Paper Assignment
Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles

Markovich Alexander

March 24, 2019

Point A

How do authors change the NN to make it capable to estimate uncertainty for regression tasks?

Authors use a neural network(NN) which outputs two values (instead of one value) in the final layer, corresponding to the predicted mean $\mu(x)$ and variance $\sigma^2(x) > 0$. Such modification allows treating predictions as random variables with the predicted mean and variance. As a result, we can minimize the negative log-likelihood criterion.

What is the distribution on the outputs, as defined by the NN architecture and loss?

In the case of a regression problem, we model Gaussian distribution. NN predicts the mean and variance of the distribution. The loss is a proper scoring rule – log-likelihood criterion.
In the case of multiclass classification, we model multinomial distribution and use softmax cross entropy loss, which is equivalent to the log-likelihood criterion.

What distribution on the outputs would be induced by an ensemble of such NNs?

For classification, distribution corresponds to averaging the predicted probabilities. For regression, the prediction is a mixture of Gaussian distributions.

Point B

What are adversarial examples?

Adversarial examples are perturbed inputs designed to fool machine learning models. For example, let we have an image of a panda. We can recognize panda on an image with high confidence, but NN, for example, says that is gibbon.

What is the purpose of using them to train the ensemble?

For each NN in an ensemble, using of adversarial examples improve the robustness of model to misspecification and out-of-distribution examples. In additional, adversarial examples smooth the predictive distribution. More formal, we increase the likelihood of the target around a ϵ-neighbourhood of the observed training examples.

Can an object with an unchanged prediction be an adversarial example?

Yes, because such an object will smooth prediction around of object and our model will not be overconfident.

Point C

Let's imagine that somebody collected a dataset with many out-of-domain images or images with wrong labels. How can the proposed uncertainty estimation method be applied to clean the dataset from such objects?

As shown in the article, the proposed method is applicable to out-of-domain objects. For such objects, the value of entropy is far from zero. Thus, we can filter predictions (and therefore objects) with entropy exceeding a certain threshold.