Machine Learning 1
Lecture 5.2 - Supervised Learning
Bayesian Linear Regression - Bayesian Model Comparison

Erik Bekkers

(Bishop 3.4)
Bayesian Model Selection

- Given $L$ models $\{M_i\}_{i=1}^L$ with prior belief $p(M_i)$

- Update prior knowledge with observations on the data $D$:
  \[
p(M_i|D) = \frac{p(D|M_i) p(M_i)}{p(D)}
  \]

- Predictive distribution / mixture distribution / model average:
  \[
p(t'|x', D) = \sum_{i=1}^L p(t'|x', M_i) p(M_i|D)
  \]

- Approximation: Use most probable model for predictions
  \[
  M^* = \arg\max_{M_i} p(M_i|D) = \arg\max_{M_i} \left[ p(D|M_i) p(M_i) \right]
  \]

  \[
p(t'|x', D, M^*)
  \]
Bayesian Model Comparison

- Model selection

\[ M^* = \arg \max_{M_i} p(M_i | D) = \arg \max_{M_i} p(D | M_i) p(M_i) \]

- Comparing two models \( M_1 \) and \( M_2 \):

\[
\frac{p(M_1 | D)}{p(M_2 | D)} = \frac{p(D | M_1)}{p(D | M_2)}
\]

- When quotient of priors \( \frac{p(M_1)}{p(M_2)} \) is known or close to 1, then we need

\[
\frac{p(D | M_1)}{p(D | M_2)} \]

Bayes factor

- Model evidence / marginal likelihood:

\[
p(D | M_i) = \int p(D | \mathbf{w}, M_i) p(\mathbf{w} | M_i) d\mathbf{w}
\]
Approximated Model Evidence

- Model evidence / marginal likelihood for single parameter $w$
  \[ p(D|M_i) = \int p(D|w, M_i)p(w|M_i)dw \]

- Note that $p(D|M_i)$ is the normalization constant of $p(w|D, M_i)$

- If posterior $p(w|D, M_i)$ is sharply peaked at $w_{\text{MAP}}$ with width $\Delta w_{\text{posterior}}$
  \[ p(w|M_i) = 1/\Delta w_{\text{prior}} \]
  \[ p(D|M_i) = \int p(D|w, M_i)p(w|M_i)dw \approx \frac{p(D|w_{\text{MAP}}, M_i)}{\Delta w_{\text{prior}}} \Delta w_{\text{post}} \]

- $\ln p(D|M_i) \approx \ln p(D|w_{\text{MAP}}, M_i) + \ln \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$

Figure: model evidence (Bishop 3.12)
Approximated Model Evidence

- \( \ln p(D|M_i) \approx \ln p(D|w_{\text{MAP}}, M_i) + \ln \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \)

- if \( \Delta w_{\text{posterior}} < \Delta w_{\text{prior}} \) then \( \ln \left( \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)^M < 0 \)

- \( M \) parameters: \( w \in \mathbb{R}^M \)

\[
p(D|M_i) = \int p(D|w, M_i)p(w|M_i)dw \approx p(D|w_{\text{MAP}}, M_i) \left( \frac{\Delta w_{\text{post}}}{\Delta w_{\text{prior}}} \right)^M
\]

\[
\ln p(D|M_i) \approx \ln p(D|w_{\text{MAP}}, M_i) + M \ln \frac{\Delta w_{\text{post}}}{\Delta w_{\text{prior}}}
\]

- Model evidence favors models of medium complexity!

**Figure:** model evidence (Bishop 3.12)
Model evidence: medium complexity

- 3 models: $M_1$ is simplest, $M_3$ is most complex

- Generate datasets $D$ from $p(D | M_i)$

  1. sample model parameters from model prior: 
     $w \sim p(w | M_i)$
  2. Sample dataset 
     $D \sim p(D | w, M_i)$

- Note: 
  
  $$\int p(D | M_i) \, dD = 1$$

- dataset $D_0$: model $M_2$ has highest model evidence

**Figure**: model evidence (Bishop 3.12)