Model Selection

Ciira wa Maina

Dedan Kimathi University of Technology

29th June 2016

Introduction

- In machine learning applications we seek to fit models that explain our data
  In science, a central task is to develop and compare models to account for the data that are gathered[1] - *David J. MacKay*
- Often we limit our analysis to a family of models governed by some parameters
- Within this family we seek the "Best" model
- This model generalises well to unseen test data

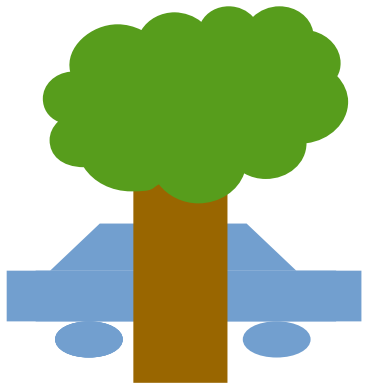
[1] MacKay, D. J. (1992). Bayesian interpolation. Neural computation, 4(3), 415-447.

Introduction: Parsimony

- Simpler explanations are to be prefered- Occam's Razor
- The scientific method:
  The existence of simple laws is, then, apparently, to be regarded as a quality of nature; and accordingly we may infer that it is justifiable to prefer a simple law to a more complex one that fits our observations slightly better.[2]

[2] D. Wrinch and H. Jeffreys. XLII. On certain fundamental principles of scientific inquiry. Philosophical Magazine Series 6, 42(249):369390, 1921.

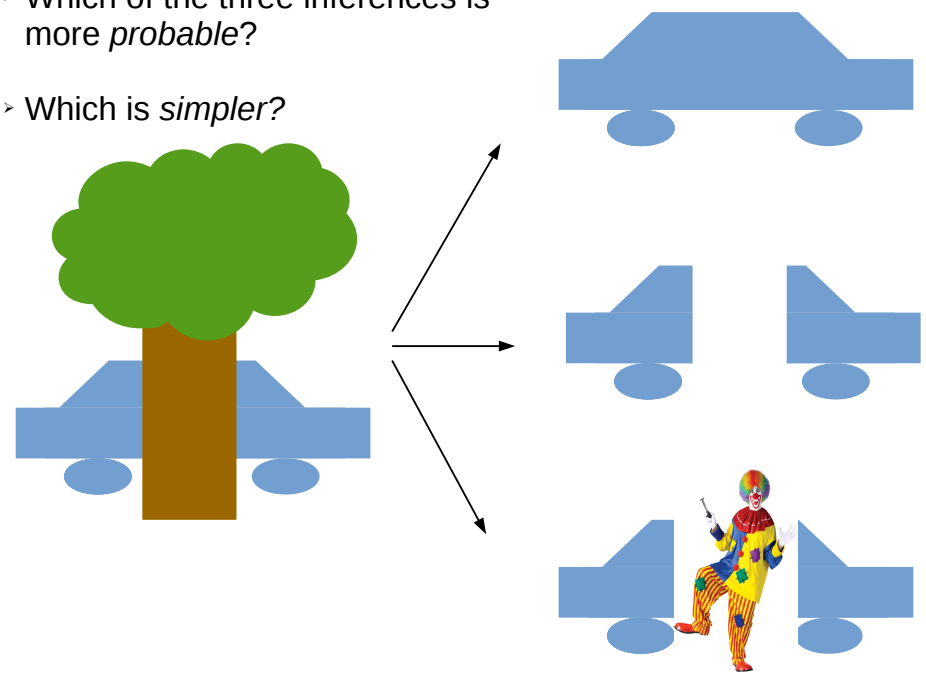
Introduction: Parsimony



[Image credit: Andreas Damianou]

Introduction: Parsimony

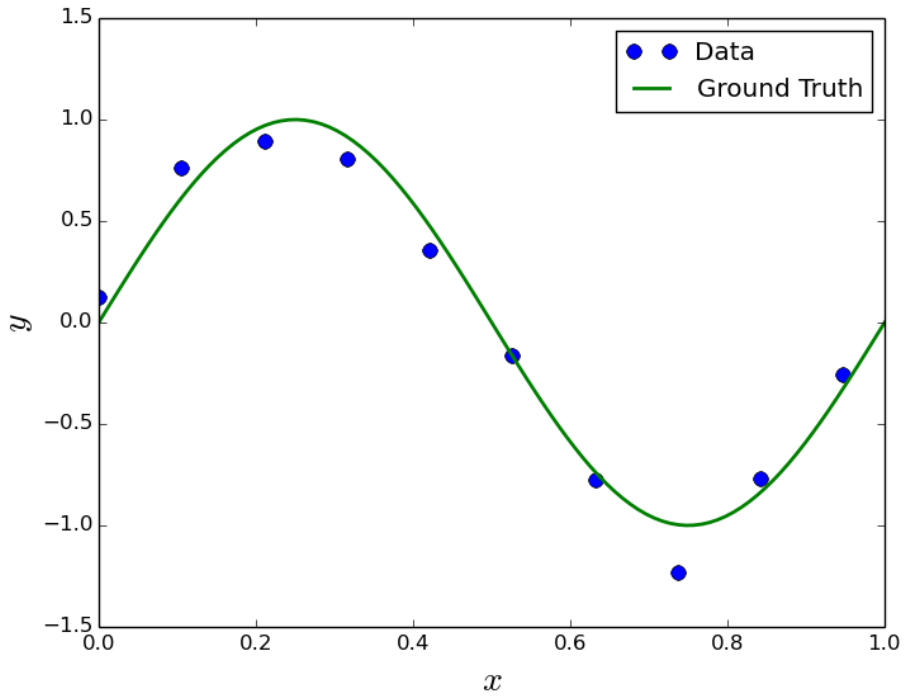
- Which of the three inferences is more *probable*?
- Which is *simpler?*



[Image credit: Andreas Damianou]

Introduction

- Consider the regression problem where we observe an input variable x and wish to predict a target variable y.
- Suppose the training data are as shown



Introduction

- Suppose we assume that we can use a polynomial to model the relationship between x and y

$$\begin{aligned} y &= w_0 + w_1 x + w_2 x^2 + \ldots + w_P x^P \\ &= \sum_{i=0}^{P} w_i x^i \end{aligned}$$

- The model has a set of parameters $\mathbf{w} = [w_0, \ldots, w_P]^T$ and we can write

$$y = f(x, \mathbf{w})$$

- Given a polynomial order P, we learn the parameters $\mathbf{w}^*$ that best explain the training data

Introduction

- By finding the $\mathbf{w}^*$ that minimizes the square error

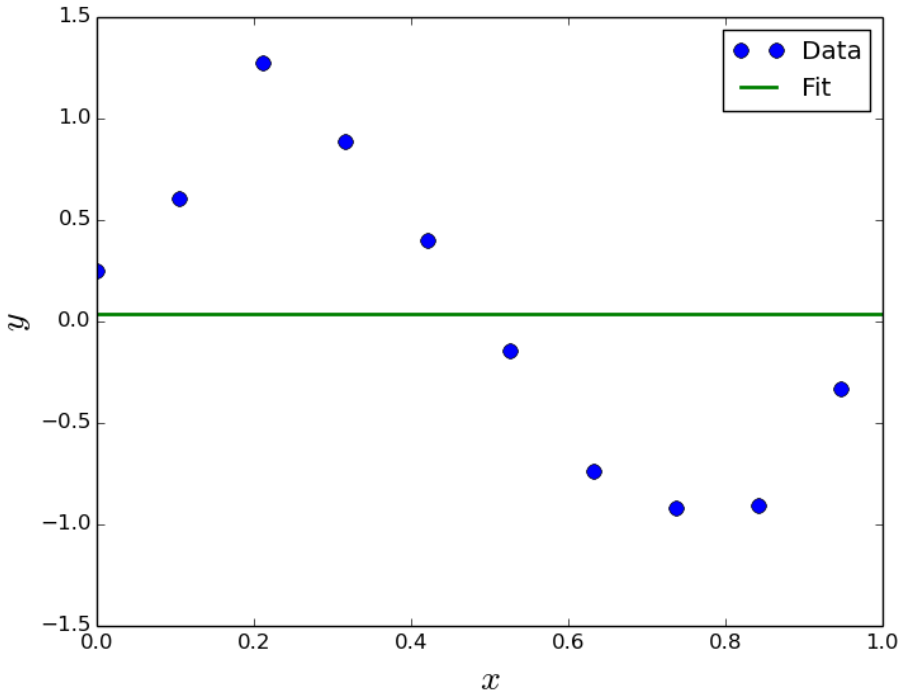
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(f(x_n, \mathbf{w}) - y_n)^2$$

  we obtain the polynomial fit of the data $f(x, \mathbf{w}^*)$
- $E(\mathbf{w})$ is quadratic in $\mathbf{w}$ and has a unique minimum

Introduction

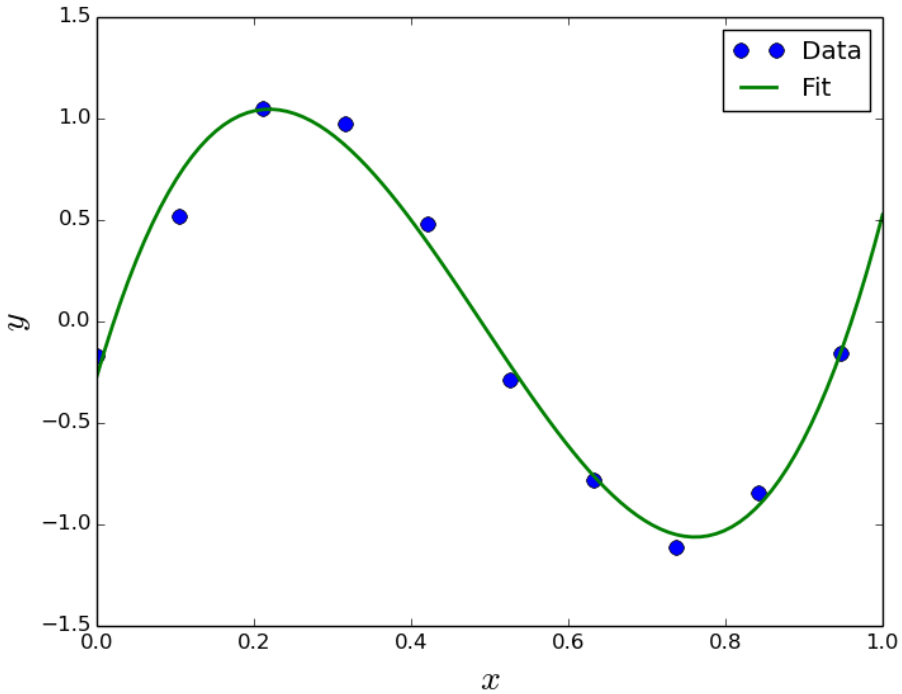
- But what about the order P
- This parameter governs the complexity of the model
- High values of P are more flexible but harder to fit and may suffer numerical instability especially when data are limited

Introduction



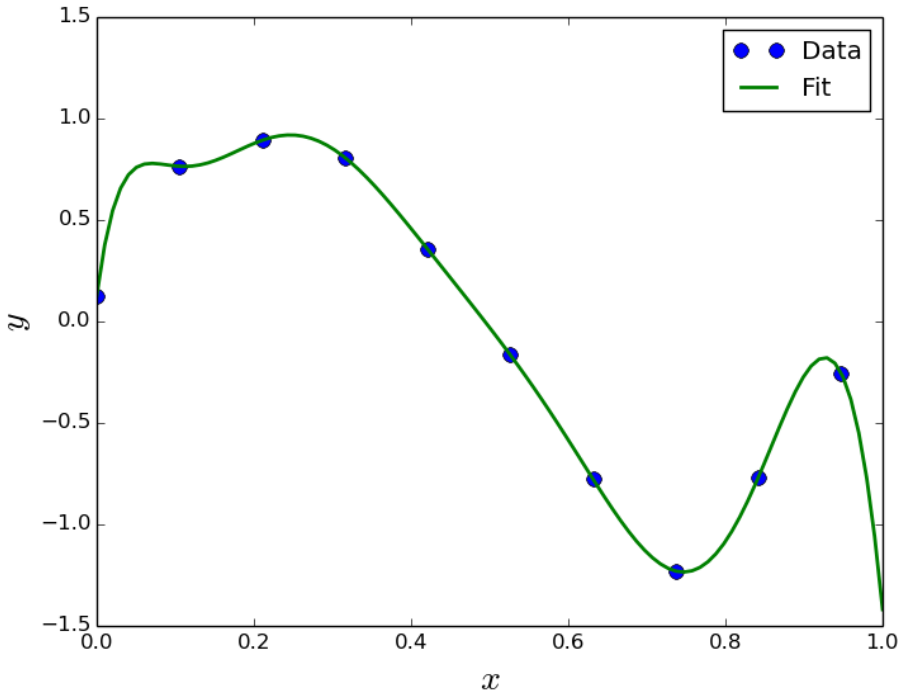
Fit for $P = 0$

Introduction



Fit for $P = 3$

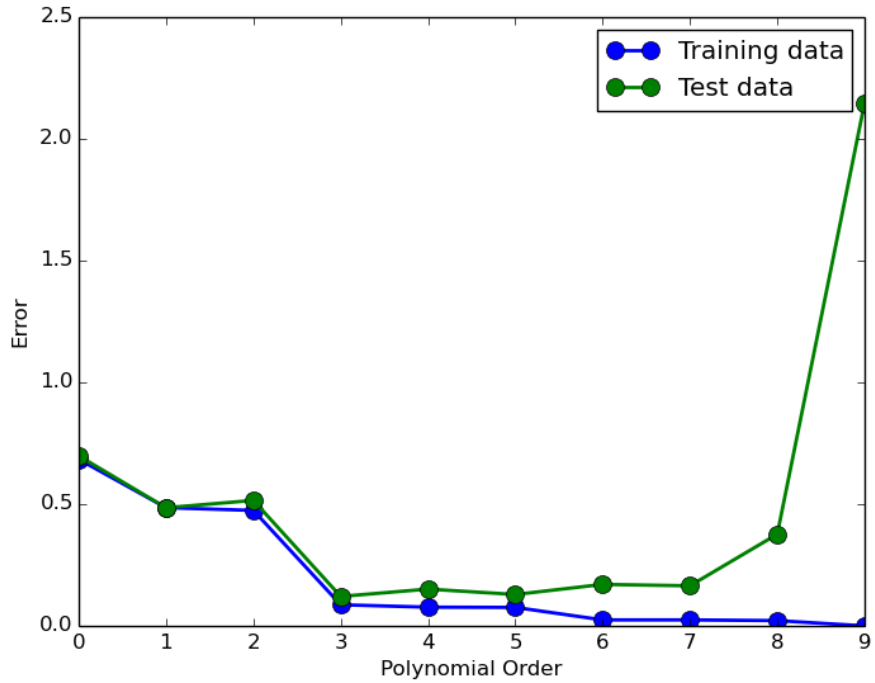
Introduction



Fit for $P = 9$

Model Assessment and Selection

- To determine P, we can monitor the error on a hold-out test set



Model Assessment and Selection

- If data are plentiful, we can use a data-driven approach
- Divide the data into three parts: training, validation and testing sets
- The training set is used to fit the model
- The validation set is used for model selection
- The test set is used to estimate the performance on unseen data

K-fold Cross-Validation

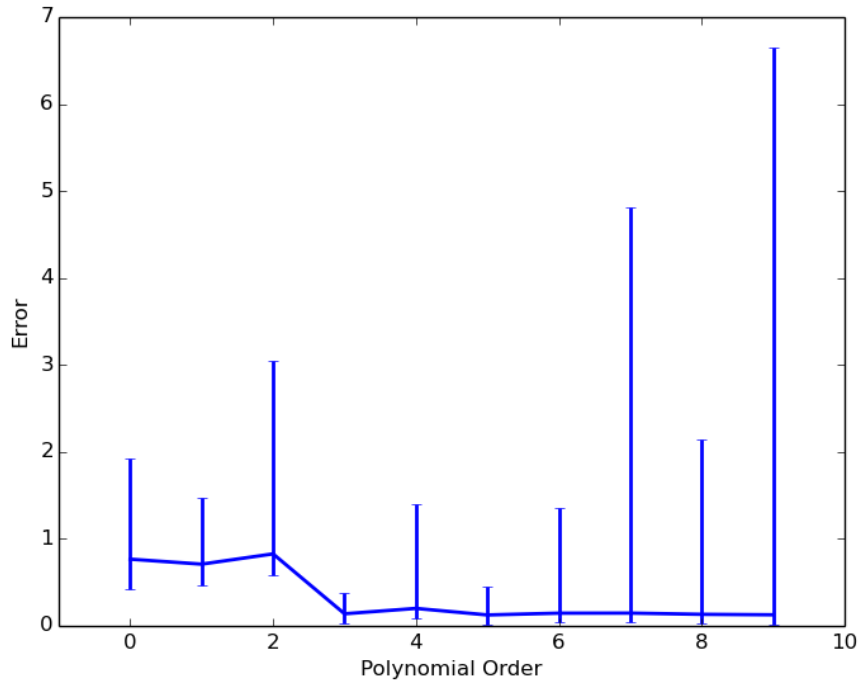
- An alternative approach is to divide the data into K (almost) equal sized parts
- Use $K - 1$ parts to fit the model and test the model on the remaining part
- If K is set to the number of data points, we have Leave-one-out cross-validation
- One drawback is the increased computational burden

K-fold Cross-Validation: Polynomial Regression

- For the polynomial regression problem, we follow these steps
  1. Divide the data into K parts
  2. For of the K folds $k = 1, 2, \ldots, K$
     - 2.1 Use the remaining $K - 1$ parts to fit polynomials of different order
     - 2.2 Determine the sum-of-squares error using the k part

K-fold Cross-Validation: Polynomial Regression

- The CV error as a function of polynomial order.
- The plot show 95% confidence intervals
- The uncertainty around $P = 3$ is least



Probabilistic Approach

- There are two main steps in data modelling[3]
  1. Assume one of our models is correct and fit model to data, repeat this for all models
  2. Compare the models
- To place the model comparison problem in a probabilistic setting, under each model $\mathcal{M}_i$, the parameters $\mathbf{w}$ are associated with a prior probability

$$p(\mathbf{w}|\mathcal{M}_i)$$

- For a given model and parameter setting, the probability of a given dataset D is $p(D|\mathbf{w}, \mathcal{M}_i)$

[3] MacKay, D. J. (1992). Bayesian interpolation. Neural computation, 4(3), 415-447.

Probabilistic Approach

- The first step of inference involves estimating the model parameters $\mathbf{w}$. Using Bayes' rule, the posterior is given by

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

- This posterior distribution allows us to perform inference about the parameters.
- We can compute the Most probable value of $\mathbf{w}$

$$\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}} p(\mathbf{w}|D, \mathcal{M}_i)$$

- Error bars can be determined from the curvature of the posterior at this maximum
- In this step the denominator $p(D|\mathcal{M}_i)$ plays no role and is ignored.

Probabilistic Approach: Model comparison

- The posterior probability of a given model is given by

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(D|\mathcal{M}_i)p(\mathcal{M}_i)$$

- Two models can be compared by computing the ratio

$$\frac{p(\mathcal{M}_i|D)}{p(\mathcal{M}_j|D)} = \frac{p(D|\mathcal{M}_i)p(\mathcal{M}_i)}{p(D|\mathcal{M}_j)p(\mathcal{M}_j)}$$

Probabilistic Approach: Bayes factor

- If we assume that all models are equiprobable, we use the **model evidence** $p(D|\mathcal{M}_i)$ to rank the models.
- To compare two models i and j we compute the ratio

$$B_{ij} = \frac{p(D|\mathcal{M}_i)}{p(D|\mathcal{M}_j)}$$

- If the ratio is greater than one, model i is preferred over model j
- B_{ij} is known as the Bayes factor.

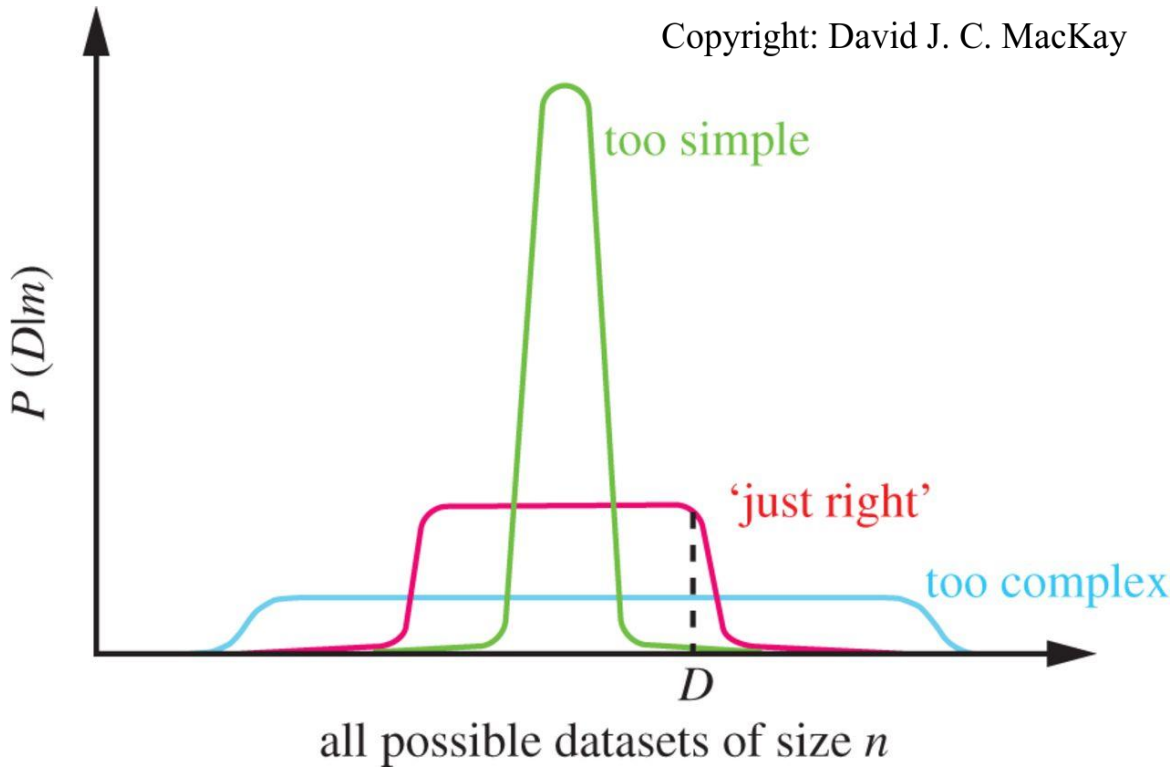
Evaluating Model Evidence

- To perform model comparison in a Bayesian setting, we must evaluate the model evidence.
- We must compute the integral

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

- In some cases, this can be performed analytically.
- Otherwise we can use sampling methods.

Evaluating Model Evidence



[Image credit: Andreas Damianou]

Model Comparison in Polynomial Regression

- To illustrate model comparison we return to the polynomial regression problem
- To proceed we place the problem in a probabilistic setting.
- We have a data set $D = \{x_i, y_i\}_{i=1}^{N}$ where x_i are the input variables and y_i are the target variable.
- We model the target variables as

$$y_i = f(x_i, \mathbf{w}) + \epsilon$$

- ϵ is zero mean Gaussian noise with variance σ_ϵ^2
- $\mathbf{w}$ is a vector of model parameters. The polynomial coefficients.

Model Comparison in Polynomial Regression

- The probability of the data D given the parameters for a given model $\mathcal{M}_i$ is

$$p(D|\mathbf{w}, \beta, \mathcal{M}_i) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left(-\frac{\beta}{2}\sum_{i=1}^{N}(f(x_i, \mathbf{w}) - y_i)^2\right)$$

  where $\beta = \frac{1}{\sigma_\epsilon^2}$ is the noise precision
- Here we make the dependance on the model explicit.
- Different models correspond to different polynomial order.

Model Comparison in Polynomial Regression

- $p(D|\mathbf{w}, \beta, \mathcal{M}_i)$ is known as the likelihood and we can obtain a maximum likelihood estimate of the parameters. In practice we maximize the log likelihood.
- We have

$$\log p(D|\mathbf{w}, \beta, \mathcal{M}_i) = \frac{N}{2}\log(\beta) - \frac{N}{2}\log(2\pi) - \frac{\beta}{2}\sum_{i=1}^{N}(f(x_i, \mathbf{w}) - y_i)^2$$

- Maximizing the log likelihood is equivalent to minimising the sum-of-squares error.

Bayesian Polynomial Regression

- For a Bayesian treatment of the regression problem, we set a prior over the parameters $\mathbf{w}$
- This prior governs the types of interpolants we will obtain.
- If the magnitudes of the polynomial coefficients are restricted to small values, the model is inflexible and results in flat interpolants
- If the coefficients are allowed to be too large, then the model can be too flexible and oscillate wildly to pass all data points.
- We seek a middle ground

Bayesian Polynomial Regression

- We select the following prior

$$p(\mathbf{w}|\alpha, \mathcal{M}_i) = \left(\frac{\alpha}{2\pi}\right)^{(P+1)/2} \exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right)$$

  where α is the precision of the coefficients. P is the polynomial order.
- When α is small, coefficients can take large values
- When α is large, coefficients are assumed to take small values

Bayesian Polynomial Regression - Inference

- The inference step seeks most probable value of $\mathbf{w}$
- The posterior distribution of $\mathbf{w}$ is

$$p(\mathbf{w}|D, \alpha, \beta, \mathcal{M}_i) = \frac{p(D|\mathbf{w}, \beta, \mathcal{M}_i)p(\mathbf{w}|\alpha, \mathcal{M}_i)}{p(D|\alpha, \beta, \mathcal{M}_i)}$$

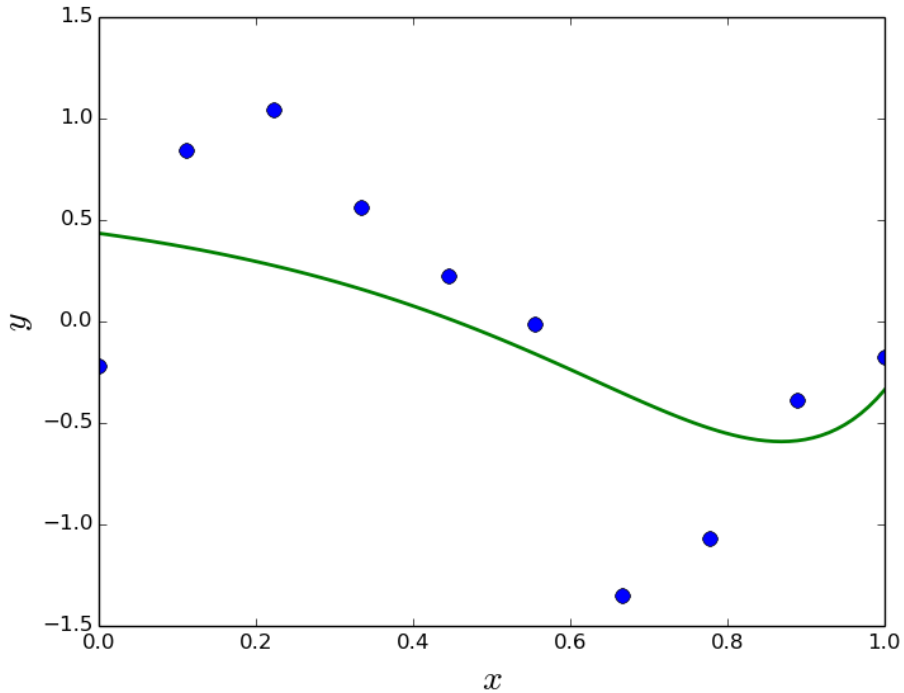
- We can show that this is a Gaussian with

$$\begin{aligned} \mu &= \beta \Sigma \Phi \mathbf{y} \\ \Sigma &= [\alpha \mathbf{I} + \beta \Phi^T \Phi]^{-1} \end{aligned}$$

Where

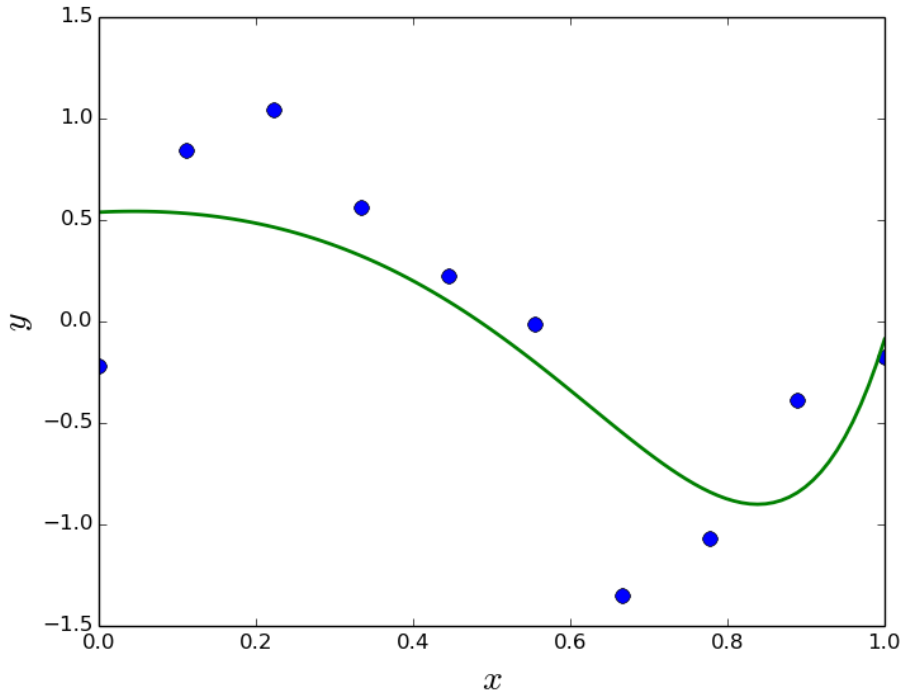
$$\Phi = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^P \\ 1 & x_2 & x_2^2 & \dots & x_2^P \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^P \end{bmatrix}$$

Inferred Polynomial



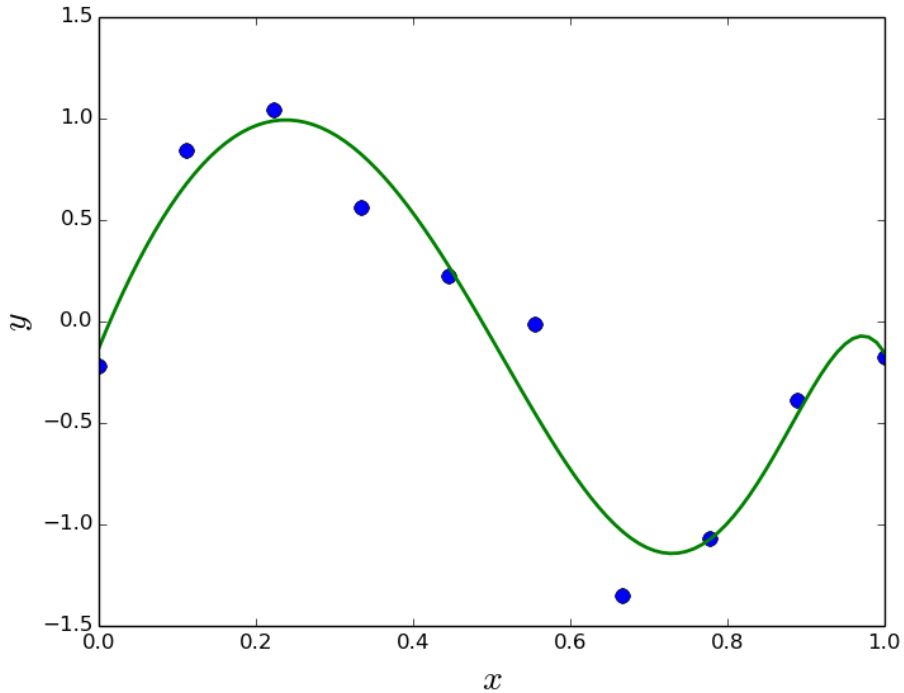
Fit for $P = 9$, $\alpha = 10$

Inferred Polynomial



Fit for $P = 9$, $\alpha = 1$

Inferred Polynomial



Fit for $P = 9$, $\alpha = 0.001$

Bayesian Polynomial Regression- Evaluation of Model Evidence

- The denominator in the expression for the posterior is the model evidence and is evaluated by integrating over $\mathbf{w}$

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

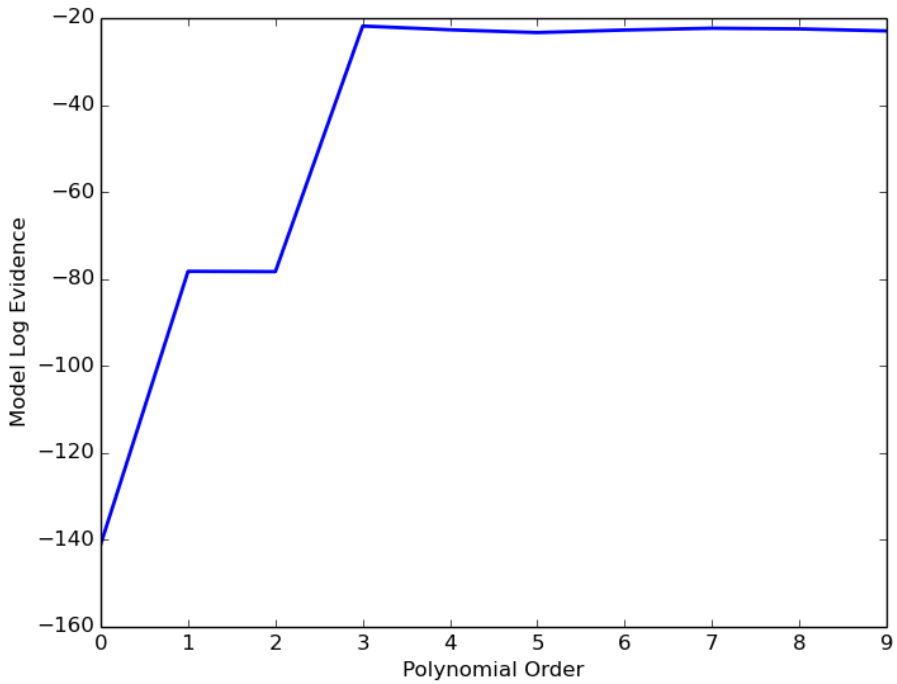
Evaluation of Model Evidence

- The integral can be evaluated in closed form since both terms in the integrand are quadratic in $\mathbf{w}$
- Completing the square and using the Gaussian normalizing coefficient yeilds the evidence[4]
- We can evaluate it for several values of polynomial order.
- This Bayesian approach uses all the data without the need to have a separate train and test set

[4] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Evaluation of Model Evidence

- We see a preference for the polynomial of order 3.



The Bayesian Information Criterion

- We have seen that the Bayes factor comparing two models i and j is given by

$$B_{ij} = \frac{p(D|\mathcal{M}_i)}{p(D|\mathcal{M}_j)}$$

- The log marginal likelihood $p(D|\mathcal{M}_i)$ can be approximated using the Laplace approximation to yeild[5]

$$\log p(D|\mathcal{M}_i) \approx \log p(D|\hat{\theta}_{ML}, \mathcal{M}_i) - \frac{d_i}{2} \log(N)$$

where $\hat{\theta}_{ML}$, is the maximum likelihood parameter estimate and d_i is the number of free parameters in model $\mathcal{M}_i$

The Bayesian Information Criterion

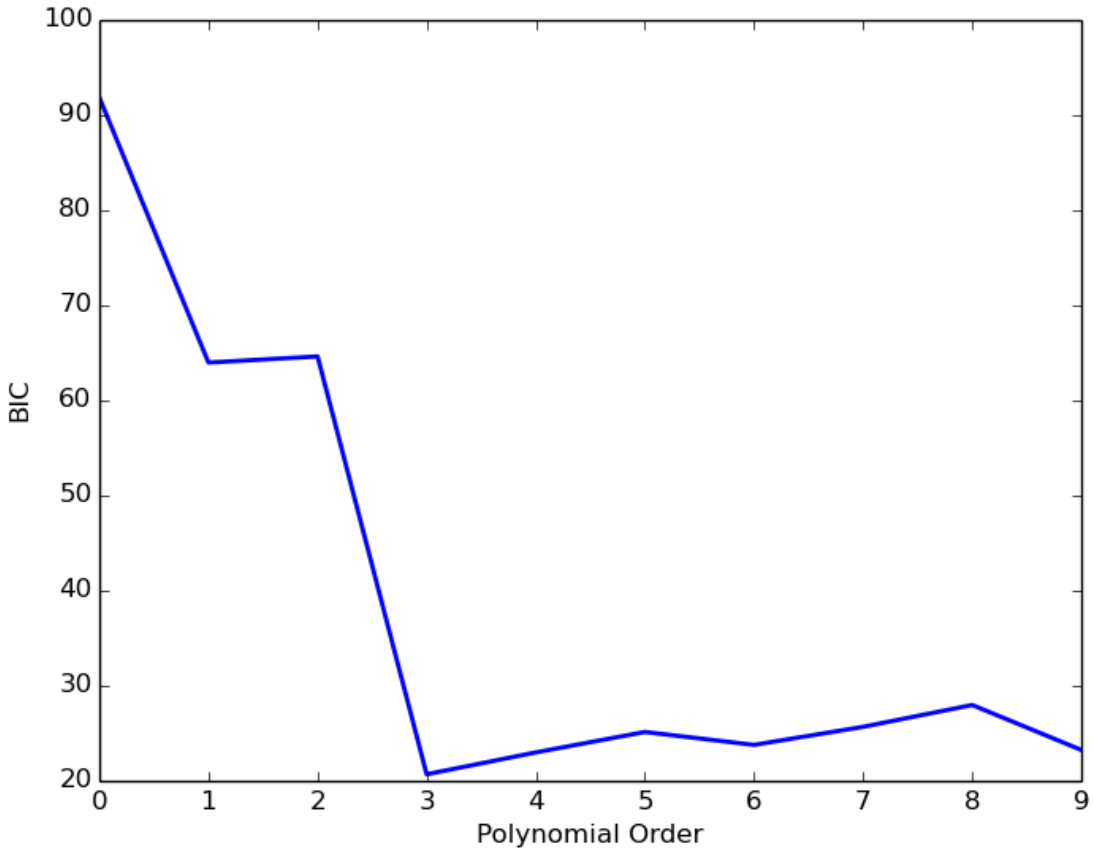
- The Bayesian Information Criterion for a model $\mathcal{M}_i$ is defined as

$$BIC_i = -2 \log p(D|\hat{\theta}_{ML}, \mathcal{M}_i) + d_i \log(N)$$

- The BIC statistic penalizes complex models
- It includes a penalty term that depends on the number of free parameters in a model
- We chose the model with the minimum BIC
- This is equivalent to chosing the model with the largest posterior probability

The Bayesian Information Criterion

- For the polynomial regression problem, the minimum BIC corresponds to $P = 3$



Conclusion

- We have seen a number of approaches to model selection
- When we have sufficient data we can hold out some test data or use cross-validation
- Bayesian approaches provide a principled approach to model selection but can be computationally intensive
- The Bayesian Information Criterion is a useful approximation to the model evidence

References

- Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning (Vol. 1)*. Springer, Berlin: Springer series in statistics.

Thank You