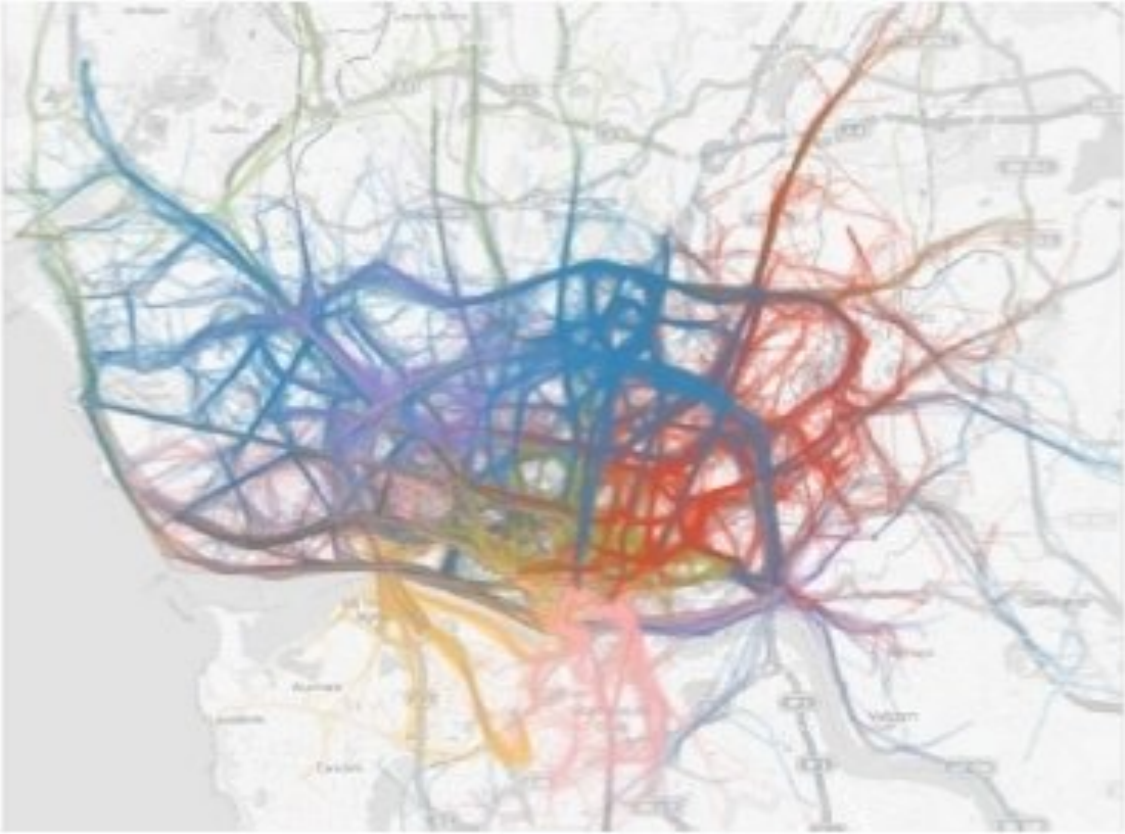
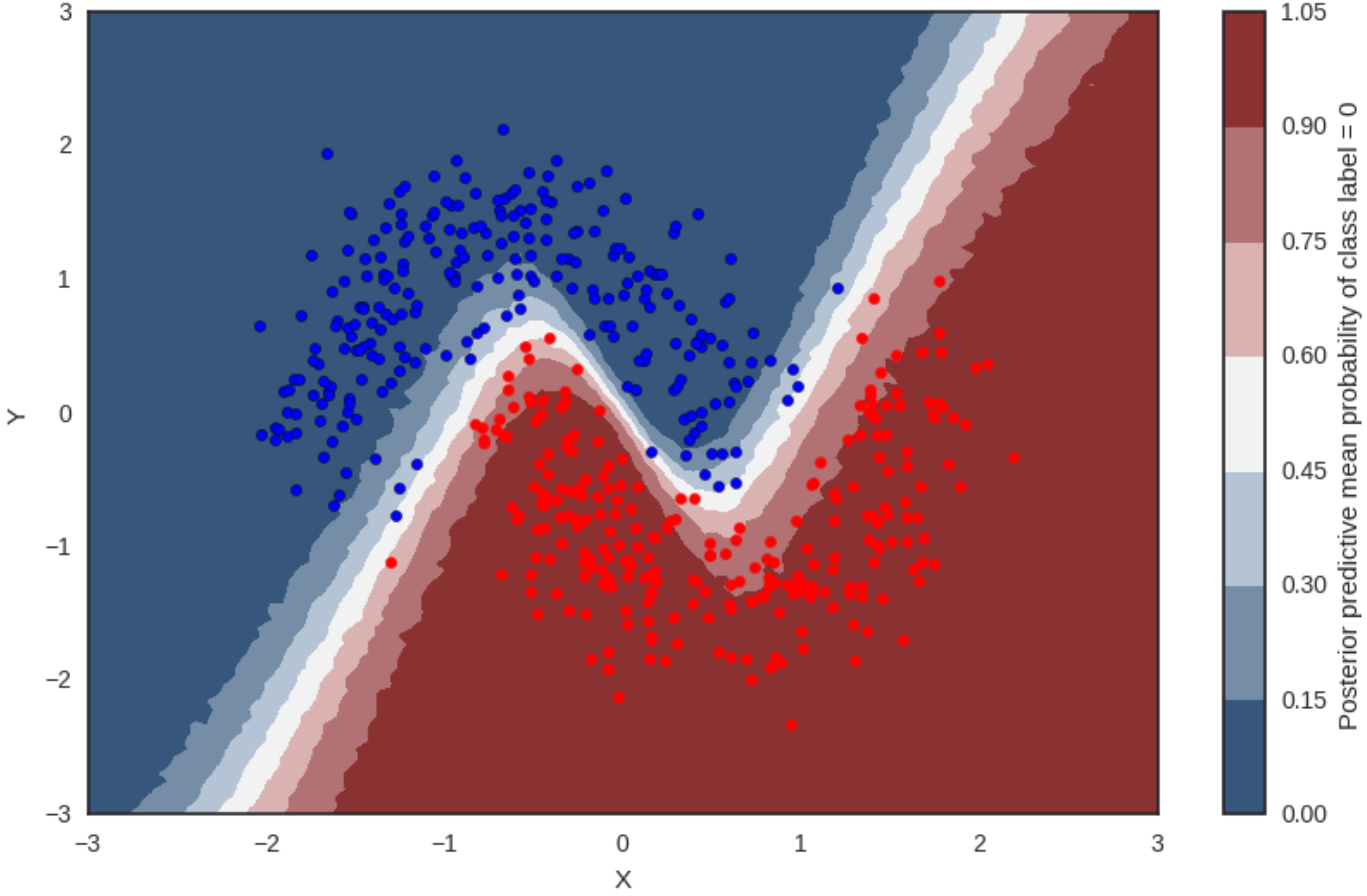
Don't just sample optimize

Peadar Coyle



Analysis of 1.7M taxi trajectories, in Stan

[Kucukelbir et al., 2016]



Bayesian Neural Networks - Thomas Wiecki - PyMC3 Docs

Challenges in Bayesian Inference

- 1. **Tradeoffs**. How do we formalize statistical and computational tradeoffs for inference?

- 2. **Software.** How do we design efficient and flexible software for generative models?

Why do we need Variational Inference?

- Inferring hidden variables
- Unlike MCMC:
  - Deterministic
  - Easy to gauge convergence
  - Requires dozens of iterations
- Doesn't require conjugacy
- Slightly hairier math

Background

<u>Given</u>

- Data set $\mathbf{x}$
- Generative model $p(\mathbf{x}, \mathbf{z})$ with latent variable $\mathbf{z} \in \mathbb{R}^d$

<u>Goal</u>

- Infer posterior $p(\mathbf{z}|\mathbf{x})$

That is the key problem in Bayesian inference

Let's look at the posterior

- We can write the conditional or posterior distribution as
$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}$$
- The denominator in the marginal distribution is called the *marginal distribution of observations (*also called the *evidence*) and it is calculated by *marginalizing* out the latent variables from the joint distribution

$$p(\mathbf{x}) = \int_z p(\mathbf{z}, \mathbf{x}) d\mathbf{z}$$

- Often this integral is intractable


APPROXIMATE
ALL THE THINGS
memegenerator.net

What do we approximate?

- We create a **variational distribution** over the latent variables $q\left(z_{1:m} \middle| \nu\right)$
- We want to find settings of ν
- So that q is close to p
- When p == q this is plain Expectation Maximization

What does closeness mean?

- We measure the closeness of distributions using Kullback-Leibler Divergence

$$\mathbb{E}_q[\log \frac{q(Z)}{p(Z|x)}]$$

- If q and p are high we're happy
- If KL = 0 , then the distributions are equal
- If q is low we don't care. If q isn't high but p isn't we pay a price
- http://bit.ly/2oROYAw

We can do some math...

$$-(\mathbb{E}_q[\log p(z|x)] - \mathbb{E}_q[\log q(z)]) + \log p(x)$$

Constant

ELBO (in brackets)

Negative of ELBO (evidence lower bound) + a constant is equal to KL divergence

Key points

- **Minimizing KL divergence** is the same as **maximizing ELBO**

- This allows us to change a **sampling problem** into an **optimization problem**

Whats new in PyMC3

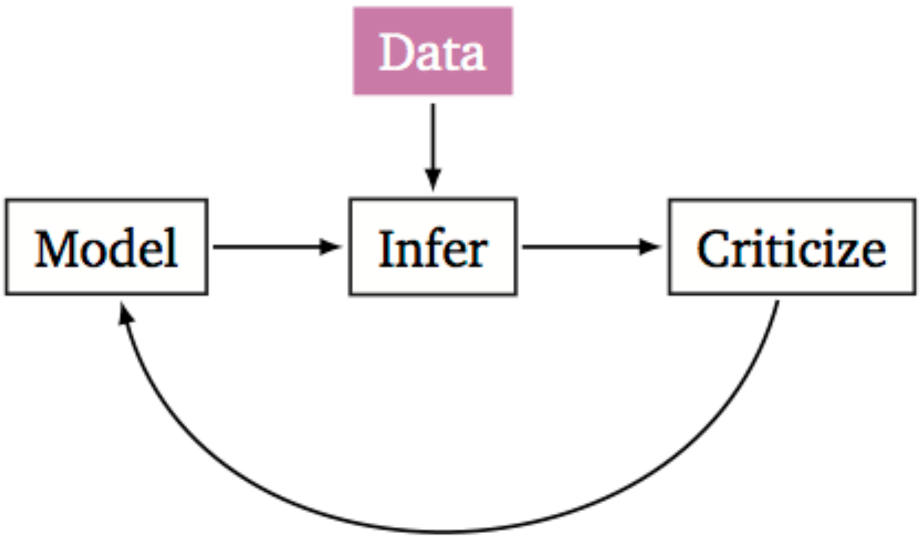
- Release of the first stable version in early 2017
- Variational Inference
- Advanced Hamiltonian Monte Carlo samplers
- Easy optimization for finding the MAP point.
- Theano support for fast compilation

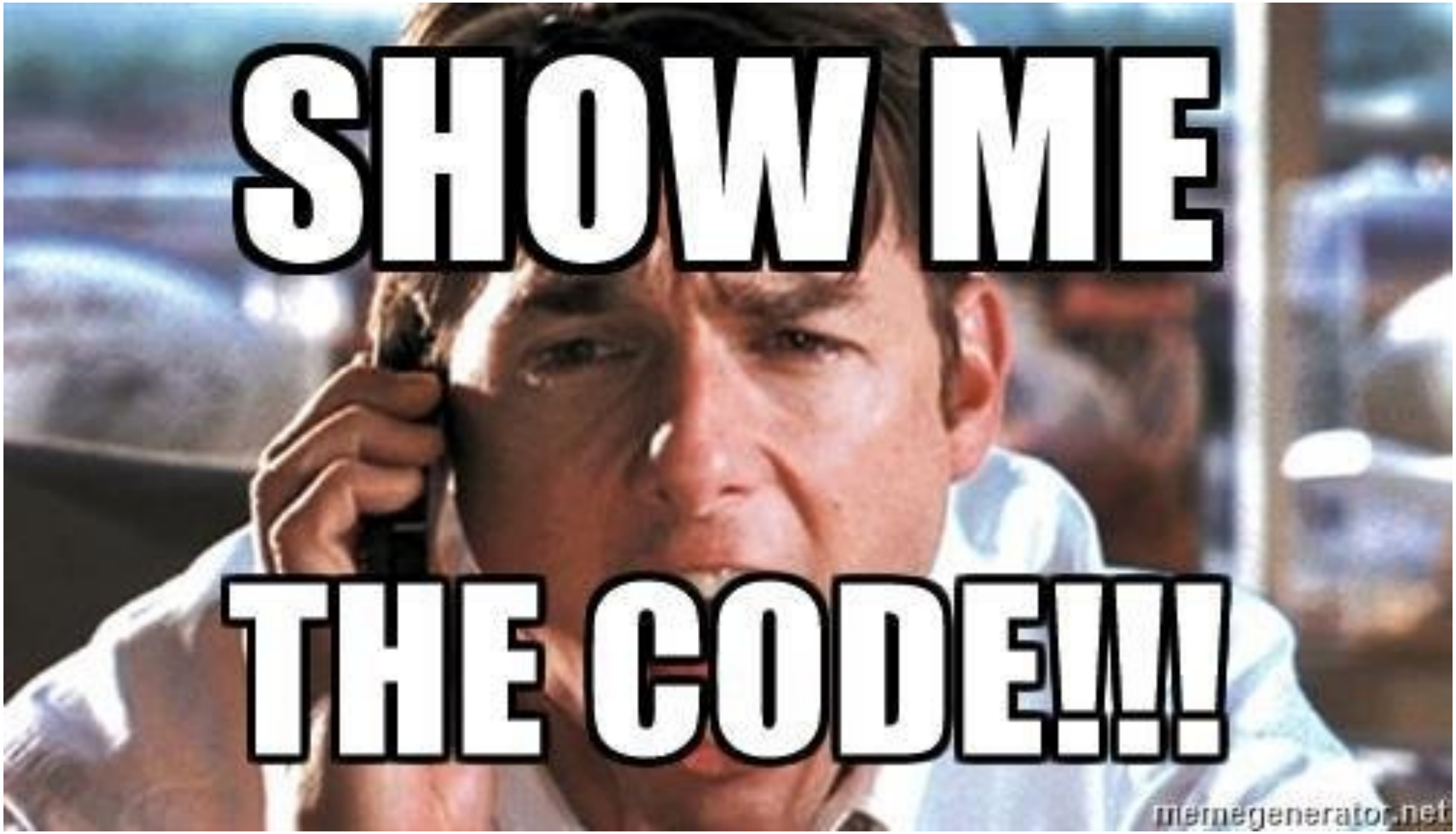
What else is new

- Gaussian process kernels
- New variants of Variational Inference (including Operator)
- Speed improvements
- API and documentation improvements
- Bayesian Methods for Hackers - in PyMC3 too

First gather data from some real-world phenomena. Then cycle through Box's loop:

1. Build a probabilistic model of the phenomena.
2. Reason about the phenomena given model and data.
3. Criticize the model, revise and repeat.




SHOW ME
THE CODE!!!
memegenerator.net