Patterns of Scalable Bayesian Inference
Background (Session 1)

Jerónimo Arenas-García

Universidad Carlos III de Madrid

jeronimo.arenas@gmail.com

June 14, 2017

Motivation. Bayesian Learning principles

- Main property of Bayesian methods: accounting for uncertainty

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

- Typically we have to average over the posterior. Two kinds of approximation methods:
  - Monte Carlo sampling
  - Variational methods

Bayesian methods in the big data scenario

- As the available training data becomes increasingly large, $p(\theta|x)$ would concentrate around the maximum of $p(x|\theta)$
  - $\hat{\theta}_{\mathrm{MAP}} \to \hat{\theta}_{\mathrm{ML}}$
  - $\int f(\theta)p(\theta|x)d\theta \to f(\hat{\theta}_{\mathrm{MAP}})$
- This will not be the case in many big data scenarios, as the number of parameters and the model complexity itself grows also with the available data
- When recurring traditional Bayesian methods in big data applications we need to be aware of
  - Computation and memory issues: sequential and parallel methods
  - It may be better to give up some desired properties (such as asymptotic unbiasedness) in favor of better scalability
  - The tradeoff between scalability and computational complexity

The book deals with scalable methods for Monte Carlo sampling and variational methods

Outline

1. Exponential families
2. Markov Chain Monte Carlo Sampling
3. Mean field variational inference
4. Stochastic gradient optimization

The exponential families of distributions (I)

- x is a random vector, $x \in \mathcal{X} \subset \Re^d$
- θ is a (random) parameter vector, $\theta \in \Theta \subset \Re^m$

- **Definition**: $\{p(\cdot|\theta)\}$ is a exponential family if the probability density function of the family can be written down as

$$p(x|\theta) = h(x) \exp\left[\langle \eta(\theta), t(x) \rangle - \log Z(\eta(\theta))\right]$$

  - $\langle \eta(\theta), t(x) \rangle = \sum_{k=1}^{K} \eta_k(\theta) t_k(x)$
  - $\eta_k(\theta)$ are the **natural parameters**
  - $t_k(x)$ are the **natural statistics**
  - $t_k(x)$ are sufficient statistics, i.e., $\theta \perp x | t_k(x)$
  - $A(\theta) = \log Z(\eta(\theta))$ ensures normalization $\forall \theta$

The exponential families of distributions (II)

- We assume that $\mathcal{X}$ does not depend on θ
- Regular family: If Θ is an open set
- The family is minimum if $\nexists a \neq 0$ such that $\langle a, t(x) \rangle$ is constant

Exercise 1: Natural parameter form for the Bernouilli distribution

Exercise 2: Natural parameter form for the Gaussian distribution

The exponential families of distributions (III)

Gaussian	$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\\| x-\mu \\|^2/(2\sigma^2)}$	$x \in \mathbb{R}$
Bernoulli	$p(x) = \alpha^x\,(1-\alpha)^{1-x}$	$x \in \{0,1\}$
Binomial	$p(x) = \binom{n}{x}\,\alpha^x\,(1-\alpha)^{n-x}$	$x \in \{0,1,2,\ldots,n\}$
Multinomial	$p(x) = \frac{n!}{x_1!x_2!\ldots x_n!}\,\prod_{i=1}^{n}\alpha_i^{x_i}$	$x_i \in \{0,1,2,\ldots,n\}\,,\ \sum_i x_i = n$
Exponential	$p(x) = \lambda\, e^{-\lambda x}$	$x \in \mathbb{R}^+$
Poisson	$p(x) = \frac{e^{-\lambda}}{x!}\,\lambda^x$	$x \in \{0,1,2,\ldots\}$
Dirichlet	$p(x) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\,\prod_i x_i^{\alpha_i - 1}$	$x_i \in [0,1]\,,\ \sum_i x_i = 1$

Sampling of a exponential family distribution

If x follows an exponential family distribution with natural parameters $\eta_k(\theta)$ and natural statistics $t_k(x)$, then a collection of samples from the distribution, $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ follows an exponential family distribution with the same natural parameters and natural statistics

$$t'_k(\mathbf{x}) = \sum_{i=1}^{n} t_k(x_i)$$

- Again, $t'_k(\mathbf{x})$ are sufficient statistics for the joint distribution
- This is an important property for big data scenarios

The role of $Z(\eta(\theta))$

To ensure proper normalization of the distribution:

$$\int p(x|\theta)dx = \exp[-A(\eta)] \int h(x) \exp\left[\eta^\top t(x)\right] dx = 1$$

Since $A(\eta) = \log Z(\eta)$, we have

$$Z(\eta) = \int h(x) \exp\left[\eta^\top t(x)\right] dx,$$

i.e., for $t(x) = x$, we would have the Laplace Transform of h(x).

Derivatives of $A(\eta(\theta))$

$$A(\eta) = \log Z(\eta) = \log \int h(x) \exp \left[\eta^\top t(x) \right] dx,$$

Exercise: Show that $\nabla_\eta A(\eta) = \mathbb{E}\{t(x)\}$

Exercise: Verify that the previous result holds for the Gaussian distribution

Proposition: $\nabla^2_\eta A(\eta) = Cov\{t(x)\}$

Moment generating function of exponential families

- For a general random variable x, its moment generating function is defined as $M_x(s) = \mathbb{E}\{e^{sx}\}$
- If it exists, it allows an easy calculation of moments, since

$$\mathbb{E}\{e^{sx}\} = 1 + \mathbb{E}\{sx\} + \frac{\mathbb{E}\{(sx)^2\}}{2!} + \frac{\mathbb{E}\{(sx)^3\}}{3!} + \dots$$

  Therefore, $\nabla_s^k M_x(s) \mid_{s=0} = \mathbb{E}\{x^k\}$
- For an exponential family, cumulants of $t(x)$ can be obtained from

$$M_T(s) = \mathbb{E}\{e^{s^\top t(x)}\} = e^{A(\eta+s)-A(\eta)}$$

Other properties

- For a regular exponential family, we define the score with respect the natural parameters as:

$$v(x,\eta) = \nabla_\eta \log p(x|\eta) = \nabla_\eta \left[\langle \eta, t\rangle - \log Z(\eta)\right]$$

$$= t(x) - \nabla_\eta \log Z(\eta) = t(x) - \mathbb{E}\{t(x)\}$$

- For a regular exponential family, Fisher information with respect to the natural parameters is

$$I(\eta) = \mathbb{E}\{v(x,\eta)v^\top(x,\eta)\} = \mathbb{E}\left\{(t(x) - \mathbb{E}\{t(x)\})(t(x) - \mathbb{E}\{t(x)\})^\top\right\}$$

$$= Cov[t(x)] = \nabla_\eta^2 \log Z(\eta)$$

Conjugate priors in Bayesian statistics

$$p(\theta|x, \alpha') = \frac{p(\theta|\alpha)p(x|\theta)}{p(x)}$$

If $p(\theta|\alpha)$ and $p(\theta|x, \alpha')$ are parametric forms of the same family, then we say that the prior $p(\theta|\alpha)$ is conjugated with the likelihood function $p(x|\theta)$

- In general, $\alpha' = \alpha'(\alpha, x)$
- The definition is general ...
- ... but if $p(x|\theta)$ is from a regular exponential family, then it is always possible to find a conjugated prior (which is also an exponential family)

Conjugate priors in Bayesian statistics: an example

- Dirichlet prior:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\Pi_i \Gamma(\alpha_i)} \Pi_i \theta_i^{\alpha_i - 1}$$

- Multinomial likelihood:

$$p(x|\theta) = \frac{(\sum_i x_i)!}{x_1! x_2! \ldots x_n!} \Pi_i \theta_i^{x_i}$$

- Posterior:

$$p(\theta|x, \alpha) \propto \Pi_i \theta_i^{(x_i + \alpha_i - 1)}$$

Therefore, the posterior is Dirichlet with parameters $\alpha_i + x_i$, implying that the Dirichlet and the Multinomial are conjugated.

Conjugate pairs

Prior		Conditional	
Gaussian	$e^{-\|\mu-\mu_0\|^2/(2\sigma^2)}$	Gaussian	$e^{-\|x-\mu\|^2/(2\sigma^2)}$
Beta	$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\,\alpha^{r-1}\,(2-\alpha)^{s-1}$	Bernoulli	$\alpha^x\,(1-\alpha)^{1-x}$
Dirichlet	$\frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)}\prod \theta_i^{\alpha_i-1}$	Multinomial	$\frac{(\sum x_i)!}{\prod x_i!}\prod \theta_i^{x_i}$
Inv. Wishart		Gaussian (cov)	

Note: Conjugacy is mutual, e.g.

Dirichlet $\rightarrow$ Multinomial $\rightarrow$ Dirichlet

Multinomial $\rightarrow$ Dirichlet $\rightarrow$ Multinomial