Bayesian probabilistic propagation

Polina Barabanshchikova

MIPT

October 11, 2022

Motivation

1. Backpropagation (BP) based optimization requires tuning of hyperparameters.
2. In classic BP we can only obtain point estimates of the weights. As a result, the predictions do not account for uncertainty.
3. On the other hand, Bayesian learning suffers from the lack of scalability in both network architecture and data size.

Problem statement

Probabilistic model

Given data $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$, made up of D-dimensional feature vectors and corresponding scalar target variables, we assume that $y_n = f(x_n; W) + \varepsilon_n$, where $f(; W)$ is the output of a multi-layer neural network with weights given by W and $\varepsilon_n \sim \mathcal{N}(0, \gamma^{-1})$.
Prior distributions:

$$p(W|\lambda) = \prod_{w \in W} \mathcal{N}(w|0, \lambda^{-1}),$$

$$p(\lambda) = \Gamma(\lambda|\alpha_0^\lambda, \beta_0^\lambda),$$
$$p(\gamma) = \Gamma(\gamma|\alpha_0^\gamma, \beta_0^\gamma).$$

Theory

Likelihood for the weights W and the noise precision γ is

$$p(\mathbf{y}|W, \mathbf{X}, \gamma) = \prod_{n=1}^{N} \mathcal{N}(y_n | f(x_n; W), \gamma^{-1}).$$

The posterior distribution for W, γ, λ

$$p(W, \gamma, \lambda | \mathcal{D}) = \frac{p(\mathbf{y}|W, \mathbf{X}, \gamma) p(W|\lambda) p(\lambda) p(\gamma)}{p(\mathbf{y}|\mathbf{X})}.$$

Probabilistic backpropagation (PBP) approximates the exact posterior with a factored distribution given by

$$q(W, \gamma, \lambda) = \prod_{w \in W} \mathcal{N}(w | m_w, v_w) \times \Gamma(\gamma | \alpha^{\gamma}, \beta^{\gamma}) \Gamma(\lambda | \alpha^{\lambda}, \beta^{\lambda}).$$

Method description

Stages of PBP

1. In the first phase, the input data is propagated forward through the network. PBP sequentially approximates the marginal posterior distributions of each weight with a collection of one-dimensional Gaussians that match their marginal means and variances. At the end of this phase, PBP computes the logarithm of the marginal probability of the target variable.

Method description

Stages of PBP

1. In the first phase, the input data is propagated forward through the network. PBP sequentially approximates the marginal posterior distributions of each weight with a collection of one-dimensional Gaussians that match their marginal means and variances. At the end of this phase, PBP computes the logarithm of the marginal probability of the target variable.

2. In the second phase, the gradients of this quantity with respect to the means and variances of the approximate Gaussian posterior are propagated back. These derivatives are used to update the means and variances of the posterior approximation.

Method description

Update rule

Let $f(w)$ encode an arbitrary likelihood function for the single weight w and let our current beliefs regarding the scalar w be captured by a distribution $q(w)$. After seeing the data, our beliefs about w are updated according to Bayes' rule:

$$s(w) = Z^{-1} f(w) q(w),$$

where Z is the normalization constant.
We approximate this posterior with a distribution q^{new} that has the same form as q. The parameters of q^{new} are chosen to minimize the KL divergence between s and q^{new}.

Method description

Update rule (Example)

Assume that $q(w) = \mathcal{N}(w|m, v)$. In this case, the parameters of the new Gaussian beliefs $q^{new}(w) = \mathcal{N}(w|m^{new}, v^{new})$ that minimize the KL divergence between s and q^{new} can be obtained by

$$m^{new} = m + v\frac{\partial \log Z}{\partial m},$$

$$v^{new} = v - v^2\left[\left(\frac{\partial \log Z}{\partial m}\right)^2 - 2\frac{\partial \log Z}{\partial v}\right].$$

Method description

Update rule (Example)

Assume that $q(w) = \mathcal{N}(w|m, v)$. In this case, the parameters of the new Gaussian beliefs $q^{new}(w) = \mathcal{N}(w|m^{new}, v^{new})$ that minimize the KL divergence between s and q^{new} can be obtained by

$$m^{new} = m + v\frac{\partial \log Z}{\partial m},$$

$$v^{new} = v - v^2\left[\left(\frac{\partial \log Z}{\partial m}\right)^2 - 2\frac{\partial \log Z}{\partial v}\right].$$

Remark: Z is approximated during forward pass. Then its derivative is used to update the parameters of marginal distributions.

Experiments

Probabilistic Backpropagation

Dataset	N	d	Avg. Test RMSE and Std. Errors VI	BP	PBP	Avg. Test LL and Std. Errors VI	PBP	Avg. Running Time in Secs VI	BP	PBP
Boston Housing	506	13	4.320±0.2914	3.228±0.1951	**3.014±0.1800**	-2.903±0.071	**-2.574±0.089**	1035	677	**13**
Concrete Compression Strength	1030	8	7.128±0.1230	5.977±0.2207	**5.667±0.0933**	-3.391±0.017	**-3.161±0.019**	1085	758	**24**
Energy Efficiency	768	8	2.646±0.0813	**1.098±0.0738**	1.804±0.0481	-2.391±0.029	**-2.042±0.019**	2011	675	**19**
Kin8nm	8192	8	0.099±0.0009	**0.091±0.0015**	0.098±0.0007	**0.897±0.010**	0.896±0.006	5604	2001	**156**
Naval Propulsion	11,934	16	0.005±0.0005	**0.001±0.0001**	0.006±0.0000	**3.734±0.116**	3.731±0.006	8373	2351	**220**
Combined Cycle Power Plant	9568	4	4.327±0.0352	4.182±0.0402	**4.124±0.0345**	-2.890±0.010	**-2.837±0.009**	2955	2114	**178**
Protein Structure	45,730	9	4.842±0.0305	**4.539±0.0288**	4.732±0.0130	-2.992±0.006	**-2.973±0.003**	7691	4831	**485**
Wine Quality Red	1599	11	0.646±0.0081	0.645±0.0098	**0.635±0.0079**	-0.980±0.013	**-0.968±0.014**	1195	917	**50**
Yacht Hydrodynamics	308	6	6.887±0.6749	1.182±0.1645	**1.015±0.0542**	-3.439±0.163	**-1.634±0.016**	954	626	**12**
Year Prediction MSD	515,345	90	9.034±NA	8.932±NA	**8.879± NA**	-3.622±NA	**-3.603± NA**	142,077	65,131	**6119**

Table 1. Characteristics of the analyzed data sets, average test performance in RMSE and log likelihood, and average running time.

Literature

[1] José Miguel Hernández-Lobato and Ryan P. Adams. *Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks*. 2015. DOI: 10.48550/ARXIV.1502.05336. URL: https://arxiv.org/abs/1502.05336.

Questions

1. Assume that the current posterior distribution for γ is $q(\gamma) = \Gamma(\alpha^\gamma, \beta^\gamma)$. After seeing new data, we update the posterior and approximate it by $q^{new}(\gamma)$. To which family of distributions does q^{new} belong?

a) Gaussian
b) Gamma
c) Uniform
d) Depends on new data

2. What is computed at the end of the PBP forward pass instead of the prediction error?