Hasso Plattner Institut

Introduction to Probabilistic Machine Learning

Ralf Herbrich, Rainer Schlosser

Tutorial 2 – Recap Unit 1 & 2

Overview

1. **Course Setup**
2. Recap: Probabilities and Distributions (Unit 1)
3. Recap: Main Concepts of Unit 2
4. Example: Bayesian Inference
5. Example: Multiplication of Normal Distributions
6. Example: Decision-Making
7. Hints for Exercise 1 (to be handed in Monday May 5)

Course Overview

Week	Topic Lecture	Tutorial	Exercises
07.04. & 08.04.	1 Probability Theory	Intro Julia	
14.04. & 15.04.	2 Inference Methods and Decision-Making	no tutorial	**Exercise 1**
21.04. & 22.04.	**no lecture**	**Theory Unit 1 & 2**	(14.04. – 05.05.)
28.04. & 29.04.	3 Graphical Models: Independence	Theory Unit 3	
05.05. & 06.05.	4 Graphical Models: Exact Inference	Theory Unit 4	Exercise 2
12.05. & 13.05.	5 Graphical Models: Approximate Inference	Theory Unit 5	(05.05. – 19.05.)
19.05. & 20.05.	6 Bayesian Ranking	Theory Unit 6	Exercise 3
26.05. & 27.05.	7 Linear Basis Function Models	Theory Unit 7	(19.05. – 02.06.)
02.06. & 03.06.	8 Bayesian Regression	Theory Unit 8	Exercise 4
09.06. & 10.06.	**no lecture**	9 Bayesian Classification	(02.06. – 23.06.)
16.06. & 11.06.	10 Non-Bayesian Classification Learning	Theory Unit 9 & 10	
23.06. & 24.06.	11 Gaussian Processes	Theory Unit 11	Exercise 5
30.06. & 01.07.	12 Information Theory	Theory Unit 12	(23.06. – 07.07.)
07.07. & 08.07.	13 Real-World Applications		

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Overview

1. Course Setup
2. **Recap: Probabilities and Distributions (Unit 1)**
3. Recap: Main Concepts of Unit 2
4. Example: Bayesian Inference
5. Example: Multiplication of Normal Distributions
6. Example: Decision-Making
7. Hints for Exercise 1 (to be handed in Monday May 5)

Recap: Expectation and Variance

Expected Value (of a random variable):

- **Discrete**: $E[X] = \sum_{x \in \Omega_X} x \cdot P(x)$ and $E[g(X)] = \sum_{x \in \Omega_X} g(x) \cdot P(x)$
- **Continuous**: $E[X] = \int_{\mathbb{R}} x \cdot p_X(x)\, dx$ and $E[g(x)] = \int_{\mathbb{R}} g(x) \cdot p_X(x)\, dx$

Recap: Expectation and Variance

Expected Value (of a random variable):

- **Discrete**: $E[X] = \sum_{x \in \Omega_X} x \cdot P(x)$ and $E[g(X)] = \sum_{x \in \Omega_X} g(x) \cdot P(x)$
- **Continuous**: $E[X] = \int_{\mathbb{R}} x \cdot p_X(x)\, dx$ and $E[g(x)] = \int_{\mathbb{R}} g(x) \cdot p_X(x)\, dx$

Variance (of a random variable):

- **General**: $\text{var}[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$
- **Discrete**: $\text{var}[X] = \sum_{x \in \Omega_X} [(x - E[X])^2] \cdot P(x)$
- **Continuous**: $\text{var}[X] = \int_{\mathbb{R}} [(x - E[X])^2] \cdot p_X(x)\, dx$
- **2nd moment**: $E[X^2] = \sum_{x \in \Omega_X} x^2 \cdot P(x)$ and $E[X^2] = \int_{\mathbb{R}} x^2 \cdot p_X(x)\, dx$

Recap: Classes of Probability Distributions

- Some **classes of probability distributions** can be fully described with a small number of parameters $\boldsymbol{\theta} \in \mathbb{R}^d$
  - **Advantages**:
    1. **Storage Efficiency**: Only d real numbers for whole function!
    2. **Compute Efficiency**: Only $O(d)$ computation for rules of probability!
  - **Disadvantages**:
    1. Too restrictive to represent true phenomenon in real data
    2. Function classes often not closed under Bayes' rule

- Examples?

Recap: Classes of Probability Distributions

- Some **classes of probability distributions** can be fully described with a small number of parameters $\boldsymbol{\theta} \in \mathbb{R}^d$
  - **Advantages**:
    1. **Storage Efficiency**: Only d real numbers for whole function!
    2. **Compute Efficiency**: Only $O(d)$ computation for rules of probability!
  - **Disadvantages**:
    1. Too restrictive to represent true phenomenon in real data
    2. Function classes often not closed under Bayes' rule

- Examples
  - **Discrete**: Bernoulli (d=1), Binomial (d=2), Poisson (d=1), Hypergeometric (d=3), Uniform (d=2), . . .
  - **Continuous**: Normal (d=2), Exponential (d=1), Gamma (d=2), Beta (d=2), Pareto (d=2), Weibull (d=2), Uniform (d=2), Triangular (d=3), . . .

Recap: Probability Distributions: Binomial & Beta Distribution

- **Binomial Distribution**. *The sum of* n *independent Bernoulli random variables with the same success probability* π *has a* Binomial distribution *with*

$$p_X(k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

- **Rarely** used in Machine Learning practice but in Machine Learning theory (for modelling the distribution of the *number* of prediction errors)
- **Properties**:

$$E[X] = n\pi$$
$$\mathrm{var}[X] = n\pi(1-\pi)$$

- **Beta Distribution**. *The conjugate distribution to the Binomial distribution* $\mathrm{Binomial}(n, \pi)$ *is the* Beta distribution *defined by*

$$p(\pi) = \frac{1}{B(\alpha,\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \qquad E(\pi) = \frac{\alpha}{\alpha+\beta}, \qquad \alpha, \beta > 0$$

  - The parameters α and β are the counts of “pseudo-observations” of positive and negative examples!

Jacob Bernoulli (1655 – 1705)

Recap: Probability Distributions: Normal

- **Normal Distribution**. *A continuous random variable X is said to have a standard normal distribution if the density is given by*

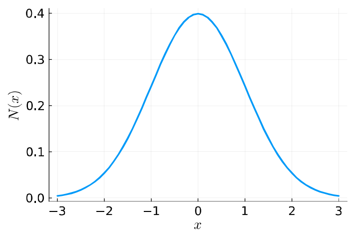
$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Properties**:

$$E[X] = \mu$$
$$\text{var}[X] = \sigma^2$$

- **Importance**. The Normal distribution plays a fundamental role in ML!
  - **Data Modelling**: The limit distribution for the sum of a large number of indepedent and identically distributed random variables.
  - **Machine Learning**: The most common belief distribution for the parameters of prediction functions!
  - **Information Theory**: The distribution function with the most uncertainty ("entropy") when fixing mean and variance of the random variable.

Carl Friedrich Gauss (1777 - 1855)



Overview

1. Course Setup
2. Recap: Probabilities and Distributions (Unit 1)
3. **Recap: Main Concepts of Unit 2**
4. Example: Bayesian Inference
5. Example: Multiplication of Normal Distributions
6. Example: Decision-Making
7. Hints for Exercise 1 (to be handed in Monday May 5)

Recap Unit 2: Overview of Concepts

Data	Regression	Prediction		Classification
Functions/Models	Characterizations by Parameters		Beliefs	Confidence
Bayesian Inference	Bayes Rule	Prior	Likelihood	Posterior
Conjugacy	Distribution Classes	Sigmoid	Normal CDF	
Representations of Gaussian Densities		Multiplication of Densities		
Maximum Likelihood	Maximum A Posterior Estimation		Point Estimates	
Loss Function	Decision Making	Actions	Expected Loss	
Optimization	Gradient Decent	Deep Learning		Updates
Reinforcement Learning				

Recap Unit 2: Overview of Concepts and Focus

Data	Regression	Prediction		Classification
Functions/Models	**Characterizations by Parameters**		**Beliefs**	**Confidence**
Bayesian Inference	**Bayes Rule**	**Prior**	**Likelihood**	**Posterior**
Conjugacy	**Distribution Classes**	Sigmoid	Normal CDF	

Representations of Gaussian Densities **Multiplication of Densities**

Maximum Likelihood Maximum A Posterior Estimation Point Estimates

Loss Function	Decision Making	Actions	Expected Loss	
Optimization	Gradient Decent	Deep Learning		Updates
Reinforcement Learning				

Bayesian Inference in a Nutshell

a) We want to **identify models/functions**, e.g.:

to predict, to classify, to express a probability distribution

b) We describe models/functions through a few **parameters** (θ).

Bayesian Inference in a Nutshell

a) We want to **identify models/functions**, e.g.:

   to predict, to classify, to express a probability distribution

b) We describe models/functions through a few **parameters** (θ).

c) Our information: We have **initial beliefs** for θ as well as observed **data**.

d) We want to identify values of θ that **fit** to our information.

e) We want to obtain a **probability distribution** for suitable θ values.

Bayesian Inference in a Nutshell

a) We want to **identify models/functions**, e.g.:

   to predict, to classify, to express a probability distribution

b) We describe models/functions through a few **parameters** (θ).

c) Our information: We have **initial beliefs** for θ as well as observed **data**.

d) We want to identify values of θ that **fit** to our information.

e) We want to obtain a **probability distribution** for suitable θ values.

f) We **combine** initial beliefs for θ and how likely data can occur under θ.

g) We use **suitable classes** of probability distributions & apply **Bayes‘ Rule**.

h) Obtained distributions for θ provide point **estimates** & **confidence** measures.

Overview

1. Course Setup
2. Recap: Probabilities and Distributions (Unit 1)
3. Recap: Main Concepts of Unit 2
4. **Example: Bayesian Inference**
5. Example: Multiplication of Normal Distributions
6. Example: Decision-Making
7. Hints for Exercise 1 (to be handed in Monday May 5)

Application: Coin Toss - How do we start?

Example 1: Consider a coin with unknown success probability. Estimate the success probability $\theta \in [0,1]$ using Bayesian inference methods.

a) Do we want to derive a point estimate for θ?

b) Do we want to estimate potential values of θ via a probability distribution?

Application: Coin Toss - How do we start?

Example 1: Consider a coin with unknown success probability. Estimate the success probability $\theta \in [0,1]$ using Bayesian inference methods.

a) Do we want to derive a point estimate for θ?

b) Do we want to estimate potential values of θ via a probability distribution?

c) Can we generate data that provides helpful information?

d) Do we have experts here with educated guesses/beliefs?

e) How do we combine these concepts?

f) How do we start? What do we have to choose? What is fixed?

We shall use Bayes‵ Rule

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}}$$

x	Data, i.e., coin toss results (outcomes x; k successes, n throws)
$p(\theta\|x)$	Probability distribution for θ inferred (based on prior & outcomes x)
$p(\theta)$	Initial belief about θ given as a probability distribution
$p(x\|\theta)$	Probability distribution for outcomes x under a fixed θ
$p(x)$	Probability distribution for outcomes x (unknown?)

We shall use Bayes‘ Rule

Bayes’ Rule for Random Variables

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

x	Data, i.e., coin toss results (outcomes x; k successes, n throws)
$p(\theta\|x)$	Probability distribution for θ inferred (based on prior & outcomes x)
$p(\theta)$	Initial belief about θ given as a probability distribution
$p(x\|\theta)$	Probability distribution for outcomes x under a fixed θ
$p(x)$	Probability distribution for outcomes x (unknown?)

We want: $p(\theta|x)$ for all values of θ, x will be fixed (realized outcomes)

$p(\theta|x)$ is determined by the other 3 components!

Choices: We have to come up with likelihood $p(x|\theta)$ and prior $p(\theta)$!

Note, $p(x)$ is then determined as normalizing constant for $p(\theta|x)$

Remaining Choices: Prior & Likelihood

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}}$$

$p(\theta)$ Choose an arbitrary distribution as **initial belief** for θ

Candidates:	U(0,1)	Uniform (continuous)
	Exp(2)	Exponential
	$p(\theta) = \begin{cases} 2 \cdot (1-\theta) & \theta \in [0,1] \\ 0 & \theta \notin [0,1] \end{cases}$	Triangular
	. . .	

Remaining Choices: Prior & Likelihood

Bayes' Rule for Random Variables

$$p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)}$$

$p(\theta)$ Choose an arbitrary distribution as **initial belief** for θ

Candidates:	U(0,1)	Uniform (continuous)
	Exp(2)	Exponential
	$p(\theta)=\begin{cases} 2\cdot(1-\theta) & \theta\in[0,1] \\ 0 & \theta\notin[0,1] \end{cases}$	Triangular
	. . .	

$p(x|\theta)$ Choose a **likelihood** distribution for x given θ (look at use case)

Candidates: Binomial, Bernoulli (outcomes x are discrete)

$p(x)$ Determined by $p(x)=\int_0^1 p(x\,|\,\theta)\cdot p(\theta)\ d\theta \quad\Rightarrow\quad \int_0^1 p(\theta\,|\,x)\ d\theta = \frac{1}{p(x)}\cdot\int_0^1 p(x\,|\,\theta)\cdot p(\theta)\ d\theta = 1$

How to Determine the Remaining Choices?

Bayes' Rule for Random Variables

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

$p(\theta)$ Distribution as **initial belief** for θ?

How to Determine the Remaining Choices?

Bayes' Rule for Random Variables

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

$p(\theta)$ Distribution as **initial belief** for θ

Let's try: $p(\theta) = \begin{cases} 2 \cdot (1-\theta) & \theta \in [0,1] \\ 0 & \theta \notin [0,1] \end{cases}$ **Triangular**

Known characteristics: EW, Var, … $E(\theta) = \int_0^1 \theta \cdot 2 \cdot (1-\theta)\ d\theta = \left[\theta^2 - \frac{2}{3}\theta^3\right]_0^1 = \frac{1}{3}$

$p(x|\theta)$ **Likelihood** distribution for x=(n,k) given θ in our use case?

How to Determine the Remaining Choices?

Bayes' Rule for Random Variables

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

$p(\theta)$ Distribution as **initial belief** for θ

Let's try: $p(\theta) = \begin{cases} 2 \cdot (1-\theta) & \theta \in [0,1] \\ 0 & \theta \notin [0,1] \end{cases}$ **Triangular**

Known characteristics: EW, Var, … $E(\theta) = \int_0^1 \theta \cdot 2 \cdot (1-\theta)\; d\theta = \left[\theta^2 - \frac{2}{3}\theta^3\right]_0^1 = \frac{1}{3}$

$p(x|\theta)$ **Likelihood** distribution for x=(n,k) given θ (look at use case)

Should be: $p(n,k \mid \theta) = \binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k}$ **Binomial**

Known characteristics: EW, Var, …

How to Determine the Remaining Choices?

Bayes' Rule for Random Variables

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

$p(\theta)$ Distribution as **initial belief** for θ

Let's try: $p(\theta) = \begin{cases} 2 \cdot (1-\theta) & \theta \in [0,1] \\ 0 & \theta \notin [0,1] \end{cases}$ **Triangular**

Known characteristics: EW, Var, … $E(\theta) = \int_0^1 \theta \cdot 2 \cdot (1-\theta)\ d\theta = \left[\theta^2 - \frac{2}{3}\theta^3\right]_0^1 = \frac{1}{3}$

$p(x|\theta)$ **Likelihood** distribution for x=(n,k) given θ (look at use case)

Should be: $p(n,k \mid \theta) = \binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k}$ **Binomial**

Known characteristics: EW, Var, …

$p(\theta|x)$ Then: $p(\theta \mid n,k) = \begin{cases} \binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k} \cdot 2 \cdot (1-\theta) / p(n,k) & \theta \in [0,1] \\ 0 & \theta \notin [0,1] \end{cases}$

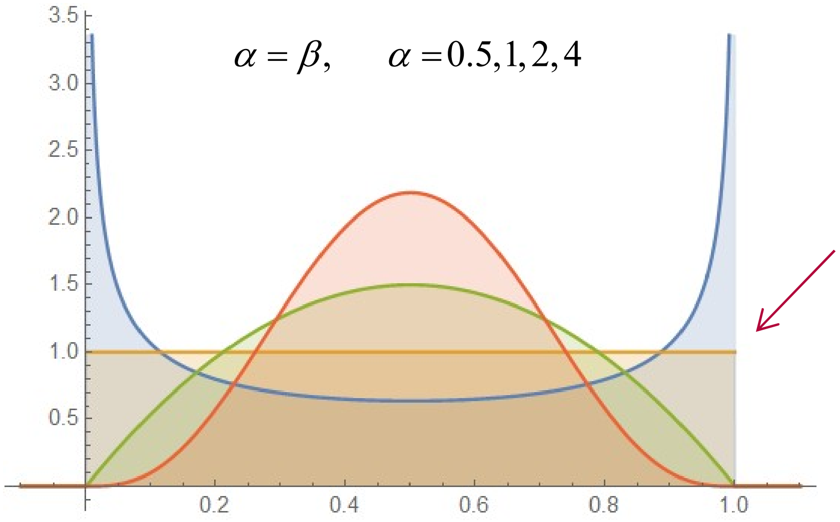
$p(x)$ Then: $p(n,k) = \int_0^1 \binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k} \cdot 2 \cdot (1-\theta)\ d\theta$ **unknown characteristics** ☹

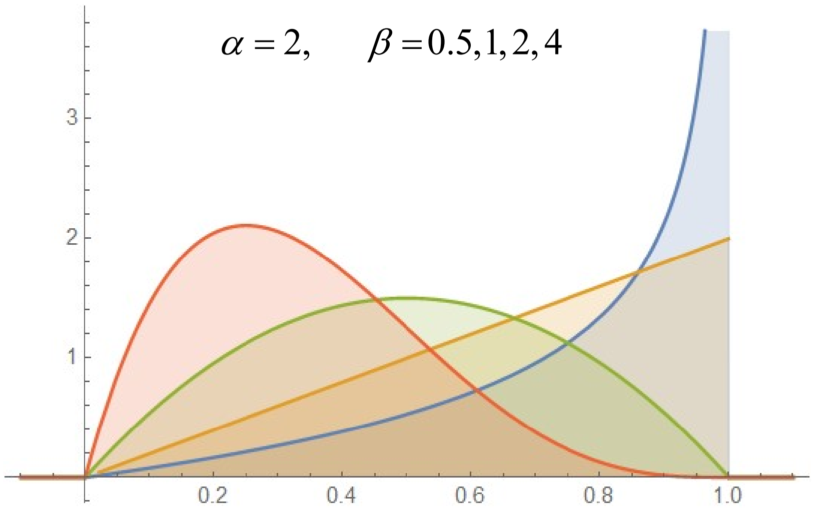
Let`s try Beta Distributions!

$$p(\theta) = \begin{cases} \frac{1}{B(\alpha,\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} & \theta \in (0,1) \\ 0 & \theta \notin (0,1) \end{cases}, \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad E(\theta) = \frac{\alpha}{\alpha+\beta}, \quad \alpha,\beta > 0, \quad \theta \in (0,1)$$

Let's try Beta Distributions!

$$p(\theta)=\begin{cases}\dfrac{1}{B(\alpha,\beta)}\cdot\theta^{\alpha-1}\cdot(1-\theta)^{\beta-1} & \theta\in(0,1)\\ 0 & \theta\notin(0,1)\end{cases},\quad B(\alpha,\beta)=\frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)},\quad E(\theta)=\frac{\alpha}{\alpha+\beta},\quad \alpha,\beta>0,\quad \theta\in(0,1)$$





Let's try Beta Distributions!

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}}$$

$p(\theta)$ **Beta Distribution** as **initial belief** for θ

Let's try: $$p(\theta) = \begin{cases} \frac{1}{B(\alpha,\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} & \theta \in [0,1] \\ 0 & \theta \notin [0,1] \end{cases}, \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Known characteristics: EW, Var, … $$E(\theta) = \frac{\alpha}{\alpha+\beta}$$

$p(x|\theta)$ **Likelihood** distribution for x=(n,k) given θ (look at usecase)

Should be: $$p(n,k\,|\,\theta) = \binom{n}{k} \cdot \theta^{k} \cdot (1-\theta)^{n-k}$$ **Binomial**

$p(\theta|x)$ Then?

$p(x)$ Then?

Let‘s try Beta Distributions!

Bayes’ Rule for Random Variables

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

$p(\theta)$ **Beta Distribution** as **initial belief** for θ

Let’s try: $$p(\theta) = \begin{cases} \frac{1}{B(\alpha,\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} & \theta \in [0,1] \\ 0 & \theta \notin [0,1] \end{cases}, \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Known characteristics: EW, Var, … $$E(\theta) = \frac{\alpha}{\alpha+\beta}$$

$p(x|\theta)$ **Likelihood** distribution for x=(n,k) given θ (look at usecase)

Should be: $$p(n,k\,|\,\theta) = \binom{n}{k} \cdot \theta^{k} \cdot (1-\theta)^{n-k}$$ **Binomial**

$p(\theta|x)$ Then: $$p(\theta\,|\,n,k) = \begin{cases} \binom{n}{k} \cdot \theta^{k} \cdot (1-\theta)^{n-k} \cdot \frac{1}{B(\alpha,\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} \,/\, p(n,k) & \theta \in [0,1] \\ 0 & \theta \notin [0,1] \end{cases}$$

$p(x)$ Then: $$p(n,k) = \int_0^1 p(n,k\,|\,\theta) \cdot p(\theta)\; d\theta$$ **known characteristics??**

Beta Distributions x Binomial Distribution = ??

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}}$$

$p(x)$ Then: $p(n,k) = \int_0^1 p(n,k\,|\,\theta) \cdot p(\theta)\ d\theta$

$p(\theta|x)$ Then: $p(\theta\,|\,n,k) = \binom{n}{k} \cdot \underbrace{\theta^k \cdot (1-\theta)^{n-k}} \cdot \frac{1}{B(\alpha,\beta)} \cdot \underbrace{\theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}} \,/\, p(n,k) \qquad ,\theta \in [0,1]$

$= ??$

Beta Distributions x Binomial Distribution = ??

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}}$$

$p(x)$ Then: $p(n,k) = \int_0^1 p(n,k\,|\,\theta) \cdot p(\theta)\ d\theta$

$p(\theta|x)$ Then: $p(\theta\,|\,n,k) = \binom{n}{k} \cdot \underbrace{\theta^k \cdot (1-\theta)^{n-k}} \cdot \frac{1}{B(\alpha,\beta)} \cdot \underbrace{\theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}} \,/\, p(n,k) \quad , \theta \in [0,1]$

$$= \frac{1}{const(n,k)} \cdot \theta^{\alpha+k-1} \cdot (1-\theta)^{\beta+n-k-1}$$

$$= ??$$

Beta Distributions x Binomial Distribution = ??

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)}}$$

$p(x)$ Then: $p(n,k) = \int_0^1 p(n,k\,|\,\theta)\cdot p(\theta)\; d\theta$

$p(\theta|x)$ Then: $p(\theta\,|\,n,k) = \binom{n}{k}\cdot \underbrace{\theta^k \cdot (1-\theta)^{n-k}} \cdot \frac{1}{B(\alpha,\beta)} \cdot \underbrace{\theta^{\alpha-1}\cdot(1-\theta)^{\beta-1}} \,/\, p(n,k) \qquad ,\theta\in[0,1]$

$$= \frac{1}{const(n,k)}\cdot\theta^{\alpha+k-1}\cdot(1-\theta)^{\beta+n-k-1}$$

i.e., update #success & #failures

$$= \frac{1}{B(\tilde\alpha,\tilde\beta)}\cdot\theta^{\tilde\alpha-1}\cdot(1-\theta)^{\tilde\beta-1}, \qquad \tilde\alpha := \alpha + k \qquad \tilde\beta := \beta + n - k$$

Hence?

Beta Distributions x Binomial Distribution = ??

$p(x)$ Then: $p(n,k)=\int_0^1 p(n,k\,|\,\theta)\cdot p(\theta)\;d\theta$

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)}}$$

$p(\theta|x)$ Then: $p(\theta\,|\,n,k)=\binom{n}{k}\cdot\underbrace{\theta^k\cdot(1-\theta)^{n-k}}\cdot\frac{1}{B(\alpha,\beta)}\cdot\underbrace{\theta^{\alpha-1}\cdot(1-\theta)^{\beta-1}}\,/\,p(n,k) \qquad ,\theta\in[0,1]$

$$=\frac{1}{const(n,k)}\cdot\theta^{\alpha+k-1}\cdot(1-\theta)^{\beta+n-k-1}$$

i.e., update #success & #failures

$$=\frac{1}{B(\tilde{\alpha},\tilde{\beta})}\cdot\theta^{\tilde{\alpha}-1}\cdot(1-\theta)^{\tilde{\beta}-1}, \qquad \tilde{\alpha}:=\alpha+k \qquad \tilde{\beta}:=\beta+n-k$$

Hence: **Without any computations, we get that $p(\theta|x)$ is Beta distributed**

and we have, e.g., $E(\theta\,|\,x=(n,k),\alpha,\beta)=\frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}}=\frac{\alpha+k}{\alpha+k+\beta+n-k}=\frac{\alpha+k}{\alpha+\beta+n}$

Back to the Coin Toss Example (1 Binomial Update)

Bayes’ Rule for Random Variables

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

Round	Toss Result	Prior $p(\theta)$	Likelihood $p(x\|\theta)$	Posterior $p(\theta\|x)$
0	/	Beta(1,1)		
1	1 (k=1)	Beta(1,1)	Binomial(1,1)	Beta(2,1)
2	1 (k=2)	Beta(1,1)	Binomial(2,2)	Beta(3,1)
3	0 (k=2)	Beta(1,1)	Binomial(3,2)	Beta(3,2)
4	0 (k=2)	Beta(1,1)	Binomial(4,2)	Beta(3,3)
5	1 (k=3)	Beta(1,1)	Binomial(5,3)	Beta(4,3)
…				
n	0/1	Beta(1,1)	Binomial(n,k)	**Beta(1+k,1+n-k)**

Back to the Coin Toss Example (n Bernoulli Updates)

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}}$$

Round	Toss Result	Prior $p(\theta)$	Likelihood $p(x|\theta)$	Posterior $p(\theta|x)$
0	/	Beta(1,1)		
1	1 (k=1)	Beta(1,1)	Binomial(1,1)	Beta(2,1)
2	1 (k=2)	Beta(2,1)	Binomial(1,1)	Beta(3,1)
3	0 (k=2)	Beta(3,1)	Binomial(1,0)	Beta(3,2)
4	0 (k=2)	Beta(3,2)	Binomial(1,0)	Beta(3,3)
5	1 (k=3)	Beta(3,3)	Binomial(1,1)	Beta(4,3)
…				
n	0/1	**last Posterior**	**Bernoulli(1,Toss)**	**Beta(1+k,1+n-k)**

Coin Toss Example: Asymptotics

$p(\theta|x)$ Then: $p(\theta \mid n,k) = \binom{n}{k} \cdot \theta^{k} \cdot (1-\theta)^{n-k} \cdot \frac{1}{B(\alpha,\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} \,/\, p(n,k) \qquad ,\theta \in [0,1]$

$$= \frac{1}{B(\tilde{\alpha},\tilde{\beta})} \cdot \theta^{\tilde{\alpha}-1} \cdot (1-\theta)^{\tilde{\beta}-1}, \qquad \tilde{\alpha} := \alpha + k \qquad \tilde{\beta} := \beta + n - k$$

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}}$$

Expectation of $p(\theta|x)$ tends to?

Variance of posterior $p(\theta|x)$ tends to?

Coin Toss Example: Asymptotics

Bayes' Rule for Random Variables

$$p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)}$$

$p(\theta|x)$ Then: $p(\theta\,|\,n,k) = \binom{n}{k}\cdot\theta^{k}\cdot(1-\theta)^{n-k}\cdot\frac{1}{B(\alpha,\beta)}\cdot\theta^{\alpha-1}\cdot(1-\theta)^{\beta-1}\,/\,p(n,k) \qquad ,\theta\in[0,1]$

$$= \frac{1}{B(\tilde{\alpha},\tilde{\beta})}\cdot\theta^{\tilde{\alpha}-1}\cdot(1-\theta)^{\tilde{\beta}-1}, \qquad \tilde{\alpha} := \alpha + k \qquad \tilde{\beta} := \beta + n - k$$

Expectation of $p(\theta|x)$ tends to true success probability

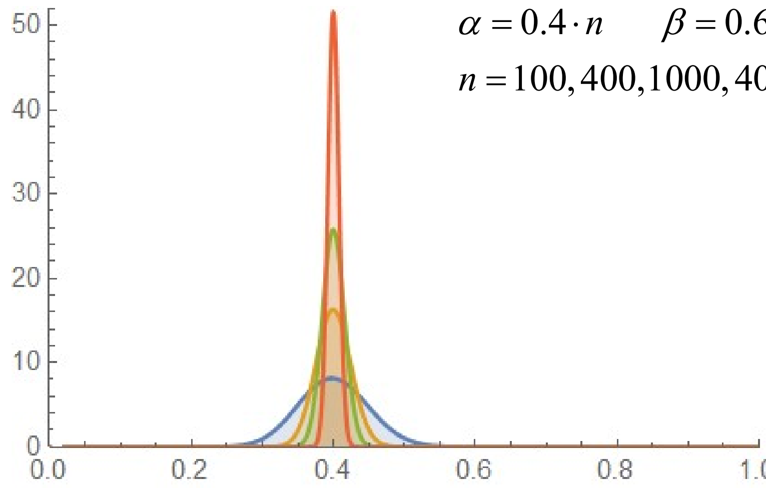
$$E(\theta) = \frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}} = \frac{\alpha+k}{\alpha+\beta+n} \xrightarrow{n\to\infty} \theta_{true}$$

Variance of posterior $p(\theta|x)$ tends to 0

$$Var(\theta) = \frac{\tilde{\alpha}\cdot\tilde{\beta}}{(\tilde{\alpha}+\tilde{\beta}+1)\cdot(\tilde{\alpha}+\tilde{\beta})^2} \xrightarrow{\tilde{\alpha}+\tilde{\beta}\to\infty} 0$$

$\alpha = 0.4\cdot n \qquad \beta = 0.6\cdot n,$

$n = 100, 400, 1000, 4000$



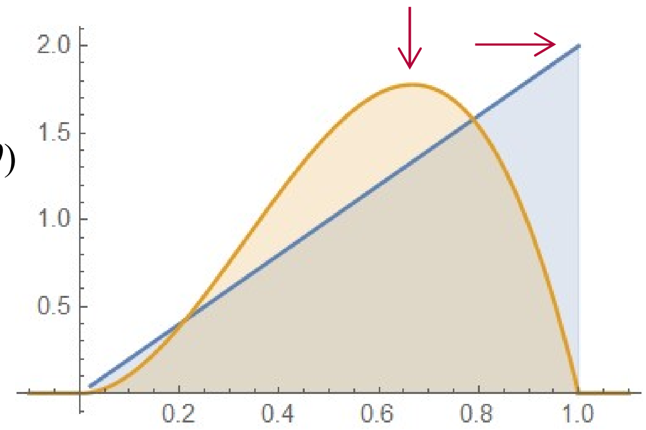
Point Estimates of the Success Probability

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)}}$$

$p(\theta|x)$ Then: $p(\theta\,|\,n,k) = \binom{n}{k}\cdot\theta^k\cdot(1-\theta)^{n-k}\cdot\frac{1}{B(\alpha,\beta)}\cdot\theta^{\alpha-1}\cdot(1-\theta)^{\beta-1}\,/\,p(n,k) \quad ,\theta\in[0,1]$

$$= \frac{1}{B(\tilde\alpha,\tilde\beta)}\cdot\theta^{\tilde\alpha-1}\cdot(1-\theta)^{\tilde\beta-1}, \quad \tilde\alpha := \alpha+k \quad \tilde\beta := \beta+n-k$$



Maximum A Posterior Estimator: $\hat\theta_{MAP} = \underset{\theta\in[0,1]}{\arg\max}\{p(\theta\,|\,n,k)\} = \underset{\theta\in[0,1]}{\arg\max}\{p(k\,|\,n,\theta)\cdot p(\theta)\}$

- Includes data & prior
- Prior can be "uninformative", cf. Uniform(0,1) / Beta(1,1)

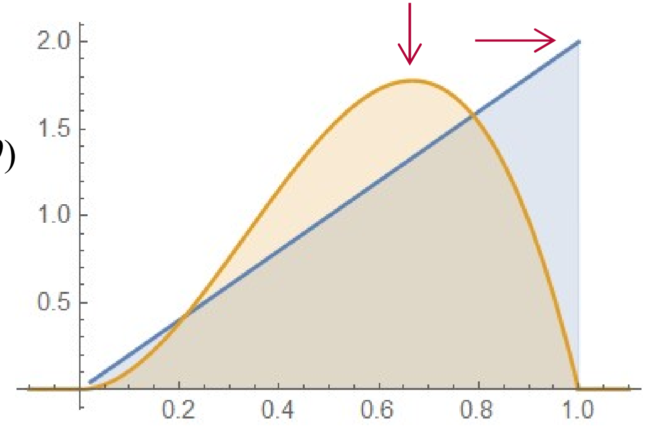
Point Estimates of the Success Probability

$p(\theta|x)$ Then: $p(\theta\,|\,n,k) = \binom{n}{k}\cdot\theta^{k}\cdot(1-\theta)^{n-k}\cdot\frac{1}{B(\alpha,\beta)}\cdot\theta^{\alpha-1}\cdot(1-\theta)^{\beta-1}\,/\,p(n,k) \qquad ,\theta\in[0,1]$

$$=\frac{1}{B(\tilde\alpha,\tilde\beta)}\cdot\theta^{\tilde\alpha-1}\cdot(1-\theta)^{\tilde\beta-1}, \qquad \tilde\alpha:=\alpha+k \qquad \tilde\beta:=\beta+n-k$$

Bayes' Rule for Random Variables

$$\boxed{p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)}}$$



Maximum A Posterior Estimator: $\hat\theta_{MAP} = \underset{\theta\in[0,1]}{\arg\max}\left\{p(\theta\,|\,n,k)\right\} = \underset{\theta\in[0,1]}{\arg\max}\left\{p(k\,|\,n,\theta)\cdot p(\theta)\right.$

- Includes data & prior
- Prior can be "uninformative", cf. Uniform(0,1) / Beta(1,1)

Maximum Likelihood Estimator: $\hat\theta_{ML} = \underset{\theta\in[0,1]}{\arg\max}\left\{p(k\,|\,n,\theta)\right\} = \frac{k}{n}$

- Includes data only

Recap: Maximum Likelihood Estimator

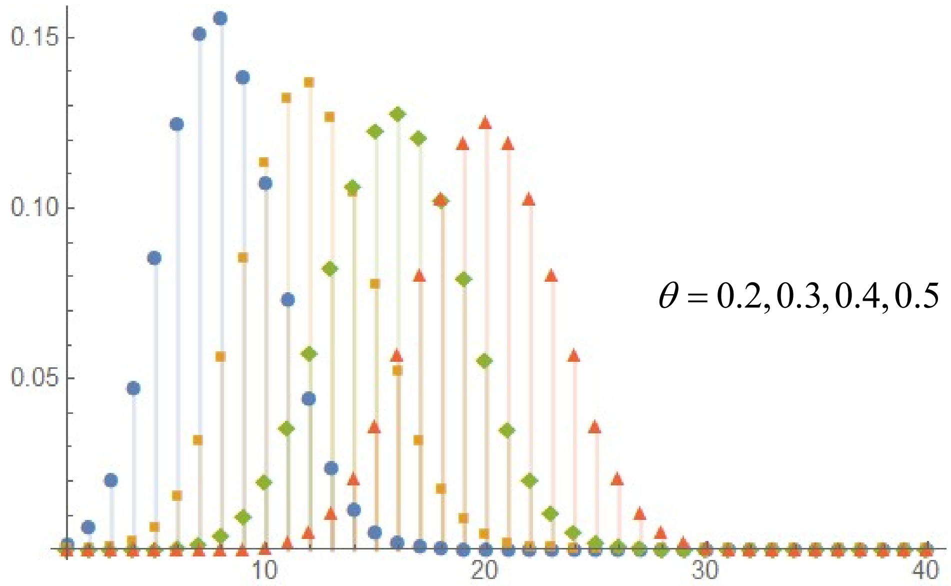
Bayes' Rule for Random Variables

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

$p(x|\theta)$ **Likelihood** distribution for data x=(n,k) given θ (look at usecase)

$$p(X_n = k \mid \theta) = \binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k}$$ **Binomial Distribution**

Maximum Likelihood: $\hat{\theta}_{ML} = \underset{\theta \in [0,1]}{\arg\max} \{ p(k \mid n, \theta) \}$ (highest probability for k)



$\theta = 0.2, 0.3, 0.4, 0.5 \qquad (n = 40)$

Overview

1. Course Setup
2. Recap: Probabilities and Distributions (Unit 1)
3. Recap: Main Concepts of Unit 2
4. Example: Bayesian Inference
5. **Example: Multiplication of Normal Distributions**
6. Example: Decision-Making
7. Hints for Exercise 1 (to be handed in Monday May 5)

Recap: Probability Distributions: Conjugacy

- **Bayes Rule for Random Variables**. *For any probability distribution p over two random variables X and Θ, it holds*

Likelihood $\quad p(x,\theta)$

$$\text{Posterior} \rightarrow p(\theta|x) = \frac{\boxed{p(x|\theta) \cdot p(\theta)}}{p(x)}$$

Prior

- **Conjugacy**. *A family $\{p(x,\theta)\}_{x,\theta}$ is conjugate if the posterior $p(\theta|x)$ is part of the same family as the prior $p(\theta)$ for any value of x.*

Likelihood $p(x\|\theta)$	**Prior** $p(\theta)$	**Posterior** $p(\theta\|x)$
$\mathrm{Ber}(x;\theta)$	$\mathrm{Beta}(\theta;\alpha,\beta)$	$\mathrm{Beta}\big(\theta;\alpha+x,\beta+(1-x)\big)$
$\mathrm{Bin}(x;n,\theta)$	$\mathrm{Beta}(\pi;\alpha,\beta)$	$\mathrm{Beta}\big(\theta;\alpha+x,\beta+(n-x)\big)$
$\mathcal{N}(x;\theta,\sigma^2)$	$\mathcal{N}(\theta;m,s^2)$	$\mathcal{N}\left(\theta;x\cdot\frac{s^2}{s^2+\sigma^2}+m\cdot\frac{\sigma^2}{s^2+\sigma^2},s^2\cdot\frac{\sigma^2}{s^2+\sigma^2}\right)$

- **Big Advantage**: Computing the exact posterior is computationally efficient!

Howard Raiffa
(1924 – 2016)

Robert Osher Schlaifer
(1914 – 1994)

Recap: Normal Distribution: Representations

- **Two Parameterizations (for different purposes)**:
  - **Scale-Location Parameters**

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

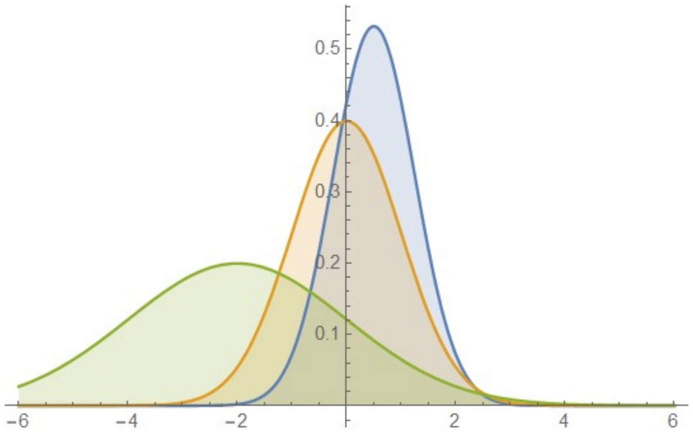
  - **Natural Parameters**

$$\mathcal{G}(x;\tau,\rho) = \sqrt{\frac{\rho}{2\pi}}\cdot\exp\left(-\frac{\tau^2}{2\rho}\right)\cdot\exp\left(\tau\cdot x-\rho\cdot\frac{x^2}{2}\right)$$

- **Conversions**

Two divisions only!

$$\mathcal{N}(x;\mu,\sigma^2) = \mathcal{G}\left(x;\frac{\mu}{\sigma^2},\frac{1}{\sigma^2}\right). \qquad \mathcal{G}(x;\tau,\rho) = \mathcal{N}\left(x;\frac{\tau}{\rho},\frac{1}{\rho}\right)$$



Recap: Normal Distributions and the Product Rule

- **Theorem (Multiplication)**. *Given two one-dimensional Gaussian distributions* $\mathcal{G}(x;\tau_1,\rho_1)$ *and* $\mathcal{G}(x;\tau_2,\rho_2)$ *we have*

Gaussian density

$$\mathcal{G}(x;\tau_1,\rho_1)\cdot\mathcal{G}(x;\tau_2,\rho_2)=\mathcal{G}(x;\tau_1+\tau_2,\rho_1+\rho_2)\cdot\mathcal{N}(\mu_1;\mu_2,\sigma_1^2+\sigma_2^2)$$

Additive updates!

- **Theorem (Division)**. *Given two one-dimensional Gaussian distributions* $\mathcal{G}(x;\tau_1,\rho_1)$ *and* $\mathcal{G}(x;\tau_2,\rho_2)$ *we have*

Gaussian density

$$\frac{\mathcal{G}(x;\tau_1,\rho_1)}{\mathcal{G}(x;\tau_2,\rho_2)}=\mathcal{G}(x;\tau_1-\tau_2,\rho_1-\rho_2)\cdot\frac{1}{\mathcal{N}\left(\frac{\tau_1-\tau_2}{\rho_1-\rho_2};\frac{\tau_2}{\rho_2},\frac{1}{\rho_1-\rho_2}+\frac{1}{\rho_2}\right)}$$

Subtractive updates!

Normal Distribution x Normal Distribution = ??

Let`s collect key facts:

- **Conjugacy** can be extremely important
- *Likelihood and Prior* have to be such that *Prior & Posterior* are of **same type**
- This particularly works for **Normal Distributions**
- Normal Distributions can also be represented via **Natural Parameters**
- For **multiplication** of densities natural parameters just have to be **added**

Multiplication of Normal Distributions

Approach: Multiply two Normal Distributions N1, N2 (scale parameters)

(1) ??

(2) ??

(3) ??

Multiplication of Normal Distributions

Approach: Multiply two Normal Distributions N1, N2 (scale parameters)

(1) Translate them into Representation G1, G2 (natural parameters)

(2) Multiply the Gaussians efficiently using the Multiplication Theorem

(3) Obtain a Gaussian G3 and the normalization factor

(4) And then ??

Multiplication of Normal Distributions

Approach: Multiply two Normal Distributions N1, N2 (scale parameters)

(1) Translate them into Representation G1, G2 (natural parameters)

(2) Multiply the Gaussians efficiently using the Multiplication Theorem

(3) Obtain a Gaussian G3 and the normalization factor

(4) Translate back the Gaussian G3 into a Normal Distribution N3

i.e., obtain N3 = N1 x N2 in standard scale parameters!

Task: Reproduce the formula for the Normal Posterior, see Unit 2, slide 9

Multiplication of Normal Distributions (Unit 2, slide 9)

Consider: $\underbrace{N(\theta; m, s^2)}_{Prior},\ \underbrace{N(x; \theta, \sigma^2)}_{Likelihood}$

(1) Translate N to G:

Multiplication of Normal Distributions (Unit 2, slide 9)

Consider: $\underbrace{N(\theta; m, s^2)}_{Prior}, \underbrace{N(x; \theta, \sigma^2)}_{Likelihood} = \underbrace{N(\theta; x, \sigma^2)}_{Likelihood}$

$$N(x; \mu, \sigma^2) = p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right)$$

(1) Translate N to G:

Multiplication of Normal Distributions (Unit 2, slide 9)

Consider: $\underbrace{N(\theta; m, s^2)}_{Prior}, \; \underbrace{N(x; \theta, \sigma^2)}_{Likelihood} = \underbrace{N(\theta; x, \sigma^2)}_{Likelihood}$

(1) Translate N to G:

$$\underbrace{N(\theta; \underbrace{m}_{\mu_1}, \underbrace{s^2}_{\sigma_1^2})}_{Prior} = G\left(\theta; \underbrace{\frac{m}{s^2}}_{\tau_1}, \underbrace{\frac{1}{s^2}}_{\rho_1}\right), \quad \underbrace{N(\theta; \underbrace{x}_{\mu_2}, \underbrace{\sigma^2}_{\sigma_2^2})}_{Likelihood} = G\left(\theta; \underbrace{\frac{x}{\sigma^2}}_{\tau_2}, \underbrace{\frac{1}{\sigma^2}}_{\rho_2}\right)$$

(2) Multiplication Theorem:

Multiplication of Normal Distributions (Unit 2, slide 9)

Consider: $\underbrace{N(\theta; m, s^2)}_{Prior}$, $\underbrace{N(x; \theta, \sigma^2)}_{Likelihood} = \underbrace{N(\theta; x, \sigma^2)}_{Likelihood}$

(1) Translate N to G: $\underbrace{N(\theta; \underbrace{m}_{\mu_1}, \underbrace{s^2}_{\sigma_1^2})}_{Prior} = G\left(\theta; \underbrace{\frac{m}{s^2}}_{\tau_1}, \underbrace{\frac{1}{s^2}}_{\rho_1}\right)$, $\underbrace{N(\theta; \underbrace{x}_{\mu_2}, \underbrace{\sigma^2}_{\sigma_2^2})}_{Likelihood} = G\left(\theta; \underbrace{\frac{x}{\sigma^2}}_{\tau_2}, \underbrace{\frac{1}{\sigma^2}}_{\rho_2}\right)$

(2) Multiplication Theorem: $\underbrace{G(\theta; \tau_1, \rho_1)}_{Prior} \cdot \underbrace{G(\theta; \tau_2, \rho_2)}_{Likelihood} = \underbrace{G(\theta; \tau_1 + \tau_2, \rho_1 + \rho_2)}_{Posterior} \cdot \underbrace{const}_{Normalization}$

$$N\left(\underbrace{\frac{\tau_1}{\rho_1}}_{\mu_1 = m}; \underbrace{\frac{\tau_2}{\rho_2}}_{\mu_2 = x}, \underbrace{\frac{1}{\rho_1}}_{\sigma_1^2 = s^2} + \underbrace{\frac{1}{\rho_2}}_{\sigma_2^2 = \sigma^2}\right)$$

(3) Translate G back to N:

Multiplication of Normal Distributions (Unit 2, slide 9)

Consider: $\underbrace{N(\theta;m,s^2)}_{Prior}, \underbrace{N(x;\theta,\sigma^2)}_{Likelihood} = \underbrace{N(\theta;x,\sigma^2)}_{Likelihood}$

(1) Translate N to G: $\underbrace{N(\theta;\underbrace{m}_{\mu_1},\underbrace{s^2}_{\sigma_1^2})}_{Prior} = G\left(\theta;\underbrace{\frac{m}{s^2}}_{\tau_1},\underbrace{\frac{1}{s^2}}_{\rho_1}\right), \quad \underbrace{N(\theta;\underbrace{x}_{\mu_2},\underbrace{\sigma^2}_{\sigma_2^2})}_{Likelihood} = G\left(\theta;\underbrace{\frac{x}{\sigma^2}}_{\tau_2},\underbrace{\frac{1}{\sigma^2}}_{\rho_2}\right)$

(2) Multiplication Theorem: $\underbrace{G(\theta;\tau_1,\rho_1)}_{Prior}\cdot\underbrace{G(\theta;\tau_2,\rho_2)}_{Likelihood} = \underbrace{G(\theta;\tau_1+\tau_2,\rho_1+\rho_2)}_{Posterior}\cdot\underbrace{const}_{Normalization}$

$$N\left(\underbrace{\frac{\tau_1}{\rho_1}}_{\mu_1=m};\underbrace{\frac{\tau_2}{\rho_2}}_{\mu_2=x},\underbrace{\frac{1}{\rho_1}}_{\sigma_1^2=s^2}+\underbrace{\frac{1}{\rho_2}}_{\sigma_2^2=\sigma^2}\right)$$

(3) Translate G back to N: $G(\theta;\tau_1+\tau_2,\rho_1+\rho_2) = N\left(\theta;\frac{\tau_1+\tau_2}{\rho_1+\rho_2},\frac{1}{\rho_1+\rho_2}\right)$

(4) Obtain final Formula:

Multiplication of Normal Distributions (Unit 2, slide 9)

Consider: $\underbrace{N(\theta; m, s^2)}_{Prior}, \underbrace{N(x;\theta,\sigma^2)}_{Likelihood} = \underbrace{N(\theta; x, \sigma^2)}_{Likelihood}$

(1) Translate N to G: $\underbrace{N(\theta; \underbrace{m}_{\mu_1}, \underbrace{s^2}_{\sigma_1^2})}_{Prior} = G\left(\theta; \underbrace{\frac{m}{s^2}}_{\tau_1}, \underbrace{\frac{1}{s^2}}_{\rho_1}\right), \underbrace{N(\theta; \underbrace{x}_{\mu_2}, \underbrace{\sigma^2}_{\sigma_2^2})}_{Likelihood} = G\left(\theta; \underbrace{\frac{x}{\sigma^2}}_{\tau_2}, \underbrace{\frac{1}{\sigma^2}}_{\rho_2}\right)$

(2) Multiplication Theorem: $\underbrace{G(\theta;\tau_1,\rho_1)}_{Prior} \cdot \underbrace{G(\theta;\tau_2,\rho_2)}_{Likelihood} = \underbrace{G(\theta;\tau_1+\tau_2,\rho_1+\rho_2)}_{Posterior} \cdot \underbrace{const}_{Normalization}$

$$N\left(\underbrace{\frac{\tau_1}{\rho_1}}_{\mu_1=m}; \underbrace{\frac{\tau_2}{\rho_2}}_{\mu_2=x}, \underbrace{\frac{1}{\rho_1}}_{\sigma_1^2=s^2} + \underbrace{\frac{1}{\rho_2}}_{\sigma_2^2=\sigma^2}\right)$$

(3) Translate G back to N: $G(\theta;\tau_1+\tau_2,\rho_1+\rho_2) = N\left(\theta; \frac{\tau_1+\tau_2}{\rho_1+\rho_2}, \frac{1}{\rho_1+\rho_2}\right)$

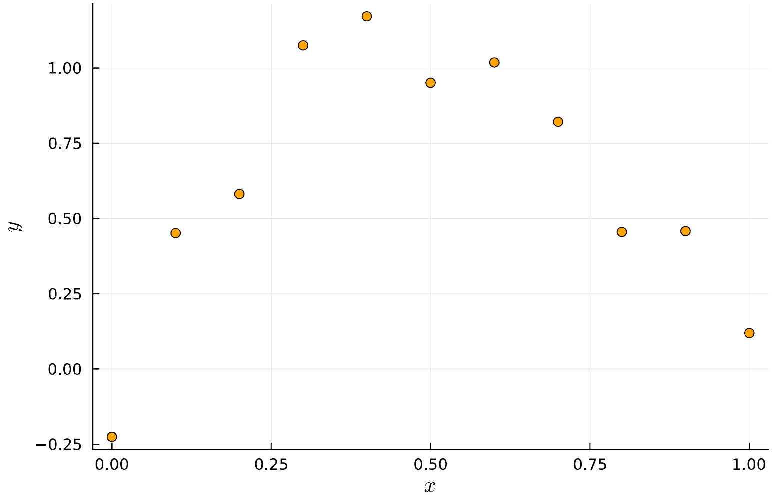
(4) Obtain final Formula: $= N\left(\theta; \frac{\frac{m}{s^2}+\frac{x}{\sigma^2}}{\frac{1}{s^2}+\frac{1}{\sigma^2}}, \frac{1}{\frac{1}{s^2}+\frac{1}{\sigma^2}}\right) = N\left(\theta; x\cdot\frac{s^2}{\sigma^2+s^2} + m\cdot\frac{\sigma^2}{\sigma^2+s^2}, s^2\cdot\frac{\sigma^2}{\sigma^2+s^2}\right)$

Overview

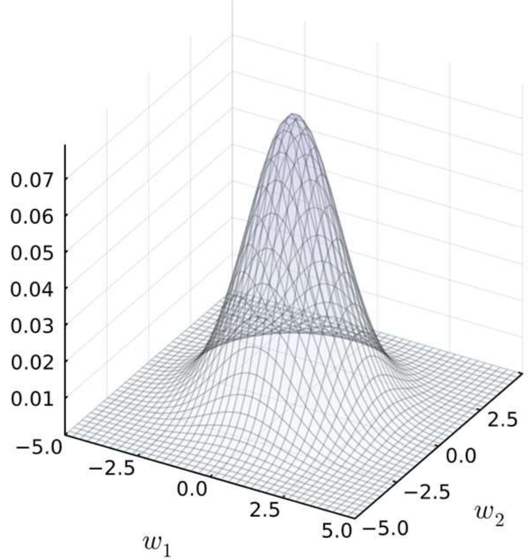
1. Course Setup
2. Recap: Probabilities and Distributions (Unit 1)
3. Recap: Main Concepts of Unit 2
4. Example: Bayesian Inference
5. Example: Multiplication of Normal Distributions
6. **Example: Decision-Making**
7. Hints for Exercise 1 (to be handed in Monday May 5)

Data and Assumed Regression Model (Unit 2, slide 4)

Training Data:



Prior for Combinations of Weights:



$$p(f) = N(w_1; 0,1) \cdot N(w_2; 0,1)$$

Assumed Regression Model (with 2 feature weights):

$$Y_i \sim N(y; w_1 x_i + w_2 x_i^2, \sigma^2)$$

Likelihood (Unit 2, slide 4)

Likelihood (of observed Data under the assumed Model for certain weights):

$$p(D \mid f) = p((y,x)_{i=1}^{n} \mid w_1, w_2) \qquad Y_i \sim N\left(y; w_1 x_i + w_2 x_i^2, \sigma^2\right)$$

$$= P_{w_1,w_2}\left(Y_1 = y_1 \mid x_1 \wedge Y_2 = y_2 \mid x_2 \wedge \ldots \wedge Y_n = y_n \mid x_n\right)$$

Likelihood (Unit 2, slide 4)

Likelihood (of observed Data under the assumed Model for certain weights):

$$p(D \mid f) = p((y,x)_{i=1}^{n} \mid w_1, w_2) \qquad Y_i \sim N\left(y; w_1 x_i + w_2 x_i^2, \sigma^2\right)$$

$$= P_{w_1,w_2}\left(Y_1 = y_1 \mid x_1 \wedge Y_2 = y_2 \mid x_2 \wedge \ldots \wedge Y_n = y_n \mid x_n\right)$$

$$= P_{w_1,w_2}\left(Y_1 = y_1 \mid x_1\right) \cdot P_{w_1,w_2}\left(Y_2 = y_2 \mid x_2\right) \cdot \ldots \cdot P_{w_1,w_2}\left(Y_n = y_n \mid x_n\right)$$

$$= N\left(y_1; w_1 x_1 + w_2 x_1^2, \sigma^2\right) \cdot \ldots \cdot N\left(y_n; w_1 x_n + w_2 x_n^2, \sigma^2\right)$$

Likelihood (Unit 2, slide 4)

Likelihood (of observed Data under the assumed Model for certain weights):

$$p(D \mid f) = p((y,x)_{i=1}^{n} \mid w_1, w_2) \qquad Y_i \sim N\left(y; w_1 x_i + w_2 x_i^2, \sigma^2\right)$$

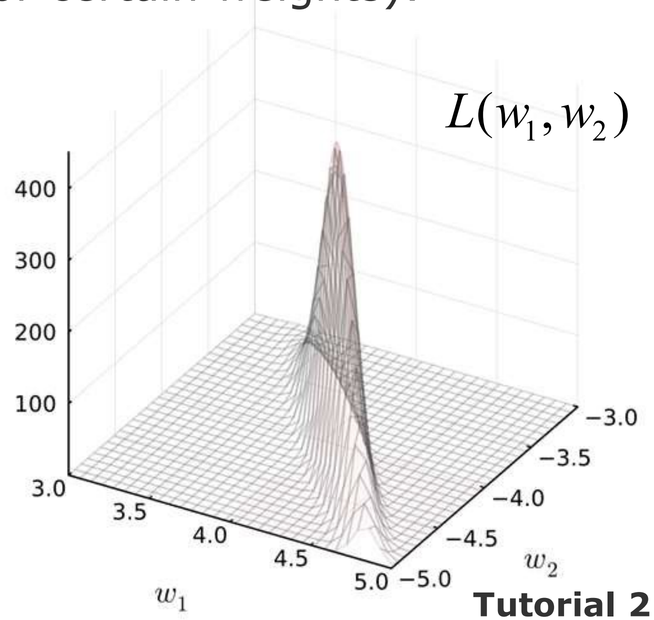
$$= P_{w_1,w_2}\left(Y_1 = y_1 \mid x_1 \wedge Y_2 = y_2 \mid x_2 \wedge \ldots \wedge Y_n = y_n \mid x_n\right)$$

$$= P_{w_1,w_2}\left(Y_1 = y_1 \mid x_1\right) \cdot P_{w_1,w_2}\left(Y_2 = y_2 \mid x_2\right) \cdot \ldots \cdot P_{w_1,w_2}\left(Y_n = y_n \mid x_n\right)$$

$$= N\left(y_1; w_1 x_1 + w_2 x_1^2, \sigma^2\right) \cdot \ldots \cdot N\left(y_n; w_1 x_n + w_2 x_n^2, \sigma^2\right)$$

$$= \prod_i N(y_i; w_1 x_i + w_2 x_i^2, \sigma^2)$$

$$=: L(w_1, w_2)$$



Posterior (Unit 2, slide 5)

Posterior Distribution (of weights based on Data and Prior):

$$p(f \mid D) = \frac{p(D \mid f) \cdot p(f)}{p(D)} = \frac{p(D \mid f) \cdot p(f)}{\int p(D \mid f) \cdot p(f) df}$$

$$\propto p(D \mid f) \cdot p(f) = L(w_1, w_2) \cdot p(f)$$

Posterior (Unit 2, slide 5)

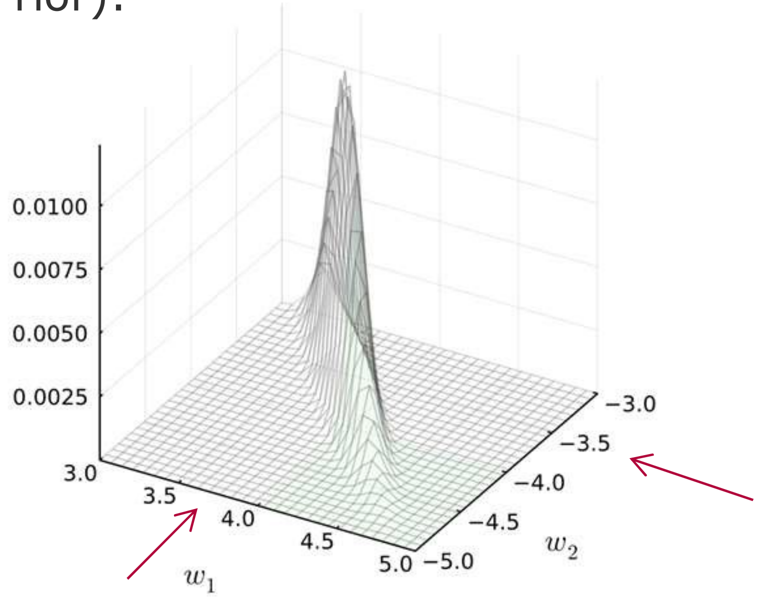
Posterior Distribution (of weights based on Data and Prior):

$$p(f \mid D) = \frac{p(D \mid f) \cdot p(f)}{p(D)} = \frac{p(D \mid f) \cdot p(f)}{\int p(D \mid f) \cdot p(f) df}$$

$$\propto p(D \mid f) \cdot p(f) = L(w_1, w_2) \cdot p(f)$$

$$= \underbrace{\prod_i N(y_i; w_1 x_i + w_2 x_i^2, \sigma^2)}_{p(D|f)} \cdot \underbrace{N(w_1; 0,1) \cdot N(w_2; 0,1)}_{p(f)}$$

$$\propto p(w_1, w_2 \mid D)$$



$$MAP: \ (w_1, w_2) = (3.75, -3.5)$$

Predictive Distribution (Unit 2, slide 5)

Predictive Distribution (based on Assumed Model, posterior-weighted):

$$p(y \mid x, D) = \int_{-\infty}^{+\infty} p(y \mid x, f) \cdot p(f \mid D) df \qquad (\textit{distribution of } y \textit{ given new } x)$$

Predictive Distribution (Unit 2, slide 5)

Predictive Distribution (based on Assumed Model, posterior-weighted):

$$p(y \mid x, D) = \int_{-\infty}^{+\infty} p(y \mid x, f) \cdot p(f \mid D) df \qquad (\textit{distribution of } y \textit{ given new } x)$$

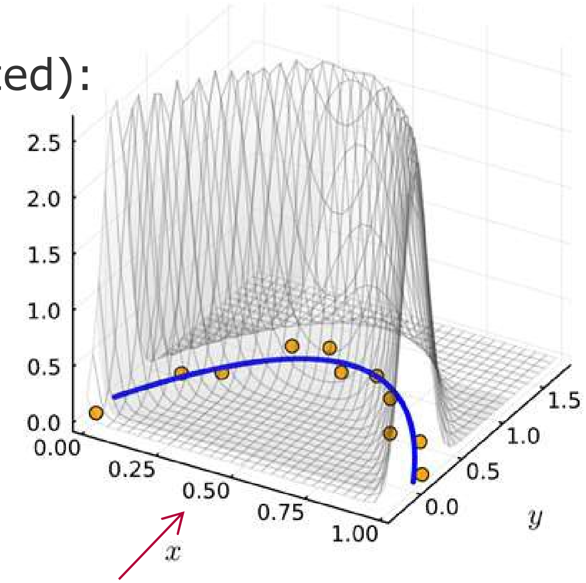
$$= \int_{-\infty}^{+\infty} \underbrace{N(y; w_1 x + w_2 x^2, \sigma^2)}_{p(y|x,f)} \cdot \underbrace{p(w_1, w_2 \mid D)}_{p(f|D)} dw_1 dw_2$$

$$= \int_{-\infty}^{+\infty} \underbrace{N(y; w_1 x + w_2 x^2, \sigma^2)}_{p(y|x,f)} \cdot \underbrace{\prod_i N(y_i; w_1 x_i + w_2 x_i^2, \sigma^2)}_{p(D|f)} \cdot \underbrace{N(w_1; 0,1) \cdot N(w_2; 0,1)}_{p(f)} dw_1 dw_2$$

$$= N(y; \mu(x), \sigma_D^2)$$

Predictive Distribution (Unit 2, slide 5)

Predictive Distribution (based on Assumed Model, posterior-weighted):



$$p(y \mid x, D) = \int_{-\infty}^{+\infty} p(y \mid x, f) \cdot p(f \mid D) df \qquad (\textit{distribution of } y \textit{ given new } x)$$

$$= \int_{-\infty}^{+\infty} \underbrace{N(y; w_1 x + w_2 x^2, \sigma^2)}_{p(y|x,f)} \cdot \underbrace{p(w_1, w_2 \mid D)}_{p(f|D)} dw_1 dw_2$$

$$= \int_{-\infty}^{+\infty} \underbrace{N(y; w_1 x + w_2 x^2, \sigma^2)}_{p(y|x,f)} \cdot \underbrace{\prod_i N(y_i; w_1 x_i + w_2 x_i^2, \sigma^2)}_{p(D|f)} \cdot \underbrace{N(w_1; 0,1) \cdot N(w_2; 0,1}_{p(f)}$$

$$= N(y; \mu(x), \sigma_D^2)$$

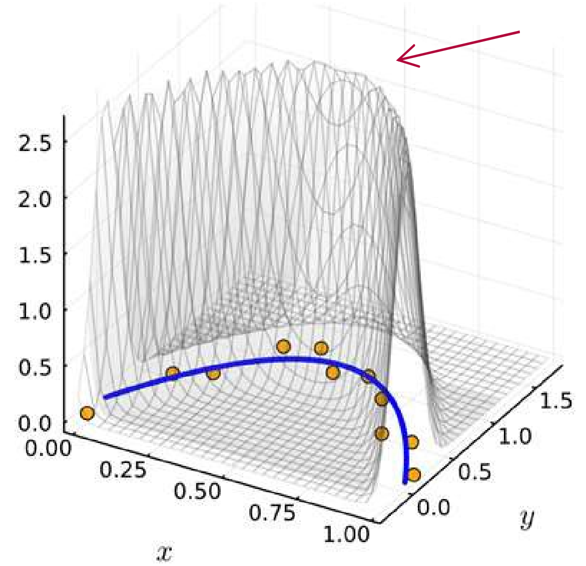
$$\overset{\substack{x=0.5 \\ \sigma_D^2=0.02}}{=} N(y; w_1 x + w_2 x^2, 0.02) \overset{\substack{w_1=3.75 \\ w_2=-3.5}}{=} N(y; 1, 0.02)$$

Decision Making: Min Expected Loss (Unit 2, slide 19-21)

What is the **Best Action a** (Supply) to meet uncertain **Demand y**?

Input (Temp.): $x = 0.5$ Pred. Output (Demand): $N(y;1,0.02)$

Loss: Ideas?



Decision Making: Min Expected Loss (Unit 2, slide 19-21)

What is the **Best Action a** (Supply) to meet uncertain **Demand y**?

Input (Temp.): $x = 0.5$

Pred. Output (Demand): $N(y; 1, 0.02)$

Loss:

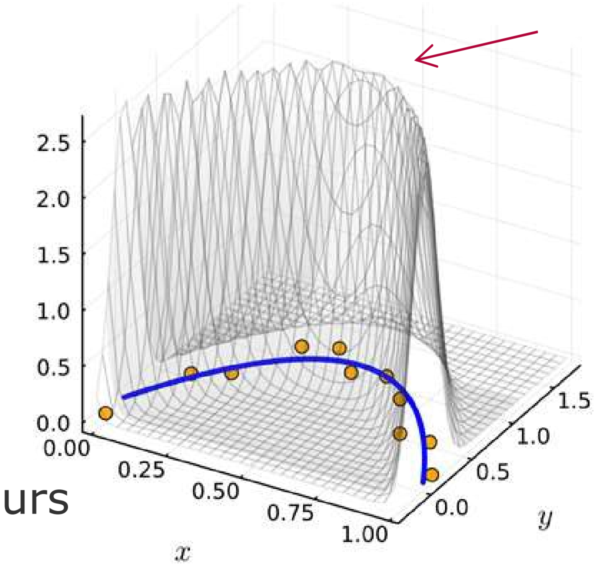
$$l(y,a) = \begin{cases} (y-a)^2 & , y > a \\ a - y & , y \le a \end{cases}$$

angry customers & workers

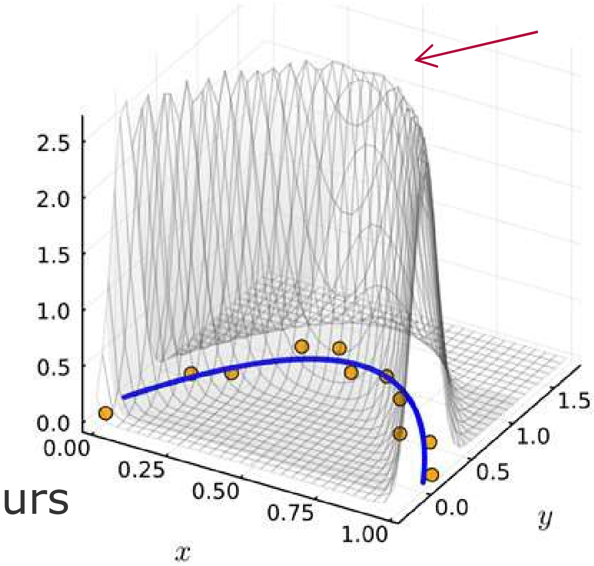
not enough to do & wasted hours

Expected Loss:

$$EL(a) = \int_0^2 l(y,a) \cdot N(y; 1, 0.02)\ dy + a$$



Decision Making: Min Expected Loss (Unit 2, slide 19-21)



What is the **Best Action a** (Supply) to meet uncertain **Demand y**?

Input (Temp.): $x = 0.5$ Pred. Output (Demand): $N(y;1,0.02)$

Loss:
$$l(y,a) = \begin{cases} (y-a)^2 & , y > a \\ a - y & , y \le a \end{cases}$$
angry customers & workers ($y > a$)
not enough to do & wasted hours ($y \le a$)

Expected Loss: $$EL(a) = \int_0^2 l(y,a) \cdot N(y;1,0.02)\ dy + a$$

Cost of Best Decision: $EL(a^*) = EL(1.16) = 1.44$

Cost of Decision *a=E(y)*: $EL(E(y)) = EL(1) = 2.06 \quad (+43\%)$

Overview

1. Course Setup
2. Recap: Probabilities and Distributions (Unit 1)
3. Recap: Main Concepts of Unit 2
4. Example: Bayesian Inference
5. Example: Multiplication of Normal Distributions
6. Example: Decision-Making
7. **Hints for Exercise 1 (to be handed in Monday May 5)**

Exercise 1 (until May 5)

- Part I: Discrete Uniform & Gaussians in Julia
- Part II: Generic Distributions in Julia
- Part III: Show Identities for Representations of Gaussians and verify the Multiplication Theorem ☺
- Part IV: Conjugacy of the Beta Distribution

Summary

- Recap I: Recap Expectation and Variance of Random Variables
- Recap II: Beta Distribution (Conjugacy, Discrete Likelihood)
- Recap III: Normal Distribution (Conjugacy, Continuous Likelihood)
- Recap IV: Decision-Making based on derived Posteria

See you next Week!