmentation
11.07.2016	Object Detection 1
18.07.2016	Object Detection 2

What is there to learn?

- Given
  - Training data : $\mathcal{D} = \{x_1, \ldots, x_n\}$
    - For example coin tosses $x_i \in \{0, 1\}$
  - Training data: $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
    - Images y_i with labels x_i, for example face/non-face
- So what is there to learn? What do we want?
- Unsupervised Learning: density estimate of $\mathcal{D}$
  - Score new examples x with $p(x)$ (or (x, y) with $p(x, y)$)
- Supervised learning. Predict with $f(y) = x$. Need $p(x|y)$ then predict with

$$\hat{x} = \underset{x' \in \mathcal{X}}{\operatorname{argmin}} \, \mathbb{E}_{p(x|y)}[\Delta(x', x)]$$

- Supervised learning. Just predict $f(y) = x$, do not need density $p(x|y)$

Learning Methods – Overview

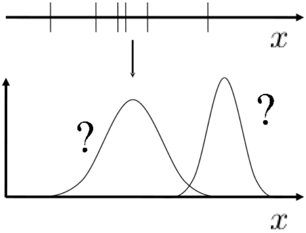
- Given
  - Training data : $\mathcal{X} = \{x_1, \ldots, x_n\}$
  - Choose a model class : $p(x \mid \theta), \theta \in \Theta$ (e.g. Gaussian, or 2x2 MRF)
- Problem : Find $\hat{\theta}$ such that $p(x \mid \hat{\theta})$ "best fits" the data $\mathcal{X}$
- What does "best fits" mean?

Parametric Models - Learning

- For parametric distributions we write

$$p(x \mid \theta)$$

- x can be discrete/continuous and scalar/multi-variate
- In the Gaussian case: $\theta = (\mu, \sigma^2)$
- Which θ to use?



Parameter Estimation

- **Point Estimates**
  - Try to estimate *one* value of θ
  - Several possible choices of estimators
  - Usually simpler (compared to Bayesian estimation)
  - Commonly used: Maximum Likelihood, Maximum-A-Posteriori
- **Bayesian Estimation**
  - Specify all knowledge about θ in a prior distribution $p(\theta)$
  - Integrate out the variable θ

$$p(x \mid \mathcal{D}) = \int_\theta p(x \mid \theta) p(\theta \mid \mathcal{D}) d\theta$$

  - Often intractable due to the integral

Let's discuss both options. Running example: Gaussian distribution.

Maximum Likelihood Estimator

- Aim to estimate one single θ
- Likelihood of the data

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; \mathcal{D}) = p(\mathcal{D} \mid \theta)$$

- Assume that the data is independent and identically distributed (iid)

$$\mathcal{L}(\theta) = p(\mathcal{D} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

- Now choose θ such that it maximizes the likelihood

$$\theta_{ML} = \operatorname*{argmax}_{\theta} \mathcal{L}(\theta; \mathcal{D})$$

Maximum Likelihood Estimator

- Maximum Likelihood

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \, \mathcal{L}(\theta; \mathcal{D})$$

is equivalent with

- minimizing the negative log-Likelihood

$$\begin{aligned} \theta_{ML} &= \underset{\theta}{\operatorname{argmax}} \log \mathcal{L}(\theta) = \underset{\theta}{\operatorname{argmax}} \, L(\theta) \\ &= \underset{\theta}{\operatorname{argmin}} - L(\theta) \\ &= \underset{\theta}{\operatorname{argmin}} - \sum_{i=1}^{n} \log p(x_i \mid \theta) \end{aligned}$$

- Numerically more stable

Kullback-Leibler divergence

- Measure of difference of probability distributions
- Discrete:

$$D_{KL}(q\|p) = \sum_i q_i \log \frac{q_i}{p_i}$$

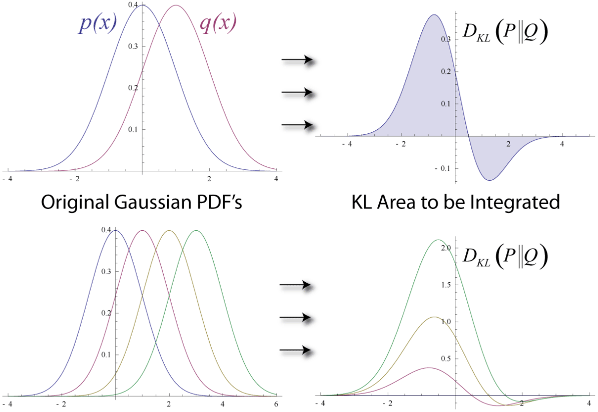
- Continuous:

$$D_{KL}(q\|p) = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)}$$

- In general non-symmetric:

$$D_{KL}(q\|p) \neq D_{KL}(p\|q)$$

Kullback-Leibler divergence



Let $q(x)$ denote the empirical distribution: $q(x) = \frac{1}{N} \sum_{i=1}^{N} [x = x_i]$

$$
\begin{aligned}
&\operatorname*{argmin}_{\theta} D_{KL}(q(x) \| p(x \mid \theta)) \\
&= \operatorname*{argmin}_{\theta} \int_x q(x) \log \frac{q(x)}{p(x \mid \theta)} \\
&= \operatorname*{argmin}_{\theta} \int_x q(x) \log q(x) - \int_x q(x) \log p(x \mid \theta) \\
&= \operatorname*{argmax}_{\theta} \int_x q(x) \log p(x \mid \theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \log p(x_i \mid \theta) \\
&= \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} p(x_i \mid \theta)
\end{aligned}
$$

Maximum Likelihood and Kullback-Leibler Divergence

- Maximum Likelihood is equivalent to minimizing KL divergence with empirical distribution

$$q(x) = \frac{1}{N}\sum_{i=1}^{N}[x = x_i]$$

Remember:

- Two choices to find $p(x \mid \theta)$
  - Point Estimates (eg Maximum Likelihood)
  - Bayesian Estimation
- Now apply the two to the Gaussian distribution

ML Estimate for Gaussian Distribution

- Maximum Likelihood for Gaussian distribution

$$\underset{\theta}{\operatorname{argmin}} - \log \mathcal{L}(\theta) = \underset{\theta}{\operatorname{argmin}} - \sum_{i=1}^{n} \log p(x_i \mid \mu, \sigma)$$

- Let's compute ...
- Is available in analytic form

$$\frac{\partial L}{\partial \mu} \stackrel{!}{=} 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\partial L}{\partial \sigma} \stackrel{!}{=} 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

ML: Multivariate Gaussian Distribution

$$
\begin{aligned}
L(\mu, \Sigma \mid \mathcal{D}) &= \sum_{i=1}^{N} \log p(x_i \mid \mu, \Sigma) \\
&= -\frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) - \frac{N}{2} \log \det(2\pi\Sigma)
\end{aligned}
$$

- Taking the derivative w.r.t. μ

$$
\nabla_\mu L(\mu, \Sigma) = \sum_{i=1}^{N} \Sigma^{-1}(x_i - \mu)
$$

- We realize μ_{ML} to be the sample mean:

$$
\mu_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i
$$

ML: Multivariate Gaussian Distribution

$$
\begin{aligned}
L(\mu, \Sigma \mid \mathcal{D}) &= -\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^\top\Sigma^{-1}(x_i-\mu) - \frac{N}{2}\log\det(2\pi\Sigma) \\
&= -\frac{1}{2}\mathrm{trace}(\Sigma^{-1}\underbrace{\sum_{i=1}(x_i-\mu)(x_i-\mu)^\top}_{:=M}) + \frac{N}{2}\log\det(2\pi\Sigma^{-1})
\end{aligned}
$$

- Taking the derivative w.r.t. Σ^{-1}:

$$
\frac{\partial}{\partial\Sigma^{-1}}L = -\frac{1}{2}M + \frac{N}{2}\Sigma
$$

- We realize Σ_{ML} to be the sample covariance:

$$
\Sigma_{ML} = \frac{1}{N}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^\top
$$

Bayesian Estimation for Gaussian Distribution

- Likelihood

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i \mid \mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

- Let us choose the following prior

$$p(\mu) = \mathcal{N}(\mu \mid \mu_0, \sigma_0^2)$$

- Now we can apply Bayes rule

$$p(\mu \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mu)p(\mu)}{\int_\mu p(\mathcal{D} \mid \mu)p(\mu)d\mu}$$

Bayesian Estimation for Gaussian Distribution

- Applying Bayes rule we obtain

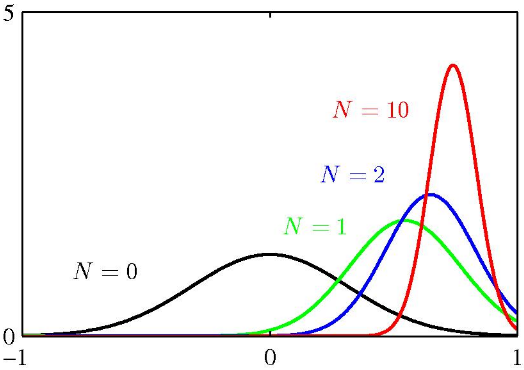
$$p(\mu \mid \mathcal{D}) = \mathcal{N}(\mu \mid \mu_n, \sigma_n^2)$$

which is again Gaussian.

- Parameters are a bit involved:

$$\begin{aligned} \mu_n &= \frac{\sigma^2}{n\sigma_0 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\mu_{ML} \\ \frac{1}{\sigma_n^2} &= \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \end{aligned}$$

Posterior for the Mean



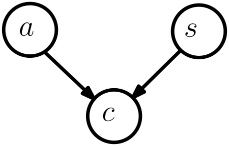
- Showing $p(\mu \mid \mathcal{D})$ for increasing size of $\mathcal{D}$
- True distribution is $p(x) \sim \mathcal{N}(x \mid \mu = 0.8, \sigma^2 = 0.1)$

Conjugate Priors

- For this case
  - the likelihood was Gaussian
  - the prior was Gaussian
  - the posterior was Gaussian
- This was no luck, but a conjugate prior
- "Def": For a given likelihood a prior is conjugate if the posterior is of the same parametric form as the prior
- In general very hard

Maximum Likelihood for Belief Networks

Lung Cancer

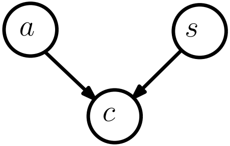


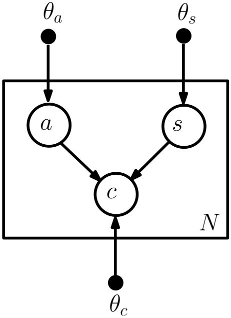
- Patient
  - has lung cancer $c \in \{0,1\}$
  - was exposed to asbestos $a \in \{0,1\}$
  - is a smoker $s \in \{0,1\}$
- Given the following relationship

$$p(a,s,c) = p(c \mid a,s)p(a)p(s)$$

- What are the parameters to learn? conditional probability table (CPT)

Yet another drawing convention: Plate notation





- Replicating data points
- The parts in the box factorize
- Priors over parameters for all points

Lung Cancer

$$p(a, s, c) = p(c \mid a, s)p(a)p(s)$$

- Observe patients: $\mathcal{D} = \{(a_1, s_1, c_1), (a_2, s_2, c_2), \ldots\}$
- The log-likelihood

$$\begin{aligned}
\log \mathcal{L}(\theta; \mathcal{D}) &= \sum_i \log p(a_i, s_i, c_i) \\
&= \sum_i \log p(c_i \mid a_i, s_i) + \log p(a_i) + \log p(s_i)
\end{aligned}$$

Lung Cancer

$$\log \mathcal{L} = \sum_i \log p(c_i \mid a_i, s_i) + \log p(a_i) + \log p(s_i)$$

- Observe patients: $(a_1, s_1, c_1), (a_2, s_2, c_2), \ldots$
- Now count:
- Denote $n(a=1, s=1, c=1) = |\{i \mid a_i = 1, s_i = 1, c_i = 1\}|$
- Similarly $n(a=0, s=1, c=1), \ldots, n(a=0, s=0, c=0)$
- All terms in the log-Likelihood with $p(c \mid a=1, s=0)$

$$\begin{aligned} & n(a=1, s=0, c=1) \log p(c=1 \mid a=1, s=0) \\ + \quad & n(a=1, s=0, c=0) \log(1 - p(c=1 \mid a=1, s=0)) \end{aligned}$$

Lung Cancer

- Use shorthand $\theta = p(c = 1 \mid a = 1, s = 0)$

$$n(a = 1, s = 0, c = 1) \log \theta + n(a = 1, s = 0, c = 0) \log(1 - \theta)$$

- Differentiating wrt. θ

$$\frac{n(a = 1, s = 0, c = 1)}{\theta} - \frac{n(a = 1, s = 0, c = 0)}{(1 - \theta)} = 0$$

- Therefore

$$\theta = \frac{n(a = 1, s = 0, c = 1)}{n(a = 1, s = 0, c = 1) + n(a = 1, s = 0, c = 0)}$$

- Maximum Likelihood solution simply corresponds to counting!

Formal derivation of ML

- That was too informal, is that general?
- Yes! Belief network can be written as factorization:

$$p(x) = \prod_{i=1}^{K} p(x_i \mid \mathrm{pa}(x_i))$$

- Recall: Maximizing the Likelihood corresponds to minimizing the KL divergence between the empirical distribution $q(x)$ (the training data) and $p(x)$ (our model)

$$\begin{aligned} KL(q\|p) &= -\left\langle \sum_{i=1}^{K} \log p(x_i \mid \mathrm{pa}(x_i)) \right\rangle_{q(x)} + const \\ &= -\sum_{i=1}^{K} \langle \log p(x_i \mid \mathrm{pa}(x_i)) \rangle_{q(x_i, \mathrm{pa}(x_i))} + const \end{aligned}$$

Formal derivation of ML

$$
\begin{aligned}
\underset{p}{\operatorname{argmin}}\, KL(q\|p) &= \underset{p}{\operatorname{argmin}} -\sum_{i=1}^{K} \langle \log p(x_i \mid \mathrm{pa}(x_i)) \rangle_{q(x_i, \mathrm{pa}(x_i))} + const \\
&= \underset{p}{\operatorname{argmin}} \sum_{i=1}^{K} \big(\langle \log q(x_i \mid \mathrm{pa}(x_i)) \rangle_{q(x_i, \mathrm{pa}(x_i))} \\
&\quad - \langle \log p(x_i \mid \mathrm{pa}(x_i)) \rangle_{q(x_i, \mathrm{pa}(x_i))} \big) \\
&= \underset{p}{\operatorname{argmin}} \sum_{i=1}^{K} \langle KL(q(x_i \mid \mathrm{pa}(x_i)) \| p(x_i \mid \mathrm{pa}(x_i))) \rangle_{q(x_i, \mathrm{pa}(x_i))}
\end{aligned}
$$

- Thus the following choice is maximizing ML (minimizing KL)

$$p(x_i \mid \mathrm{pa}(x_i)) = q(x_i \mid \mathrm{pa}(x_i))$$

ML solution

- We should set p as follows

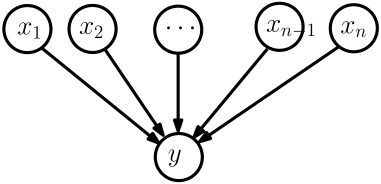
$$p(x_i \mid \mathrm{pa}(x_i)) = q(x_i \mid \mathrm{pa}(x_i))$$

- for given empirical distribution

$$p(x_i = s \mid pa(x_i) = t) \propto \sum [x_i = s, pa(x_i) = t]$$

That's it – that's all

- ML corresponds to counting, is there more?
- What may be the problem with this BN?



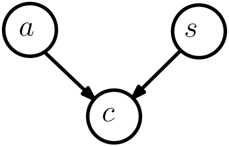
- CPT contains 2^n entries
- Solution: parametrize CPT with fewer variables

Conditional probability functions

- Instead of storing all 2^n entries of the CPT, one could fit a function
- For example
$$p(y = 1 \mid x, w) = \frac{1}{1 + \exp(-x^\top w)}$$
- Now the parameters are w of size n
- This also acts as regularization, fewer degrees of freedom
- How to find ML solution w_{ML}?

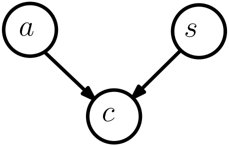
Bayesian Learning of Belief Networks

Bayesian Learning of BN



- The Bayesian approach:
  - Define a prior on the parameters $p(\theta)$
  - Then compute $p(\theta \mid \mathcal{D})$

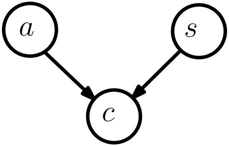
Bayesian Learning of BN – Parameters



- $p(a = 1 \mid \theta_a) = \theta_a$
- $p(s = 1 \mid \theta_s) = \theta_s$
- $p(c = 1 \mid a = 0, s = 1, \theta_c) = \theta_c^{0,1}$
- $p(c = 1 \mid a = 1, s = 1, \theta_c) = \theta_c^{1,1}$
- ...
- In total we have parameters

$$\theta_a, \theta_s, \underbrace{\theta_c^{0,0}, \theta_c^{1,0}, \theta_c^{0,1}, \theta_c^{1,1}}_{\theta_c}$$

Bayesian Learning of BN – Prior Assumptions



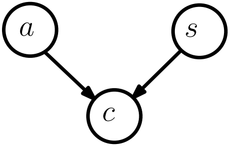
- We model the prior over θ as

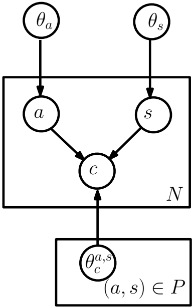
$$p(\theta_a, \theta_s, \theta_c) = p(\theta_a)p(\theta_s)p(\theta_c)$$

- Several other choices – our model freedom
- For example we could choose

$$p(\theta_c) = p(\theta_c^{0,0})p(\theta_c^{1,0})p(\theta_c^{0,1})p(\theta_c^{1,1})$$

Plate notation



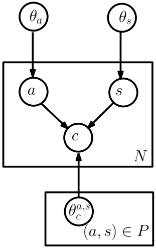


- This situation in plate notation
- Prior $p(\theta_c) = \prod_{a,s \in P} p(\theta_c^{a,s})$, with $P = \{(0,0), (1,0), (0,1), (1,1)\}$

Bayesian Learning of BN – Prior Assumptions

- Bayes rule

$$p(\theta_a, \theta_s, \theta_c \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta_a, \theta_s, \theta_c) p(\theta_a, \theta_s, \theta_c)}{p(\mathcal{D})}$$



Bayesian Learning

$$
\begin{aligned}
p(\theta_a, \theta_s, \theta_c \mid \mathcal{D}) &\propto p(\mathcal{D} \mid \theta_a, \theta_s, \theta_c) p(\theta_a, \theta_s, \theta_c) \\
&= (p(\theta_a) \prod_n p(a_n \mid \theta_a))(p(\theta_s) \prod_n p(s_n \mid \theta_s)) \\
&\quad (p(\theta_c) \prod_n p(c_n \mid s_n, a_n, \theta_c)) \\
&\propto p(\theta_a \mid \mathcal{V}_a) p(\theta_s \mid \mathcal{V}_s) p(\theta_c \mid \mathcal{V}_c)
\end{aligned}
$$

- Prior $p(\theta)$ factorizes $\Rightarrow$ posterior factorizes
- Each part can be optimized in parallel

First look at $p(\theta_a \mid \mathcal{D}_a)$

- Now look at one parameter only: $p(a = 1 \mid \theta_a) = \theta_a$
- Likelihood contribution of this parameter (Binomial distribution)

$$\prod_i p(a_i \mid \theta_a) = \theta_a^{n(a=1)}(1 - \theta_a)^{n(a=0)}$$

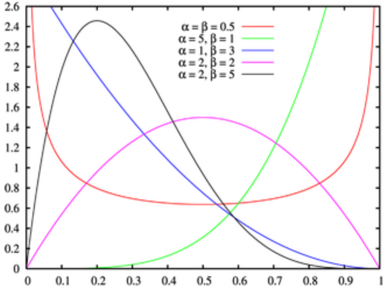
- and the posterior ($\propto$ prior $\times$ likelihood)

$$p(\theta_a \mid \mathcal{D}_a) \propto p(\theta_a) \times \theta_a^{n(a=1)}(1 - \theta_a)^{n(a=0)}$$

- This suggests to set prior to the Beta-distribution (why?)

$$p(\theta_a) = B(\theta_a \mid \alpha_a, \beta_a) = \frac{1}{B(\alpha_a, \beta_a)}\theta_a^{\alpha_a - 1}(1 - \theta_a)^{\beta_a - 1}$$

The Beta distribution



- Some examples of the Beta distribution

$$B(x \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Posterior distribution

- Because of conjugacy
  - Prior – Beta distribution
  - Likelihood – Binomial distribution
  - Posterior – Beta distribution
- Posterior parameters:

$$p(\theta_a \mid \mathcal{D}_a) = B(\theta_a \mid \alpha_a + n(a=1), \beta_a + n(a=0))$$

- and thus (remember: $p(a=1 \mid \theta_a) = \theta_a$)

$$p(a=1 \mid \mathcal{D}_a) = \int_{\theta_a} p(\theta_a \mid \mathcal{D}_a)\theta_a = \frac{\alpha_a + n(a=1)}{\alpha_a + n(a=1) + \beta_a + n(a=0)}$$

Limits: No data – Infinite amount of data

- No data limit ($N \to 0$)

$$\begin{aligned} p(a = 1 \mid \mathcal{D}_a) &= \frac{\alpha_a + n(a=1)}{\alpha_a + n(a=1) + \beta_a + n(a=0)} \\ &\to \frac{\alpha_a}{\alpha_a + \beta_a} \end{aligned}$$

- CPT entry corresponds to the prior (mean of Beta distribution)
- Infinite data limit ($N \to \infty$)

$$\begin{aligned} p(a = 1 \mid \mathcal{D}_a) &= \frac{\alpha_a + n(a=1)}{\alpha_a + n(a=1) + \beta_a + n(a=0)} \\ &\to \frac{n(a=1)}{n(a=1) + n(a=0)} \end{aligned}$$

- CPT entry corresponds to ML solution

Example

- Assume we have observed the following seven patients

a	s	c
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

- Let us use a flat prior for $p(a = 1)$ that is $\alpha_a = \beta_a = 1$

Example – marginal posterior

From last slide:

$$p(a = 1 \mid \mathcal{D}_a) = \frac{\alpha_a + n(a=1)}{\alpha_a + n(a=1) + \beta_a + n(a=0)}, \alpha_a = \beta_a = 1$$

$$p(a = 1 \mid \mathcal{D}_a) = \frac{1 + n(a=1)}{2 + N} = \frac{5}{9} \approx 0.556$$

- Different to the Maximum Likelihood setting, that is $4/7 = 0.571$
- Bayesian result is "pulling" towards the prior (of 0.5)

a	s	c
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

More states than binary

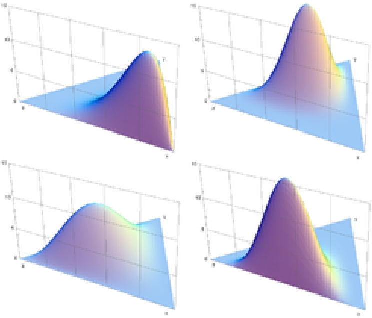
- So far only binary variables (binomial,Beta)
- Now let a variable v take values in $\{1, \ldots, I\}$
- Therefore the posterior

$$
\begin{aligned}
p(\theta \mid \mathcal{D}) &\propto p(\theta) \prod_{n=1}^{N} \prod_{i=1}^{I} \theta_i^{[v_n = i]} \\
&= p(\theta) \prod_{i=1}^{I} \theta_i^{\sum_{n=1}^{N} [v_n = i]}
\end{aligned}
$$

- Suggests a Dirichlet prior

$$
p(\theta) \propto \prod_{i=1}^{I} \theta_i^{u_i - 1}
$$

The Dirichlet distribution



- Some examples of the Dirichlet distribution

$$Dirichlet(x \mid u) = \frac{\Gamma(\sum_i u_i)}{\prod_i \Gamma(u_i)} \prod_{i=1}^{I} x_i^{u_i - 1}$$

Maximum Likelihood Learning of Undirected Models

Maximum Likelihood for MRF

- Markov Network defined on cliques

$$p(\mathcal{X} \mid \theta) = \frac{1}{Z(\theta)} \prod_{c} \Phi_c(\mathcal{X}_c \mid \theta_c)$$

- Partition function

$$Z(\theta) = \sum_{\mathcal{X}} \prod_{c} \Phi_c(\mathcal{X}_c \mid \theta_c)$$

- Training data $\mathcal{D} = \{\mathcal{X}^1, \ldots, \mathcal{X}^N\}$
- Log-Likelihood

$$L(\theta; \mathcal{D}) = \sum_{n} \sum_{c} \log \Phi_c(\mathcal{X}_c^n \mid \theta_c) - N \log Z(\theta)$$

Comments

- For Belief Networks, the posterior decomposed into different parts (due to independence of the prior)
- Here this is not the case (in general)
- Difficulty is the unknown partition function $Z(\theta)$

Optimizing

- If there is no closed form solution of θ,
  we can can try to optimize θ numerically!
- For example: gradient descent
  - Init at θ^0
  - Update $\theta^t = \theta^{t-1} + \epsilon \frac{\partial}{\partial \theta} L(\theta^{t-1})$

Likelihood Gradient

- The Log-Likelihood (repeated from last slide)

$$L(\theta; \mathcal{D}) = \sum_n \sum_c \log \Phi_c(\mathcal{X}_c^n \mid \theta_c) - N \log Z(\theta)$$

- and its gradient?

$$\begin{aligned}\frac{\partial}{\partial \theta_c} L(\theta) &= N \left\langle \frac{\partial}{\partial \theta_c} \log \Phi_c(\mathcal{X}_c \mid \theta_c) \right\rangle_{q(\mathcal{X})} \\ &\quad - N \left\langle \frac{\partial}{\partial \theta_c} \log \Phi_c(\mathcal{X}_c \mid \theta_c) \right\rangle_{p(\mathcal{X}_c \mid \theta)}\end{aligned}$$

- Empirical distribution/Training Data $q(x) = \frac{1}{N} \sum_{i=1}^{N} [x = x_i]$
- Last term depends on $p(\mathcal{X}_c \mid \theta)$ (model average)

Model Average

- In order to compute the gradient we need to compute

$$\left\langle \frac{\partial}{\partial \theta_c} \log \Phi_c(\mathcal{X}_c \mid \theta_c) \right\rangle_{p(\mathcal{X}_c|\theta)}$$

- Either we can compute it
  - Tree graphical models
- Or we have to approximate it
  - Sampling, Variational Approximation
- Or we could choose a different score for estimating θ
  - Pseudo-Likelihood, Max-Margin, Moment Matching, ...

Next Lectures

... Computer Vision, finally!

- Human Body Models
- Stereo, Optical Flow
- Image Segmentation
- Object Detection