Bayesian Learning and Spatio-temporal Statistics

Part I

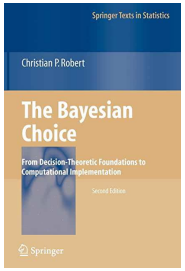
François Septier

http://www.univ-ubs.fr/septier/

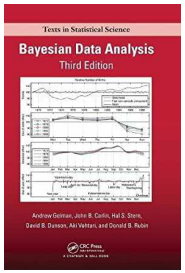
francois.septier@univ-ubs.fr

Outline

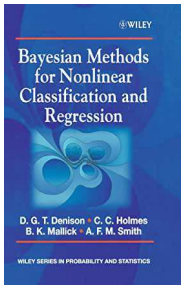
Some reference books



Robert, C. P. (2007). *The bayesian choice*.
Springer, second edition edition

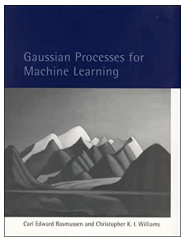


Gelman, A., Carlin, J., Stern, H., Dunson, D. B., Vehtari, A., and Rubin, D. (2003). *Bayesian Data Analysis*.
Chapman and Hall/CRC, third edition

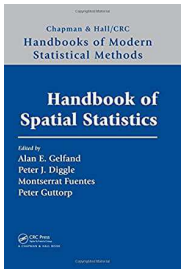


Denison, D. G. T., Holmes, C. C., and Mallick, B. K. (2002). *Bayesian Methods for Nonlinear Classification and Regression*.
Wiley Series in Probability and Statistics

Some reference books

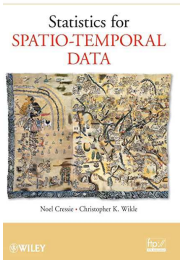


Rasmussen, C. and Williams, C. (2006). *Gaussian processes for machine learning*.
The MIT Press



Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P., editors

(2010). *Handbook Of Spatial Statistics*.
Chapman & Hall/CRC



Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*.
Wiley

Outline

An information revolution

- We are in an era of abundant data:
  - **Society**: the web, social networks, mobile networks, government, digital archives
  - **Science**: large-scale scientific experiments, biomedical data, climate data, scientific literature
  - **Business**: e-commerce, electronic trading, advertising, personalization
- We need tools for modeling, searching, visualizing, and understanding large data sets.

Learning: the view from different fields

- **Engineering**: signal processing, system identification, adaptive and optimal control, information theory, robotics,...
- **Computer Science**: Artificial Intelligence, computer vision, information retrieval, natural language processing, data mining,...
- **Statistics**: estimation, learning theory, data science, inference from data,...
- **Cognitive Science and Psychology**: perception, movement control, reinforcement learning, mathematical psychology, computational linguistics,...
- **Computational Neuroscience**: neuronal networks, neural information processing, ...
- **Economics**: decision theory, game theory, operational research, e-commerce, choice modeling,...

Different fields, Convergent ideas

- The **same set of ideas and mathematical tools** have emerged in many of these fields, albeit with different emphases.
- **Machine learning** *is an interdisciplinary field focusing on both the mathematical foundations and practical applications of systems that learn, reason and act.*

Modeling vs toolbox views of Machine Learning

- **Machine Learning is a toolbox of methods for processing data**: feed the data into one of many possible methods; choose methods that have good theoretical or empirical performance; make predictions and decisions
- **Machine Learning is the science of learning models from data**: define a space of possible models; learn the parameters and structure of the models from data; make predictions and decisions

Modeling tools

Our modeling tools should:

- Faithfully represent **uncertainty** in our model structure and parameters and **noise** in our data
- Be automated and **adaptive**
- Exhibit **robustness**
- **Scale well** to large data sets

Probabilistic Modeling

Our modeling tools should:

- A model describes data that one could observe from a system
- If we use the mathematics of probability theory to express all forms of uncertainty and noise associated with our model...
- ...then *inverse probability* (i.e. Bayes rule) allows us to infer unknown quantities, adapt our models, make predictions and learn from data.

Probabilistic Modelling

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

Thomas Bayes

(1702-1761)

- Bayes rule tells us how to do inference about hypotheses from data.
- Learning and prediction can be seen as forms of inference.

Outline

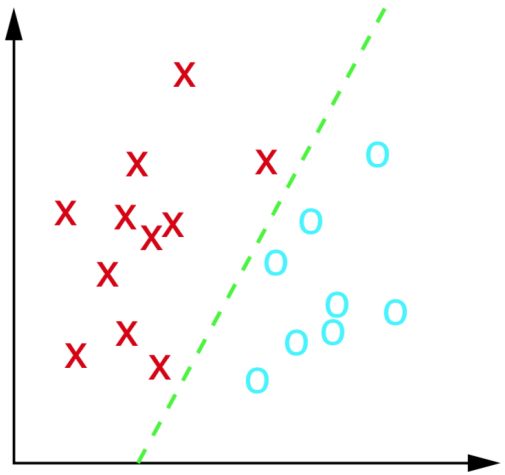
Some canonical machine learning problems

- Linear **classification**
- Polynomial **Regression**
- **Clustering** with Gaussian Mixtures (Density Estimation)

Linear Classification

Data: N data points $\mathcal{D} = \left\{\boldsymbol{x}^{(i)}, y^{(i)}\right\}_{i=1}^{N}$ with

$$\boldsymbol{x}^{(i)} \in \mathbb{R}^D$$
$$y^{(i)} \in \{+1; -1\}$$



Model:

$$P(y^{(i)} = +1 | \boldsymbol{w}, \boldsymbol{x}^{(i)}) = \begin{cases} 1 & \text{if } \sum_{d=1}^{D} w_d x_d^{(i)} + w_0 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

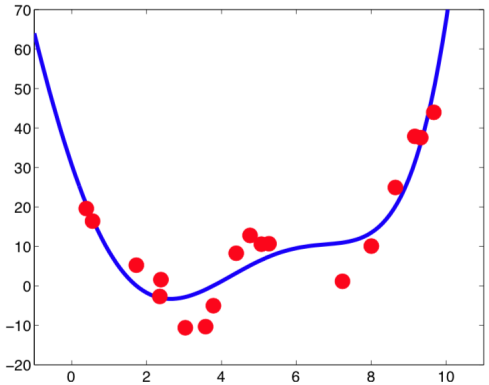
Parameters: $\boldsymbol{w} \in \mathbb{R}^{D+1}$

Goal: Infer $\boldsymbol{w}$ from the data $\mathcal{D}$ to predict the future labels, i.e. $P(y^* | \mathcal{D}, \boldsymbol{x}^*)$

Polynomial Regression

Data: N data points $\mathcal{D} = \left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{N}$ with

$$x^{(i)} \in \mathbb{R}$$
$$y^{(i)} \in \mathbb{R}$$



Model: (of order m)

$$y^{(i)} = w_0 + w_1 x^{(i)} + w_2 (x^{(i)})^2 + \cdots + w_m (x^{(i)})^m + \varepsilon$$

where

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Parameters: $\boldsymbol{\theta} = (w_0, w_1, \ldots, w_m, \sigma) \in \mathbb{R}^{(m+1)} \times \mathbb{R}^+$

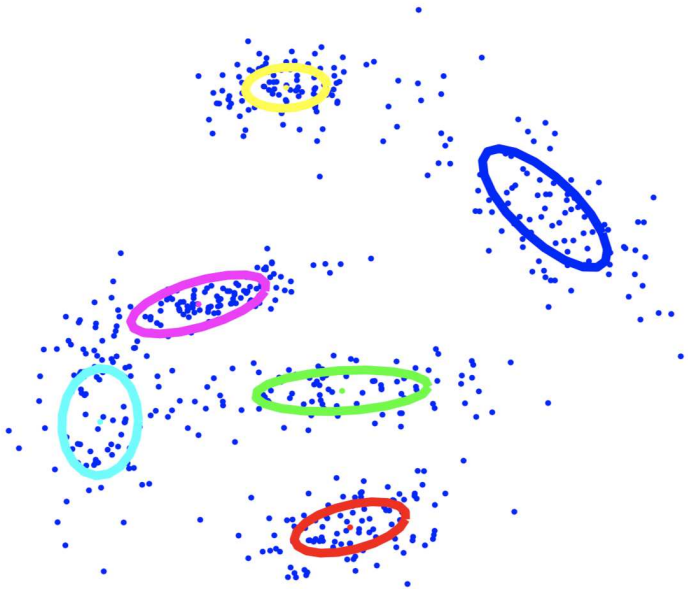
Goal: Infer $\boldsymbol{\theta}$ from the data $\mathcal{D}$ to predict the future outputs, i.e. $P(y^*|\mathcal{D}, x^*, m)$

Clustering with Gaussian Mixtures

Data: N data points $\mathcal{D} = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{N}$ with

$$\boldsymbol{x}^{(i)} \in \mathbb{R}^D$$

Model: (with m components)



$$\boldsymbol{x}^{(i)} \sim \sum_{k=1}^{m} \pi_k p_k(\boldsymbol{x}^{(i)})$$

where

$$p_k(\boldsymbol{x} = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Parameters: $\boldsymbol{\theta} = ((\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, \ldots, (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \pi_1, \ldots, \pi_m)$

Goal: Infer $\boldsymbol{\theta}$ from the data $\mathcal{D}$ to predict the density $p(\boldsymbol{x}^*|\mathcal{D}, m)$ and to infer which points belong to the same cluster.

The Bayesian Approach to Machine Learning (Or Anything)

1. We formulate our knowledge about the situation probabilistically:
   - We define a **model** that expresses qualitative aspects of our knowledge (eg, forms of distributions, independence assumptions). The model will have some unknown **parameters**.
   - We specify a **prior** probability distribution for these unknown parameters that expresses our beliefs about which values are more or less likely, before seeing the data.
2. We gather data.
3. We compute the **posterior** probability distribution for the parameters, given the observed data.
4. We use this posterior distribution to:
   - Reach scientific conclusions, properly accounting for uncertainty.
   - Make predictions by averaging over the posterior distribution.
   - Make decisions so as to minimize posterior expected loss.

Finding the Posterior Distribution

The **posterior distribution** for the model parameters given the observed data is found by combining the prior distribution with the likelihood for the parameters given the data.
This is done using **Bayes' Rule**:

$$P(\text{parameters}|\text{data}) = \frac{P(\text{parameters})P(\text{data}|\text{parameters})}{P(\text{data})}$$

The denominator is just the required normalizing constant, and can often be filled in at the end, if necessary. So as a proportionality, we can write

$$P(\text{parameters}|\text{data}) \propto P(\text{parameters})P(\text{data}|\text{parameters})$$

which can be written schematically as

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

We make predictions by integrating with respect to the posterior:

$$P(\text{new data}|\text{data}) = \int_{\text{parameters}} P(\text{new data}|\text{parameters})P(\text{parameters}|\text{data})$$

Inference at a Higher Level: Comparing Models

So far, we've assumed we were able to start by making a definite choice of model. What if we're unsure which model is right?

We can compare models based on the **marginal likelihood** (aka, the evidence) for each model, which is the probability the model assigns to the observed data. This is the normalizing constant in Bayes' Rule that we previously ignored:

$$P(\text{data}|\mathcal{M}_1) = \int_{\text{parameters}} P(\text{data}|\text{parameters}, \mathcal{M}_1) P(\text{parameters}|\mathcal{M}_1)$$

Here, $\mathcal{M}_1$ represents the condition that model $\mathcal{M}_1$ is the correct one (which previously we silently assumed). Similarly, we can compute $P(\text{data}|\mathcal{M}_1)$, for some other model (which may have a different parameter space).

We might choose the model that gives higher probability to the data, or average predictions from both models with weights based on their marginal likelihood, multiplied by any prior preference we have for $\mathcal{M}_1$ versus $\mathcal{M}_2$.

Bayesian Modeling

Everything follows from two simple rules:

Sum Rule: $P(x) = \sum_y P(x, y)$

Product Rule: $P(x, y) = P(x|y)p(y) = p(y|x)p(x)$

$$P(\theta|\mathcal{D}, \mathcal{M}) = \frac{P(\mathcal{D}|\theta, \mathcal{M})P(\theta|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})}$$

$P(\mathcal{D}\|\theta, \mathcal{M})$	likelihood of parameters θ in model $\mathcal{M}$
$P(\theta\|\mathcal{M})$	prior probability of parameters θ in $\mathcal{M}$
$P(\theta\|\mathcal{D}, \mathcal{M})$	posterior of θ given data $\mathcal{D}$ in $\mathcal{M}$

Prediction:

$$P(x^*|\mathcal{D}, \mathcal{M}) = \int P(x^*|\theta, \mathcal{D}, \mathcal{M})P(\theta|\mathcal{D}, \mathcal{M})d\theta$$

Model Comparison:

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})}$$

$$P(\mathcal{D}|\mathcal{M}) = \int P(\mathcal{D}|\theta, \mathcal{M})P(\theta|\mathcal{M})d\theta$$

Example: Bayesian Regression

- **Objective:** Determine the relationship between some response variable $\boldsymbol{Y}$ and a set of D predictor variables $\boldsymbol{X} = (x_1, \ldots, x_D)$
- Most common assumption $\rightsquigarrow$ Relation through some deterministic function f and some additive (zero-mean) random error component ε

$$\boldsymbol{Y} = f(\boldsymbol{X}) + \boldsymbol{\varepsilon}$$

- In most situations the predictor variables, $\boldsymbol{X}$ are assumed to be observed without error so they are not considered as random

$$\Rightarrow f(\boldsymbol{x}) = \mathbb{E}\left[\boldsymbol{Y} | \boldsymbol{X} = \boldsymbol{x}\right]$$

- Our aim is to determine $f(\cdot)$ so that we can uncover the true relationship between the response $\boldsymbol{y}^*$ at predictor location $\boldsymbol{x}^*$ given by $\boldsymbol{y}^* = f(\boldsymbol{x}^*)$ *(generally only interested in estimating f over some range of plausible predictor values)*
- The true regression function is unknown and we have no way of determining its analytic form exactly, even if one actually exists $\Rightarrow$ we must content ourselves with finding approx. to it which are close to the truth by making use of the observed dataset $\mathcal{D} = \left\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right\}_{i=1}^{N}$

Example: Bayesian Regression

- Without prior knowledge as to the exact form of f we need to approximate it
  - One solution is to make direct linear assumptions, i.e.:

$$f(\boldsymbol{x}) \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D$$

  - more flexible models make use of **basis functions**:

$$f(\boldsymbol{x}) \approx \sum_{i=1}^{k} \beta_i \boldsymbol{B}_i(\boldsymbol{x})$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ is the set of coefficients corresponding to basis functions $\boldsymbol{B} = (\boldsymbol{B}_1, \ldots, \boldsymbol{B}_k)$.
Basis functions are nonlinear transformations of the data vector $\boldsymbol{x}$ - some examples include:

- polynomial functions
- radial basis functions
- etc

Remark: Linear model is just a special just case where $k = D + 1$, $\boldsymbol{B}_1(\boldsymbol{x}) = 1$ and $\boldsymbol{B}_i(\boldsymbol{x}) = x_{i-1}$ for $i = 2, \ldots, D + 1$.

Example: Bayesian Regression

Lets assume that we want to model our N data points $\mathcal{D} = \left\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right\}_{i=1}^{N}$ as:

$$\boldsymbol{y}^{(j)} = \sum_{i=1}^{k} \beta_i \boldsymbol{B}_i(\boldsymbol{x}^{(j)}) + \boldsymbol{\varepsilon}^{(j)} \qquad \text{for } j = 1, \ldots, N$$

where $\boldsymbol{\varepsilon}^{(j)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and for some specific choice of basis functions $\Rightarrow$ this is our choice of model $\mathcal{M}$.

The mechanisms of the Bayesian approach:

1. Assign priors to all the unknown parameters $\boldsymbol{\theta} = (\sigma^2, \beta_1, \ldots, \beta_k) \Rightarrow p(\boldsymbol{\theta}|\mathcal{M})$;

Example: Bayesian Regression

Lets assume that we want to model our N data points $\mathcal{D} = \left\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right\}_{i=1}^{N}$ as:

$$\boldsymbol{y}^{(j)} = \sum_{i=1}^{k} \beta_i \boldsymbol{B}_i(\boldsymbol{x}^{(j)}) + \boldsymbol{\varepsilon}^{(j)} \qquad \text{for } j = 1, \ldots, N$$

where $\boldsymbol{\varepsilon}^{(j)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and for some specific choice of basis functions $\Rightarrow$ this is our choice of model $\mathcal{M}$.

The mechanisms of the Bayesian approach:

1. Assign priors to all the unknown parameters $\boldsymbol{\theta} = (\sigma^2, \beta_1, \ldots, \beta_k) \Rightarrow p(\boldsymbol{\theta}|\mathcal{M})$;
2. write down the likelihood of the data given the parameters $p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})$;

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) &= p(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)}|\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}, \boldsymbol{\theta}, \mathcal{M}) \\ &= \prod_{j=1}^{N} \mathcal{N}(\boldsymbol{y}^{(j)}; \sum_{i=1}^{k} \beta_i \boldsymbol{B}_i(\boldsymbol{x}^{(j)}), \sigma^2 \boldsymbol{I}) \end{aligned}$$

Example: Bayesian Regression

Lets assume that we want to model our N data points $\mathcal{D} = \left\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right\}_{i=1}^{N}$ as:

$$\boldsymbol{y}^{(j)} = \sum_{i=1}^{k} \beta_i \boldsymbol{B}_i(\boldsymbol{x}^{(j)}) + \boldsymbol{\varepsilon}^{(j)} \qquad \text{for } j = 1, \ldots, N$$

where $\boldsymbol{\varepsilon}^{(j)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and for some specific choice of basis functions $\Rightarrow$ this is our choice of model $\mathcal{M}$.

The mechanisms of the Bayesian approach:

1. Assign priors to all the unknown parameters $\boldsymbol{\theta} = (\sigma^2, \beta_1, \ldots, \beta_k) \Rightarrow p(\boldsymbol{\theta}|\mathcal{M})$;
2. write down the likelihood of the data given the parameters $p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})$;

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) &= p(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)}|\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}, \boldsymbol{\theta}, \mathcal{M}) \\ &= \prod_{j=1}^{N} \mathcal{N}(\boldsymbol{y}^{(j)}; \sum_{i=1}^{k} \beta_i \boldsymbol{B}_i(\boldsymbol{x}^{(j)}), \sigma^2 \boldsymbol{I}) \end{aligned}$$

3. determine the posterior distribution of the parameters given the data using Bayes' theorem.

$$p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}$$

Example: Bayesian Regression

Lets assume that we want to model our N data points $\mathcal{D} = \left\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right\}_{i=1}^{N}$ as:

$$\boldsymbol{y}^{(j)} = \sum_{i=1}^{k} \beta_i \boldsymbol{B}_i(\boldsymbol{x}^{(j)}) + \boldsymbol{\varepsilon}^{(j)} \qquad \text{for } j = 1, \ldots, N$$

where $\boldsymbol{\varepsilon}^{(j)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and for some specific choice of basis functions $\Rightarrow$ this is our choice of model $\mathcal{M}$.
Once this posterior distribution of the parameters given the data using Bayes' theorem obtained

$$P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) = \frac{P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})P(\boldsymbol{\theta}|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})}$$

we can:

- predict the response $\boldsymbol{y}^*$ at new predictor $\boldsymbol{x}^*$ and more importantly be able to characterize its full distribution (with uncertainty) as:

$$p(\boldsymbol{y}^*|\boldsymbol{x}^*, \mathcal{D}, \mathcal{M}) = \int p(\boldsymbol{y}^*|\boldsymbol{x}^*, \boldsymbol{\theta}, \mathcal{D}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) d\boldsymbol{\theta}$$

Example: Bayesian Regression

Lets assume that we want to model our N data points $\mathcal{D} = \left\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\right\}_{i=1}^{N}$ as:

$$\boldsymbol{y}^{(j)} = \sum_{i=1}^{k} \beta_i \boldsymbol{B}_i(\boldsymbol{x}^{(j)}) + \boldsymbol{\varepsilon}^{(j)} \qquad \text{for } j = 1, \ldots, N$$

where $\boldsymbol{\varepsilon}^{(j)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ and for some specific choice of basis functions $\Rightarrow$ this is our choice of model $\mathcal{M}$.
Once this posterior distribution of the parameters given the data using Bayes' theorem obtained

$$P(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) = \frac{P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})P(\boldsymbol{\theta}|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})}$$

we can:

- predict the response $\boldsymbol{y}^*$ at new predictor $\boldsymbol{x}^*$ and more importantly be able to characterize its full distribution (with uncertainty) as:

$$p(\boldsymbol{y}^*|\boldsymbol{x}^*, \mathcal{D}, \mathcal{M}) = \int p(\boldsymbol{y}^*|\boldsymbol{x}^*, \boldsymbol{\theta}, \mathcal{D}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) d\boldsymbol{\theta}$$

- Compute the **evidence** of the chosen model (used for model comparison)

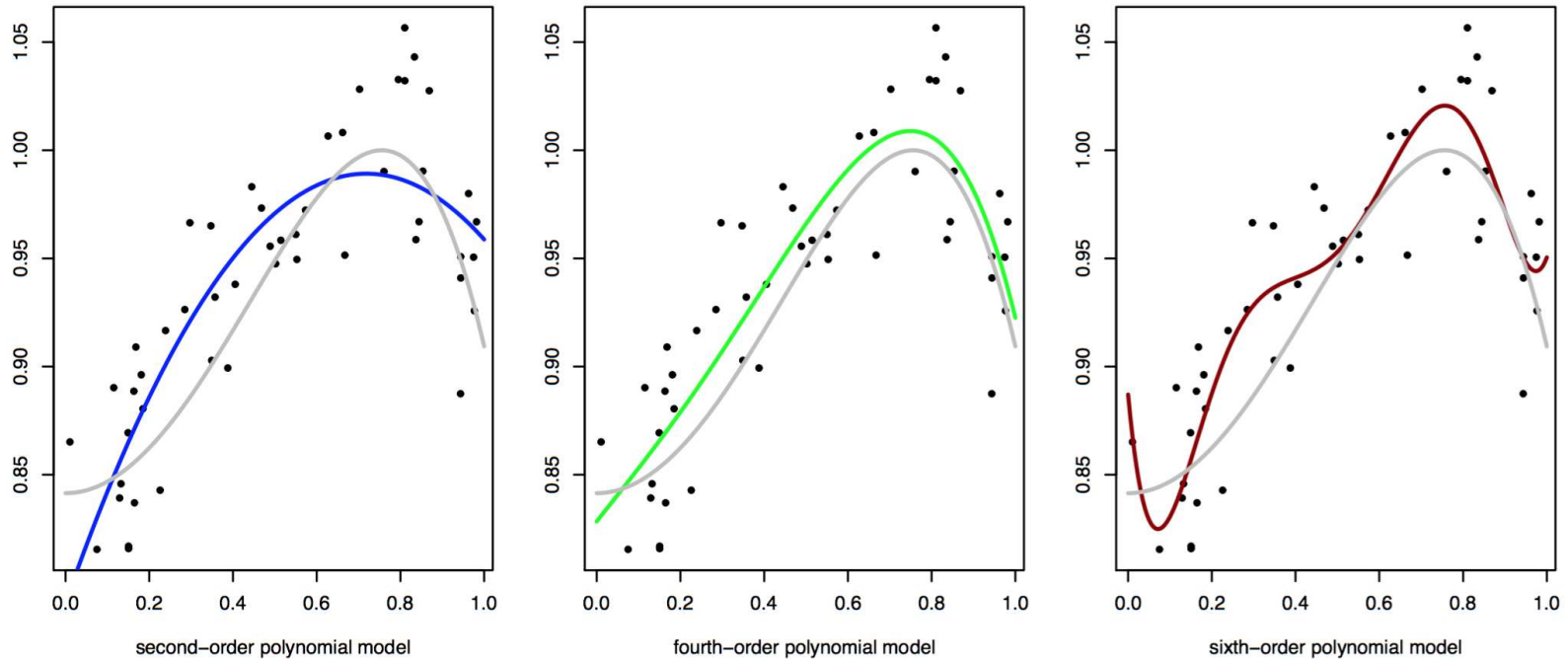
$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}$$

Bayesian Model Selection and Occam's Razor

Here are the least-squares fits of polynomial models for y having the form (for $p = 2, 4, 6$) of

$$y = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \text{noise}$$

where the noise has $\mathcal{N}(0, 0.03^2)$ distribution.



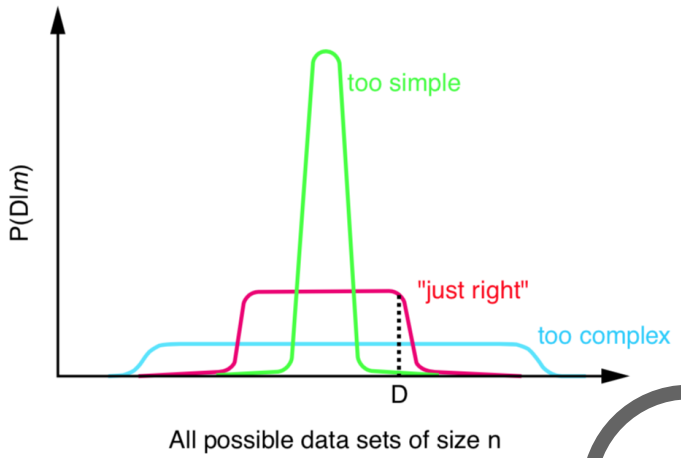
Extra basis function will allow us to explain the training data better but will not necessarily lead to better predictions by the model as the extra basis function may give rise to the model **overfitting** the data

Bayesian Model Selection and Occam's Razor

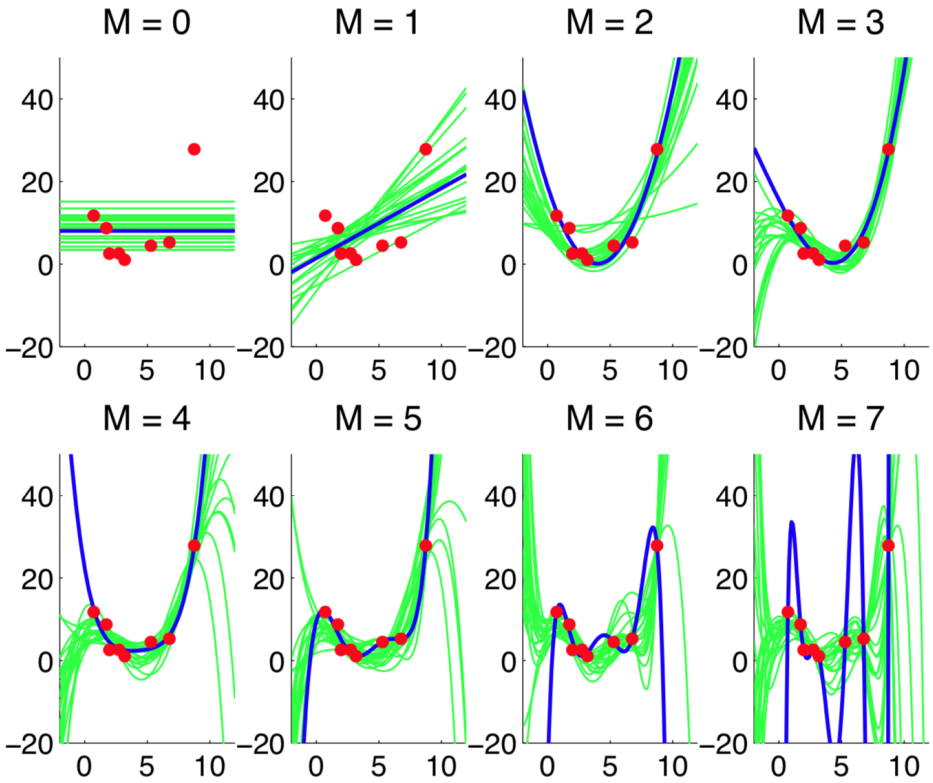
- Combating overfitting is vital to the performance of flexible models
- Regularization techniques are generally employed to penalize too complex (flexible) models
- The Bayesian framework contains a natural penalty against over complex models, sometimes called **Occam's Razor**, which simply states

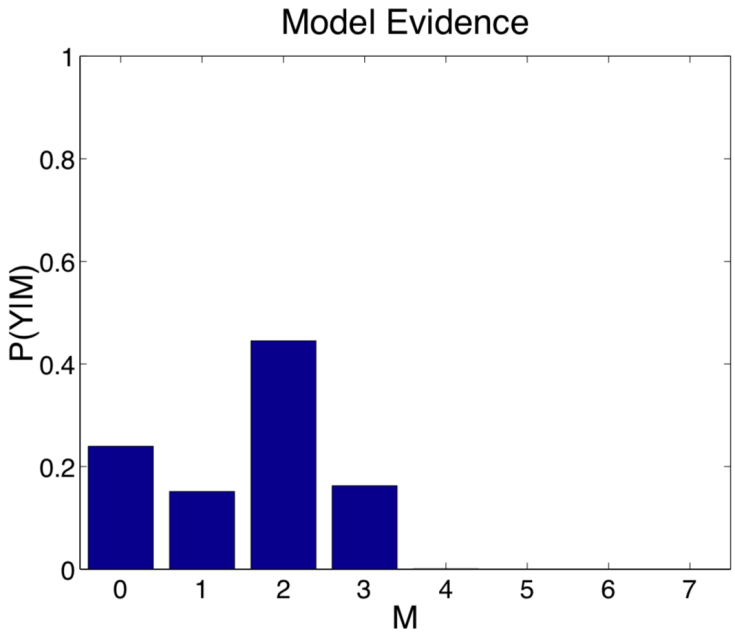
"a simpler theory is to be favored over a more complex one, all other being equal"

Suppose that there are 3 competing models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ where model $\mathcal{M}_1$ covers only a subspace of that covered by $\mathcal{M}_2$ (i.e $\mathcal{M}_1 \subset \mathcal{M}_2$) and $\mathcal{M}_2 \subset \mathcal{M}_3$. As each marginal likelihood $p(\mathcal{D}|\mathcal{M}_i)$ must integrate to one over the data-space, if data $\mathcal{D}$ are observed, the simpler model $\mathcal{M}_2$ is preferred as it has higher probability than the more complex one $\mathcal{M}_3$.



Bayesian Model Selection and Occam's Razor





For example, for quadratic polynomials ($m = 2$): $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2)$ and parameters $\boldsymbol{\theta} = (\sigma^2, \beta_1, \ldots, \beta_2)$

On Choosing Priors

- **Objective priors**: distribution derived from the assumed model rather than assessed from expert opinions or could be non-informative priors that attempt to capture ignorance and have good frequentist properties.
- **Hierarchical priors**: multiple levels of priors:

$$p(\theta) = \int p(\theta|\alpha)p(\alpha)d\alpha$$
$$\int p(\theta|\alpha)\left(\int p(\alpha|\nu)p(\nu)d\nu\right)d\alpha$$

- **Empirical priors**: learn some of the parameters of the prior from the data ("Empirical Bayes")
- **Subjective priors**: priors should capture our beliefs about reasonable hypotheses before observing the data as well as possible. They are subjective but not arbitrary.

Subjective Priors

Priors should capture our **beliefs and knowledge** about the range of reasonable hypotheses as well as possible.

Otherwise we (or our learning machine) will make inferences and decisions which are **not coherent** with our (its) beliefs and knowledge. How do we know our beliefs?

- Think about the problems domain.
- Generate data from the prior. Does it match expectations?

Even very vague prior beliefs can be useful, since the data will concentrate the posterior around reasonable models.

The key ingredient of Bayesian methods is not the prior, it's the idea of averaging over different possibilities.

How to choose stupid priors

- Choose an improper or uninformative prior so that your marginal likelihoods are meaningless.
- Alternatively, choose very dogmatic narrow priors that can't adapt to the data.
- Choose a prior that is very hard to compute with.
- After choosing your prior, don't sample from your model to see whether simulated data make sense.
- Never question the prior. Don't describe your prior in your paper, so that your work is not reproducible.

Bayesian solution

The posterior distribution of the parameters given the data using Bayes' theorem:

$$P(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = \frac{P(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})P(\boldsymbol{\theta}|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})}$$

we can:

- predict the response $\boldsymbol{y}^*$ at new predictor $\boldsymbol{x}^*$ and more importantly be able to characterize its full distribution (with uncertainty) as:

$$p(\boldsymbol{y}^*|\boldsymbol{x}^*,\mathcal{D},\mathcal{M}) = \int p(\boldsymbol{y}^*|\boldsymbol{x}^*,\boldsymbol{\theta},\mathcal{D},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})d\boldsymbol{\theta}$$

- Compute the **evidence** of the chosen model (used for model comparison)

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}$$

Unfortunately, in most of the cases, a closed form expression of the integrals involved in these expressions cannot be derived

$\Rightarrow$ One can resort to **approximations methods**: Laplace approximation, Variational approximation, Simulation-based techniques (Monte-Carlo sampling)

Some numerical examples with Python

2 python notebooks:

1. **Bayesian Coin flips**
2. **Bayesian Linear Regression**

References I

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley.

Denison, D. G. T., Holmes, C. C., and Mallick, B. K. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics.

Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P., editors (2010). *Handbook Of Spatial Statistics*. Chapman & Hall/CRC.

Gelman, A., Carlin, J., Stern, H., Dunson, D. B., Vehtari, A., and Rubin, D. (2003). *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition.

Rasmussen, C. and Williams, C. (2006). *Gaussian processes for machine learning*. The MIT Press.

Robert, C. P. (2007). *The bayesian choice*. Springer, second edition edition.