Evidence and Occam's razor

Based on David J.C. MacKay: Information Theory and Learning Algorithms, chapters 24,27, and 28

Arto Klami

18th March 2004

Contents

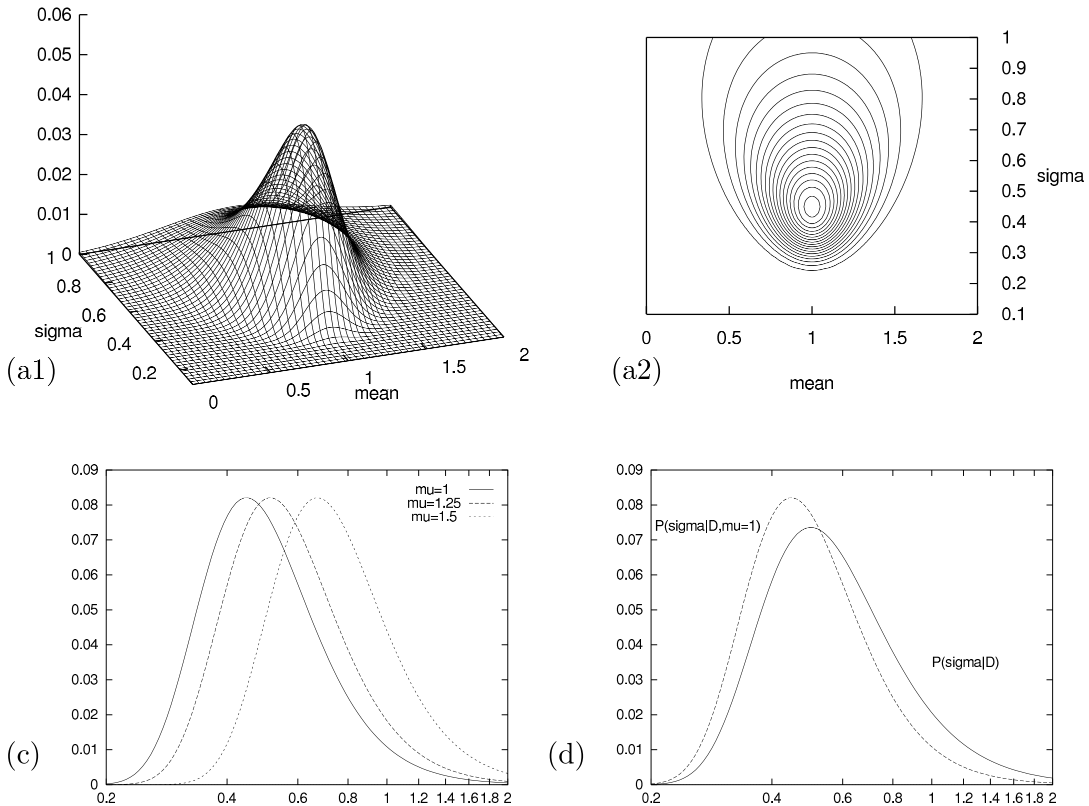
- Tools:
  Exact marginalization
  Laplace's approximation
- Occam's razor:
  Idea
  Two stages of modeling
  Evidence and Occam factor
  Minimum Description Length (MDL)
  Connection to cross-validation

Exact marginalization

$$p(x|H) = \int p(x, y|H)dy$$

- "..is a macho activity enjoyed by those who are fluent in definite integration" (MacKay)
- The concept is necessary:
  $p(x|H)$ is not the same as $p(x|\hat{y}, H)$, where $\hat{y}$ is some fixed value
- In practice possible only for some simple distributions (Gaussian) and conjugate priors, still quite difficult
- Discrete distributions: sum over all values
  Also possible in graphs etc. (Chapters 25, 26)
- Low-dimensional distributions can be discretized

Marginalization vs Point estimates



Laplace's approximation

- The goal is to approximate normalization constant Z of an unnormalized probability distribution, $Z = \int p(x)dx$
- Idea: Approximate the distribution by a Gaussian at the mode
- Taylor's expansion of the logarithm:

$$\ln p(x) = \ln p(x_0) - \frac{1}{2}(x - x_0)^T A(x - x_0) + ...$$

- Needs only the posterior mode and matrix of second derivatives (Hessian matrix, $A_{ij} = -\frac{\partial^2}{\partial x_i \partial x_j} \ln p(x)|_{x=x_0}$)
- Easy to compute Z because the normalization constant of the Gaussian is known

Laplace's approximation 2/2

- Problem or opportunity:
  depends on the basis, i.e., non-linear transformation changes the approximation (Exercise)
  $\longrightarrow$ find a parameterization that gives approximately normal distribution
- Approximates only one mode of multimodal distributions

Occam's razor - Idea

- "Accept the simplest explanation that fits the data"
- Machine learning needs to grasp the same intuition
- Bayesian way of thinking? We could prefer simpler models by giving them larger prior
- It turns out that we do not need to make such prior assumptions. Instead, the Occam's razor is automatically achieved by Bayesian inference

Two stages of inference

- Model fitting and model comparison
- Fitting: posterior $= \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \propto$ likelihood $\times$ prior
- Comparison: posterior $\propto$ evidence $\times$ prior
- Evidence does what Occam's razor asks for

Evidence

- Posterior ratio of hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \frac{P(H_1)}{P(H_2)}$$

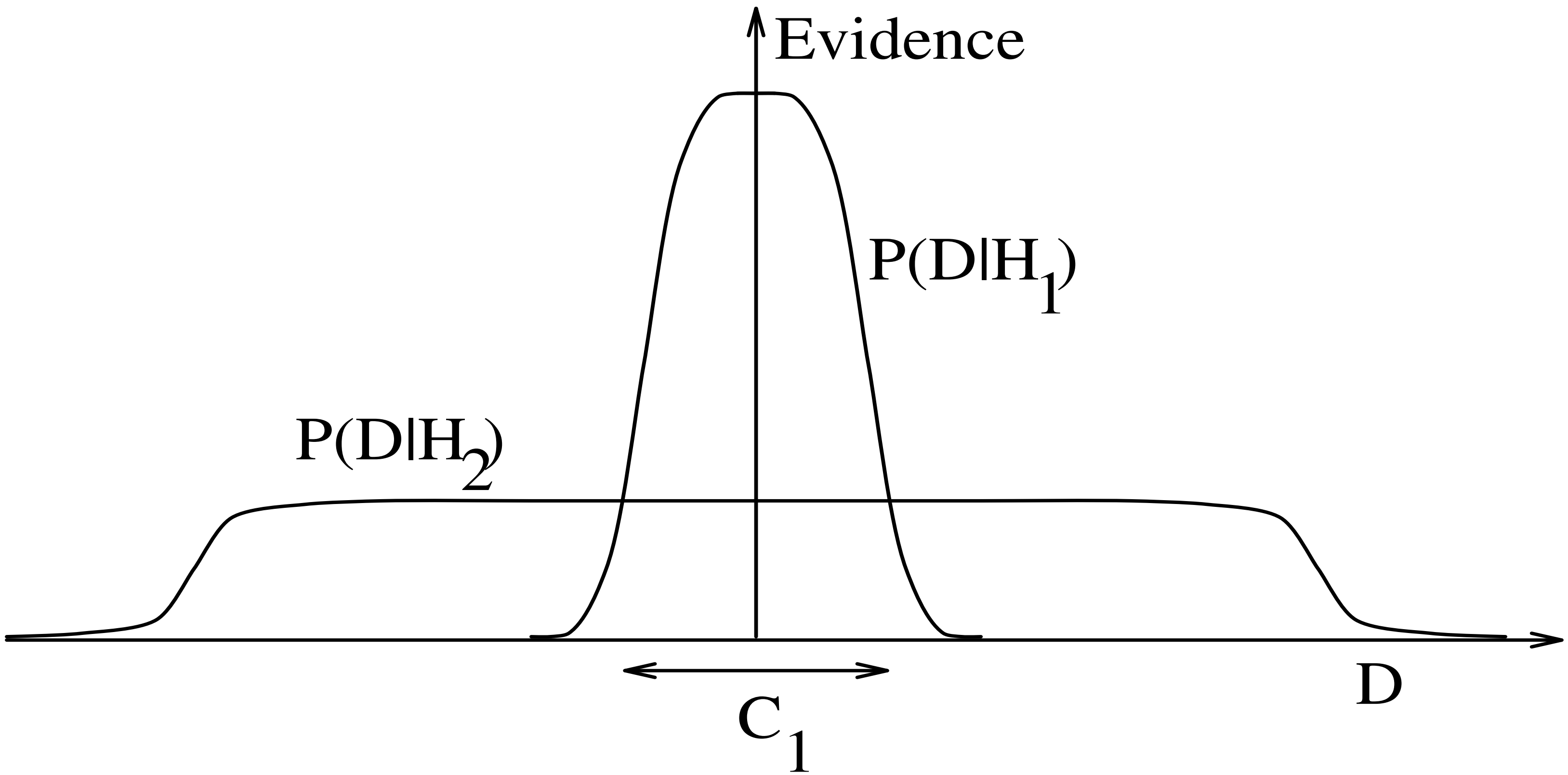
- $P(D|H) = \int P(D|w, H)P(w|H)dw$ is called the evidence of the model

- Evidence is the average probability of generating the data by randomly selecting parameter values

- Simple model: a few data sets, high evidence

- Complex model: numerous data sets, small evidence

Evidence — an illustration



What to do with evidence

- MacKay: Always average over different models, weighting each model by $P(H|D)$
- In practice we often need to select one model
- Interpreting the Bayes factor $B = \frac{P(D|H_1)}{P(D|H_2)}$:

Jeffreys (1961)		Kass, Raftery (1995)	
B	Evidence against H_2	B	Evidence against H_2
1 - 3.2	Worth mentioning	1 - 3	Worth mentioning
3.2 - 10	Substantial	3 - 20	Positive
10 - 100	Strong	20 - 150	Strong
> 100	Decisive	> 150	Very strong

Computing evidence

- Exact evidence – often impossible

$$P(D|H) = \int P(D|\boldsymbol{w}, H)P(\boldsymbol{w}|H)d\boldsymbol{w}$$

- Laplace's method:

$$\begin{array}{rccc} P(D|H) \approx & P(D|\boldsymbol{w}_{\mathsf{MP}}, H) & \times & P(\boldsymbol{w}_{\mathsf{MP}}|H)\sigma_{\boldsymbol{w}|D} \\ \text{Evidence} \approx & \text{Best fit likelihood} & \times & \text{Occam factor} \end{array}$$

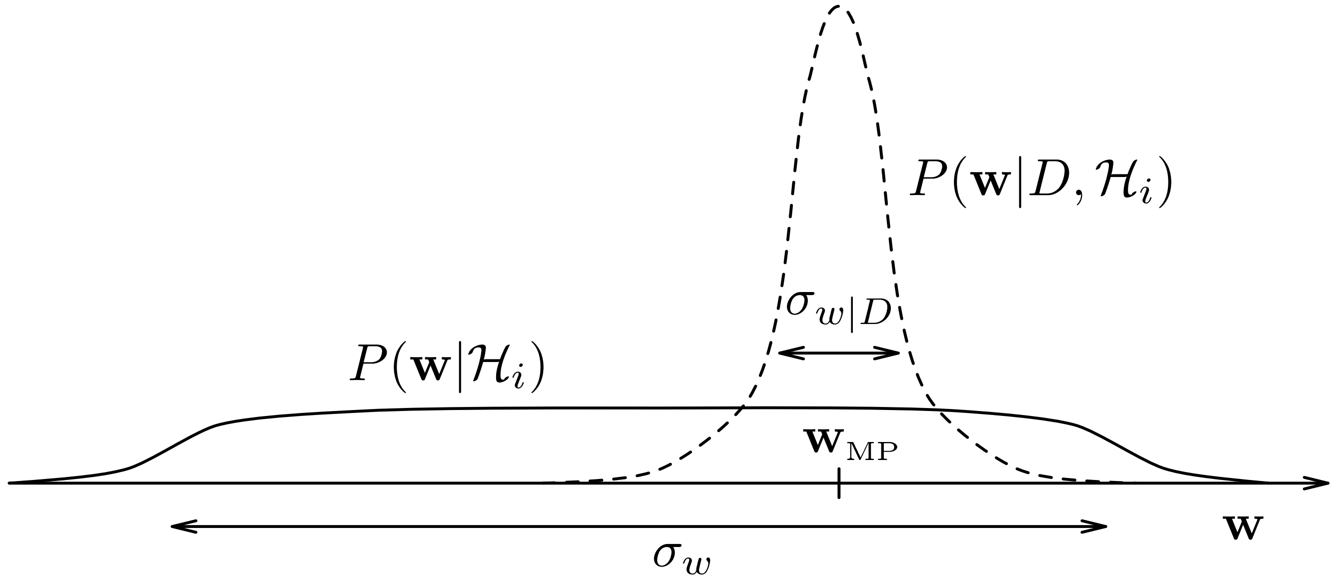
- Normalization constant $\propto \sigma_{\boldsymbol{w}|D}$, the standard deviation of the posterior distribution

- Only MAP-estimate and error bars (Hessian) required

Occam factor

- Occam factor: $P(\boldsymbol{w}_{\mathsf{MP}}|H)\sigma_{\boldsymbol{w}|D}$
- Interpretation: Assume flat prior, then $P(\boldsymbol{w}_{\mathsf{MP}}|H) = 1/\sigma_{\boldsymbol{w}}$
  $\rightarrow$ Occam factor is ratio of posterior and prior widths
- The factor by which hypothesis space collapses when the data arrive
- Logarithm of the factor measures the amount of information gained about parameters when the data arrive

Occam factor — an illustration



Occam factor - Problems

- The prior has to be proper
- The factor depends on the prior
- Consider two identical models with different priors:
  The one with better fitting prior has larger evidence
- Should tweaking the prior lead to higher evidence?
- Conclusion: be careful with Occam factor

Minimum description length and Occam's razor

- Instead of probabilities, consider message lengths required to communicate events without loss
- Message lengths correspond to probabilities by $L(x) = -\log_2 P(x)$
- Communicate data with two-part message: the model and the data given the model $L(D, H) = L(H) + L(D|H)$
- Sending the model means identifying what model to use and then sending the parameters of the model
- Corresponds to the Bayesian analysis:

$$L(D, H) = -\log P(H) - \log(P(D|H)\delta D) = -\log P(H|D) + const$$

Evidence and cross-validation

- Evaluating the evidence has a relation to cross-validation
- De-compose the log-evidence into

  $$\log P(D|H) = \log P(x_1|H) + \log P(x_2|x_1, H) + ... + \log P(x_n|x_1, ..., x_{n-1}, H)$$

- Leave-one-out cross-validation measures the expectation of the last term $\log P(x_n|x_1, ..., x_{n-1}, H)$ under data re-orderings
- Evidence, on the other hand, measures how well the whole data is predicted by the model, starting from scratch

Conclusions

- Bayesian inference consists of model fitting and comparison
- Occam's razor: prefer simpler models — automatically embodied by evidence of the model
- Computing the evidence in difficult — in practice some approximations have to be used

Exercises

- Exercise 27.1, page 342: Laplace's approximation for Poisson distribution in two bases. Compare the resulting approximations to the unnormalized posterior, and study the differences in approximation accuracy.
- Exercise 28.1, page 354: Evaluate the evidences of two competing models. For H_1, assume uniform prior for m. Discretizing the problem is probably the easiest way of computing the evidence. Why Laplace's approximation would not be good here? How would you interpret the results?