How to be a Bayesian

Václav Šmídl,

Winter school of machine learning,
Czech Technical University
vasek.smidl@gmail.com

January 20, 2020

Overview

Extract from *Hierarchical Bayesian Models*, FJFI summer

- Lecture 1: How to be a Bayesian
- Lecture 2: Approximations and computational tools
- Lecture 3: Application to Deep Active Learning

Overview

Extract from *Hierarchical Bayesian Models*, FJFI summer

Lecture 1: How to be a Bayesian

Lecture 2: Approximations and computational tools

Lecture 3: Application to Deep Active Learning

Lecture 1:

Bayesian theory

- philosophy
- calculus

Examples:

- Linear regression
- Model averaging

Who is Bayesian

Bayesian $=$

Who is Bayesian

Bayesian = someone who uses **probability** calculus to quantify **uncertainty**.

Who is Bayesian

Bayesian $=$ someone who uses **probability** calculus to quantify **uncertainty**.

Justification: Uncertainty and randomness have the same effect on decision-making.

Who is Bayesian

Bayesian = someone who uses **probability** calculus to quantify **uncertainty**.

Justification: Uncertainty and randomness have the same effect on decision-making.

Gravitational acceleration:

constant	$g = 9.80665$
range	$g = 9.80665 \pm 0.00001$
distribution	$g \sim \mathcal{N}(9.80665, 0.00001)$
	(std $= 0.00001$)

Who is Bayesian

Bayesian = someone who uses **probability** calculus to quantify **uncertainty**.

Justification: Uncertainty and randomness have the same effect on decision-making.

Gravitational acceleration:

constant	$g = 9.80665$
range	$g = 9.80665 \pm 0.00001$
distribution	$g \sim \mathcal{N}(9.80665, 0.00001)$
	(std $= 0.00001$)

?? Is gravitational acceleration a random quantity?

Probability of an event

Probability=Frequency of an event:

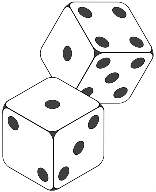
$$P(x) = \frac{\#\,\text{realizations}}{\#\,\text{trials}}$$

[1] Book: **The Theory That Would Not Die:** How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy

Probability of an event

Probability=Frequency of an event:

$$P(x) = \frac{\#\,\text{realizations}}{\#\,\text{trials}}$$



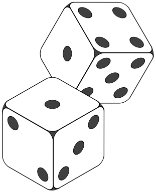
$P(x = 1) = \frac{1}{6}$

[1]Book: **The Theory That Would Not Die:** How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy

Probability of an event

Probability=Frequency of an event:

$$P(x) = \frac{\#\,\text{realizations}}{\#\,\text{trials}}$$



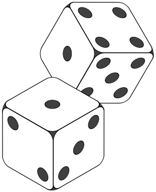
$P(x = 1) = \frac{1}{6}$

[1]Book: **The Theory That Would Not Die:** How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy

Probability of an event

Probability=Frequency of an event:

$$P(x) = \frac{\#\,\text{realizations}}{\#\,\text{trials}}$$



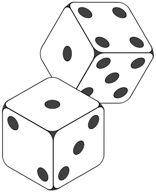
$P(x = 1) = \frac{1}{6}$

[1] Book: **The Theory That Would Not Die:** How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy

Probability of an event

Frequentist:
Probability=Frequency of an event:

$$P(x) = \frac{\#\,\text{realizations}}{\#\,\text{trials}}$$



$P(x = 1) = \frac{1}{6}$

Bayesian:

Frequency:

$$P(\text{Sparta beats Slavia}) = \frac{133}{294} \approx 45\%$$

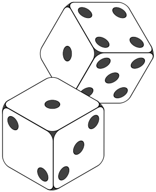
[1] Book: **The Theory That Would Not Die:** How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy

Probability of an event

Frequentist:

Probability=Frequency of an event:

$$P(x) = \frac{\#\,\text{realizations}}{\#\,\text{trials}}$$



$P(x = 1) = \frac{1}{6}$

Bayesian:

Frequency:

$$P(\text{Sparta beats Slavia}) = \frac{133}{294} \approx 45\%$$

Degree (state) of belief:

$$P(x|d) = \frac{P(d|x)P(x)}{\sum_x P(d|x)P(x)}$$

$P(\text{Sparta vs. Slavia} = 1) = 1/1.8$

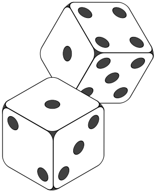
[1]Book: **The Theory That Would Not Die:** How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy

Probability of an event

Frequentist:

Probability=Frequency of an event:

$$P(x) = \frac{\#\,\text{realizations}}{\#\,\text{trials}}$$



$P(x=1) = \frac{1}{6}$

Bayesian:

Frequency:

$$P(\text{Sparta beats Slavia}) = \frac{133}{294} \approx 45\%$$

Degree (state) of belief:

$$P(x|d) = \frac{P(d|x)P(x)}{\sum_x P(d|x)P(x)}$$

$P(\text{Sparta vs. Slavia} = 1) = 1/1.8$

Same probability calculus

Different [1] role of prior $P(x)$, applications and methods

[1] Book: **The Theory That Would Not Die:** How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy

Probability calculus: discrete

Random variables:

$$X \in \{x_1, \ldots, x_M\}$$
$$Y \in \{y_1, \ldots, y_L\}$$

Joint probability

$$P(X = x_i, Y = y_j) = \frac{n_{i,j}}{N}$$

where N ($N \to \infty$) is the number of realizations and $n_{i,j}$ is the number of trials where $X = x_i$, $Y = y_j$.

Rules:

1. sum rule

$$P(X = x_i) = \sum_{j=1}^{L} P(X = x_i, Y = y_i),$$

2. product rule

$$P(X, Y) = p(Y|X)p(X)$$

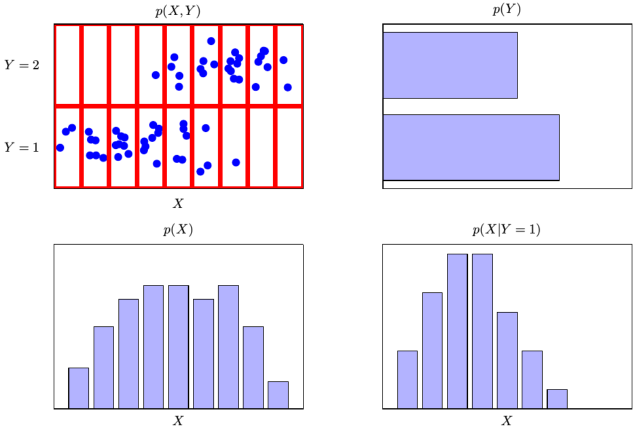
All you need is rules: Rules of probability

1. Product rule (Chain rule)

$$P(X, Y) = P(X|Y)P(Y), \\ = P(X)P(Y|X)$$

2. Sum rule (Marginalization)

$$P(X) = \sum_Y P(X, Y) \\ P(Y) = \sum_X P(X, Y)$$



Cancer example

- Approximately 1% of women aged 40-50 have breast cancer.
- A woman with breast cancer has a 90% chance of a positive test.
- A woman without cancer has a 10% chance of a false positive result.

What is the probability a woman has breast cancer given that she just had a positive test?

Cancer example

- Approximately 1% of women aged 40-50 have breast cancer.
- A woman with breast cancer has a 90% chance of a positive test.
- A woman without cancer has a 10% chance of a false positive result.

What is the probability a woman has breast cancer given that she just had a positive test?

- X =1 if a woman has cancer
- Y =1 if the test is positive

We want to know

$$P(X=1|Y=1) = \frac{P(Y|X)P(X)}{P(Y)}$$

Cancer example

- Approximately 1% of women aged 40-50 have breast cancer.
- A woman with breast cancer has a 90% chance of a positive test.
- A woman without cancer has a 10% chance of a false positive result.

What is the probability a woman has breast cancer given that she just had a positive test?

- X =1 if a woman has cancer
- Y =1 if the test is positive

We want to know

$$P(X=1|Y=1) = \frac{P(Y|X)P(X)}{P(Y)}$$

$$\begin{aligned}
&P(Y=1|X=1) = 0.9, \\
&P(X=1) = 0.01, \\
&P(Y) = \textstyle\sum_X P(Y|X)P(X) = \\
&\qquad P(Y|X=1)P(X=1) + \\
&\qquad P(Y|X=0)P(X=0) \\
&\quad = 0.9 * 0.01 + 0.1 * 0.99 = 0.108
\end{aligned}$$

$$P(X=1|Y=1) = \tfrac{0.009}{0.108} = 8.3\%$$

Probability calculus: continuous

Random variable: $x \in \langle -\infty, \infty \rangle$
Probability that it is in an interval $\langle a, b \rangle$ is

$$p\left(x \in \langle a, b \rangle\right) = \int_a^b p(x)dx,$$

where $p(x)$ probability density function

$$p(x) \geq 1, \qquad \int p(x)dx = 1,$$

Cumulative function

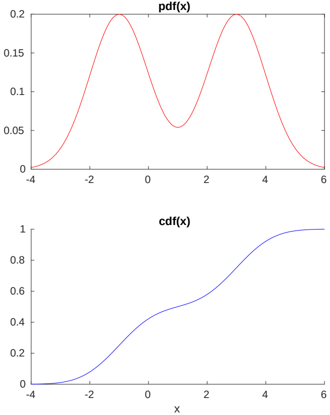
$$P(y) = \int_{-\infty}^{y} p(x)dx$$

Expected value:

$$\mathrm{E}_{p(x)}(g(x)) = \int g(x)p(x)dx,$$

Quantiles:

$$Q(p) = \inf\left\{x : p \leq P(x)\right\}.$$



Probability calculus: multivariate continuous

Joint probability distribution $p(x, y)$

1. sum rule

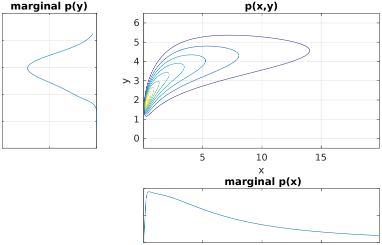
$$p(x) = \int p(x, y) dy,$$

2. product rule

$$p(x, y) = p(y|x)p(x)$$

3. change of variables:

$$x = f(y), \text{ with } p_x(x)$$
$$p_y(y) = p_x(f(y))|f'(y)|.$$



Multivariate Normal distribution

Multivariate normal distribution:
$x = [x_1, x_2]$

$$p(x) = \mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$$
$$\propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right),$$

Marginals

$$p(x_1) = \mathcal{N}(\mu_1, \Sigma_{11}), \quad p(x_1) = \mathcal{N}(\mu_2, \Sigma_{22}),$$

Conditional:

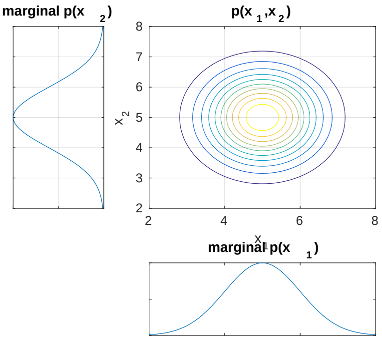
$$p(x_1|x_2) = \mathcal{N}(\overline{\mu}, \overline{\Sigma}),$$
$$\overline{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$
$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Matrix N., Generalized N., GP ...

Example:

$$\mu = [5; 5]; \qquad \Sigma_{11} = \Sigma_{22} = 1.$$

$$\Sigma_{12} = 0$$



Multivariate normal distribution:
$x = [x_1, x_2]$

$$p(x) = \mathcal{N}\left(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$$
$$\propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right),$$

Marginals

$$p(x_1) = \mathcal{N}(\mu_1, \Sigma_{11}), \quad p(x_1) = \mathcal{N}(\mu_2, \Sigma_{22}),$$

Conditional:

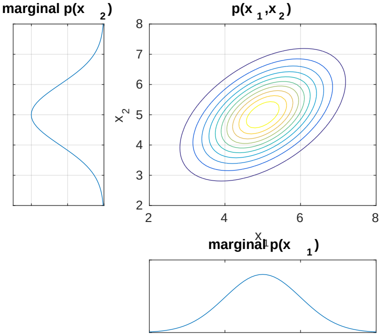
$$p(x_1|x_2) = \mathcal{N}(\overline{\mu}, \overline{\Sigma}),$$
$$\overline{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$
$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Matrix N., Generalized N., GP ...

Example:

$$\mu = [5; 5]; \qquad \Sigma_{11} = \Sigma_{22} = 1.$$

$$\Sigma_{12} = 0.5$$



Multivariate Normal distribution

Multivariate normal distribution:
$x = [x_1, x_2]$

$$p(x) = \mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$$
$$\propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right),$$

Marginals

$$p(x_1) = \mathcal{N}(\mu_1, \Sigma_{11}), \quad p(x_1) = \mathcal{N}(\mu_2, \Sigma_{22}),$$

Conditional:

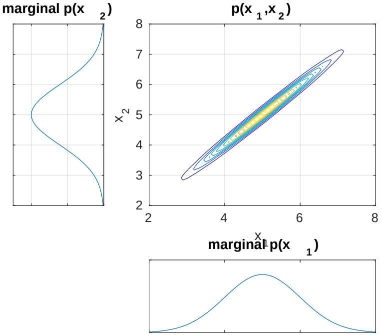
$$p(x_1|x_2) = \mathcal{N}(\overline{\mu}, \overline{\Sigma}),$$
$$\overline{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$
$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Matrix N., Generalized N., GP ...

Example:

$$\mu = [5; 5]; \qquad \Sigma_{11} = \Sigma_{22} = 1.$$

$$\Sigma_{12} = 0.99$$



Bayes Rule

From chain rule:

$$P(X|Y)P(Y) = P(Y|X)P(X).$$

$$P(X|Y) = \frac{P(Y|X)P(X).}{P(Y)}$$

Bayes Rule

From chain rule:

$$P(X|Y)P(Y) = P(Y|X)P(X).$$

$$P(X|Y) = \frac{P(Y|X)P(X).}{P(Y)}$$

Application: θ is a parameter, D is a random observation

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.$$

Bayes Rule

From chain rule:

$$P(X|Y)P(Y) = P(Y|X)P(X).$$
$$P(X|Y) = \frac{P(Y|X)P(X).}{P(Y)}$$

Application: θ is a parameter, D is a random observation

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.$$

Philosophical issue:

Frequentists: parameter is NOT a random quantity, $p(\theta)$ should not exist.

Bayesian: $p(\theta|D)$ is our degree of belief in parameter values.

Example: curve fitting

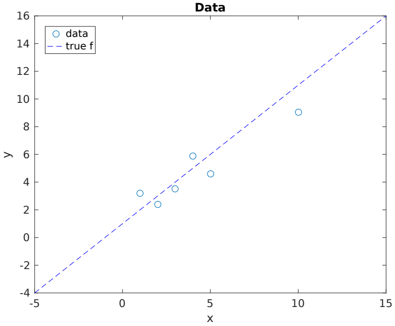
Fit by a linear function:

$$\begin{array}{llll} y_1 & = ax_1 & +b1, & +e_1 \\ y_2 & = ax_2 & +b1 & +e_2, \\ \vdots & \vdots & \vdots & \vdots \end{array}$$

In matrix notation $\theta = [a, b]^T$:

$$\mathbf{y} = X\theta + \mathbf{e},$$

Minimize $\sum_i e_i^2 = \mathbf{e}^T\mathbf{e}$:



Example: curve fitting

Fit by a linear function:

$$
\begin{array}{llll}
y_1 & = a x_1 & + b 1, & + e_1 \\
y_2 & = a x_2 & + b 1 & + e_2, \\
\vdots & \vdots & \vdots & \vdots
\end{array}
$$

In matrix notation $\theta = [a, b]^T$:

$$\mathbf{y} = X\theta + \mathbf{e},$$

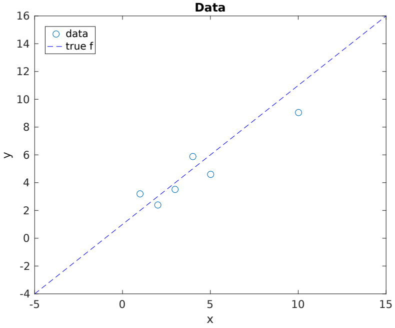
Minimize $\sum_i e_i^2 = \mathbf{e}^T\mathbf{e}$:

$$\frac{d(\mathbf{e}^T\mathbf{e})}{d\theta} = 0.$$

$$\frac{d}{d\theta}((\mathbf{y} - X\theta)^T(\mathbf{y} - X\theta)) = 0$$

$$\frac{d}{d\theta}(\mathbf{y}^T\mathbf{y} - \theta^T X^T\mathbf{y} - \mathbf{y}^T X\theta + \theta^T X^T X\theta) = 0$$

$$-X^T\mathbf{y} + X^T X\theta = 0$$



Example: curve fitting

Fit by a linear function:

$$\begin{array}{llll} y_1 & = a x_1 & + b1, & + e_1 \\ y_2 & = a x_2 & + b1 & + e_2, \\ \vdots & \vdots & \vdots & \vdots \end{array}$$

In matrix notation $\theta = [a, b]^T$:

$$\mathbf{y} = X\theta + \mathbf{e},$$

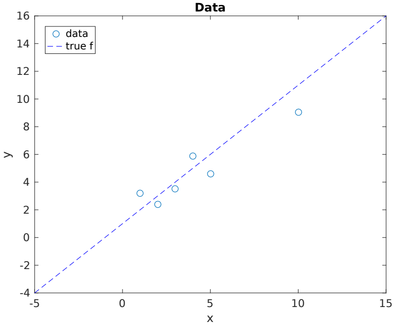
Minimize $\sum_i e_i^2 = \mathbf{e}^T \mathbf{e}$:

$$\frac{d(\mathbf{e}^T\mathbf{e})}{d\theta} = 0.$$

$$\frac{d}{d\theta}((\mathbf{y} - X\theta)^T(\mathbf{y} - X\theta)) = 0$$

$$\frac{d}{d\theta}(\mathbf{y}^T\mathbf{y} - \theta^T X^T \mathbf{y} - \mathbf{y}^T X\theta + \theta^T X^T X\theta) = 0$$

$$-X^T\mathbf{y} + X^T X\theta = 0$$



Solution:

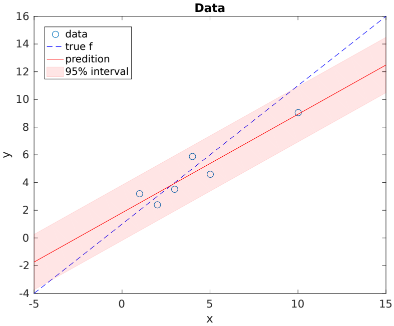
$$\hat{\theta} = (X^T X)^{-1} X^T \mathbf{y}.$$

Prediction

Prediction with LS estimate:

$$\hat{y} = X\hat{\theta} + e.$$

Known variance of e.
Why it does not extrapolate well?



Prediction

Prediction with LS estimate:

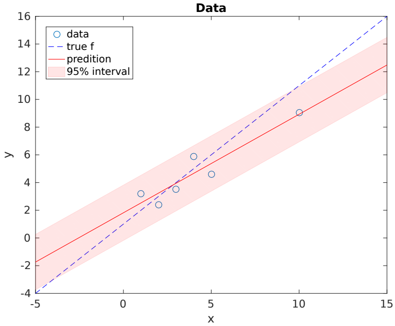
$$\hat{y} = X\hat{\theta} + e.$$

Known variance of e.
Why it does not extrapolate well?

Bayesian explanation
Prediction

$$\hat{y} \sim p(y'|\hat{\theta}),$$

assumes **certainty** in estimate of θ.



Prediction

Prediction with LS estimate:

$$\hat{y} = X\hat{\theta} + e.$$

Known variance of e.
Why it does not extrapolate well?

Bayesian explanation
Prediction

$$\hat{y} \sim p(y'|\hat{\theta}),$$

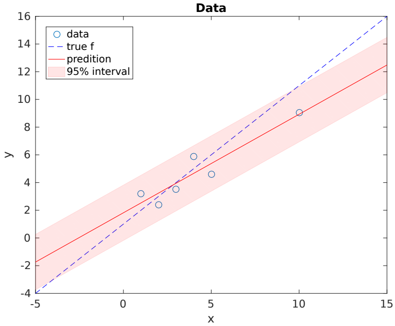
assumes **certainty** in estimate of θ.

- All that is certain is the data!

$$\hat{y} \sim p(y'|y, X)$$

- Working out the rules:

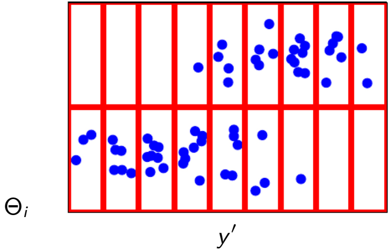
$$p(y'|y, X) = \int p(y'|\theta)p(\theta|y, X)d\theta$$



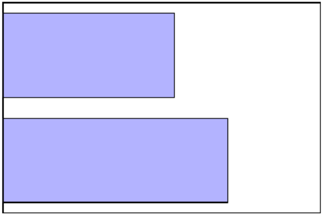
Intuition behind marginalizaton

Definitely not exact math! $\theta \in \{\Theta_1, \Theta_2\}$

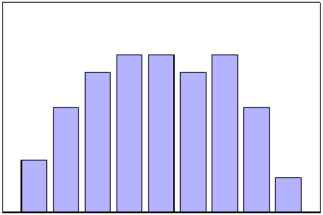
$p(y', \theta | y, X)$



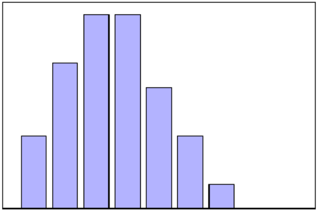
$p(\theta | y, X)$



$p(y' | y, X)$



$p(y' | \theta = \Theta_2)$



Bayesian Prediction

- Bayesian prediction:

$$p(y'|y, X) = \int p(y'|\theta)p(\theta|y, X)d\theta$$

- Posterior probability

$$p(\theta|y, X) \propto p(y|\theta, X)p(\theta)$$

for choices:

$$\begin{aligned} p(y|\theta, X) &= \mathcal{N}(X\theta, 1), \\ \log p(y|\theta, X) &= -\frac{1}{2}(y - X\theta)^\top (y - X\theta) + c, \end{aligned}$$

Bayesian Prediction

- Bayesian prediction:

$$p(y'|y, X) = \int p(y'|\theta)p(\theta|y, X)d\theta$$

- Posterior probability

$$p(\theta|y, X) \propto p(y|\theta, X)p(\theta)$$

for choices:

$$\begin{aligned} p(y|\theta, X) &= \mathcal{N}(X\theta, 1), \\ \log p(y|\theta, X) &= -\frac{1}{2}(y - X\theta)^\top(y - X\theta) + c, \end{aligned}$$

- Solution

$$\begin{aligned} p(\theta|y, X) &= \mathcal{N}(\hat{\theta}, S_n), \\ \hat{\theta} &= (X'X)^{-1}X'y, \quad S_n = (X'X)^{-1}. \end{aligned}$$

Bayesian Prediction

- Bayesian prediction:

$$p(y'|y, X) = \int p(y'|\theta)p(\theta|y, X)d\theta$$

- Posterior probability

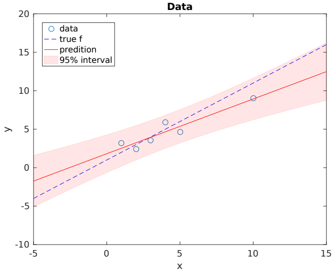
$$p(\theta|y, X) \propto p(y|\theta, X)p(\theta)$$

for choices:

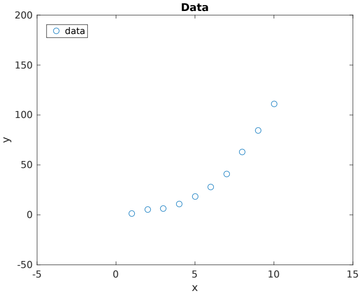
$$p(y|\theta, X) = \mathcal{N}(X\theta, 1),$$
$$\log p(y|\theta, X) = -\frac{1}{2}(y - X\theta)^\top (y - X\theta) + c,$$

- Solution

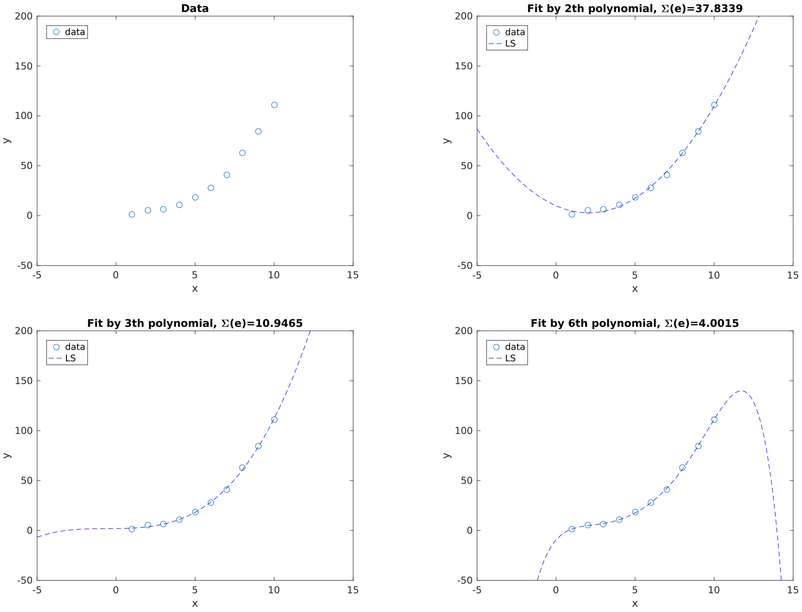
$$\begin{aligned} p(\theta|y, X) &= \mathcal{N}(\hat{\theta}, S_n), \\ \hat{\theta} &= (X'X)^{-1}X'y, \quad S_n = (X'X)^{-1}. \\ y' &= {\color{green}X\hat{\theta}} + {\color{red}\sqrt{1 + [1, x]S_n[1, x]^\top}} e \end{aligned}$$



Challenge: curve fitting



Challenge: curve fitting



What is wrong with minimization?

1. The error of the fit is minimized
   - over-fitting,
2. Model complexity is not taken into account
3. How the humans decide?

What is wrong with minimization?

1. The error of the fit is minimized
   - over-fitting,
2. Model complexity is not taken into account
3. How the humans decide?

- Potentially many answers
  - penalization / regularization terms,
  - information criteria
  - cross validation testing / training data,

What is wrong with minimization?

1. The error of the fit is minimized
   - over-fitting,
2. Model complexity is not taken into account
3. How the humans decide?

- Potentially many answers
  - penalization / regularization terms,
  - information criteria
  - cross validation testing / training data,

- Bayesian answer:
  - admit that the model order is **unknown**.

Bayesian Model Selection

- **Unknown** quantity: model order r has distribution $p(r|y, X)$
- Known data: $\mathbf{y}, X$ with model $p(\mathbf{y}|\theta, X, r) = N(X\theta, 1)$,

Looking for $p(r|\mathbf{y}, X)$:

1. Bayes rule

$$p(r|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, r)p(r)}{\sum_r p(\mathbf{y}|X, r)p(r)}, \qquad p(r) = ?$$

Bayesian Model Selection

- **Unknown** quantity: model order r has distribution $p(r|y, X)$
- Known data: $\mathbf{y}, X$ with model $p(\mathbf{y}|\theta, X, r) = N(X\theta, 1)$,

Looking for $p(r|\mathbf{y}, X)$:

1. Bayes rule
$$p(r|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, r)p(r)}{\sum_r p(\mathbf{y}|X, r)p(r)}, \qquad p(r) =?$$
2. Marginalization
$$p(\mathbf{y}|X, r) = \int p(\mathbf{y}, \theta|X, r)d\theta$$

Bayesian Model Selection

- **Unknown** quantity: model order r has distribution $p(r|y, X)$
- Known data: $\mathbf{y}, X$ with model $p(\mathbf{y}|\theta, X, r) = N(X\theta, 1)$,

Looking for $p(r|\mathbf{y}, X)$:

1. Bayes rule
$$p(r|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, r)p(r)}{\sum_r p(\mathbf{y}|X, r)p(r)}, \qquad p(r) =?$$
2. Marginalization
$$p(\mathbf{y}|X, r) = \int p(\mathbf{y}, \theta|X, r)d\theta$$
3. Chain rule
$$p(\mathbf{y}, \theta|X, r) = p(\mathbf{y}|\theta, X, r)p(\theta|r), \qquad p(\theta|r) =?$$

Bayesian Model Selection

- **Unknown** quantity: model order r has distribution $p(r|y,X)$
- Known data: $\mathbf{y}, X$ with model $p(\mathbf{y}|\theta, X, r) = N(X\theta, 1)$,

Looking for $p(r|\mathbf{y}, X)$:

1. Bayes rule

$$p(r|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, r)p(r)}{\sum_r p(\mathbf{y}|X, r)p(r)}, \qquad p(r) = 1/r_{max}$$

2. Marginalization

$$p(\mathbf{y}|X, r) = \int p(\mathbf{y}, \theta|X, r) d\theta$$

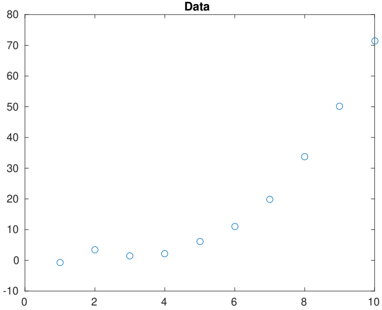
3. Chain rule

$$p(\mathbf{y}, \theta|X, r) = p(\mathbf{y}|\theta, X, r)p(\theta|r), \qquad p(\theta|r) = N(0, \alpha I)$$

Solution:

$$p(r|\mathbf{y}, X, \alpha) \propto \left|X^T X + \alpha I\right|^{-1/2} \exp\left(-\frac{1}{2}\hat{\theta}\left(X^T X + \alpha I\right)\hat{\theta}\right)$$

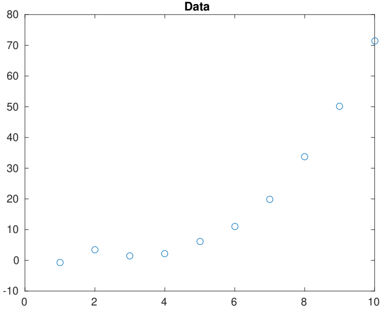
Application of the polynomial



α	1e-8	1e-6	1e-4	"best"
$P(x=2)$	44%	8%	1%	44%
$P(x=3)$	55%	92%	99%	55%
$P(x=4)$	0%	0%	0%	0%

How to choose α?

Application of the polynomial

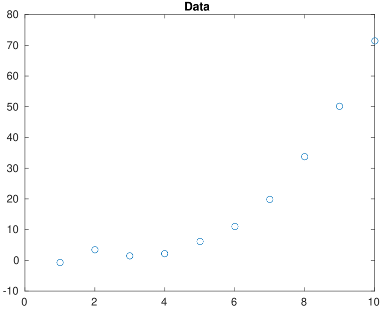


α	1e-8	1e-6	1e-4	"best"
$P(x=2)$	44%	8%	1%	44%
$P(x=3)$	55%	92%	99%	55%
$P(x=4)$	0%	0%	0%	0%

How to choose α?

- assume α an unknown **hyperparametr**
- **uncertainty =>** **hierarchical** prior $p(\alpha) = \Gamma(\gamma, \delta)$.
- solve $p(r|y, X) = \int p(r|y, x, \alpha)p(\alpha)d\alpha$

Application of the polynomial

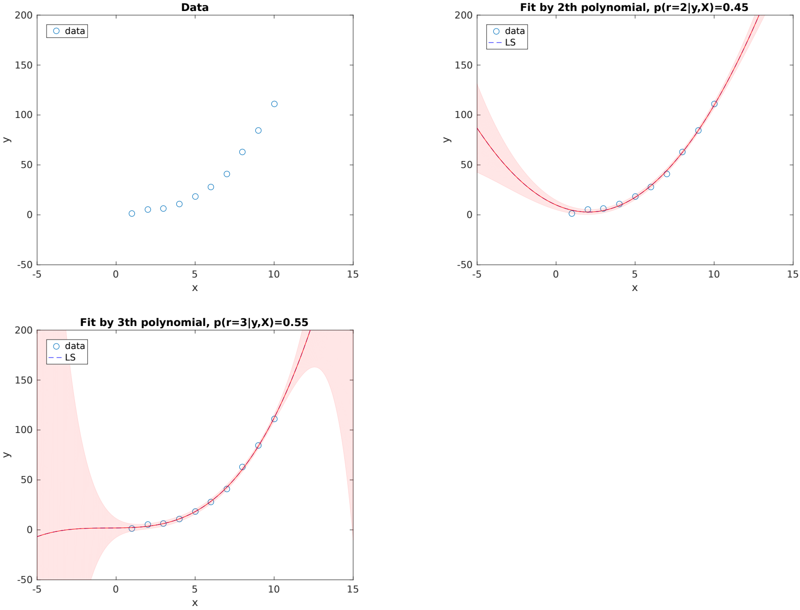


α	1e-8	1e-6	1e-4	"best"
$P(x=2)$	44%	8%	1%	44%
$P(x=3)$	55%	92%	99%	55%
$P(x=4)$	0%	0%	0%	0%

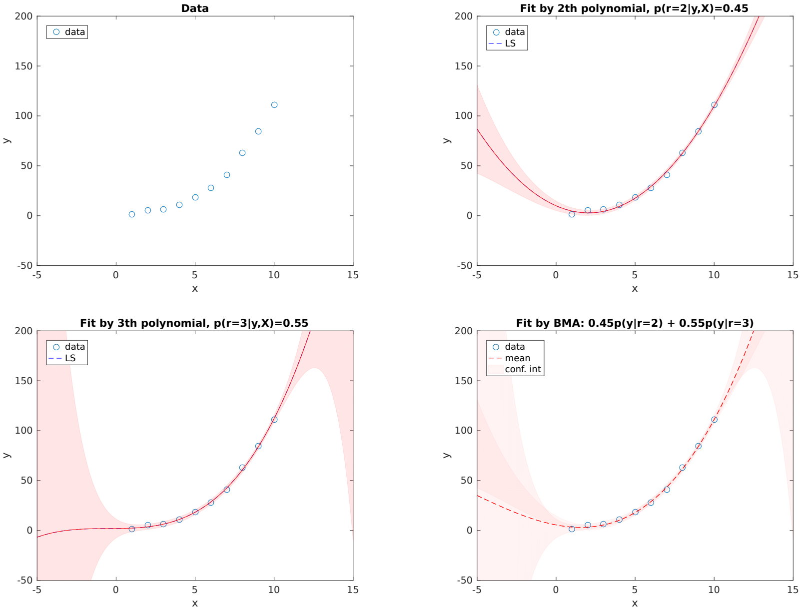
How to choose α?

- assume α an unknown **hyperparametr**
- **uncertainty => hierarchical** prior $p(\alpha) = \Gamma(\gamma, \delta)$.
- solve $p(r|y, X) = \int p(r|y, x, \alpha) p(\alpha) d\alpha$
- works for $\gamma = \delta = 0$ which is Jeffrey's improper prior $p(\alpha) \propto 1/\alpha$,
  - Recursion ends! no need for next hierarchy.

Bayesian prediction:



Bayesian prediction:



Take home message

- Bayesians represent uncertainty by probability
- Prior knowledge is problem specific
  - previously observed data
  - different source of data
  - structural information (positivity)
- Uncertainty of any kinds should be acknowledged and respected,
  - marginalize!
  - key computational difficulty