ISTA 421/521 Introduction to Machine Learning

Lecture 14: Marginal Likelihood Model Selection

Clay Morrison
clayton@sista.arizona.edu
Gould-Simpson 819
Phone 621-6609

9 October 2014

Back to Model Selection

- Recall in Chapter 1 we used Cross-Validation to estimate the generalization error of different orders of polynomial model, and selected the model order with the lowest loss.
- We found in Chapter 2 that Maximum Likelihood prefers complex models.
- In Chapter 3, we've used Marginal Likelihood to choose among different prior densities.
- We can also use Marginal Likelihood to choose models.

Marginal Likelihood for Model Selection

Marginal Likelihood for our Gaussian Model

$$p(\mathbf{t}|\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \int p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\, d\mathbf{w}$$
$$= \mathcal{N}(\mathbf{X}\boldsymbol{\mu}_0, \sigma^2\mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\mathsf{T})$$

Just as in the simulated experiment in Ch 1, generate data from a 3rd-order polynomial

Then compute the marginal likelihood for models from 1st to 7th order

For each model, use Gaussian prior on **w** with zero mean and an identity covariance matrix

For example:

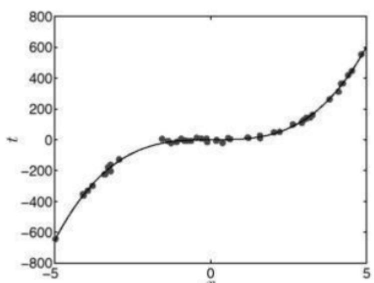
$$\boldsymbol{\mu}_0 = [0, 0]^\mathsf{T},\ \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

First-order model

$$\boldsymbol{\mu}_0 = [0, 0, 0, 0, 0]^\mathsf{T},\ \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

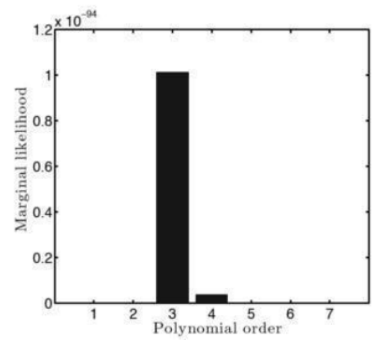
4th-order model

Results of Simulation



$$t = 5x^3 - x^2 + x$$

+ Gaussian noise: mean = 0, var = 150



Marginal likelihood for models 1st through 7th order

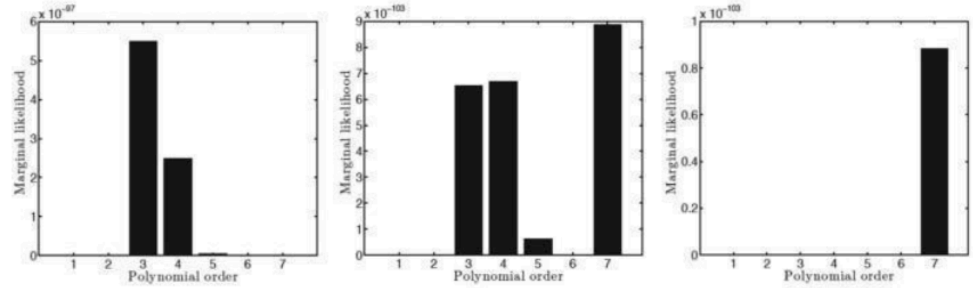
Plug in relevant prior and evaluate the density at **t**

Advantages:
- Very clear peak
- Don't have to compute CV over multiple datasets
- Get to use ***all*** of the data

Disadvantage:
- Calculating marginal likelihood is generally very hard

Results Depend on Priors

define $\mathbf{\Sigma}_0 = \sigma_0^2 \mathbf{I}$ and vary σ_0^2



(a) $\sigma_0^2 = 0.7$ (b) $\sigma_0^2 = 0.4$ (c) $\sigma_0^2 = 0.3$

By decreasing, we're saying parameters have to take smaller and smaller values

To fit our model well, one of the parameters needs to be 5: $t = 5x^3 - x^2 + x$

By decreasing σ_0^2, 5 becomes less likely and higher order models with lower parameter values become more likely.

When we talk about a **model**, we mean
the order of polynomial ***AND*** the prior specification

Quick note about (the variants of) "Empirical Bayes"

$$p(y_N|\alpha, \beta) = \int_{r=0}^{r=1} p(y_N|r)p(r|\alpha, \beta)\, dr$$

$$p(\mathbf{t}|\mathbf{X}, \boldsymbol{\mu}_0, \mathbf{\Sigma}_0) = \int p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\mu}_0, \mathbf{\Sigma}_0)\, d\mathbf{w}$$

If you do estimate your priors based on data, you should then use your model to model ***new*** data, otherwise you are overfitting.

Also, to be "truly" Bayesian about model selection, put a prior over possible models (e.g., for polynomial order, use a prior over the non-negative integers) and then integrate over your uncertainty in the order!

Sick of Linear Regression?