Bristol Machine Learning Reading Group

Introduction to Variational Inference

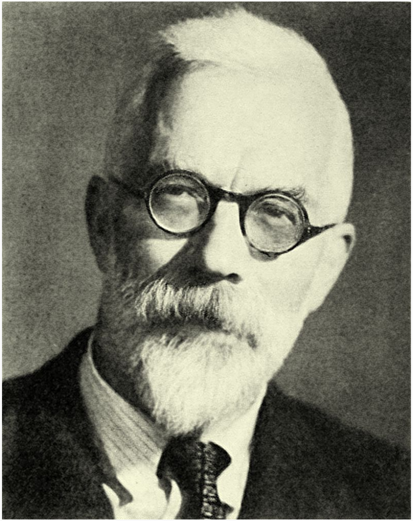
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

November 25, 2016

http://www.carlhenrik.com

Introduction

Ronald Aylmer Fisher



TODAY

$$p(Y) = \int p(Y, X) dX \qquad\qquad p(X|Y) = \frac{p(Y, X)}{p(Y)}$$

- *"Being Bayesian"* implies not making a point estimate, only deductively impossible scenarios should be given zero probability
- Learning: maximise evidence of data
- Decision/Reasoning: posterior distribution

The evidence is the key-quantity in machine learning as it includes all possible knowledge

In practice

$$p(Y) = \int p(Y, X) dX$$

$$p(X|Y) = p(Y|X)\frac{p(X)}{p(Y)}$$

In practice

- We can usually formulate joint distribution
  - most commonly as likelihood times prior
- reaching posterior is hard
  - as evidence is challenging to compute

Laplace quote

"Nature laughs at the difficulties of integration"
– Simon Laplace

Pachinko

YouTube

Two paths

$$p(Y) \approx \sum_i p(Y, X_i)$$

$$X_i \sim p(X)$$

Stochastic

+ correct in limit
- now evidence of approximation

$$p(Y) = L(q(X)) + D(q(X))$$

$$q(X) \approx p(X|Y)$$

Deterministic

+ know how good approximation is
- will never be correct

Variational Inference

Formalise

$$\log\, p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{X} = \log \int p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X}$$

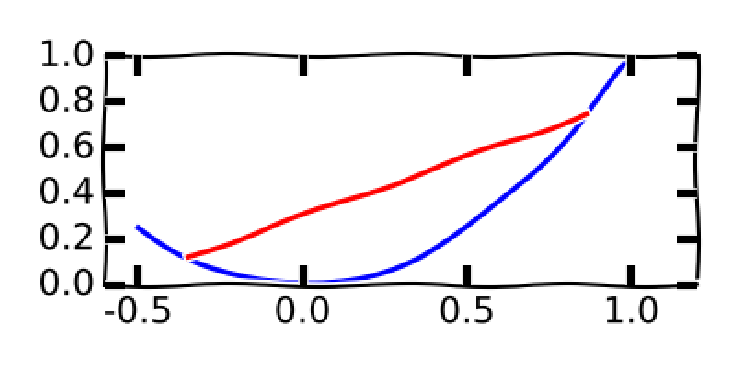
$$\log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X}$$

Jensen Inequality

Convex Function

$$\lambda f(x_0) + (1-\lambda) f(x_1) \geq f(\lambda x_0 + (1-\lambda) x_1)$$
$$x \in [x_{min}, x_{max}]$$
$$\lambda \in [0, 1]]$$

Jensen Inequality



$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

$$\int f(x)p(x)\mathrm{d}x \geq f\left(\int xp(x)\mathrm{d}x\right)$$

Jensen Inequality in Variational Bayes

$$\int \log(x)p(x)\mathrm{d}x \leq \log\left(\int xp(x)\mathrm{d}x\right)$$

moving the log inside the the integral is a lower-bound on the integral

Variational Bayes cont.

$$
\begin{aligned}
\log p(\mathbf{Y}) &= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}) \mathrm{d}\mathbf{X} = \\
&\geq \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} + \int \mathrm{d}\mathbf{X} p(\mathbf{Y}) \\
&= -\mathrm{KL}\left(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})\right) + \log\ p(\mathbf{Y})
\end{aligned}
$$

Variational Bayes cont.

$$\begin{aligned}
\log p(\mathbf{Y}) &= \log \int \frac{q(\mathbf{X})}{q(\mathbf{X})} p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})\mathrm{d}\mathbf{X} = \\
&\geq \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})}{q(\mathbf{X})}\mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})}\mathrm{d}\mathbf{X} + \int \mathrm{d}\mathbf{X} p(\mathbf{Y}) \\
&= -\mathrm{KL}\left(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})\right) + \log\, p(\mathbf{Y})
\end{aligned}$$

- if $q(\mathbf{X})$ is the true posterior we have an equality, therefore match the distributions
- i.e. $\operatorname{argmin}_q \mathrm{KL}\left(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})\right)$

$\Rightarrow$ variational distributions are approximations to intractable posteriors

ELBO

$$
\begin{aligned}
\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) &= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \mathrm{d}\mathbf{X} \\
&= \int q(\mathbf{X}) log \frac{q(\mathbf{X})}{p(\mathbf{X},\mathbf{Y})} \mathrm{d}\mathbf{X} + log\ p(\mathbf{Y}) \\
&= H(q(\mathbf{X})) - \mathbb{E}_{q(\mathbf{X})}\left[log\ p(\mathbf{X},\mathbf{Y})\right] + log\ p(\mathbf{Y})
\end{aligned}
$$

ELBO

$$
\begin{aligned}
log\ p(\mathbf{Y}) &= \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) + \underbrace{\mathbb{E}_{q(\mathbf{X})}\left[log\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X}))}_{\text{ELBO}} \\
&\geq \mathbb{E}_{q(\mathbf{X})}\left[log\ p(\mathbf{X},\mathbf{Y})\right] - H(q(\mathbf{X})) = \mathcal{L}(q(\mathbf{X}))
\end{aligned}
$$

Evidence Lower BOund

- if we maximise the ELBO we,
  - find an approximate posterior
  - get an approximation to the marginal likelihood
- *maximising* $p(\mathbf{Y})$ is learning
- finding $p(\mathbf{X}|\mathbf{Y}) \approx q(\mathbf{X})$ is prediction

ELBO

```
% Define block styles
\usetikzlibrary{shapes,arrows}
\tikzstyle{astate} = [circle, draw, text centered, font=\foo
\tikzstyle{rstate} = [circle, draw, text centered, font=\foo
\tikzstyle{bstate} = [circle, draw, text centered, font=\foo

\begin{tikzpicture}[->,>=stealth', shorten >=1pt, auto, node
\node [astate] (X) at (0,1.5) {X};
\node [rstate] (Y) at (0,0) {Y};
\node [astate] (X2) at (1.5,1.5) {X};
\node [rstate] (Y2) at (1.5,0) {Y};
\node [bstate] (T) at (2.3,1.5) {$\theta$};
\path (X) edge (Y);
\end{tikzpicture}
```

Why is this useful?

Why is this a sensible thing to do?

- Taking the expectation of a log is usually easier than the expectation
- We are allowed to choose the distribution to take the expectation over

– Ryan Adams in Talking Machines[1]

[1] Talking Machines - Season Two Episode Five

Approximate Distribution

Mean Field Approximation

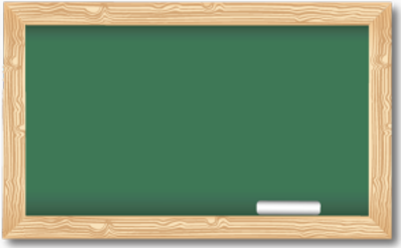
$$q(\mathbf{X}) = \prod_i q_i(X_i)$$

- Introduced in statistical physics[2]
- Approximates the marginals of the posterior

[2]Peterson, C., and Anderson, J. R. (1987) A mean field theory learning algorithm for neural networks

Examples

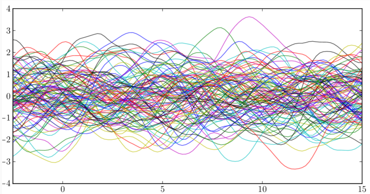
Ising Model

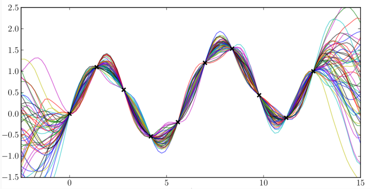


Gaussian Process

Why?

- Not directly applicable to variational bayes
- Introduces variational compression by augumentation
- Exemplifies well what VB is in practice





Gaussian Process 101

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{X},\mathbf{X}) & k(\mathbf{X},\mathbf{x}_*) \\ k(\mathbf{x}_*,\mathbf{X}) & k(\mathbf{x}_*,\mathbf{x}_*) \end{bmatrix} \right)$$

$$p(\mathbf{f}_*|\mathbf{x}_*,\mathbf{X},\mathbf{f},\boldsymbol{\theta}) = \mathcal{N}(k(\mathbf{x}_*,\mathbf{X})^{\mathrm{T}} K(\mathbf{X},\mathbf{X})^{-1}\mathbf{f},$$
$$k(\mathbf{x}_*,\mathbf{x}_*) - k(\mathbf{x}_*,\mathbf{X})^{\mathrm{T}} K(\mathbf{X},\mathbf{X})^{-1} K(\mathbf{X},\mathbf{x}_*))$$

Joint Distribution

$$p(\mathbf{Y},\mathbf{F},\mathbf{X}) = p(\mathbf{Y}|\mathbf{F})(\mathbf{F}|\mathbf{X})p(\mathbf{X})$$
$$= p(\mathbf{X})\prod_{j=1}^{d} p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})$$

Learning Task

$$p(\mathbf{Y}) = \int p(\mathbf{Y}|\mathbf{F})(\mathbf{F}|\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X}\mathrm{d}\mathbf{F}$$

we can analytically integrate out $\mathbf{F}$ but $\mathbf{X}$ appears non-linearly w.r.t. $\mathbf{Y}$ rendering this intractable

Gaussian Process

$$
\begin{aligned}
\mathcal{L}_{\mathcal{A},\mathcal{B}} &= \int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}|\mathbf{F}) p(\mathbf{F}|\mathbf{X})) p(\mathbf{X})}{q(\mathbf{X})} \right) \\
&= \int_{\mathbf{F},\mathbf{X}} q(\mathbf{X})(\mathbf{Y}|\mathbf{F}) p(\mathbf{F}|\mathbf{X}) - \int_{\mathbf{X}} q(\mathbf{X}) \log \frac{q(\mathbf{X})}{p(\mathbf{X})} \\
&= \tilde{\mathcal{L}} - \mathrm{KL}\left(q(\mathbf{X}) \,\|\, p(\mathbf{X})\right)
\end{aligned}
$$

Gaussian Process

$$\tilde{\mathcal{L}} = \int_{\mathbf{F},\mathbf{X}} q(\mathbf{X}) \log p(\mathbf{Y}|\mathbf{F}) p(\mathbf{F}|\mathbf{X})$$

$$f \sim \mathcal{GP}(\mathbf{y}, k(\cdot,\cdot)) \Rightarrow p(\mathbf{F}|\mathbf{X}) = \prod_{j=1}^{d} \mathcal{N}(\mathbf{f}_{:,j}|\mathbf{0}, \mathbf{K})$$

$$k\left(\mathbf{x}_{:,i}, \mathbf{x}_{:,j}\right) = \sigma e^{-\frac{1}{2}\sum_{q=1}^{Q} w_q \left(x_{q,i} - x'_{q,j}\right)^2}$$

Gaussian Process

Add another set of samples from the same prior

$$p(\mathbf{U}|\mathbf{Z}) = \prod_{j=1}^{d} \mathcal{N}(\mathbf{u}_{:,j}|\mathbf{0}, \mathbf{K})yy$$

Conditional distribution

$$p(\mathbf{f}_{:,j}, \mathbf{u}_{:,j}|\mathbf{X}, \mathbf{Z}) = p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z})p(\mathbf{u}_{:,j}|\mathbf{Z})$$

$$= \mathcal{N}\left(\mathbf{f}_{:,j}|\mathbf{K}_{fu}(\mathbf{K}_{uu})^{-1}\mathbf{u}_{:,j}, \mathbf{K}_{ff} - \mathbf{K}_{fu}(\mathbf{K}_{uu})^{-1}\mathbf{K}_{uf}\right)\mathcal{N}\left(\mathbf{u}_{:,j}|\mathbf{0}, \mathbf{K}_{uu}\right),$$

New Augmented Model

$$p(\mathbf{Y}, \mathbf{F}, \mathbf{U}, \mathbf{X}|\mathbf{Z}) = p(\mathbf{X}) \prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}) p(\mathbf{u}_{:,j}|\mathbf{Z})$$

- we have done nothing to the model, just added *halucinated* observations
- however, $\mathbf{U}$ and $\mathbf{X}_u$ are not random but variational parameters

Gaussian Process

Variational distributions are approximations to intractable posteriors,

$$q(\mathbf{U}) \approx p(\mathbf{U}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{F})$$
$$q(\mathbf{F}) \approx p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{Y})$$
$$q(\mathbf{X}) \approx p(\mathbf{X}|\mathbf{Y})$$

If U is sufficient statistics of F this means,

$$p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{Y}) = p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z})$$

Gaussian Process

$$\begin{aligned}\tilde{\mathcal{L}} &= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{F},\mathbf{U}|\mathbf{X},\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})} \\ &= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j})p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X},\mathbf{Z})p(\mathbf{u}_{:,j}|\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})q(\mathbf{X})}\end{aligned}$$

Gaussian Process

$$\tilde{\mathcal{L}} = \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{F},\mathbf{U}|\mathbf{X},\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})}$$
$$= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j})p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X},\mathbf{Z})p(\mathbf{u}_{:,j}|\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})q(\mathbf{X})}$$

Assume that U is sufficient statistics for F

$$q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) = p(\mathbf{F}|\mathbf{U},\mathbf{X},\mathbf{Z})q(\mathbf{U})q(\mathbf{X})$$

Gaussian Process

$$\tilde{\mathcal{L}} = \int_{\mathbf{X},\mathbf{F},\mathbf{U}} \prod_{j=1}^{d} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) q(\mathbf{u}_{:,j}) q(\mathbf{X})$$

$$\log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) \cancel{p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z})} p(\mathbf{u}_{:,j}|\mathbf{Z})}{\prod_{j=1}^{d} \cancel{p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z})} q(\mathbf{u}_{:,j})} =$$

$$= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} \prod_{j=1}^{p} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) q(\mathbf{u}_{:,j}) q(\mathbf{X}) \log \frac{\prod_{j=1}^{p} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) p(\mathbf{u}_{:,j}|\mathbf{Z})}{\prod_{j=1}^{p} q(\mathbf{u}_{:,j})}$$

$$= \mathbb{E}_{q(\mathbf{F}),q(\mathbf{X}),q(\mathbf{U})} \left[p(\mathbf{Y}|\mathbf{F}) \right] - \mathrm{KL}\left(q(\mathbf{U}) || p(\mathbf{U}|\mathbf{Z}) \right)$$

Gaussian Process

Summary

$$\mathbb{E}_{q(\mathbf{F}),q(\mathbf{X}),q(\mathbf{U})}\left[p(\mathbf{Y}|\mathbf{F})\right] - \mathrm{KL}\left(q(\mathbf{U})||p(\mathbf{U}|\mathbf{Z})\right) - \mathrm{KL}\left(q(\mathbf{X})||p(\mathbf{X})\right)$$

- Expectation tractable
- Can be computed for certain priors
- Reduces to expectations over co-variance functions know as Ψ statistics

Conclusion

Summary

- Often efficient
- Not stochastic
- Provides you with posterior and a bound on marginal likelihood
- *its fun* a lot of the work relates to multi-variate calculus tricks and substitutions

Bristol Machine Learning Reading Group

Berkeley Tea Talk Style

- *everyone* reads paper
- someone introduces paper and leads discussion

+ constant workload on everyone

- requires everyone to take this serious

Seminar Style

- everyone *skims* paper
- someone is responsible for presenting paper

+ will work

- very uneven workload

Choosing the paper

- Presenter picks freely
- Presenter picks from agreed pool
- Currator chooses papers
- Topics
  - several papers on one topic
  - cover lots of single topics

eof

Source blocks

```
import numpy as np
import matplotlib.pyplot as plt
plt.xkcd()

plt.savefig(path)
return path
```