Redes Neurais com Dropout sequer são Bayesianas?

Zuilho Rodrigues Castro Segundo
Escola de Matemática Aplicada
Fundação Getulio Vargas
Rio de Janeiro, RJ
segundozuilho@gmail.com

George Dutra
Escola de Matemática Aplicada
Fundação Getulio Vargas
Rio de Janeiro, RJ
georgedutra661@gmail.com

Abstract

Neste estudo, analisamos *papers* que abordam visões conflitantes acerca da aplicação de *Dropout* em Redes Neurais para simular um modelo Bayesiano com o objetivo de tornar possível expressar a incerteza do modelo de *Deep Learning*. Para isso, reimplementamos as aplicações dos *papers* em código *python*, buscando comparar os resultados de cada modelo e validar as conclusões dos autores.

1 Introdução

Os avanços em Deep Learning têm proporcionado uma revolução significativa em várias áreas da inteligência artificial, com aplicações crescentes em visão computacional, processamento de linguagem natural e tomada de decisões complexas. No entanto, um desafio persistente é a capacidade limitada desses modelos de expressar incerteza em suas previsões, uma métrica crucial para avaliar a confiabilidade dos resultados, especialmente sob a perspectiva Bayesiana.

O *paper* de Yarin Gal e Zoubin Ghahramani [2], "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning", propõe que o Dropout, uma técnica originalmente desenvolvida para regularização de redes neurais, pode ser reinterpretado como uma aproximação bayesiana. Eles argumentam que o uso de *Dropout* durante o treinamento permite capturar a incerteza do modelo através de múltiplas passagens (*forward pass*) com pesos aleatoriamente omitidos.

Em contrapartida, trabalhos como "Is MC Dropout Bayesian?" questionam essa interpretação. Os autores demonstram que o Dropout, na verdade, não modela adequadamente a distribuição a posteriori dos parâmetros do modelo, apresentando limitações significativas em relação a outras técnicas de inferência variacional.

Neste estudo, replicamos os experimentos desses trabalhos para analisar empiricamente como o *Dropout* se comporta em diferentes cenários, como classificação de imagens e regressão. Além disso, discutimos as implicações teóricas e práticas dessas técnicas, buscando compreender se o *Dropout* pode ser considerado uma ferramenta válida para estimar a incerteza em modelos de *Deep Learning*.

Nos próximos tópicos, abordaremos detalhadamente cada um desses trabalhos, apresentando suas metodologias, resultados e conclusões, além de discutir possíveis limitações e direções futuras de pesquisa.

2 *Dropout as Bayesian*

Como já comentado, buscando resolver os problemas inerentes aos modelos de *Deep Learning* de incapacidade de capturar a incerteza de predições, o paper de Yarin Gal e Zoubin Ghahramani [2], [3] propõe a utilização da técnica de *dropout* como uma forma de aproximação Bayesiana para modelar a incerteza. De maneira sucinta, o *dropout* pode se interpretada como uma forma de amostragem de Monte Carlo para inferência Bayesiana em redes neurais profundas.

Projeto Final - ML (2024).

Sendo utilizado como uma técnica de inferência variacional, o que estamos buscando é uma distribuição $q(\theta)$ que aproxima a distribuição a posteriori $p(\theta|\mathcal{D})$. Diante dessa visão, podemos entender que utilizar o *dropout* é equivalente a ter uma distribuição Bernoulli para os pesos da rede, onde os pesos são multiplicados por variáveis r_{ij} amostradas da distribuição.

Sendo assim, dado um conjunto de dados $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$:

$$\log p(y|x, \mathcal{D}) \approx \int q(\theta) \log p(y|x, \theta) d\theta - KL(q(\theta)||p(\theta),$$

onde KL é a divergência de Kullback-Leibler.

Agora, em posse disso, após o treinamento da rede, realizamos seguidos *foward pass* com *dropout*, e para cada um deles $\hat{y}^{(t)} = f(x, \tilde{W}^{(t)})$, onde $\tilde{W}^{(t)}$ são os pesos da rede após a aplicação do *dropout* utilizando a distribuição $q(\theta)$. Com isso, podemos aproaximar a predição final para:

$$\overline{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)},$$

e a variância pode ser calculada como:

$$Var(\overline{y}) = \frac{1}{T} \sum_{t=1}^{T} (\hat{y}^{(t)} - \overline{y})^2.$$

Como principais conclusões o *paper* mostra que conseguiu apresentar uma interpretação probabilística para a utilização do *dropout*, apresentando uma forma de baixo-custo de computar a incerteza dentro de um rede e demonstrando alguns possíveis usos da técnica.

3 *Dropout* ~~not~~ *as Baeysian*

No *paper* "Is MC Dropout Bayesian?" [7], surge uma análise crítica à ideia de que o *MC Dropout* funciona como uma aproximação para técnica Bayesianas pois ela, comparada a outras é muito pior em quesitos de estimação da posteriori. Apesar da popularidade do método citado, que possui mais de 7000 citações, os autores desse artigo apontam que acontece uma modificação do modelo Bayesiano, levando a uma aproximação incorreta.

No artigo, quanto falado de Inferência Variacional, é apontado que o principal objetivo é aproximar a real posterior $p(\theta|X, T)$, e isso é feito através da maximização da Evidence Lower Bound (ELBO):

$$ELBO(q) = \log p(X, T) - KL[q(\theta)||p(\theta||X, T)]$$

Como já citado, a predição para o *dropout* é realizada através da passagem por dversas vezes da mesma entrada pela rede. Aqui temos uma fórmula matemática para representar a predição do *MC Dropout*, sendo ela:

$$p_{\text{MC-dropout}}(y^* \mid x^*, X, T) = \sum_{z \in \{0,1\}^P} p(y^* \mid x^*, \hat{\theta} \odot z) q(z),$$

onde $\odot$ significa a operação em cada elemento.

O principal argumento apresentado no artigo é que, diferente dos demais modelos de aproximação de incertezas Bayesianos, o *MC Dropout* atribui probabilidade não zero para um número finito de valores, o que leva a uma posterior multimodal que não se assemelha a posterior real, sendo essa uma propriedade do método, e não do modelo Bayesiano. No *paper*, isso é demonstrado através de um gráfico onde a posterior do *MC Dropout* é incapaz de capturar a característica multimodal da distribuição, enquanto em contrapartida, modelos como *mean-field VI* e *structures normal VI* realizam uma melhor aproximação da posterior. Uma boa forma de interpretar a posterior do *MC*

Dropout é ver que ela é controlada pela probabilidade p de ativação do neurônio, e quando esse é igual a 1, ele coincide com o MAP e quando é 0, a massa é toda deslocada pra origem. Dessa forma, o que acontece é que estamos modificando o modelo Bayesiano com uma prior que induz esparcidade onde cada parâmetro θ_i é associaado com com uma variável z_i, levando a uma interpretação heurística do que é proposto.

O artigo conclui apresentando a ferramenta de Inferência Variacional que serve como alternativa e mostrando que apesar de muito popular o *MCDropout* falha ao demonstrar que é uma aproximação Bayesiana confiável, ainda mais no que se trata de capturar a multimodalidade da posterior.

4 Replicando os Experimentos

Nosso código pode ser encontrado em `https://github.com/ZuilhoSe/dropout-bayesian-approximation`.

4.1 *Dropout as Bayesian*

Nesse estágio, tentamos replicar os experimentos realizados no *paper* original, e escolhemos o MNIST para a tarefa de classificação e o *dataset California Housing*. A estrutura das nossas redes podem ser encontradas no repositório já citado.

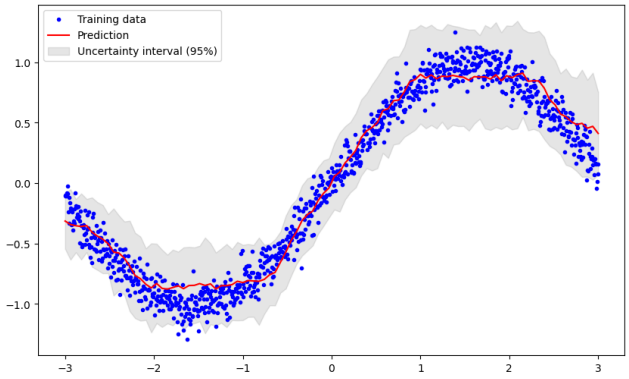

Figure 1: Dados Sintético

Para começar nossos testes, criamos dados sintéticos com ruído e *fitamos* uma rede utilizando o *MC Dropout*. Os resultados podem ser encontrados em "Fig. 1", e consegue, assim como o *paper* sugere, demonstrar uma incerteza, sendo maior em alguns pontos, devido a concentração de dados.

Em seguida, aplicando uma rede para o MNIST, obtivemos resultados que parecem fazer sentido. Como podemos ver em "Fig. 2", a rede parece ter bastante certeza sobre os valores que não podem ser assumidos pelo dados, ficando em dúvida em casos como o do 4, onde tem incerteza relativamente alta entre os valores de 4 e 9.

Prosseguindo para o *dataset* para regressão, para a rede sem a utilização do *MC Dropout* ("Fig. 3"), podemos perceber que não é possível inferir nada sobre a incerteza da rede, sabendo apenas o valor médio encontrado. No entanto, ao aplicarmos o método, encontramos ("Fig. 4") algo como duas aproximações para os dados, uma seguindo algo parecido com o que já havíamos encontrado, mas algo também mais linear. A ideia aqui é que tirando a média desses pontos vermelhos, teríamos algo mais acurado, em média para as predições. Além disso, como podemos perceber, quando mais afastado do centro, maior a variância, indicando algo esperado para a incerteza.

4.2 *Dropout ~~not~~ as Baeysian*

Nesse estágio, tentamos replicar os experimentos realizados no *paper* original, e escolhemos utilizamos uma RBF com as mesmas especificações e o *dataset California Housing*. A estrutura das nossas redes podem ser encontradas no repositório já citado.

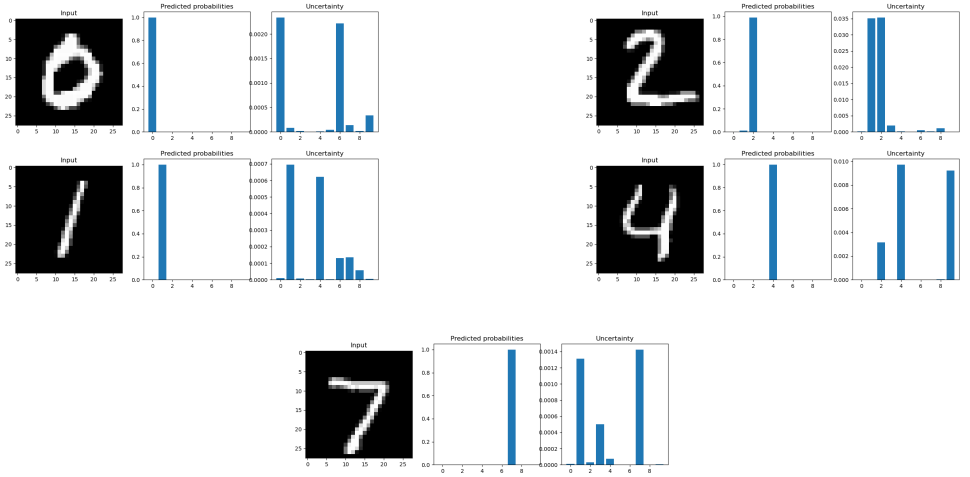

Figure 2: Incertezas para MNIST

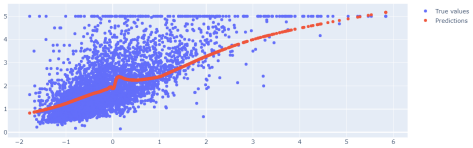

Figure 3: Dados sem MCD

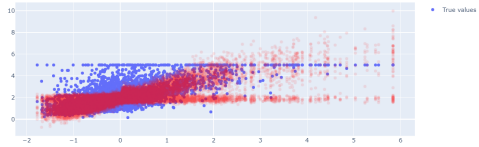

Figure 4: Dados com MCD

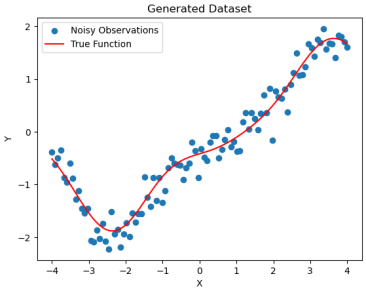

Figure 5: Dados sem MCD

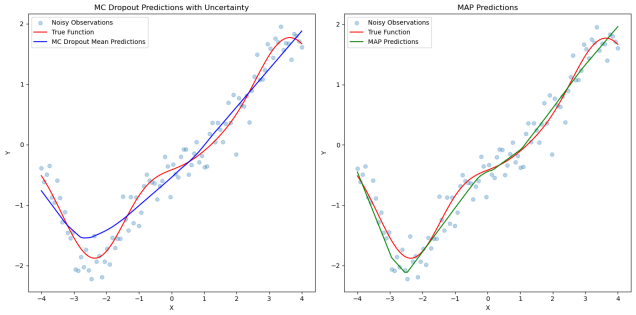

Figure 6: Dados com MCD

Gerando a rede RBF base (Fig 5), seguindo as instruções do *paper*, esperamos encontrar o mesmo resultado, com a rede utilizando *MC Dropout* sendo uma aproximação pior do que a que segue o método de Inferência Variacional, no caso o MAP.)Como podemos ver nos dados (Fig 6), a aproximação da curva com *dropout* ("esquerda") é muito mais grosseira do que a aproximação utilizando MAP. No nosso repositório tem mais um exemplo de como isso acontece para uma rede RBF, agora utilizando uma função seno.

Para realizar um teste mais complexo, decidimos utilizar, como já falado o *California Housing*. Como podemos ver na “Fig. 7”, para esse *dataset* mais complexo e que não conhecemos a distribuição real, não parece haver muita diferença, o que inclusive é indicado pelos MSE, que são 0.699 e 0.694, para o *MC Dropout* e o MAP, respectivamente.

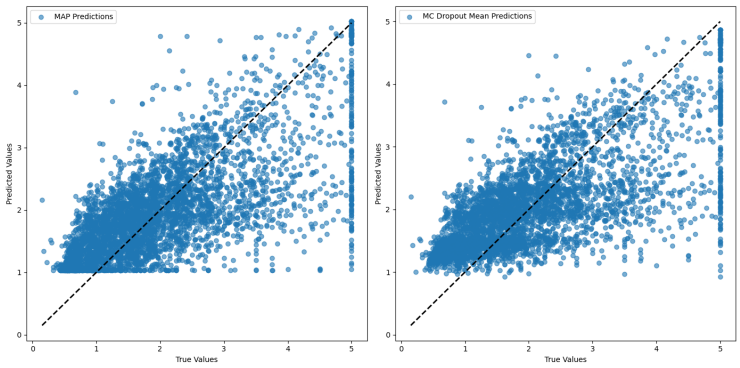

Figure 7: VI x MCD - California Housing

5 Conclusão

Analisando esses dois *papers*, e mais alguns materiais que estarão *linkados* nas referências, pudemos perceber que apesar de ser amplamente utilizado, o *MC Dropout* carece de comprovações teóricas fortes que garantam sua efetividade. O paper mencionado demonstra para alguns casos onde fica claro que o método é muito mais incompleto e dá uma aproximação muito pífia, apesar de parecer fazer sentido em alguns casos.

Um dos pontos que gostaríamos de destacar é uma discussão que encontramos no Reddit [8], citando o *paper Risk versus Uncertainty in Deep Learning: Bayes, Bootstrap and the Dangers of Dropout. Workshop on Bayesian Deep Learning* [5], que faz uma distinção interessante entre risco e incerteza, destacando que o risco é uma incerteza devido a variabilidade intrínseca dos dados, algo mais estocástico, enquanto a incerteza é algo relacionado aos parâmetros. Logo, fazendo essa distinção, podemos perceber que o que o *MC Dropout* estaria modelando seriam incertezas epistemológicas, que tem relação com os parâmetros e não com a variável e sua distribuição a posteriori real. Alguns papers no entando, que não nos aprofundamos tanto, sugerem que o *dropout* seria capaz de capturar tanto incerteza quanto o risco [1], [6].

Com isso, nossa percepção é a de que o *MC Dropout* apesar de não possuir comprovações teóricas muito fortes, através de suas mais de 7000 citações conseguiu encontrar bases empíricas de que consegue de alguma forma traduzir e modelar a incerteza. É importante notar, que o *paper* original possui premissas bem fortes, como por exemplo, que a KL é insignificante, podendo ser desconsiderada no cálculo da média e variância. Talvez isso seja o maior problema ao aplicar em datasets reais, que podem não atender as premissas.

Por fim, é importante notar que o mundo real dificilmente pode ser modelado através de uma gaussiana padrão. Os diferentes modelos possuem vantagens e desvantagens, sendo mais próximos ou menos do ideal. A principal vantagem do *MC Dropout* é estimar algum tipo de incerteza (mesmo que não aquela da posteriori real dos dados) à um baixo custo. Dependendo da aplicação, essa explicação seja suficiente. Portanto é importante ter senso crítico e compreender os métodos sendo utilizados para entender as premissas e *trade-offs* sendo realizados.

Referências

[1] Alarab, I., Prakoonwit, S. & Nacer, M.I. Illustrative Discussion of MC-Dropout in General Dataset: Uncertainty Estimation in Bitcoin. Neural Process Lett 53, 1001–1011 (2021). `https://doi.org/10.1007/s11063-021-10424-x`

[2] Gal, Yarin and Ghahramani, Zoubin. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. (2016). arXiv:1506.02142. `https://arxiv.org/abs/1506.02142`

[3] Gal, Yarin and Ghahramani, Zoubin. Dropout as a Bayesian Approximation: Appendix. (2016). arXiv:1506.02157. `https://arxiv.org/abs/1506.02157`

[4] Hron, Jiri and Matthews, Alexander G. de G. and Ghahramani, Zoubin. Variational Gaussian Dropout is not Bayesian. (2017). arXiv:1711.02989. `https://arxiv.org/abs/1711.02989`

[5] Osband, Ian. Risk versus Uncertainty in Deep Learning: Bayes, Bootstrap and the Dangers of Dropout. Workshop on Bayesian Deep Learning, NIPS (2016). https://api.semanticscholar.org/CorpusID:8985844

[6] Seoh, Ronald. Qualitative Analysis of Monte Carlo Dropout. (2020). arXiv:2007.01720. https://arxiv.org/abs/2007.01720

[7] Folgoc, Loic Le and Baltatzis, Vasileios and Desai, Sujal and Devaraj, Anand and Ellis, Sam and Martinez Manzanera, Octavio E. and Nair, Arjun and Qiu, Huaqi and Schnabel, Julia and Glocker, Ben. Is MC Dropout Bayesian? (2021). arXiv:2110.04286. https://arxiv.org/abs/2110.04286

[8] "R/machinelearning on reddit: [d] what is the current state of dropout as Bayesian approximation?" Reddit. Available at: https://www.reddit.com/r/MachineLearning/comments/7bm4b2/d_what_is_the_current_state_of_dropout_as/