Does Bayesian model averaging "overfit"?

Balaji Lakshminarayanan

Sources: [Domingos, 2000], [Minka, 2000], [Clarke, 2003], [Monteith et al., 2011]

Bayesian model averaging (BMA)

- Simple binary classification: Training data $D = \{x_n, y_n\}$, classifier $h \in H$
- BMA: prediction

$$\begin{aligned} p(y|x, D) &= \sum_h p(y|x, h) p(h|D) \\ p(h|D) &\propto p(h) \prod_n p(y_n|x_n, h) \qquad (1) \end{aligned}$$

Does Bayesian model averaging "overfit"?

"Bayesian averaging of classifiers and the over fitting problem" [Domingos, 2000]

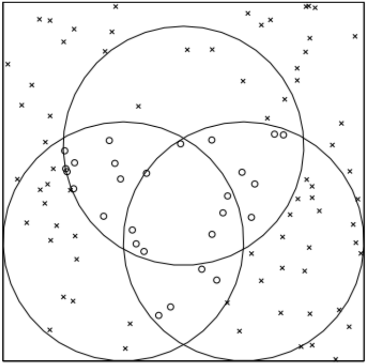
- *Bagging can be interpreted as importance sampling approximation to BMA* in (1)
- Empirical evaluation shows that bagging outperforms BMA
- *Further investigation shows this to be due to a marked tendency to overfit on the part of Bayesian model averaging, contradicting previous beliefs that it solves (or avoids) the overfitting problem.*

Does Bayesian model averaging "overfit"? (contd.)

- Say $p(y|x,h)$ is $1-\epsilon$ if h correctly predicts y
- Let h_k correctly classify r_k out of n training data points
- $p(h_k|D) \propto \epsilon^{n-r_k}(1-\epsilon)^{r_k}$
- For $n=100$, a learner that achieved 95% accuracy would be weighted as 17 times more likely than a learner that achieved an accuracy of 94%.
- *This is an example of overfitting: preferring a hypothesis that does not truly have the lowest error of any hypothesis considered, but that by chance has the lowest error on the training data*
- "Better" Bayesian inference seems to perform worse empirically .... what's going on here?

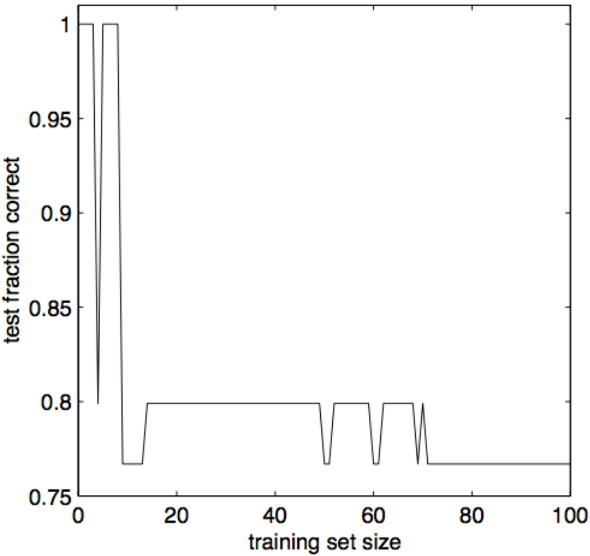
What does BMA *really* do?

"Bayesian model averaging is not model combination"
[Minka, 2000]



- Class 'o' if data point under two or more circles, 'x' otherwise
- BMA converges to top-most circle

What does BMA *really* do? (contd.)



What does BMA *really* do? (contd.)

- BMA accounts for uncertainty of model correctness by integrating over the model space and weighting each model by the probability of its being the correct model.
- Although BMA produces a combination of models, it assumes that **one and only one of the models is indeed the Data generating model** (DGM).
- BMA is "soft" model selection. In the limit of infinite data, BMA would converge to the single best model.

What does BMA *really* do? (contd.)

- BMA accounts for uncertainty of model correctness by integrating over the model space and weighting each model by the probability of its being the correct model.
- Although BMA produces a combination of models, it assumes that **one and only one of the models is indeed the Data generating model** (DGM).
- BMA is "soft" model selection. In the limit of infinite data, BMA would converge to the single best model.
- [Minka, 2000]: *"... the only flaw with BMA is the belief that it is an algorithm for model combination, when it is not."*

What does BMA *really* do? (contd.)

- Ensemble methods do more than accounting for model uncertainty. They operate on a much richer hypothesis space.
  - Approximate BMA interpretation of bagging misses the point

What does BMA *really* do? (contd.)

- Ensemble methods do more than accounting for model uncertainty. They operate on a much richer hypothesis space.
  - Approximate BMA interpretation of bagging misses the point
- "Comparing Bayes model averaging and stacking when model approximation error cannot be ignored" [Clarke, 2003]
  - If true DGM is not in the model space, BMA converges to the single best model (NOT the best combination)
  - BMA is not robust to model misspecification issues
- [Monteith et al., 2011]: Brute force Bayesian averaging over combination of models (about $3^{10} = 50K$ model combinations) outperforms bagging and stacking

Take home messages

- Even if you are a Bayesian, you still need to be mindful about model misspecification ... "Better" Bayesian inference in a misspecified model can lead to poorer empirical performance
- If DGM is a combination of models, model combination methods (eg. bagging, stacking) can outperform optimal model averaging
- Bayesian inference over additive hypothesis spaces should outperform bagging and stacking ... Surprisingly little work on computationally efficient Bayesian methods for this problem

Take home messages

- Even if you are a Bayesian, you still need to be mindful about model misspecification ... "Better" Bayesian inference in a misspecified model can lead to poorer empirical performance
- If DGM is a combination of models, model combination methods (eg. bagging, stacking) can outperform optimal model averaging
- Bayesian inference over additive hypothesis spaces should outperform bagging and stacking ... Surprisingly little work on computationally efficient Bayesian methods for this problem

PS: SMC posterior for Bayesian decision trees $\neq$ Random forests :)

Thank you!

Clarke, B. (2003).
Comparing bayes model averaging and stacking when model approximation error cannot be ignored.
The Journal of Machine Learning Research, 4:683–712.

Domingos, P. (2000).
Bayesian averaging of classifiers and the overfitting problem.
In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 223–230.

Minka, T. P. (2000).
Bayesian model averaging is not model combination.
MIT Media Lab note. `http://research.microsoft.com/en-us/um/people/minka/papers/bma.html`.

Monteith, K., Carroll, J. L., Seppi, K., and Martinez, T. (2011).
Turning bayesian model averaging into bayesian model combination.
In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 2657–2663. IEEE.