Bayesian Neural Networks

ROY team: Ilya Zharikov,
Roman Isachenko,
Artem Bochkarev

Skolkovo Institute of Science and Technology
Bayesian Methods course

May 25, 2017

Project goal

Goal

Estimate posterior distributions of the model parameters from data

Problem

Monte Carlo sampling is very slow for high-dimensional data

Probabilistic Programming:

- Uncertainty in predictions;
- Uncertainty in representations;
- Regularizations with priors;
- Transfer learning;
- Hierarchical Neural Networks.

Related work

1. Salvatier J, Wiecki T. V., Fonnesbeck C. Probabilistic programming in Python using PyMC3. // *PeerJ Computer Science*. 2016.
2. Blundell C. et al. Weight Uncertainty in Neural Network // *Proceedings of The 32nd International Conference on Machine Learning*. 2015.
3. Kucukelbir A. et al. Automatic Differentiation Variational Inference // *arXiv preprint arXiv:1603.00788*. – 2017.

Problem Statement

Inference problem

Bayes' theorem states: $\mathbb{P}(\boldsymbol{\theta} \mid \mathbf{X}) = \dfrac{\mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathbf{X})}$

Maximum A Posteriori (MAP) estimation

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \left[\ln \mathbb{P}(\boldsymbol{\theta} \mid \mathbf{X})\right] = \arg\max_{\boldsymbol{\theta}} \left[\ln \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta}) + \ln \mathbb{P}(\boldsymbol{\theta})\right]$$

Monte Carlo approach:

- Metropolis-Hastings sampling;
- Gibbs sampling;
- No-U-Turn Sampling (NUTS).

Variational Inference

Goal

Approximate posterior distribution $p(X, \theta)$ by function $q(\theta)$ from parametric family.

$$\ln p(\mathbf{X}) = \mathrm{KL}(q||p) + \mathrm{ELBO}(q)$$

$$\Updownarrow \qquad\qquad \Updownarrow$$

$$\int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{X})} \mathrm{d}\boldsymbol{\theta} \qquad \int q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}$$

Minimization of $\mathbf{KL(q||p)}$ $\Leftrightarrow$ Maximization of $\mathbf{ELBO(q)}$

Automatic Differentiation Variational Inference (ADVI)

- Automatic transformation of constrained variables $\zeta = T(\boldsymbol{\theta})$;
  Example: $\theta \in \mathbb{R}_+ \Rightarrow \zeta = T(\theta) = \log \theta$, then $\zeta \in \mathbb{R}$.

- $q(\zeta) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is diagonal;

$$\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* = \underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}}{\arg\max}\, \mathrm{ELBO}(q)$$

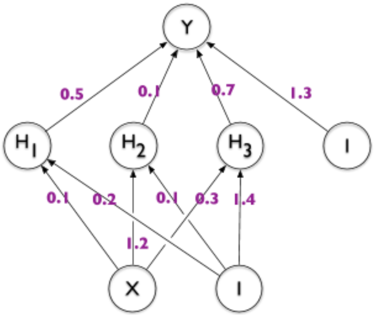
- Stochastic optimization;

- Reparametrization trick to apply automatic differentiation;

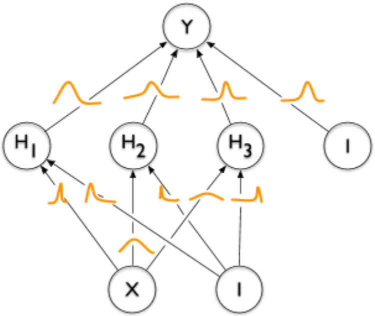
- Adaptive step-size.

Difference

Neural Networks
Predict values of parameters by fitting complex model on the huge dataset



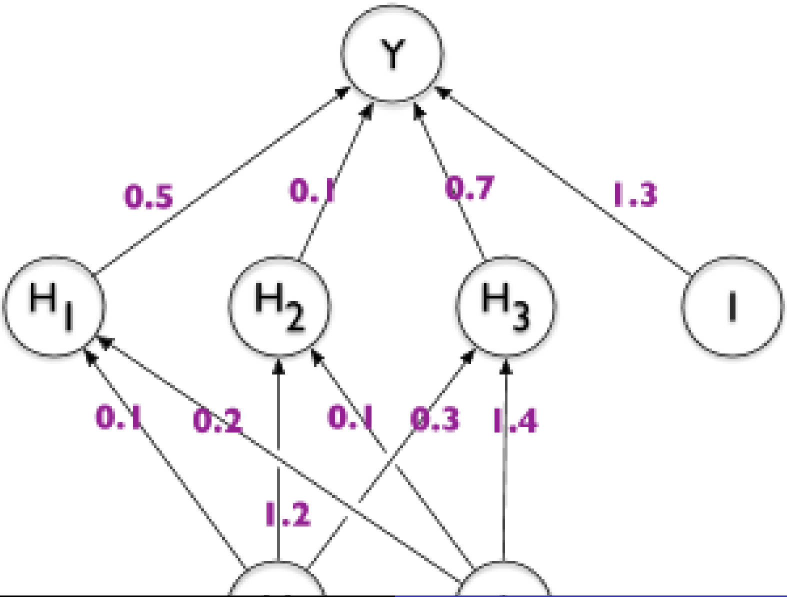
Bayesian Neural Networks
Predict the parameters of the weights distributions from the dataset



---

http://bit.ly/2rMQuDq

Deep Learning



Experiments

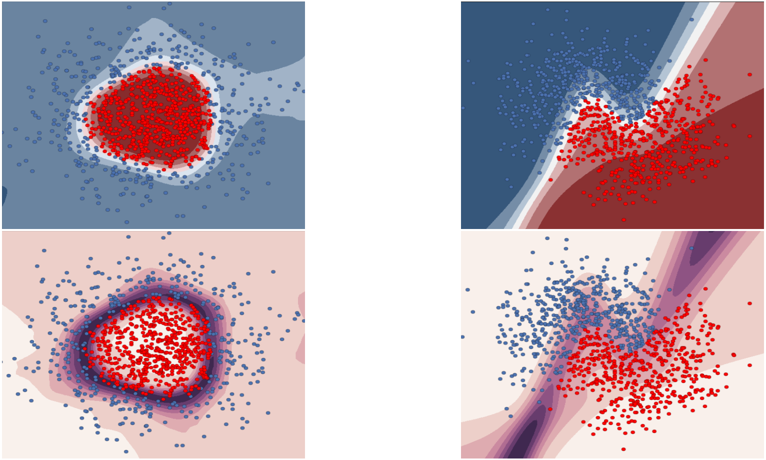
- Synthetic datasets: moons, circles
- Real data: MNIST

Goals:

- investigate influence of different priors on the predictions
- visualize uncertainties in predictions
- analyze the model behaviour

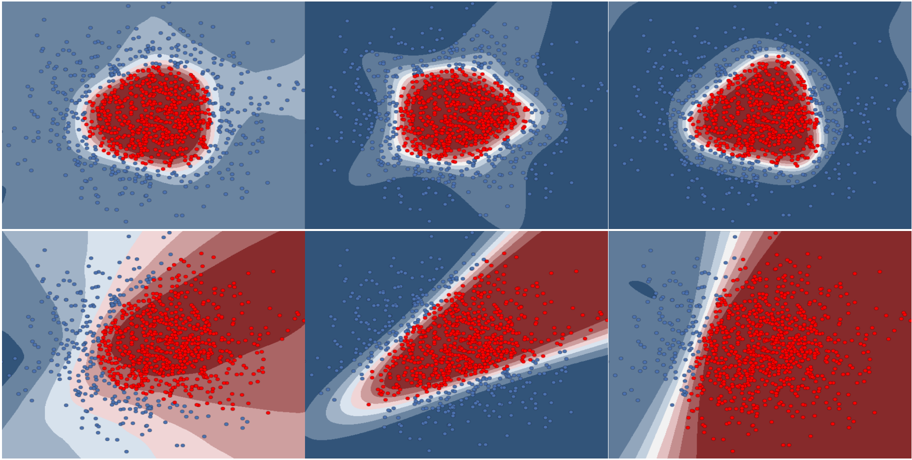
Experiments

The best priors for moons and circles are gauss and cauchy with hyperprior respectively.



Experiments

Including hyperpriors is almost always the best choice if we are unsure about prior parameters or data is noisy and hard.



Experiments

Laplace prior does indeed provide sparser solutions compared to cauchy or gauss.