

UNIVERSITY OF AMSTERDAM Informatics Institute



Machine Learning 1

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

Erik Bekkers

(Bishop 3.4)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Bayesian Model Selection

- Given L models $\{\mathcal{M}_i\}_{i=1}^L$ with prior belief $p(\mathcal{M}_i)$
- Update prior knowledge with observations on the data D: $p(\mathcal{M}_i|D) = \frac{p(D|\mathcal{M}_i)p(\mathcal{M}_i)}{p(D)}$
- Predictive distribution / mixture distribution / model average: $p(t'|\mathbf{x}', D) = \sum_{i=1}^{L} \rho(t'|\mathbf{x}', M; \rho(M; D))$
- Approximation: Use most probable model for predictions Mar print

Bayesian Model Comparison

Model selection

$$\mathcal{M}^* = \underset{\mathcal{M}_i}{\operatorname{arg\,max}} p(\mathcal{M}_i | D) = \underset{\mathcal{M}_i}{\operatorname{arg\,max}} p(D | \mathcal{M}_i) p(\mathcal{M}_i)$$

- Comparing two models \mathcal{M}_1 and \mathcal{M}_2 : $\frac{p(\mathcal{M}_1|D)}{p(\mathcal{M}_2|D)} = \frac{p(\mathcal{D}\mathcal{M}_q)p(\mathcal{M}_r)}{p(\mathcal{D}\mathcal{M}_2)p(\mathcal{M}_2)}$
- When quotient of priors $\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}$ is known or close to 1, then we need



 $\frac{p(D|\mathcal{M}_1)}{p(D|\mathcal{M}_2)}$ Bays Jacks

Model evidence / marginal likelihood:

$$p(D|\mathcal{M}_i) = \int p(D|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i) d\mathbf{w}$$

Approximated Model Evidence

- Model evidence / marginal likelihood for single parameter w $p(D|\mathcal{M}_i) = \int p(D|w, \mathcal{M}_i) p(w|\mathcal{M}_i) dw$
- Note that $p(D|M_i)$ is the normalization constant of $p(w|D,M_i)$
- If posterior $p(w|D, \mathcal{M}_i)$ is sharply peaked at w_{MAP} with width $\Delta w_{\text{posterior}}$

$$p(w|\mathcal{M}_{i}) = 1/\Delta w_{\text{prior}}$$

$$p(D|\mathcal{M}_{i}) = \int p(D|w, \mathcal{M}_{i})p(w|\mathcal{M}_{i})dw \approx \frac{\rho(D|w_{\text{MAP}}, \mathcal{M}_{i})}{\Delta w_{\text{prior}}} \Delta w_{\text{post}}$$

$$\ln p(D|\mathcal{M}_{i}) \approx \ln p(D|w_{\text{MAP}}, \mathcal{M}_{i}) + \ln \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$$

$$peraddes Complexity Figure: model evidence (Bishop 3.12)$$

Approximated Model Evidence

- $\ln p(D|\mathcal{M}_i) \approx \ln p(D|w_{\text{MAP}}, \mathcal{M}_i) + \ln \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \mathcal{M}$ • if $\Delta w_{\text{posterior}} < \Delta w_{\text{prior}}$ then $\ln \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} < 0$
- M parameters: $\mathbf{w} \in \mathbb{R}^{M}$ $p(D|\mathcal{M}_{i}) = \int p(D|\mathbf{w}, \mathcal{M}_{i})p(\mathbf{w}|\mathcal{M}_{i})d\mathbf{w} \approx \bigcap \left(\bigcup |\mathcal{W}_{M,P}, \mathcal{M}_{i} \right) \left(\underbrace{\mathcal{A}_{P}}_{\mathcal{A}_{P}} \right) d\mathbf{w}$ $\ln p(D|\mathcal{M}_{i}) \approx \ln p(\mathcal{D}|\mathcal{W}_{M,P}, \mathcal{M}_{i}) + \mathcal{M} \ln \frac{\Delta_{P}}{\mathcal{D}_{P}} d\mathbf{w}$
- Model evidence favors models of medium complexity!



Model evidence: medium complexity

- ◆ 3 models: M₁ is simplest, M₃ is most complex
- Generate datasets D from $p(D|M_i)$



 ◆ dataset D₀: model M₂ has highest model evidence

Figure: model evidence (Bishop 3.12)