Notes on "Dropout as a Bayesian Approximation."

Xavier Garcia

March 1, 2019

The purpose of these notes is to explain a few of the confusing parts of the paper "Dropout as a Bayesian Approximation". We begin by an exposition into the Bayesian framework.

1 The Bayesian Framework

In this section, we first discuss what the Bayesian framework is, and how it relates to traditional supervised learning. From here, we will then discuss why there's a need to consider uncertainty in estimates and what this means from a Bayesian point of view.

1.1 Supervised learning in the Bayesian framework.

In the traditional supervised setting, we have data $(x_1, y_1), ..., (x_N, y_N)$, where the $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ which we assume to be sampled from some distribution (X, Y). In traditional machine learning, we assume we have some loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and the goal becomess to find the function $f : \mathbb{R}^n \to \mathbb{R}$ which minimizes the loss given this loss function i.e.

$$f = \operatorname*{argmin}_f \mathbb{E}\, L(f(X), Y)$$

Normally, we choose the squared error loss

$$L(f(x), y) = (y - f(x))^2$$

for the regression task or cross entropy

$$L(f(x), y) = y \log(\sigma(f(x)) + (1 - y) \log(1 - \sigma(f(x)))$$

in the case the classification case where $y \in \{0, 1\}$, σ is the sigmoid function and $\sigma(f(x)) = \mathbb{P}(y = 1|X = x)$. From a frequentist point of view, these are natural loss functions to consider. However, when we take a step back, they can seem quite arbitrary. For example, why squared error? Why not quartic error?

Or even absolute error? The Bayesian framework addresses this question by establishing a unifying principle in tackling problems. Instead of assuming we have some loss function L, we are going to assume that we know the conditional distribution $\mathbb{P}(Y|X)$ completely depends on some unknown function f and hence so does its density

$$p_{Y|X}(y|x) = p_{Y|X}(y|x, f).$$

With this assumption, we can choose the f which maximizes likelihood i.e.

$$f = \operatorname*{argmax}_f \, \mathbb{E}\, p_{Y|X}(Y|X, f) = \operatorname*{argmax}_f \mathbb{E}\, \log p_{Y|X}(Y|X, f) \tag{1}$$

To see the connection between this viewpoint and the traditional viewpoint (i.e. the Frequentist approach), consider the following example:

Example 1. *Suppose that we assume $Y|X \sim N(f(x), 1)$. In this case, we have that*

$$\log p_{Y|X}(y|x, f) = -\frac{(y - f(x))^2}{2} - \frac{1}{2}\log(2\pi)$$

and hence

$$\begin{aligned}\operatorname*{argmax}_f \mathbb{E}\, p_{Y|X}(Y|X, f) &= \operatorname*{argmax}_f -\mathbb{E}\,\frac{1}{2}(Y - f(X))^2 - \log 2\pi \\ &= \operatorname*{argmin}_f \mathbb{E}[(Y - f(X))^2]\end{aligned}$$

In the classification case, we leave it to the reader to verify that if we assume $Y|X \sim \text{Bernouilli}(f(X))$, then maximizing likelihood is equivalent to minimizing the cross entropy. More generally, given a loss function L, we can define a distribution by

$$\log p_{Y|X}(y|x, f) = -L(f(x), y.)$$

The key point here is that loss functions are no longer arbitrary. Instead, they manifest as assumptions we've made on the distribution $Y|X$.

1.2 The need for uncertainty

In real life, we never have access to the joint distribution of (X, Y) and so we are unable to compute the expectations in the previous sections. In particular, we are forced to approximate them in the typical Monte Carlo fashion i.e.

$$\mathbb{E}L(f(X), Y) \approx \frac{1}{N}\sum_{i=1}^{N} L(f(x_i), y_i)$$

This approximation can be quite dangerous since there are multiple distributions which could have yielded this particular sample. In particular, if we define the random variables $(\tilde{X}, \tilde{Y})$ by

$$\mathbb{P}\left(\tilde{X} = x_i, \tilde{Y} = y_i\right) = \frac{1}{N}\,\forall i$$

then we have that

$$\frac{1}{N}\sum_{i=1}^{N} L(f(x_i), y_i) = \mathbb{E}[L(f(\tilde{X}), \tilde{Y})]. \tag{2}$$

Notice that the distribution of $(\tilde{X}, \tilde{Y})$ is close to that of (X, Y) and converges to that of (X, Y) by the law of large numbers. Nevertheless, because N is finite, they are distinct. We say we are overfitting if we end up approximating $(\tilde{X}, \tilde{Y})$ more so than (X, Y). This phenomenon can manifest in the form of the model possibly predicting wildly incorrect guesses for data which is vastly different from the sample.

In particular, if we get a new instance x^* that is far away from the original data $x_1, ..., x_n$, we have no guarantees that our model's prediction $f(x^*)$ is accurate. Moreover, we should express distrust over such predictions. Interestingly, this problem persists even if had the full joint distribution (X, Y). For example, if the problem is to predict where x is a picture of a dog or a cat and we present it with a picture of a tree, it's clear that whatever prediction the model presents should be viewed with concern.

These concerns about data being different from training raises many concerns: How much is "far away"? What does "far away" even mean? Can we quantify the certainty of our model's predictions for a given x? At a glance, these concerns seem only valid for the regression case. In the classification case i.e. where $y \in \{0, ..., C-1\}$ for some positive integer C, we viewed the result of f as the probabilities $\mathbb{P}(Y = i)$ for $i = 0, ..., C-1$. These probabilities should serve as a statement of our model's confidence but it's easy to imagine that that the computations of these probabilities themselves may be uncertain. Indeed, it is possible for our model to give a particular class a large probability with large uncertainty. Going back to the example where we pass a tree picture into a dog-vs-cat classifier, it's quite possible that our model confidently assigns a high probability that the tree is a cat, as opposed to giving the more reasonable answer of 50-50. Since the problem reduces to the regression case, we will strictly restrict ourselves to the continuous case.

1.3 Quantifying uncertainty

We now seek to find a rigorous definition of uncertainty. Unfortunately, no one seems to really want to pin this down rigorously, so we attempt to provide a few metrics of certainty, all of which fall within the Bayesian framework.

The reason why there is uncertainty is that given data

$$\mathcal{D} := \{(x_1, y_1), ..., (x_N, y_N)\},$$

there exists a variety of functions f which could satisfy our loss function. From the Bayesian perspective, uncertainity is another word for probability. Therefore, we could start with a prior distribution for our f, i.e. we have a way of assigning meaning to the quantity $\mathbb{P}(f)$. Given this information, the Bayesian approach is to condition on the data i.e. look at the posterior distribution

$$\mathbb{P}(f|\mathcal{D}).$$

The idea is that we use the data to update our beliefs of what the function f should be. We can then proceed as usual to recover f, namely by maximizing the probability of f given the data i.e.

$$f = \operatorname*{argmax}_f \mathbb{P}(f|\mathcal{D}).$$

We can connect this definition to the frequentist approach by using Bayes rule:

$$\begin{aligned}\mathbb{P}(f|\mathcal{D}) &= \frac{\mathbb{P}(\mathcal{D}|f)\cdot\mathbb{P}(f)}{\mathbb{P}(\mathcal{D})} \\ &\propto \mathbb{P}(\mathcal{D}|f)\cdot\mathbb{P}(f)\end{aligned}$$

This implies the following equality:

$$\operatorname*{argmax}_f \log\mathbb{P}(f|\mathcal{D}) = \operatorname*{argmax}_f \log\mathbb{P}(D|f) + \log\mathbb{P}(f). \tag{3}$$

We first remark that $\mathbb{P}(\mathcal{D}|f)$ can be written as

$$\mathbb{P}(\mathcal{D}|f) = \frac{1}{N}\sum_{i=1}^{N} p_{Y|X}(y_i|x_i, f).$$

Next, notice that if there was a notion of "uniform distribution of functions" i.e. $\mathbb{P}(f) \propto 1$, then choosing that distribution as a prior distribution on f would make the right hand side of (3) equal to objective (2). Even more interesting, however, is that we can look at the whole distribution $\mathbb{P}(f|\mathcal{D})$, rather than just looking at its mode $\operatorname{argmax}_f \mathbb{P}(f|\mathcal{D})$. For example, if we could make sense of the quantity $\operatorname{var} f|\mathcal{D}$, then we could use this as a notion of uncertainty of our model's choice.

Definition 1. *Given data $\mathcal{D}$ and a prior $\mathbb{P}(f)$ on functions as well as a new input x^*, we define the uncertainty of the prediction y^* by its distribution i.e.*

$$\mathbb{P}(y^*|x^*, \mathcal{D}) = \int p(y^*|x^*, f)\, d\mathbb{P}(f|\mathcal{D}) \tag{4}$$

This definition of uncertainty is a qualitative one, not a quantitative one. As a numerical proxy, we can compute the variance $\operatorname{var}(y^*|\mathcal{D})$ as a measurement of the uncertainity. Notice that in this definition, the randomness comes from our uncertainity in f after the conditioning.

2 Defining distributions on functions

In the previous section, we defined the Bayesian framework and realized the necessity of uncertainity i.e. the necessity of having a distribution on functions. In this section, we outline a simple way of generating distributions, namely Gaussian fields and their extension, deep Gaussian fields. Next, we'll briefly introduce variational inference.

2.1 Gaussian fields

In order to make equation (4) a little more digestible, we will further make the assumption that $y = f(x) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma)$ for some $\sigma > 0$ and ϵ independent of f. As we saw from the first section, this implies that $y|x, f \sim \mathcal{N}(f(x), \sigma)$ and in particular, the dependence of y on f only happens through $f(x)$, which is a real-valued random variable.

Before we even begin to compute any part of equation (4), we must first solve the problem of putting a distribution on f. This is a very subtle problem, and slightly too technical to deal with utmost rigor. Instead, we shall make some simplifications. Namely, we'll say we know the distribution of f if for any integer N and any $x_1, ..., x_N$, we know the joint distribution of $(f(x_1), ..., f(x_N))$. By the previous paragraph's assumption, this is a reasonable definition of a distribution of f for our purposes. With this definition, we can define a Gaussian field as follows:

Definition 2. *For a function $\mu : \mathbb{R}^n \to \mathbb{R}$ and a postive-definite function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, we define the Gaussian field as the distribution on f such that for any x, x' we have that $f(x) \sim \mathcal{N}(\mu(x), K(x, x))$ and*

$$\mathrm{cov}(f(x), f(x')) = K(x, x').$$

For the purposes of these notes, we can give this family of models as a simple interpretation: $\mu(x)$ represents our prediction for the input x, and the covariance kernel $K(x, x)$ gives us a measure of how confident the model is. The Gaussian field is a natural family of distributions, since it is completely determined by the functions μ and K, akin to how Gaussian distributions are determined by their means and variance. For clarity and for future convenience, consider the following example:

Example 2. *(Shallow Gaussian Fields)*

Suppose $c \sim \mathcal{N}(0, I_K)$ is a multivariate normal with mean 0 and covariance matrix equal to the $K \times K$ identity I_K. Furthermore, suppose we have some collection of functions $\Phi = (\Phi_1, ..., \Phi_K)$. Then, the function

$$f(x) = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} c_i \Phi_i(x)$$

is a Gaussian field with mean 0 and covariance kernel

$$\tilde{K}(x,x') = \frac{1}{K}\sum_{i=1}^{K}\Phi_i(x)\Phi_i(x') = \frac{1}{K}\Phi(x)\Phi^T(x').$$

Proof. Exercise. □

Moreover, it turns out that we can actually compute the posteror $f|\mathcal{D}$, where $\mathcal{D} := \{(x_1,y_1),...,(x_N,y_N)\}$ as in the previous sections.

Theorem 2.1. *Suppose $f \sim \mathcal{N}(\mu(\cdot),K(\cdot,\cdot))$ is a Gaussian field with mean function μ and covariance kernel K, and we have some data $\mathcal{D} = ((x_1,y_1),...,(x_N,y_N))$. Then, there exists functions $\mu_{\mathcal{D}}$ and $K_{\mathcal{D}}$ such that the distribution of $f|\mathcal{D}$ is a Gaussian field with mean $\mu_{\mathcal{D}}$ and covariance kernel $K_{\mathcal{D}}(x,x')$. Moreover, $K_{\mathcal{D}}(x,x') = 0$ for any x,x' in $\mathcal{D}$.*

We won't delve into the details of this proof, nor the exact formulas $\mu_{\mathcal{D}}$ and $K_{\mathcal{D}}$. Instead, we point out that equation (4) is now tractible with these assumptions. We outline the steps:

1. Start with a Gaussian field prior on f, say $f \sim \mathcal{N}(0,K)$ for your favorite kernel K.
2. Use your data and compute $f|\mathcal{D} \sim \mathcal{N}(\mu_{\mathcal{D}},K_{\mathcal{D}})$.
3. For a new datapoint x^*, $f(x^*) \sim \mathcal{N}(\mu_{\mathcal{D}}(x^*),K_{\mathcal{D}}(x^*,x^*))$.
4. Under the assumption that $y = f(x)+\epsilon, \epsilon \sim \mathcal{N}(0,\sigma^2)$, we have
$$y^* \sim \mathcal{N}(\mu_D(x^*),K_{\mathcal{D}}(x^*,x^*)+\sigma^2)$$
5. (Exercise) If you unravel the integral in equation (4), you will also recover this formula.

As a funny sidenote, point 5 illustrates why we prefer to use expectations as opposed to integrals in probability.

Next, notice that there's also nothing stopping us from making the Gaussian field a vector-valued random variable rather than real-valued one. The only thing that changes is that μ becomes vector-valued, and $K(x,x')$ becomes a covariance matrix.

Example 3. *Suppose $W \in \mathbb{R}^{n\times m}$ is a matrix consisting of Gaussian random variables which are jointly Gaussians, and $b \in \mathbb{R}^m$ is a multivariate Gaussian. Then, the function $f_{W,b}(x) = Wx+b$ is a vector-valued Gaussian field.*

This observation allows us to define deep Gaussian fields:

Definition 3. *Suppose we have L Gaussian fields $f_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ for $i = 1...,L$, with $f_i \sim \mathcal{N}(\mu_i(\cdot),K_i(\cdot,\cdot))$. Let $F_j = f_j \circ ... \circ f_1$. Then, F_L is called an L layer deep Gaussian field.*

In particular, $F_j|F_{j-1}$ is a Gaussian field for all j.

We point out that deep Gaussian fields are no longer Gaussian field. In particular, it is *not* determined by some global covariance kernel K and mean μ, and there is no analytically tractible expression for $\mathbb{P}(f|\mathcal{D})$. To mitigate this, we present an example of a tractible deep Gaussian field.

Example 4. *Suppose you have matrices $W_i \in \mathbb{R}^{n_i \times n_{i+1}}$, which consists of Gaussian random variables, and vectors $b_i \in R^{n_{i+1}}$ which also consists of Gaussian random variables. Define $f_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ by*

$$f_i = W_i \sigma(x) + b_i$$

for $i > 1$ and $f_1(x) = W_1 x + b_1$. Then

$$F_N : \mathbb{R}^{n_1} \to \mathbb{R}^{n_{L+1}}, F_N = f_N \circ \ldots \circ f_1$$

is a deep Gaussian field, completely determined by the distribution of

$$\omega := (W, b).$$

For the purposes of these notes, we shall restrict our usage of the term ”deep Gaussian field” to the example above. Notice that with this restriction, we've reduced the problem of studying distributions on functions to distributions on real-valued vectors ω. We use this fact and abuse notation to write the following:

$$\mathbb{P}(f) = \mathbb{P}(\omega) \text{ and } \mathbb{P}(f|\mathcal{D}) = \mathbb{P}(w|\mathcal{D})$$

With this notation, we can write objective (4) in a simpler fashinon:

$$\begin{aligned} p(y^*|x^*, \mathcal{D}) &= \int p(y^*|x^*, f) d\mathbb{P}(f|\mathcal{D}) \\ &= \int p(y^*|x^*, \omega) p(\omega|\mathcal{D}) d\omega \\ &= \mathbb{E}_{\omega \sim p(\omega|\mathcal{D})}[p(y^*|x^*, \omega)] \end{aligned}$$

While this formulation reduces the complexity to studying traditional random variables, it does not make computing the posterior tractable.

2.2 Variational Inference

Intractability of posteriors is a common problem in Bayesian statistics. A frequentist's solution to this problem is to find a distribution $q(w)$ which, when replacing $p(\omega|\mathcal{D})$ with $q(\omega)$, yields a tractable posterior. Traditionally, we assume q are normally-distributed, but this is not required. Let's now investigate the cost in using q versus $p(\omega|\mathcal{D})$:

$$
\begin{aligned}
\log p(y^*|x^*, \mathcal{D}) &= \log \int p(y^*|x^*, \omega) p(\omega, \mathcal{D}) d\omega \\
&= \log \int p(y^*|x^*, \omega) \frac{p(\omega|\mathcal{D})}{q(\omega)} q(\omega) d\omega \\
&= \log \mathbb{E}_{\omega \sim q(\omega)} \left[p(y^*|x^*, \omega) \frac{p(\omega|\mathcal{D})}{q(\omega)} \right] \\
&\geq \mathbb{E}_{\omega \sim q(\omega)} \left[\log \left(p(y^*|x^*, \omega) \frac{p(\omega|\mathcal{D})}{q(\omega)} \right) \right] \\
&= \mathbb{E}_{\omega \sim q(\omega)} [\log p(y^*|x^*, \omega)] - \mathbb{E}_{\omega \sim q(\omega)} \left[\log \frac{q(\omega)}{p(\omega|\mathcal{D})} \right] \\
&= \mathbb{E}_{\omega \sim q(\omega)} [\log p(y^*|x^*, \omega)] - D_{KL} \left(q || \, p(\,\cdot\,|\mathcal{D}) \right)
\end{aligned}
$$

where the inequality follows from Jensen's inequality and D_{KL} is the Kullback-Liebler divergence. This final quantity is called the Effective Lower BOund (ELBO) and so maximizing this quantity makes it closer to the real distribution. The important contribution of these manipulations can be summarized as follows: We can replace $p(\omega|\mathcal{D})$ with $q(\omega)$, but we incur a cost of $D_{KL}(q \,||\, p(\,\cdot\,|\mathcal{D}))$. The actual matter is a lot more complicated and subtle but it'll do for the purpose of these notes.