Introduction to Machine Learning
Topic 1: Bayesian learning

Lecture 1b

Mattias Villani

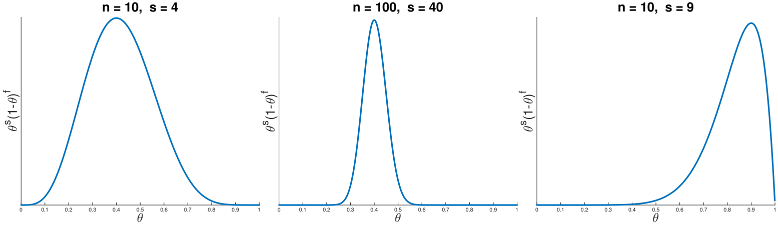
Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Overview of Lecture 1b

- Introduction to Bayesian learning
- Bernoulli model with beta prior
- Normal model with normal prior
- Multinomial model with Dirichlet prior

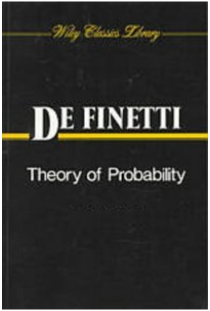
The likelihood function - Bernoulli trials

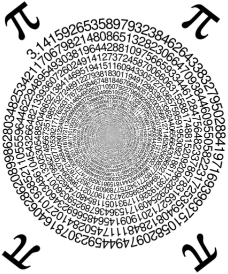
- **Bernoulli trials**:
$$x_1, ..., x_n|\theta \overset{iid}{\sim} Bern(\theta).$$
- **Likelihood** from $s = \sum_{i=1}^{n} x_i$ successes and $f = n - s$ failures.
$$p(x_1, ..., x_n|\theta) = p(x_1|\theta) \cdots p(x_n|\theta) = \theta^s (1-\theta)^f$$
- **Maximum likelihood estimator** $\hat{\theta}$ maximizes $p(x_1, ..., x_n|\theta)$.
- Given the data $x_1, ..., x_n$, we may plot $p(x_1, ..., x_n|\theta)$ as a function of θ.



Uncertainty and subjective probability

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





Bayesian learning

- **Bayesian learning** about a model parameter θ:
  - state your **prior** knowledge about θ 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.

Great theorems make great tattoos

- **Bayes' theorem**

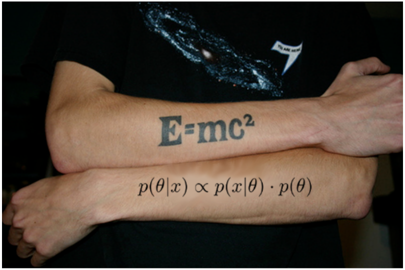
$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)}.$$

- 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**

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

- **Prior**

$$\theta \sim Beta(\alpha, \beta)$$

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

- **Posterior**

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

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

$$\theta \sim Beta(\alpha, \beta) \stackrel{x_1, ..., x_n}{\Longrightarrow} \theta | x_1, ..., x_n \sim Beta(\alpha+s, \beta+f).$$

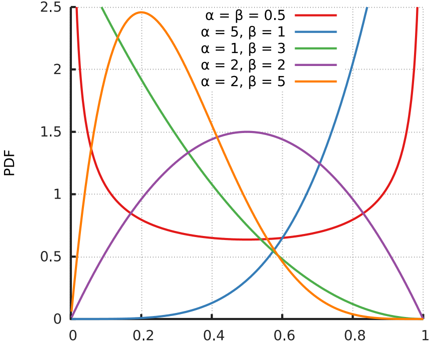
Beta distribution

- **Beta random variable**

$$X \sim Beta(\alpha, \beta)$$

- Probability density function (**pdf**)

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

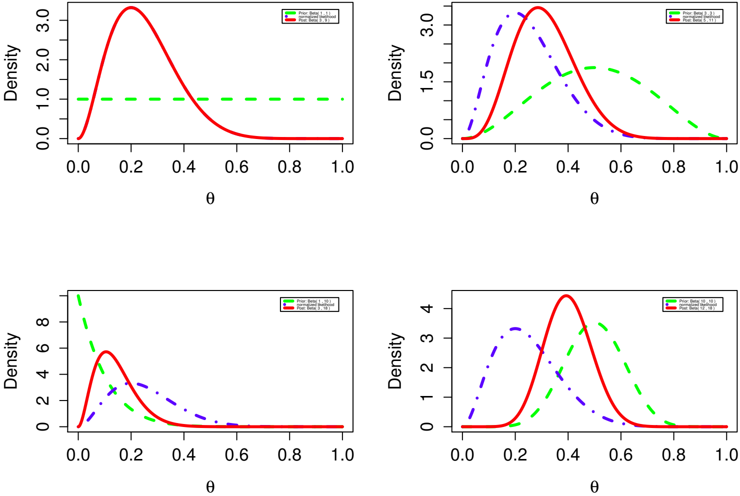


Bernoulli example: spam emails

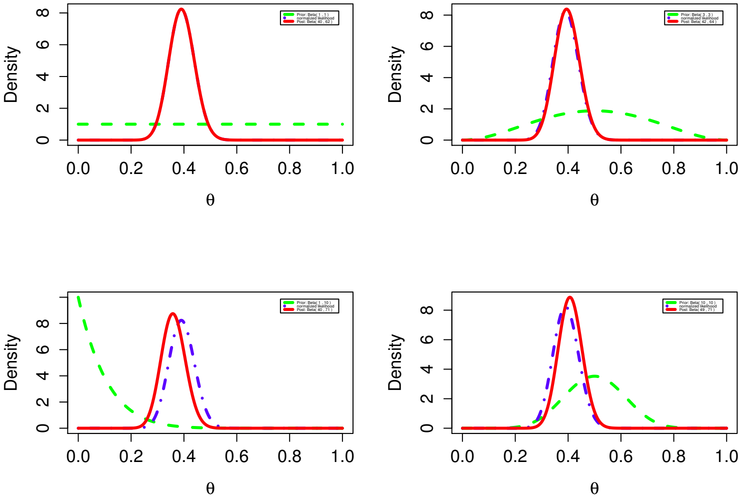
- George has gone through his collection of 4601 e-mails. He classified 1813 of them to be spam.
- Let $x_i = 1$ if i:th email is spam. Assume $x_i|\theta \overset{iid}{\sim} Bernoulli(\theta)$ and $\theta \sim \mathrm{Beta}(\alpha, \beta)$.
- Posterior

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

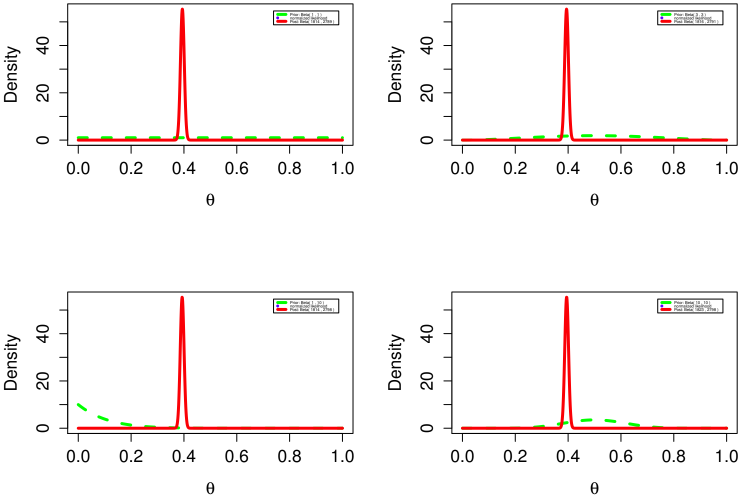
SPAM DATA (N=10): PRIOR SENSITIVITY



Spam data (n=100): prior sensitivity



Spam data (n=4601): prior sensitivity



Normal data, known variance - normal prior

- Prior

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

- Posterior

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

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

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

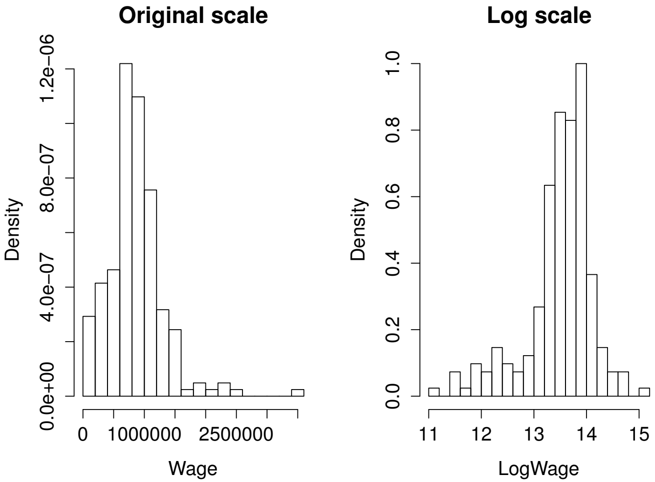
$$\text{Posterior precision} = \text{Data precision} + \text{Prior precision}$$

$$\text{Posterior mean} =$$

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

Canadian wages data

- Data on wages for 205 Canadian workers.



Canadian wages

- Model

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

- Prior

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

- Posterior

$$\theta | x_1, ..., x_n \sim N\left(\mu_n, \tau_n^2\right),$$

  where $\mu_n = w\bar{x} + (1-w)\mu_0$.
- For the Canadian wage data:

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

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

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

Marginalization

- Models with multiple parameters $\theta_1, \theta_2, ...$
- Examples: $x_i \stackrel{iid}{\sim} N(\theta, \sigma^2)$; multiple regression ...
- **Joint posterior distribution**

$$p(\theta_1, \theta_2, ..., \theta_p | y) \propto p(y | \theta_1, \theta_2, ..., \theta_p) p(\theta_1, \theta_2, ..., \theta_p).$$

... or in vector form:

$$p(\theta) \propto p(y|\theta)p(\theta).$$

- Complicated to graph the joint posterior.
- Some of the parameters may not be of direct interest (**nuisance**).
- Integrate out (**marginalize**) all nuisance parameters.
- Example: $\theta = (\theta_1, \theta_2)'$, θ_2 is a nuisance. **Marginal posterior** of θ_1

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y) d\theta_2.$$

Multinomial model with Dirichlet prior

- *Data*: $y = (y_1, ... y_K)$, where y_k counts the number of observations in the kth category. $\sum_{k=1}^{K} y_k = n$. Example: brand choices.
- **Multinomial model**:

$$p(y|\theta) \propto \prod_{k=1}^{K} \theta_k^{y_k}, \text{ where } \sum_{k=1}^{K} \theta_j = 1.$$

- ***Prior***: Dirichlet$(\alpha_1, ..., \alpha_K)$

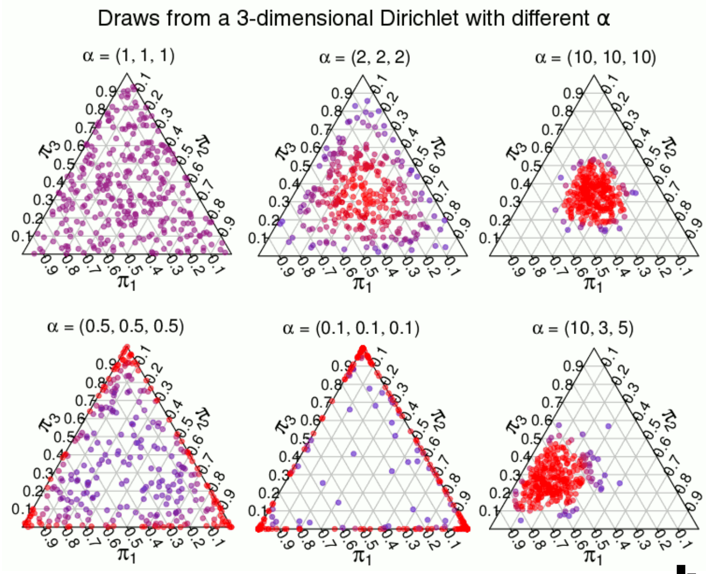
$$p(\theta) \propto \prod_{k=1}^{K} \theta_j^{\alpha_j - 1}.$$

- Moments of $\theta = (\theta_1, ..., \theta_K)' \sim Dirichlet(\alpha_1, ..., \alpha_K)$

$$\mathrm{E}(\theta_k) = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j}$$

- $\alpha_+ = \sum_{k=1}^{K} \alpha_k$ is a precision parameter. Variance of θ_k is large when α_+ is small.

DIRICHLET DISTRIBUTION



Multinomial model with Dirichlet prior

- 'Non-informative': $\alpha_1 = ... = \alpha_K = 1$ (uniform and proper).
- **Simulating** from the Dirichlet distribution:
  - Generate $x_1 \sim Gamma(\alpha_1, 1), ..., x_K \sim Gamma(\alpha_K, 1)$.
  - Compute $y_k = x_k / (\sum_{j=1}^{K} x_j)$.
  - $y = (y_1, ..., y_K)$ is a draw from the $\mathrm{Dirichlet}(\alpha_1, ..., \alpha_K)$ distribution.
- ***Prior-to-Posterior updating***:

$$\begin{aligned} \mathit{Model}: &\quad y = (y_1, ...y_K) \sim \mathrm{Multin}(n; \theta_1, ..., \theta_K) \\ \mathit{Prior}: &\quad \theta = (\theta_1, ..., \theta_K) \sim \mathrm{Dirichlet}(\alpha_1, ..., \alpha_K) \\ \mathit{Posterior}: &\quad \theta | y \sim \mathrm{Dirichlet}(\alpha_1 + y_1, ..., \alpha_K + y_K). \end{aligned}$$

Example: market shares

- A recent survey among consumer smartphones owners in the U.S. showed that among the 513 respondents:
  - 180 owned an iPhone
  - 230 owned an Android phone
  - 62 owned a Blackberry phone
  - 41 owned some other mobile phone.
- Previous survey: iPhone 30%, Android 30%, Blackberry 20% and Other 20%.
- Pr(Android has largest share | Data)
- Prior: $\alpha_1 = 15, \alpha_2 = 15, \alpha_3 = 10$ and $\alpha_4 = 10$ (prior info is equivalent to a survey with only 50 respondents)
- Posterior: $(\theta_1, \theta_2, \theta_3, \theta_4)|\mathbf{y} \sim \text{Dirichlet}(195, 245, 72, 51)$

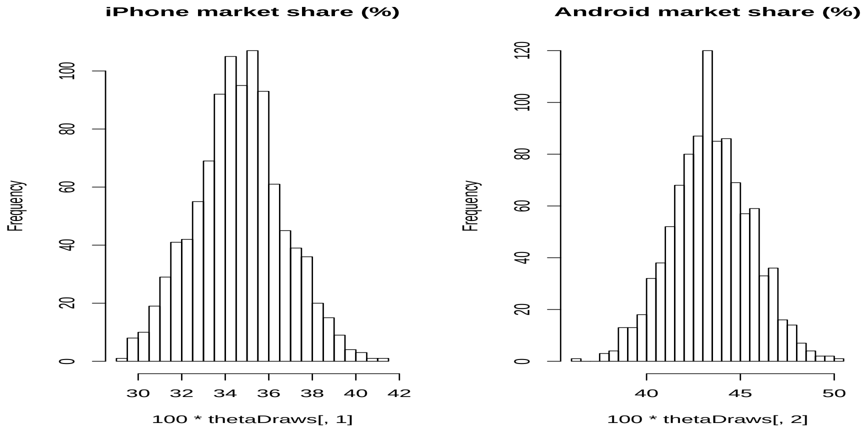
R code for market share example

```r
# Setting up data and prior
y <- c(180,230,62,41) # The cell phone survey data (K=4)
alpha <- c(15,15,10,10) # Dirichlet prior hyperparameters
nIter <- 1000 # Number of posterior draws

# Defining a function that simulates from a Dirichlet distribution
SimDirichlet <- function(nIter, param){
  nCat <- length(param)
  thetaDraws <- as.data.frame(matrix(NA, nIter, nCat)) # Storage.
  for (j in 1:nCat){
    thetaDraws[,j] <- rgamma(nIter,param[j],1)
  }
  for (i in 1:nIter){
    thetaDraws[i,] = thetaDraws[i,]/sum(thetaDraws[i,])
  }
  return(thetaDraws)
}

# Posterior sampling from Dirichlet posterior
thetaDraws <- SimDirichlet(nIter,y + alpha)
```

R code for market share example, cont

```r
# Posterior mean and standard deviation of Androids share (in %)
message(mean(100*thetaDraws[,2]))

## 43.4557585065132

message(sd(100*thetaDraws[,2]))

## 2.12277847008832

# Computing the posterior probability that Android is the largest
PrAndroidLargest <- sum(thetaDraws[,2] > max(thetaDraws[,c(1,3,4)]))/nIter
message(paste('Pr(Android has the largest market share) = ', PrAndroidLargest))

## Pr(Android has the largest market share) =  0.835
```

R code for market share example, cont



Bayesian prediction

- Example: Supervised learning. Model: $x \to y$.
- **Posterior predictive distribution**

$$p(y_{test}|x_{test}, y_{train}, x_{train}) = \int_{\mathbf{w}} p(y_{test}|\mathbf{w}, x_{test})p(\mathbf{w}|y_{train}, x_{train})d\mathbf{w}$$

  where

  - $p(y_{test}|\mathbf{w}, x_{test})$ is the predictive distribution from the model if the parameters $\mathbf{w}$ are known.
  - $p(\mathbf{w}|y_{train}, x_{train})$ is the posterior distribution of the model parameters $\mathbf{w}$

- The **parameter uncertainty** is represented in the predictive distribution by **averaging over** $p(\mathbf{w}|y_{train}, x_{train})$.
- Compute the predictive distribution by **simulation**. Iterate:
  - Simulate a random parameter draw $\tilde{\mathbf{w}}$ from posterior $p(\mathbf{w}|y_{train}, x_{train})$
  - Simulate a y_{test} from $p(y_{test}|\mathbf{w} = \tilde{\mathbf{w}}, x_{test})$.