

# BAYESIAN LEARNING - LECTURE 1

Mattias Villani

**Division of Statistics and Machine Learning  
Department of Computer and Information Science  
Linköping University**

# COURSE OVERVIEW

- ▶ Four **modules** with:
  - ▶ Lectures (Mattias Villani)
  - ▶ Computer Labs (Shutong Ding)
- ▶ Modules:
  - ▶ The **basics**, single- and multiparameter models
  - ▶ **Regression** models
  - ▶ Tackling more advanced models with **MCMC**
  - ▶ **Flexible models** and **Model Inference**
- ▶ **Examination**
  - ▶ Lab reports, 2 credits
  - ▶ Bayesian project report, 4 credits
  - ▶ Oral exam (for some students, if needed)
- ▶ **Individual Bayesian project report**
  - ▶ Perform a Bayesian analysis on real data.
  - ▶ Shutong Ding is in charge.
  - ▶ Deadline **May 22, 2015**.

# LECTURE OVERVIEW

- ▶ The likelihood function
- ▶ Bayesian inference
- ▶ Bernoulli model
- ▶ Normal model with known variance

# THE LIKELIHOOD FUNCTION - BERNOULLI TRIALS

- **Bernoulli trials:**

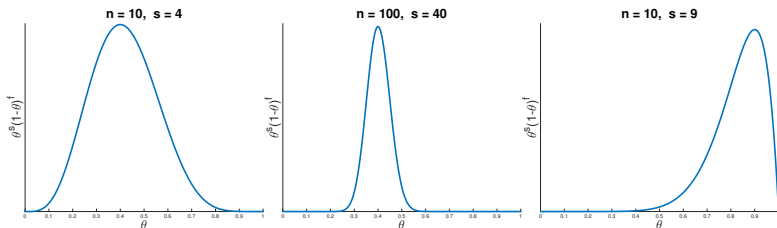
$$x_1, \dots, x_n | \theta \stackrel{iid}{\sim} \text{Bern}(\theta).$$

- **Likelihood** from  $s = \sum_{i=1}^n x_i$  successes and  $f = n - s$  failures.

$$p(x_1, \dots, x_n | \theta) = p(x_1 | \theta) \cdots p(x_n | \theta) = \theta^s (1 - \theta)^f$$

- **Maximum likelihood estimator**  $\hat{\theta}$  maximizes  $p(x_1, \dots, x_n | \theta)$ .

- Given the data  $x_1, \dots, x_n$ , we may plot  $p(x_1, \dots, x_n | \theta)$  as a function of  $\theta$ .



# THE LIKELIHOOD FUNCTION

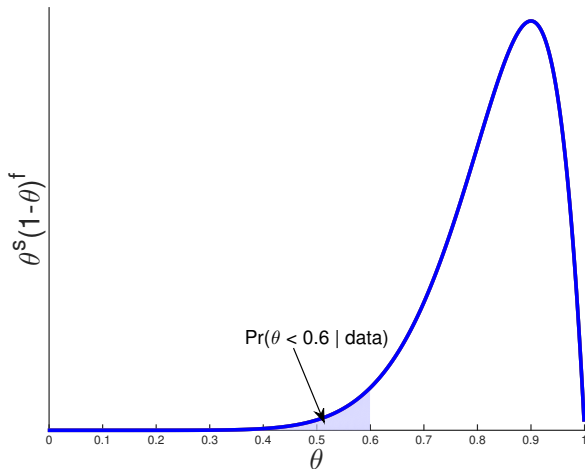
- ▶ Say it out loud:

*The likelihood function is  
the probability of the observed data  
considered as a function of the parameter.*

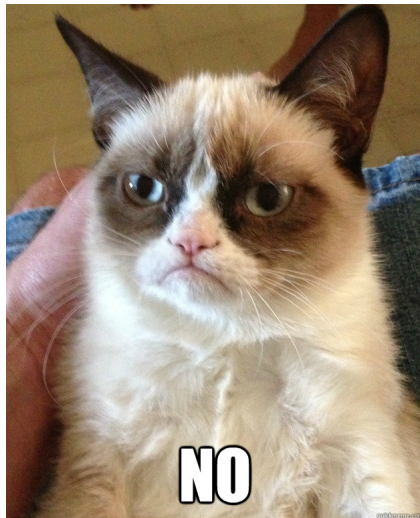
- ▶ The symbol  $p(x_1, \dots, x_n | \theta)$  plays two different roles:
- ▶ **Probability distribution** for the data.
  - ▶ The data  $x_1, \dots, x_n$ , are random.
  - ▶  $\theta$  is fixed.
- ▶ **Likelihood function** for the parameter
  - ▶ The data  $x_1, \dots, x_n$  are fixed.
  - ▶  $p(x_1, \dots, x_n | \theta)$  is function of  $\theta$ .

# PROBABILITIES FROM THE LIKELIHOOD!!

**n = 10, s = 9**

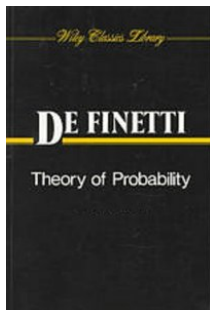


# PROBABILITIES FROM THE LIKELIHOOD!!



# UNCERTAINTY AND SUBJECTIVE PROBABILITY

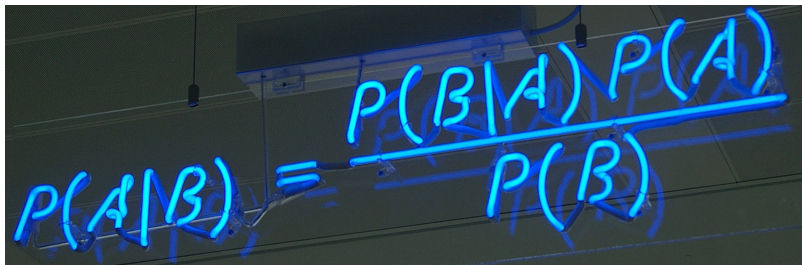
- ▶ Statements like  $\Pr(\theta < 0.6 | \text{data})$  only make sense if  $\theta$  is random.
- ▶ But  $\theta$  may be a fixed natural constant?
- ▶ **Bayesian: doesn't matter if  $\theta$  is fixed or random.**
- ▶ Do You know the value of  $\theta$  or not?
- ▶  $p(\theta)$  reflects Your knowledge/**uncertainty** about  $\theta$ .
- ▶ **Subjective probability.**
- ▶ The statement  $p(\text{10th decimal of } \pi = 9) = 0.1$  makes sense.





# BAYESIAN LEARNING

- ▶ **Bayesian learning** about a model parameter  $\theta$ :
  - ▶ state your **prior** knowledge about  $\theta$  as a probability distribution  $p(\theta)$ .
  - ▶ **collect data**  $x$  and form the **likelihood** function  $p(x|\theta)$ .
  - ▶ **combine** your prior knowledge  $p(\theta)$  with the data information  $p(x|\theta)$ .
- ▶ How to combine the two sources of information? **Bayes' theorem**.



A photograph of a chalkboard with the formula for Bayes' theorem written in blue chalk. The formula is 
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
 The chalk is bright blue, and the background is dark. There are some faint, illegible markings on the board, possibly from previous writing.

# LEARNING FROM DATA - BAYES' THEOREM

- ▶ How do we **update** from the **prior**  $p(\theta)$  to the **posterior**  $p(\theta|Data)$ ?
- ▶ **Bayes' theorem** for events  $A$  and  $B$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

- ▶ Bayes' Theorem for a model parameter  $\theta$

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

- ▶ The prior  $p(\theta)$  is the hero that converts the likelihood function  $p(Data|\theta)$  into a posterior probability density  $p(\theta|Data)$ .
- ▶ A probability distribution for  $\theta$  is extremely useful. **Decision making.**
- ▶ **No prior - no posterior - no useful inferences - no fun.**

# BAYES' THEOREM FOR MEDICAL DIAGNOSIS

- ▶  $A = \{\text{Horrible and very rare disease}\}$ ,  $B = \{\text{Positive medical test}\}$ .
- ▶  $p(A) = 0.0001$ .  $p(B|A) = 0.9$ .  $p(B|A^c) = 0.05$ .
- ▶ Probability of being sick given a positive test:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|A^c)p(A^c)} \approx 0.001797.$$

- ▶ Probably not sick, but 18 times more probable than before the test.
- ▶ Morale of the story: If you want  $p(A|B)$  then  $p(B|A)$  does not tell the whole story. The prior probability  $p(A)$  is also very important.

***“You can’t enjoy the Bayesian omelette  
without breaking the Bayesian eggs”***

*Leonard Jimmie Savage*



# THE NORMALIZING CONSTANT IS NOT IMPORTANT

- Bayes theorem

$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)} = \frac{p(Data|\theta)p(\theta)}{\int_{\theta} p(Data|\theta)p(\theta)d\theta}.$$

- The integral  $p(Data) = \int_{\theta} p(Data|\theta)p(\theta)d\theta$  can make you cry.
- $p(Data)$  is just a constant that makes  $p(\theta|Data)$  integrate to one.
- Example:  $x \sim N(\mu, \sigma^2)$

$$p(x) = (2\pi\sigma^2)^{-1/2} \exp \left[ -\frac{1}{2\sigma^2}(x - \mu)^2 \right].$$

- We may write

$$p(x) \propto \exp \left[ -\frac{1}{2\sigma^2}(x - \mu)^2 \right].$$

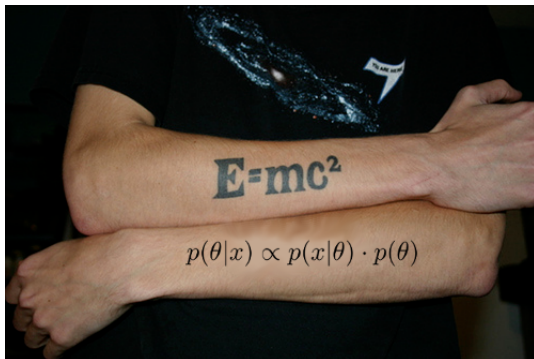
# GREAT THEOREMS MAKE GREAT TATTOOS

- All you need to know:

$$p(\theta|Data) \propto p(Data|\theta)p(\theta)$$

or

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



# BERNOULLI TRIALS - BETA PRIOR

## ► Model

$$x_1, \dots, x_n | \theta \stackrel{iid}{\sim} \text{Bern}(\theta)$$

## ► Prior

$$\theta \sim \text{Beta}(\alpha, \beta)$$

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1} \quad \text{for } 0 \leq \theta \leq 1.$$

## ► Posterior

$$\begin{aligned} p(\theta | x_1, \dots, x_n) &\propto p(x_1, \dots, x_n | \theta) p(\theta) \\ &\propto \theta^s (1 - \theta)^f \theta^{\alpha-1} (1 - \theta)^{\beta-1} \\ &= \theta^{s+\alpha-1} (1 - \theta)^{f+\beta-1}. \end{aligned}$$

- This is proportional to the  $\text{Beta}(\alpha + s, \beta + f)$  density.
- The **prior-to-posterior** mapping reads

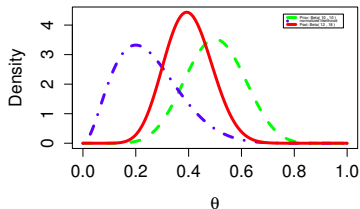
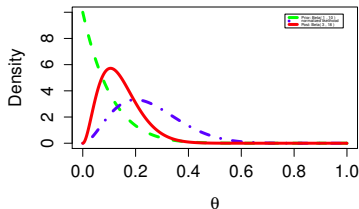
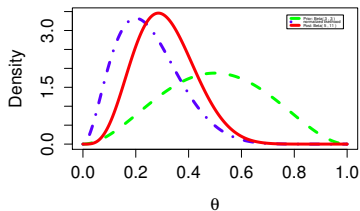
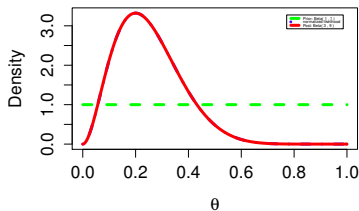
$$\theta \sim \text{Beta}(\alpha, \beta) \xrightarrow{x_1, \dots, x_n} \theta | x_1, \dots, x_n \sim \text{Beta}(\alpha + s, \beta + f).$$

# BERNOULLI EXAMPLE: SPAM EMAILS

- ▶ George has gone through his collection of 4601 e-mails. He classified 1813 of them to be spam.
- ▶ Let  $x_i = 1$  if  $i$ :th email is spam. Assume  $x_i|\theta \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$  and  $\theta \sim \text{Beta}(\alpha, \beta)$ .
- ▶ Posterior

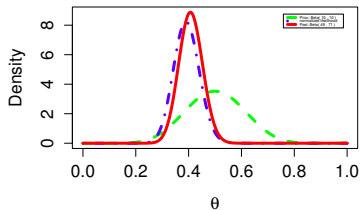
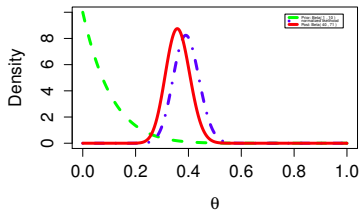
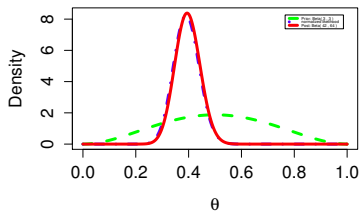
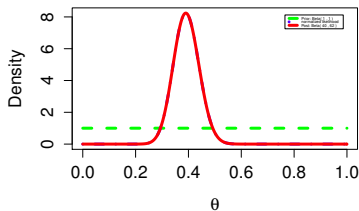
$$\theta|x \sim \text{Beta}(\alpha + 1813, \beta + 2788)$$

# SPAM DATA (N=10): PRIOR SENSITIVITY

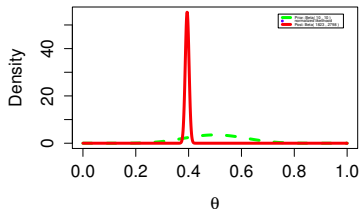
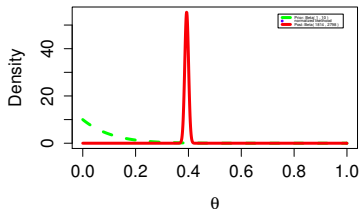
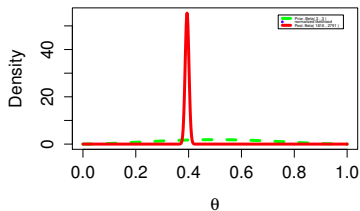
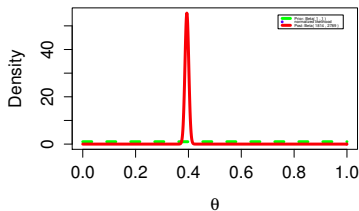




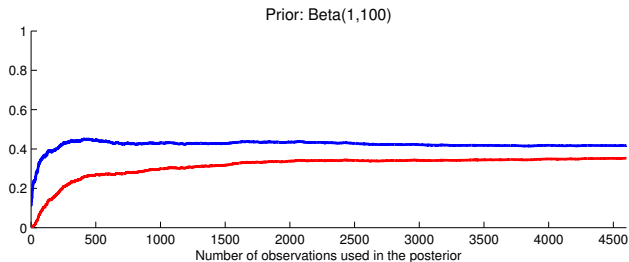
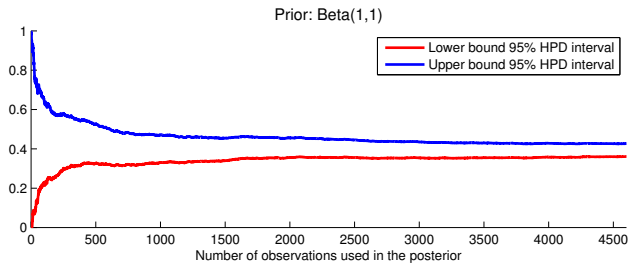
# SPAM DATA (N=100): PRIOR SENSITIVITY



# SPAM DATA (N=4601): PRIOR SENSITIVITY



# SPAM DATA: POSTERIOR CONVERGENCE



# NORMAL DATA, KNOWN VARIANCE - UNIFORM PRIOR

- ▶ Model:

$$x_1, \dots, x_n | \theta, \sigma^2 \stackrel{iid}{\sim} N(\theta, \sigma^2).$$

- ▶ Prior:

$$p(\theta) \propto c \text{ (a constant)}$$

- ▶ Likelihood

$$\begin{aligned} p(x_1, \dots, x_n | \theta, \sigma^2) &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp \left[ -\frac{1}{2\sigma^2} (x_i - \theta)^2 \right] \\ &\propto \exp \left[ -\frac{1}{2(\sigma^2/n)} (\theta - \bar{x})^2 \right]. \end{aligned}$$

- ▶ Posterior

$$\theta | x_1, \dots, x_n \sim N(\bar{x}, \sigma^2/n)$$

# NORMAL DATA, KNOWN VARIANCE - NORMAL PRIOR

► Prior

$$\theta \sim N(\mu_0, \tau_0^2)$$

► Posterior

$$\begin{aligned} p(\theta | x_1, \dots, x_n) &\propto p(x_1, \dots, x_n | \theta, \sigma^2) p(\theta) \\ &\propto N(\theta | \mu_n, \tau_n^2), \end{aligned}$$

where

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2},$$

$$\mu_n = w\bar{x} + (1 - w)\mu_0,$$

and

$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}.$$

# NORMAL DATA, KNOWN VARIANCE - NORMAL PRIOR

$$\theta \sim N(\mu_0, \tau_0^2) \xrightarrow{x_1, \dots, x_n} \theta|x \sim N(\mu_n, \tau_n^2).$$

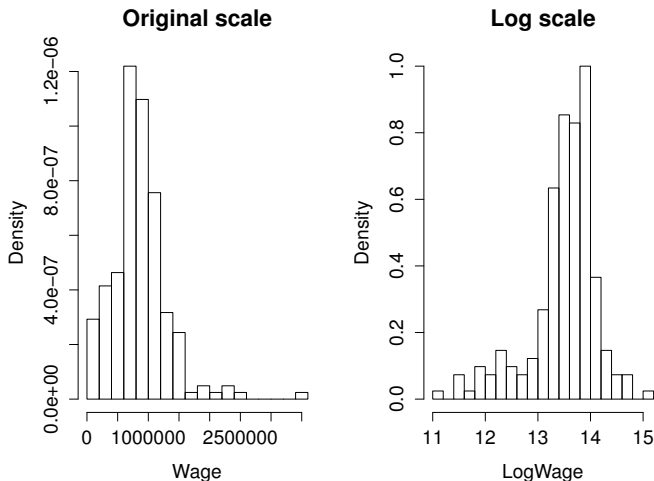
Posterior precision = Data precision + Prior precision

Posterior mean =

$$\frac{\text{Data precision}}{\text{Posterior precision}} (\text{Data mean}) + \frac{\text{Prior precision}}{\text{Posterior precision}} (\text{Prior mean})$$

# CANADIAN WAGES DATA

- Data on wages for 205 Canadian workers.



# CANADIAN WAGES

- ▶ Model

$$X_1, \dots, X_n | \theta \sim N(\theta, \sigma^2), \quad \sigma^2 = 0.4$$

- ▶ Prior

$$\theta \sim N(\mu_0, \tau_0^2), \quad \mu_0 = 12 \text{ and } \tau_0 = 10$$

- ▶ Posterior

$$\theta | x_1, \dots, x_n \sim N(\mu_n, \tau_n^2),$$

where  $\mu_n = w\bar{x} + (1 - w)\mu_0$ .

- ▶ For the Canadian wage data:

$$w = \frac{\sigma^{-2}n}{\sigma^{-2}n + \tau_0^{-2}} = \frac{2.5 \cdot 205}{2.5 \cdot 205 + 1/100} = 0.999.$$

$$\mu_n = w\bar{x} + (1 - w)\mu_0 = 0.999 \cdot 13.489 + (1 - 0.999) \cdot 12 \approx 13.489$$

$$\tau_n^2 = (2.5 \cdot 205 + 1/100)^{-1} = 0.00195$$



# CANADIAN WAGES DATA - MODEL FIT

