STCS 6701: Foundations of Graphical Models: Reading 9

Maxime TCHIBOZO (MT3390)

November 2020

1 Probabilistic Deep Learning (2020) – Durr & Sick

Part 3 – Bayesian Approaches For Probabilistic DL Models (ch.7 – Bayesian Learning; ch.8 – Bayesian Neural Networks)

Traditional Deep Learning methods are reliable when the testing data is similar to the training data. However, these Deep Learning methods are overconfident when making predictions on new data and are unable to extrapolate from the narrow context from which they were trained. They do not provide accurate estimates of uncertainty.

Probabilistic Machine Learning models account for uncertainty inherent to the data (e.g. via modeling of noise) to alleviate part of this issue. Nevertheless, an additional source of uncertainty should also be taken into account: "epistemic uncertainty", which measures the inability to perfectly estimate parameters as we cannot infinitely sample training data. The data at hand is usually drawn from a specific distribution, and estimated parameters from this data often no longer hold when new test data is not taken from the exact same distribution.

Bayesian models in Deep Learning make better predictions with little training data and in novel situations as they account for low density regions (i.e. with few training examples). They also provide confidence intervals to measure how much a set of estimated (latent) parameters can be trusted when making new predictions. Frequentist Maximum Likelihood estimators also provide confidence intervals, but these are often absurd when little data has been observed.

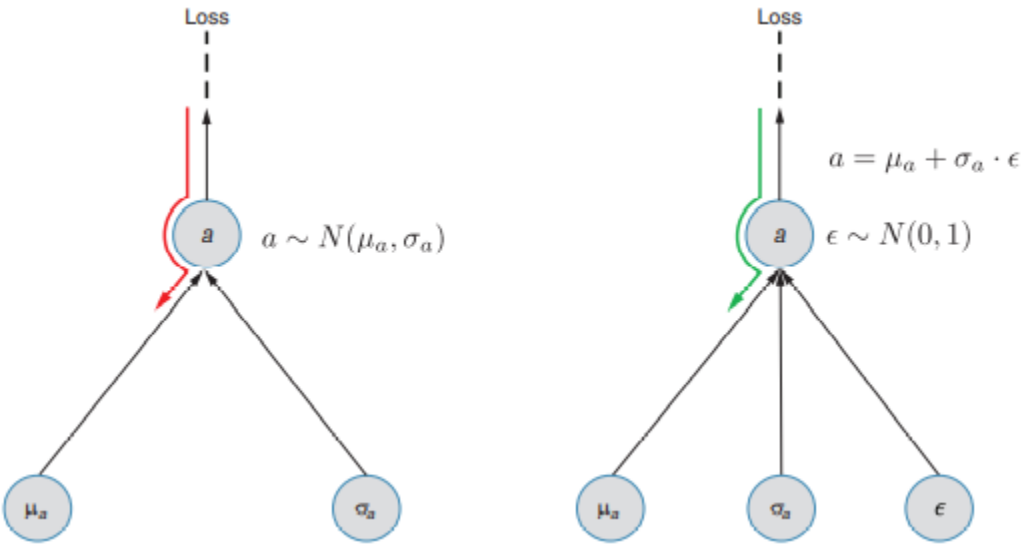
Bayesian methods are often intractable and slow; we use efficient approximation methods for Bayesian Deep Learning parameter estimation: Variational Inference (VI) and Monte Carlo dropout (MC dropout). TensorFlow Probability supports VI to fit Bayesian neural networks, and MC dropout can be done easily in Keras.

In Bayesian neural networks, each weight is replaced by a distribution which is dependent upon the surrounding nodes. The VI approximation consists in learning a parametric similar distribution (q, the variational distribution) for which conditional independence between weights is assumed. The variational parameters are then tuned to allow for q to be as close as possible to the true posterior. The prior can be chosen around small values, which would make it analogous to a regularizer and will push approximated posterior distributions to favor small parameters. To implement this in TensorFlow Probability, one would define the variables according to their assumed data generating process:

y_prob = tfd.Normal(loc = x · a + b, scale = sigma)

loss_neg_log_likelihood = -tf.reduce_sum(y_prob.log_prob(ytensor))

Derivatives are computed through sampled variables using the reparametrization trick (the noise parameter is not updated).



In the following code, we could use the following reparametrization:

a = tfd.Normal(loc=mu_a,scale=sigma_a)

equivalent to

e = tfd.Normal(loc=0.,scale=1.)

a = mu_a + sig_a*e.sample()

We can then use regular gradient descent to optimize the variational parameters. This can also be done in Keras:

model = tf.keras.Sequential([

tfp.layers.DenseReparameterization(1, input_shape=(None,1)),

tfp.layers.DenseReparameterization(2),

tfp.layers.DenseReparameterization(3)])

The distribution for each weight is usually modelled with two parameters(mean, sd). This unfortunately implies that going Bayesian with VI requires roughly twice as many parameters as for standard neural networks.

MC dropout is a smart workaround this problem.

Whereas ordinary neural networks with dropout randomly drop parameters during training and use all parameters for testing, MC dropout keeps the dropout rate during testing. The model prediction is then taken to be an average over several repeated predictions on the test data with dropout. This can also give frequentist distributions of the predicted output, thus providing confidence bounds.

This book was valuable for its examples of code with TensorFlow Probability. However, having a 300 page book to detail this seems excessive; most of this information was already contained in papers from past readings.