SPR Day 08

Bayesian Linear Regression: Prediction and Model Selection

Readings: Bishop PRML

Sec 3.3.1–3.3.2 Predictive distribution

Sec. 3.4 Model comparison

Sec 3.5 Evidence

Outline:

(1) Posterior predictive

(2) Evidence: A measure of model quality

(3) Model selection methods
Fixed Validation / Cross Validation / Evidence

(4) Estimating hyperparameters α and β

(1) Posterior predictive for Linear Regression

Recall our model:

Prior: $p(w \mid \alpha) = \mathcal{N}(w \mid 0, \alpha^{-1} I)$

Likelihood: $p(\mathbf{t} \mid w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid w^T \phi(x_n), \beta^{-1})$

iid assumption says if we saw a new observation at x_*, then: $p(t_*, \mathbf{t} \mid w, \beta) = p(\mathbf{t} \mid w, \beta)\, p(t_* \mid w, \beta)$

Then by linear-Gaussian rules:

Joint distribution $p(\mathbf{t}_{1:N}, w \mid \alpha, \beta)$ is Gaussian

Expanded Joint distribution $p(t_*, \mathbf{t}_{1:N}, w \mid \alpha, \beta)$ is Gaussian

Marginal of above $p(t_*, \mathbf{t}_{1:N} \mid \alpha, \beta)$ is Gaussian

Conditional of above $p(t_* \mid \mathbf{t}_{1:N}, \alpha, \beta)$ is Gaussian

Formulas for Predictive Posterior

$$p(t_* | t_{1:N}, \alpha, \beta) = \mathcal{N}\Big(t_* \,\Big|\, m_N^T \phi(x_*), \underbrace{\frac{1}{\beta} + \phi(x_*)^T S_N \phi(x_*)}_{\sigma_N^2(x_*)}\Big)$$

where

	shape		
m_N	$(M,1)$	is	posterior mean vector
S_N	(M,M)	is	posterior covariance matrix
$\phi(x_*)$	$(M,1)$	is	feature vector at test point x_*
β	$(1,1)$	is	scalar likelihood precision

Useful Properties

- Average over many weights w, do not commit to a point estimate
- Variance may change with location of prediction x_*. Will always be at least $\frac{1}{\beta}$, can be larger.
- Variance $\sigma_N^2(x_*)$ cannot increase as more data seen.

  In HW2 we'll prove $\sigma_{N+1}^2(x_*) \leq \sigma_N^2(x_*)$

(2) Evidence calculation

$p(w, t \mid \alpha, \beta)$ is joint

$p(t \mid \alpha, \beta)$ is evidence. A PDF over $\{t_n\}_{n=1}^{N}$ observed outputs

$$p(t \mid \alpha, \beta) = \int_w \underbrace{p(t \mid w, \beta)}_{\text{likelihood}} \underbrace{p(w \mid \alpha)}_{\text{prior}} \, dw$$

Following textbook Eqns 3.77 – 3.85, we can

(1) sub in Gaussian PDF for both likelihood and prior

(2) bring constants outside integral, completing the square

(3) recongizing the remaining form is

$$\int \exp\left\{ E(m_N) + \tfrac{1}{2}(w - m_N)^T S_N^{-1} (w - m_N) \right\} dw$$

$$= e^{E(m_N)} \frac{1}{C(m_N, S_N)}$$

← normalization constant of MV Gaussian

Thus, the log of evidence PDF is:

$$\log p(t \mid \alpha, \beta) = \frac{M}{2} \log \alpha + \frac{N}{2} \log \beta$$

prior + lik norm constants

$$- \frac{\beta}{2} \|t - \Phi m_N\|^2 - \frac{\alpha}{2} \|m_N\|^2 \quad E(m_N)$$

$$+ \frac{1}{2} \log(\det S_N) - \frac{N}{2} \log(2\pi) \quad \frac{1}{C(m_N, S_N)}$$

Remember: $\|a\|^2 = a^T a = \sum_{i=1}^{I} a_i^2$ for vectors of size I

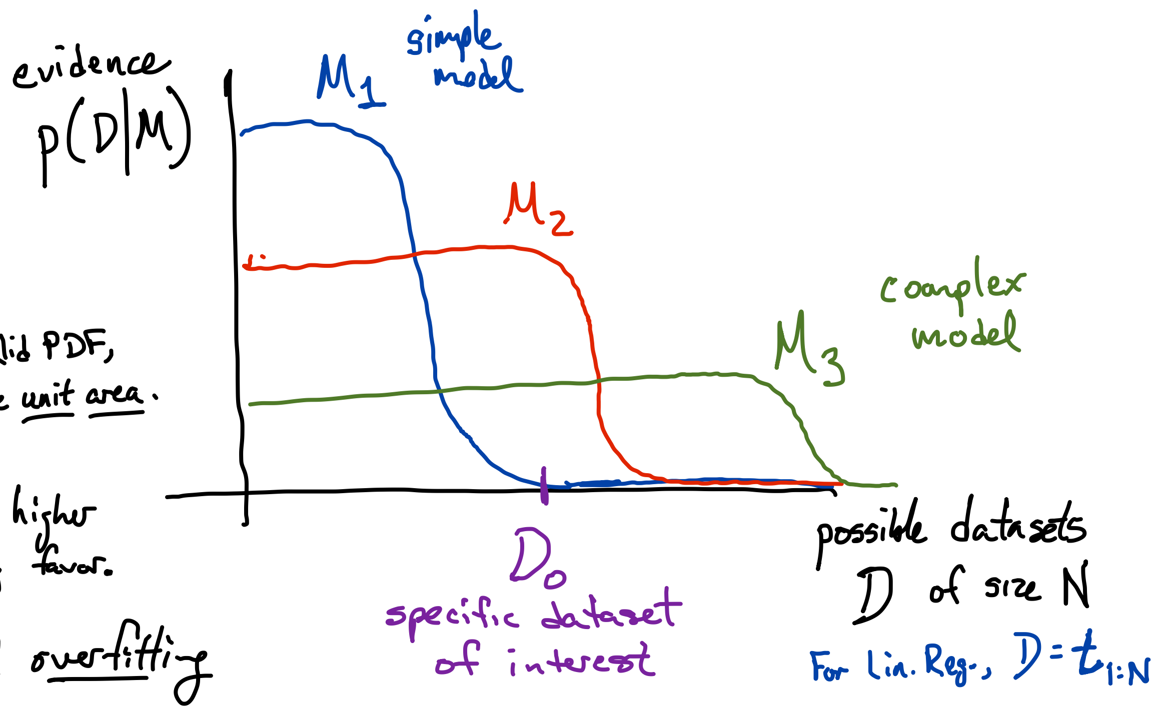
(3) Hyperparameter Selection / Model Selection

Which is better? $\alpha = 0.1, \beta = 0.1$ or $\alpha = 0.2, \beta = 0.1$?

degree-2 or degree 3 $\phi(x)$?

Can use the <u>evidence</u> for selection

Bishop PRML Fig 3.13



Evidence PDFs, like any valid PDF, are normalized to have <u>unit area</u>.

Thus, simpler models give higher density to datasets they favor.

Using evidence will avoid <u>overfitting</u>

Usual Procedure: Pick single best model M^* from L possible. Use that to make predictions.

Bayesian Procedure: Average over all models, weighted by posterior

$$p(t_* \mid \mathbf{t}) = \sum_{\ell=1}^{L} \underbrace{p(t_* \mid \mathbf{t}, M_\ell)}_{\text{predictive posterior for model } \ell} \underbrace{p(M_\ell \mid \mathbf{t})}_{\text{model posterior}}$$

Comparison of Methods for

Hyperparameter Selection

	Fixed valid. set (fraction f)	K-fold cross-validation	Bayesian evidence	
Fraction data used for training run	(1.0 – f)	(K-1) / K	**1.0**	**Higher is better** Better use of training data
Total runs/ examples seen for training	**1 run** **(1 – f) N**	K runs (K-1) * N	**1 run** **N**	**Lower is better** Faster training
Total runs/ examples seen for evaluation of fitness	1 run fN	K runs N	**1 run** **N**	**Lower is better** Faster evaluation
Fitness function	Heldout likelihood	Heldout likelihood	Evidence	

Hyperparameter estimation

Want to know good values for α, β given dataset.

$$\hat{\alpha}, \hat{\beta} = \underset{\substack{\alpha > 0 \\ \beta > 0}}{\arg\max} \log p(\mathbf{t} \mid \alpha, \beta)$$

This is "empirical Bayesian" point estimation, or Type 2 - maximum likelihood.

Can be solved via:

1) enumeration via grid search

2) gradient descent

3) coordinate descent (see E-M algorithm later in Unit 3)

4) analytical estimates (see PRML 3.92-3.95)

Guess α_0, β_0.

While not converged:

$$\lambda \leftarrow \text{EigenValues}(\beta_t \Phi^T \Phi) \qquad \alpha_{t+1} \leftarrow \frac{\gamma}{m_N^T m_N}$$

$$\gamma \leftarrow \sum_{m=1}^{M} \frac{\lambda_m}{\alpha_t + \lambda_m}$$

$$m_N \leftarrow \text{posteriorMean}(\alpha_t, \beta_t, \Phi, t] \qquad \beta_{t+1}^{-1} \leftarrow \frac{1}{N-\gamma} \sum_{n=1}^{N} (t_n - m_N^T \phi(x_n))^2$$