Essential Probability & Statistics

(Lecture for CS510 Advanced Topics in Information Retrieval)

ChengXiang Zhai

Department of Computer Science

University of Illinois, Urbana-Champaign

Prob/Statistics & Text Management

- Probability & statistics provide a principled way to quantify the uncertainties associated with natural language

- Allow us to answer questions like:
  - **Given that we observe "baseball" three times and "game" once in a news article, how likely is it about "sports"? (text categorization, information retrieval)**
  - **Given that a user is interested in sports news, how likely would the user use "baseball" in a query? (information retrieval)**

Basic Concepts in Probability

- **Random experiment: an experiment with uncertain outcome** (e.g., tossing a coin, picking a word from text)
- **Sample space: all possible outcomes, e.g.,**
  - Tossing 2 fair coins, S ={HH, HT, TH, TT}
- **Event: $E \subseteq S$, E happens iff outcome is in E, e.g.,**
  - E={HH} (all heads)
  - E={HH,TT} (same face)
  - Impossible event ({}), certain event (S)
- **Probability of Event : $1 \geq P(E) \geq 0$, s.t.**
  - P(S)=1 (outcome always in S)
  - $P(A \cup B)=P(A)+P(B)$ if $(A \cap B)=\emptyset$ (e.g., A=same face, B=different face)

Basic Concepts of Prob. (cont.)

- **Conditional Probability :$P(B|A)=P(A\cap B)/P(A)$**
  - $P(A\cap B) = P(A)P(B|A) = P(B)P(A|B)$
  - So, $P(A|B)=P(B|A)P(A)/P(B)$ (**Bayes' Rule**)
  - For **independent events**, $P(A\cap B) = P(A)P(B)$, so $P(A|B)=P(A)$
- **Total probability: If $A_1, \ldots, A_n$ form a partition of S, then**
  - $P(B)= P(B\cap S)=P(B\cap A_1)+\ldots+P(B \cap A_n)$ (why?)
  - So, $P(A_i|B)=P(B|A_i)P(A_i)/P(B)$

    $= P(B|A_i)P(A_i)/[P(B|A_1)P(A_1)+\ldots+P(B|A_n)P(A_n)]$
  - This allows us to compute $P(A_i|B)$ based on $P(B|A_i)$

Interpretation of Bayes' Rule

Hypothesis space: $H=\{H_1, \ldots, H_n\}$

Evidence: E

$$P(H_i \mid E) = \frac{P(E \mid H_i) P(H_i)}{P(E)}$$

If we want to pick the most likely hypothesis H*, we can drop P(E)

Posterior probability of H_i

Prior probability of H_i

$$P(H_i \mid E) \propto P(E \mid H_i) P(H_i)$$

Likelihood of data/evidence if H_i is true

Random Variable

- **X: S $\rightarrow \Re$ ("measure" of outcome)**
  - **E.g., number of heads, all same face?, …**
- **Events can be defined according to X**
  - **$E(X=a) = \{s_i | X(s_i)=a\}$**
  - **$E(X \geq a) = \{s_i | X(s_i) \geq a\}$**
- **So, probabilities can be defined on X**
  - **$P(X=a) = P(E(X=a))$**
  - **$P(a \geq X) = P(E(a \geq X))$**
- Discrete vs. continuous random variable (think of "partitioning the sample space")

An Example: Doc Classification

Sample Space $S=\{x_1,\ldots, x_n\}$

For 3 topics, four words, n=?

	Topic	the	computer	game	baseball
X_1:	[sport	1	0	1	1]
X_2:	[sport	1	1	1	1]
X_3:	[computer	1	1	0	0]
X_4:	[computer	1	1	1	0]
X_5:	[other	0	0	1	1]

Events · · · · · ·

$E_{sport} = \{x_i \mid topic(x_i) = \text{“sport”}\}$

$E_{baseball} = \{x_i \mid baseball(x_i) = 1\}$

$E_{baseball, \neg computer} =$
$\{x_i \mid baseball(x_i) = 1 \ \& \ computer(x_i) = 0\}$

Conditional Probabilities:

$P(E_{sport} \mid E_{baseball})$, $P(E_{baseball} \mid E_{sport})$,
$P(E_{sport} \mid E_{baseball, \neg computer})$, ...

Thinking in terms of random variables

Topic: $T \in \{\text{“sport”}, \text{“computer”}, \text{“other”}\}$,
“Baseball”: $B \in \{0,1\}$, …
$P(T=\text{“sport”} \mid B=1)$, $P(B=1 \mid T=\text{“sport”})$, ...

An inference problem:

Suppose we observe that “baseball” is mentioned, how likely the topic is about “sport”?

$P(T=\text{“sport”} \mid B=1) \propto P(B=1 \mid T=\text{“sport”}) P(T=\text{“sport”})$

But, $P(B=1 \mid T=\text{“sport”})=?$, $P(T=\text{“sport”})=?$

Getting to Statistics ...

- P(B=1|T="sport")=? (parameter estimation)
  - If we see the results of a huge number of random experiments, then

$$\hat{P}(B=1 \mid T=\text{"sport"}) = \frac{count(B=1, T=\text{"sport"})}{count(T=\text{"sport"})}$$

  - But, what if we only see a small sample (e.g., 2)? Is this estimate still reliable?
- In general, statistics has to do with drawing conclusions on the whole population based on observations of a sample (data)

Parameter Estimation

- General setting:
  - Given a (hypothesized & probabilistic) model that governs the random experiment
  - The model gives a probability of any data $p(D|\theta)$ that depends on the parameter θ
  - Now, given actual sample data $X=\{x_1,\ldots,x_n\}$, what can we say about the value of θ?
- Intuitively, take your best guess of θ -- "best" means "best explaining/fitting the data"
- Generally an optimization problem

Maximum Likelihood vs. Bayesian

- **Maximum likelihood estimation**
  - **"Best" means "data likelihood reaches maximum"**

  - **Problem: small sample**

$$\hat{\theta} = \arg\max_{\theta} P(X \mid \theta)$$

- **Bayesian estimation**
  - **"Best" means being consistent with our "prior" knowledge and explaining data well**

  - **Problem: how to define prior?**

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid X) = \arg\max_{\theta} P(X \mid \theta) P(\theta)$$

Maximum a Posteriori (MAP) estimate

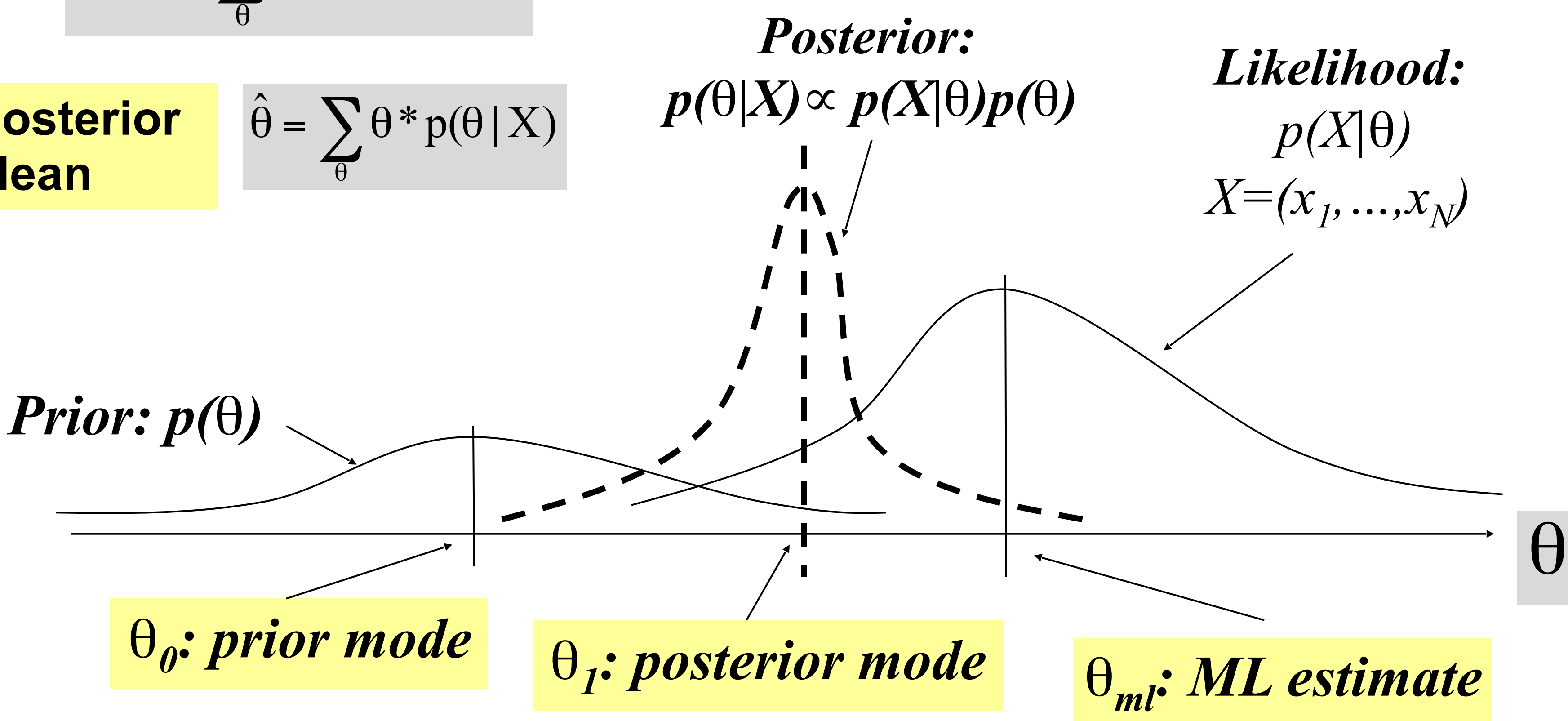
Illustration of Bayesian Estimation

Bayesian inference: f(θ)=?

$$\hat{f}(\theta) = \sum_{\theta} f(\theta) p(\theta \mid X)$$

Posterior Mean

$$\hat{\theta} = \sum_{\theta} \theta * p(\theta \mid X)$$



Maximum Likelihood Estimate

Data: a document d with counts $c(w_1)$, …, $c(w_N)$, and length |d|

Model: multinomial distribution M with parameters $\{p(w_i)\}$

Likelihood: p(d|M)

Maximum likelihood estimator: M=argmax $_M$ p(d|M)

$$p(d \mid M) = \binom{|d|}{c(w_1)\ldots c(w_N)} \prod_{i=1}^{N} \theta_i^{c(w_i)} \propto \prod_{i=1}^{N} \theta_i^{c(w_i)} \qquad where, \theta_i = p(w_i) \quad \sum_{i=1}^{N} \theta_i = 1$$

$$l(d \mid M) = \log\ p(d \mid M) = \sum_{i=1}^{N} c(w_i) \log \theta_i$$

We'll tune $p(w_i)$ to maximize l(d|M)

$$l'(d \mid M) = \sum_{i=1}^{N} c(w_i) \log \theta_i + \lambda\left(\sum_{i=1}^{N} \theta_i - 1\right)$$

Use Lagrange multiplier approach

$$\frac{\partial l'}{\partial \theta_i} = \frac{c(w_i)}{\theta_i} + \lambda = 0 \quad \Rightarrow \quad \theta_i = -\frac{c(w_i)}{\lambda}$$

Set partial derivatives to zero

$$Since \sum_{i=1}^{N} \theta_i = 1, \lambda = -\sum_{i=1}^{N} c(w_i) = -|d| \qquad So, \boxed{\theta_i = p(w_i) = \frac{c(w_i)}{|d|}}$$

ML estimate

What You Should Know

- Probability concepts:
  - sample space, event, random variable, conditional prob. multinomial distribution, etc
- Bayes formula and its interpretation
- Statistics: Know how to compute maximum likelihood estimate