MSiA 400 Lab 5

Bayesian Statistics

Huiyu Wu

10/31/2022

NORTHWESTERN
UNIVERSITY

Frequentist vs Bayesian

- Frequentist: unknown parameter is fixed
- Bayesian: unknown parameter has some distribution, this distribution reflects our belief of its uncertainty
- Coin in hand
- Parameter estimation

Frequentist approach

- MLE: maximum likelihood estimator
- $L(\theta; data) = p(\text{data}|\theta)$
- $\hat{\theta} = argmax : L(\theta; data)$

Bayesian approach

- Bayes' Rule $p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx} \propto p(y|x)p(x)$
- Want to deduce density from data $P(\text{parameters} \mid \text{data})$
- Can be expressed as $P(\text{data} \mid \text{parameters})$ times prior belief about parameters $P(\text{parameters})$

Terms

- $P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})P(\text{model})}{P(\text{data})}$
  - **Prior** $P(\text{model})$
  - **Likelihood** $P(\text{data} \mid \text{model})$
  - **Posterior** $P(\text{model} \mid \text{data})$

Coin Flipping Example

- $P(\text{heads} \mid \theta) = \theta$
- $P(\text{tails} \mid \theta) = 1 - \theta$
- $P(\textit{thhhttth}... \mid \theta) = \theta^{\#h}(1-\theta)^{\#t}$

Maximum Likelihood (ML) Estimate

- Log-likelihood: $\#h \log \theta + \#t \log(1-\theta)$
- Differentiate wrt parameter: $\frac{\#h}{\theta} - \frac{\#t}{1-\theta}$
- Equate to 0 & solve: $\theta = \frac{\#h}{\#h+\#t}$

Constrained Optimization Approach

Log-likelihood formulation:

$$\max_{\theta} \#h \log \theta_h + \#t \log \theta_t \text{ s.t. } \theta_h + \theta_t = 1$$

The Lagrangian is:

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = \#h \log \theta_h + \#t \log \theta_t - \lambda(1 - \theta_h - \theta_t)$$

KKT conditions are $\nabla \mathcal{L} = 0$ and $\theta_h + \theta_t = 1$. Start with the gradient of the Lagrangian:

$$\left(\frac{\#h}{\theta_h}, \frac{\#t}{\theta_t} \right) - \lambda(1, 1) = 0$$

Constrained Optimization Approach

This yields a system of equations:

$$\begin{cases} \frac{\#h}{\lambda} = \theta_h \\ \frac{\#t}{\lambda} = \theta_t \\ \theta_h + \theta_t = 1 \end{cases}$$

Solve for λ:

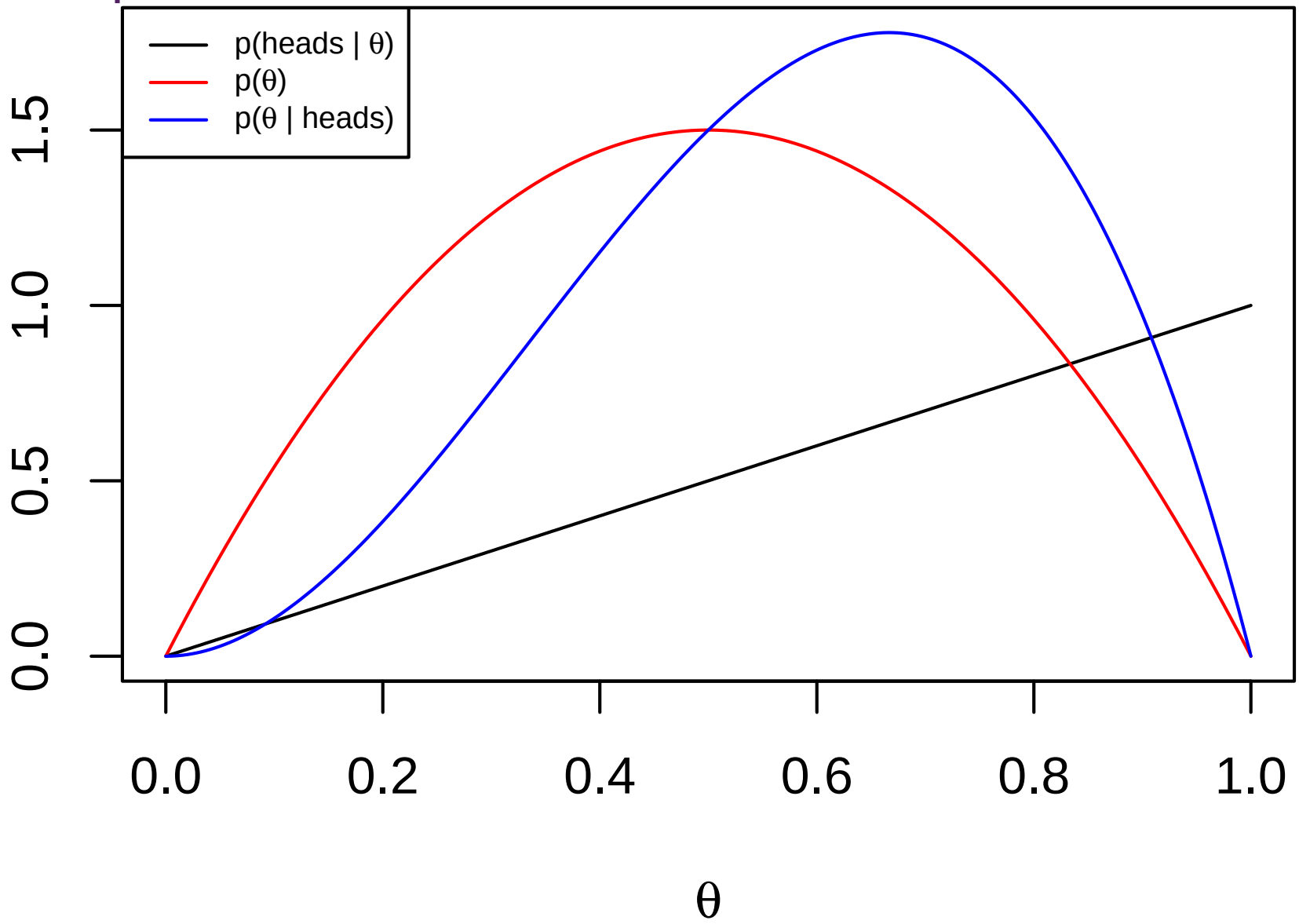
$$\frac{\#h}{\lambda} + \frac{\#t}{\lambda} = 1$$

$$\lambda = \#h + \#t$$

Solve for θ_t, θ_h:

$$\begin{cases} \frac{\#h}{\#h+\#t} = \theta_h \\ \frac{\#t}{\#h+\#t} = \theta_t \end{cases}$$

One Flip, Coin Lands on Heads



Bayes Method

- Prior: $P(\theta) = \text{Beta}(\alpha_h, \alpha_t) \propto \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1},\ \alpha_h, \alpha_t > 0$
  - Prior prediction: $E[\theta] = \frac{\alpha_h}{\alpha_h + \alpha_t}$
- Posterior: $P(\theta \mid \#h, \#t) \propto \theta^{\#h + \alpha_h - 1}(1-\theta)^{\#t + \alpha_t - 1}$
  - Posterior prediction: $E[\theta] = \frac{\#h + \alpha_h}{\#h + \alpha_h + \#t + \alpha_t}$
- Maximum a posteriori (MAP) estimate
  - Similar to ML, but incorporates prior distribution
  - MAP prediction: $\theta = \frac{\#h + \alpha_h - 1}{\#h + \alpha_h + \#t + \alpha_t - 2}$
- What about a flat prior?

Intuition

- α_h, α_t can be thought of as imaginary counts based on our prior beliefs
- The equivalent sample size is $\alpha_h + \alpha_t$
    - The larger the sum, the more confident we are in our prior

Naïve Bayes

- Classifier $f(x) = \arg\max_y P(Y = y \mid X = x,)$
- Learn distribution $P(Y = y \mid X = x)$ from data
- Assumes features are independent given class

Example: Your first NLP

```
Hello,

My name is *****, an investment manager here in Asia,USA and Canada; we represent the interests
of very wealthy Investors mainly from all parts of the world.

Due to the sensitivity of their position they hold in their Organization and the unstable investment
environment of their countries they prefer to channel/move majority of their funds into more stable
economies and developing nations where they can get good yield for their money and its safety.

Kindly let us know your acceptance to this offer we will then provide you with all necessary information
including the amount involved at this moment our Investor have up to 3 Billions dollars set aside for
investments, we will also like to know what project you have in mind or at hand that needs funding
because our investor is a very serious
one who don't have time for games.

waiting to read back from you on my private Investment email: *****@gmail.com  Next to Investor.

Regards
*****
Email: *****@gmail.com
+123456789101

--
This email has been checked for viruses by Avast antivirus software.
https://www.avast.com/antivirus
```

- Spam or not Spam?

Bag of Words

- Transform message to counts of occurences for each word:

```
words = strsplit(message, split=" ") #split message into words
bagOfWords = table(words) #get counts of each word
# sort
bagOfWords = bagOfWords[names(sort(bagOfWords,decreasing=T))]
bagOfWords[1:7]
```

```
## words
##    to    of   the their   and   for email
##     7     5     5     5     4     4     3
```

Bag of Words

- Ignores word order
  - No emphasis on title/headers
  - No compositional meaning
  - Etc.
- But, massively reduces complexity
  - In spam classification: $2N + 1$ parameters, where N is the size of your dictionary, as opposed to infinite

    $$P(\text{Spam} \mid \text{Message}) = \frac{\prod_i P(\text{Word}_i \mid \text{Spam}) P(\text{Spam})}{P(\text{Message})}$$
- Works well in practice
  - Use sum of logs to avoid underflow

Parameter Estimates

For if word is present or not in document:

- $\theta_{ij} = P(w_i \mid \text{class} = j)$
- ML: $\theta_{ij} = \frac{\#\text{ docs in class } j \text{ with word } i}{\#\text{ docs in class} j}$
  - Not useful if certain words rarely (or never) occur in a certain class
- MAP: $\theta_{ij} = \frac{\#\text{ docs in class } j \text{ with word } i+1}{\#\text{ docs in class } j + 2}$

Hidden Markov Model

- Discrete-time stochastic prosses with hidden state
- Transition probabilities $t_{jk} = P(\pi_{i+1} = k \mid \pi_i = j)$
- Emission probabilities $e_{jk} = P(x_i = k \mid \pi_i = j)$

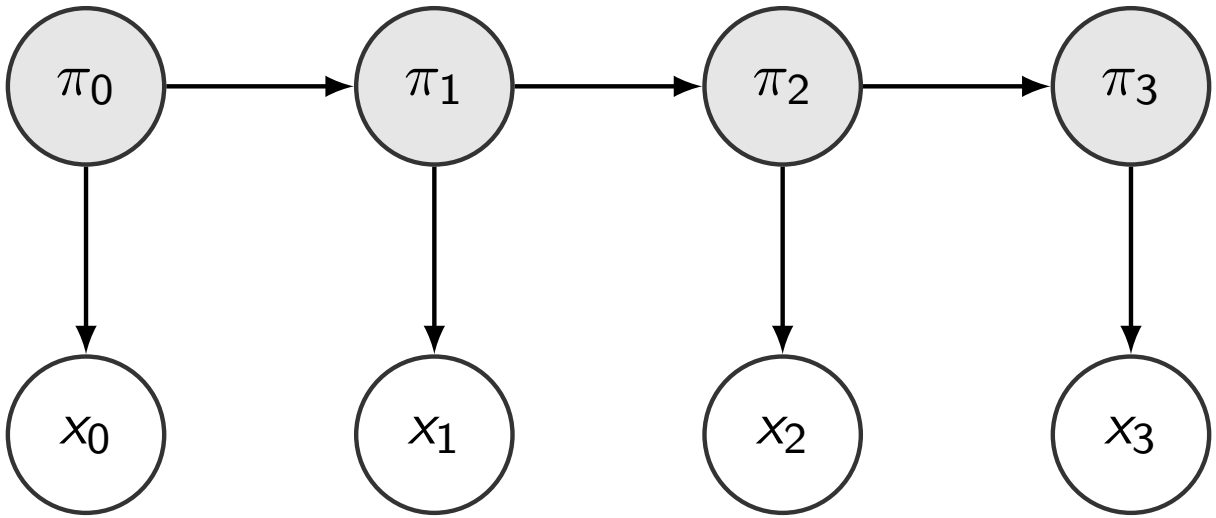


Figure 1: HMM

Learning Methods

- Expectation Maximization (EM)
  - Variational Inference
- Gibbs Sampling

Gaussian Mixture Model

$$p(x) = \sum_{k=1}^{K} \phi_k \mathcal{N}(x|\mu_k, \sigma_k)$$

$$\mathcal{N}(x|\mu_k, \sigma_k) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

$$\sum_{k=1}^{K} \phi_k = 1$$

EM for Gaussian Mixture Model

Expectation (E) Step:

$$\hat{\gamma}_{ik} = \frac{\hat{\phi}_k \mathcal{N}(x_i|\hat{\mu}_k, \hat{\sigma}_k)}{\sum_{j=1}^{K} \hat{\phi}_j \mathcal{N}(x_i|\hat{\mu}_j, \hat{\sigma}_j)}$$

Maximization (M) Step:

$$\hat{\phi}_k = \frac{1}{N} \sum_{i=1}^{N} \hat{\gamma}_{ik}$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik} x_i}{\sum_{i=1}^{N} \hat{\gamma}_{ik}}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik} \left(x_i - \hat{\mu_k}\right)^2}{\sum_{i=1}^{N} \hat{\gamma}_{ik}}$$