Week 9: Bayesian Interpolation

Topic

First week of "CSC2541: Scalable and Flexible Models of Uncertainty" course.
Readings section.
Bayesian regression

- MacKay, 1992. Bayesian interpolation.

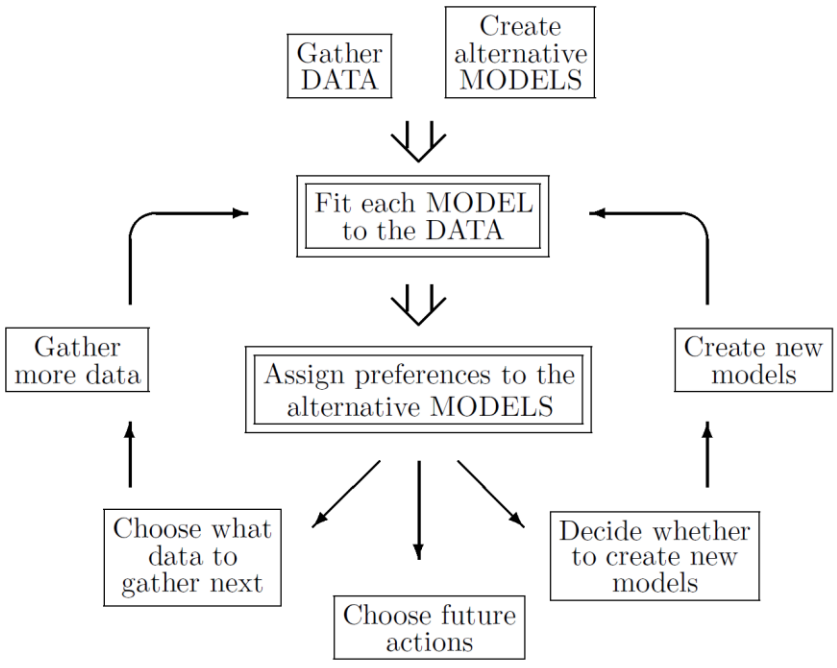
It is a rather long paper (33 pages) and I think we should be able to cover around 10-15 pages in the first week. Hopefully, we will continue it in the second week.

We covered Following sections in this week:

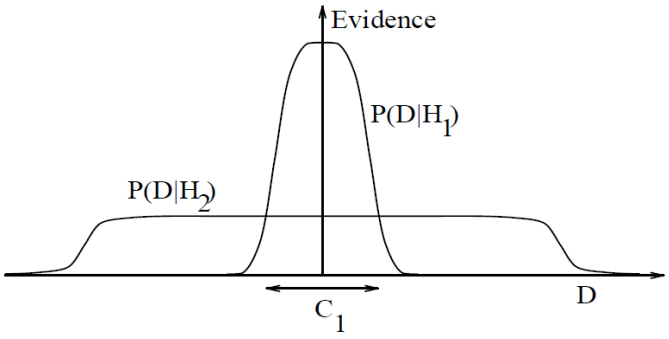
1. Data modeling and Occam's razor
2. The evidence and the Occam factor

Summary

- Abstraction of data modelling process



- Why Bayes embodies Occam's razor



Questions

You should write any questions you encounter during reading in this section. Don't hesitate to write any questions you have.

1. "The fact that Bayesian model comparison embodies Occam's razor has been rediscovered by Kashyap (1977) in the context of modeling time series; his paper includes a thorough discussion of how Bayesian model comparison is different from orthodox "Hypothesis testing." " Has anyone read this paper?!

2. "It is true that at the first level of inference, a Bayesian's results will often differ little from the outcome of an orthodox attack.". Is it actually a true statement in the world of deep learning?
3. Can somebody explain equation 2.2?
4. Why does it start with model comparison if we should theoretically use the sum rule to marginalize the unknown model?
   P(y|x,D, Hi) = \sigma p(y|x,w,Hi) p(w|Hi ,D)
5. " Patrick and Wallace discuss a practical method of assigning relative prior probabilities to alternative models by evaluating the lengths of the computer programs that decode data previously encoded under each model." What does it actually say?! **A:** (Chapter 28, Section 3 Mackay book)
6. P9: "The Occam factor also provides a penalty for models that have to be finely tuned to fit the data; the Occam factor promotes models for which the required precision of the parameters Delta(w) is coarse." What does it say?
7. What is the "Sure Thing" in p7?
8. What is the characteristic width exactly and how is it computed?

9. In the p7: "Historically, Bayesian analysis has been accompanied by methods to work out the "right" prior P(w|H) for a problem, for example, the principles of insufficient reason and maximum entropy." I don't understand the point.

10. "The log of the Occam factor can be interpreted as the amount of information we gain about the model when the data arrive" can somebody give me an intuition on this?

11. p10: "It is common for there to be degeneracies in models with many parameters; that is, several equivalent parameters could be relabeled without affecting the likelihood. In these cases, the right-hand side of equation 2.6 should be multiplied by the degeneracy of w_mp to give the correct estimate of the evidence." Can someone provide some insights?

Comments

You should write any comments you have about the paper in this section. You can highlight some parts that you find ambiguous, add some extra explanations to some parts, pointing out to the recent works that can complement this reading (add extra references so anyone can read them later.)

1. Maybe it is good to look at this paper to see the difference between the assumption of this paper and the deep learning regime.
2. One thing that I think is worth thinking about is the role of prior. Let's now focus on the first level of inference (where we have fixed the hypothesis and we want to infer the parameters based on the data). We have seen in the paper that if we have a significant amount of data (I mean the number of data points is much larger than the number of parameters), the prior does not play a significant role; in this scenario, data reduces the importance of the prior. But now let's think about the deep leaning setting, where the number of parameters is significantly larger than the number of samples. Therefore, in this regime (over-parameterized regime), prior can play an important role. **The question is what is the role of prior in this scenario? And how can we think about choosing the right prior in this over-parameterized setting** (putting prior on the weights of the network in this over-parameterized regime seems kind of arbitrary)?

That is the challenging question and there is not a unanimous answer to this question. Maybe you can look at this blog post.