Lecture 9

Model Selection

- Bayes theorem quantifies the posterior pdf of parameters describing a single model, with that model assumed to be true.

- In model selection, we formulate alternative models/scenarios and ask which ones are best supported by data.

e.g. is data well described by a Gaussian or Cauchy dist.?

e.g. is a set of points better fit by a straight line or a parabola?

- Given two models, M_1 & M_2, the "support" for one model over the other is given by the odds' ratio,

$$\boxed{O_{21} \equiv \frac{p(M_2|D,I)}{p(M_1|D,I)}}$$

where $p(M_x|D,I)$ is the posterior probability for the model M_x.

- Recall that the full Bayes theorem is,

$$p(M,\theta \mid D,I) = \frac{p(D \mid M,\theta,I)\, p(M,\theta \mid I)}{p(D \mid I)}$$

Thus, the posterior for the model M is

$$\boxed{p(M \mid D,I) = \int p(M,\vec{\theta} \mid D,I)\, d\vec{\theta}}$$

i.e. it is marginalizing over the model parameters,

- Also recall that the prior can be written as,

$$\boxed{p(M,\vec{\theta} \mid I) = p(\vec{\theta} \mid M,I)\, p(M \mid I)}$$

where $p(\vec{\theta} \mid M,I)$ are the "priors on the parameters, $\vec{\theta}$," for the given model M & information, I; while $p(M \mid I)$ is the prior probability on the model itself.

– Going back to Bayes' theorem,

$$p(M \mid D,I) = \frac{p(D \mid M,I)\, p(M \mid I)}{p(D \mid I)}$$

where

$$\boxed{E(M) \equiv p(D \mid M,I) = \int p(D \mid M,\vec{\theta},I)\, p(\vec{\theta} \mid M,I)\, d\theta}$$

is called the marginal likelihood for

is called the marginal likelihood for model M.

- $E(M)$ quantifies the probability that the data D would be observed if the model M is the correct model.

- In physics, $E(M)$ is often referred to as "evidence".

- This is also sometimes called the "global likelihood".
- This is thus the weighted average of the likelihood function with the prior for for model parameters acting as the weighting function.

- Also notice that to calculate the odds ratio, the denominator cancels out, leaving,

$$\boxed{O_{21} = \frac{E(M_2)\,p(M_2|I)}{E(M_1)\,p(M_1|I)} = B_{21}\,\frac{p(M_2|I)}{p(M_1|I)}}$$

where B_{21} is called the Bayes' factor, and is,

$$B_{21} = \frac{\int p(D|M_2,\vec{\theta}_2,I)\,p(\vec{\theta}_2|M_2,I)\,d\vec{\theta}_2}{\int p(D|M_1,\vec{\theta}_1,I)\,p(\vec{\theta}_1|M_1,I)\,d\vec{\theta}_1}$$

- Notice when two (or more) models have equal prior probability, i.e. $p(M_2|I) = p(M_1|I)$

$\Rightarrow O_{21} = B_{21}$

– The odds ratio is interpreted using the subjective Jeffrey's scale,

$O_{21} < 3$ → evidence "not worth mentioning"

$O_{21} > 10$ → "strong" evidence

$O_{21} > 100$ → "decisive" evidence.

Bayesian hypothesis testing:

– A special case of model comparison is hypothesis testing, where $M_2 = \overline{M}_1$, i.e.

$$p(M_1) + p(M_2) = 1 .$$

- If M_1 is the null hypothesis, we can ask whether the data supports the alternative hypothesis, or in other words, if we can reject the null hypothesis.

– Given that M_2 is simply a complementary hypothesis to M_1, it is not possible to compute $p(D|M_2)$.

- This inability to reject M_1 in the absence of an alternative hypothesis is very different from frequentist statistics.

- This is because Bayesian inference is based on the posterior, rather than just the likelihood, and cannot reject a hypothesis if there are no alternative explanations for our observed data.

- For example, in the coin flip example, to reject our null hypothesis that the coin is fair ($b_* = 0.5$), we need an alternative hypothesis that the coin has an unknown true heads probability b_0, and compare these two models through the odds ratio (or BF if M_1 & M_2 have equal probability).

Occam's Razor

- The expression for the odds' ratio intrinsically penalizes models for their complexity.

- When data are much more informative than priors:
  - If a parameter, or degenerate combination, is unconstrained by the data, then there is no penalty associated with including it.
  - The odds ratio can justify an additional, constrained parameter only if it is justified by an increase in the likelihood, or by the ratio of the prior probabilities

by the ratio of the prior probabilities.

– A cheaper way to calculate model evidence or selection than the full odds' ratio is the <u>Bayesian Information criterion</u>.

- The BIC, based on Gaussianity assumptions for likelihood & posterior, is given by

$$BIC \equiv -2 \ln [L^{o}(M)] + k \ln N$$

where $L^{o}(M)$ is the max. likelihood value (not the MAP) & N is no. of model parameters.

- The model with lower BIC is the preferred model.