Bayesian model choice via mixture estimation model : Poisson versus Geometric regression models

SELVESTREL Alexandre, KAMARY Kaniav

1. Introduction
2. Overview of bayesian methods for model selection
3. Overview of bayesian methods for model selection

Problem of model selection in the bayesian framework

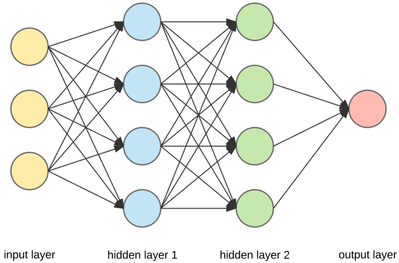


Figure 1 – A shallow neural network

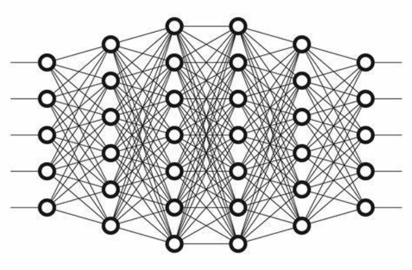


Figure 2 – A deeper neural network

Question of the activation function :

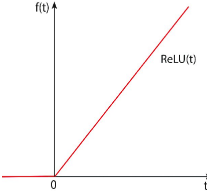


Figure 3 – ReLU function

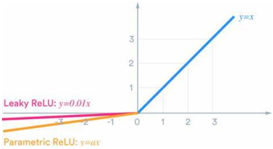


Figure 4 – Leaky ReLU function

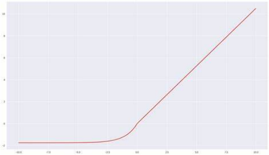
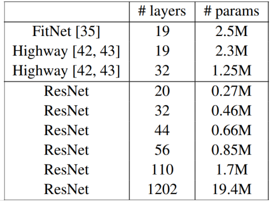


Figure 5 – SeLU function

Very high number of parameters (without regularisation)

Example of resnet : [?]

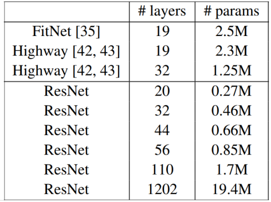


	# layers	# params
FitNet [35]	19	2.5M
Highway [42, 43]	19	2.3M
Highway [42, 43]	32	1.25M
ResNet	20	0.27M
ResNet	32	0.46M
ResNet	44	0.66M
ResNet	56	0.85M
ResNet	110	1.7M
ResNet	1202	19.4M

Figure 6 – Number of parameters for different neural networks

Very high number of parameters (without regularisation)

Example of resnet : [**?**]



	# layers	# params
FitNet [35]	19	2.5M
Highway [42, 43]	19	2.3M
Highway [42, 43]	32	1.25M
ResNet	20	0.27M
ResNet	32	0.46M
ResNet	44	0.66M
ResNet	56	0.85M
ResNet	110	1.7M
ResNet	1202	19.4M

Figure 6 – Number of parameters for different neural networks

$$\mathrm{AIC} = 2(\mathrm{Card}(\mathrm{param}) - \log(\hat{L})) \tag{1}$$

VC dimension : maximal number of point such that there exists a generally positioned data point set of that can be shattered by the model

Need for Bayesian methods

- Go further than a simple balance to strike between accuracy on the training data set and over-fitting
- More interpretability
- More understanding of our uncertainty

Bayes Factor

$$B_{01} = \frac{\frac{P(H_0|x)}{P(H_1|x)}}{\frac{P(H_0)}{P(H_1)}} = \frac{P(H_0|x)P(H_1)}{P(H_1|x)P(H_0)} \tag{2}$$

Advantage :

- Allows to clearly see the dependency on initial hypothesis (or to "eliminate" it partially...)
- Shows the importance of new data

Bayes Factor

$$B_{01} = \frac{\frac{P(H_0|x)}{P(H_1|x)}}{\frac{P(H_0)}{P(H_1)}} = \frac{P(H_0|x)P(H_1)}{P(H_1|x)P(H_0)} \qquad (2)$$

Advantage :

- Allows to clearly see the dependency on initial hypothesis (or to "eliminate" it partially...)
- Shows the importance of new data

Disadvantages :

- Just a description of the "evolution" of the probability
- No penalization nor finegrained description of uncertainty

Directly giving (hierarchical) probabilities to hypothesis

Kaniav Kamary, Kerrie Mengersen, Christian P. Robert, and Judith Rousseau. Testing hypotheses via a mixture estimation model, 2018.