Maximum Likelihood vs. Bayesian Parameter Estimation

Ronald J. Williams
CSG 220
Spring 2007

Contains numerous slides downloaded from
www.cs.huji.ac.il/course/2003/pmai/tirguls/tirgul10.ppt
(apparently authored by Nir Friedman)

Example: Binomial Experiment (Statistics 101)

- When tossed, it can land in one of two positions: *Head* or *Tail*
- We denote by θ the (unknown) probability $P(H)$.

Estimation task:

- Given a sequence of toss samples $x[1], x[2], \ldots, x[M]$ we want to estimate the probabilities $P(H)= \theta$ and $P(T) = 1 - \theta$

2

Statistical Parameter Fitting

- Consider instances ***x[1], x[2], …, x[M]*** such that
  - The set of values that x can take is known
  - Each is sampled from the same distribution
  - Each sampled independently of the rest

  (i.i.d. samples)

- Here we focus on multinomial distributions
  - Only finitely many possible values for x
  - Special case: binomial, with values H(ead) and T(ail)

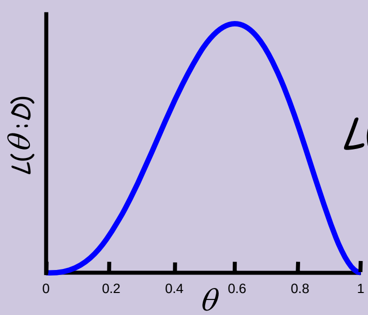
3

The Likelihood Function

- How good is a particular θ?
  It depends on how likely it is to generate the observed data

$$L(\theta : D) = P(D \mid \theta) = \prod_m P(x[m] \mid \theta)$$

- The likelihood for the sequence H,T, T, H, H is



$$L(\theta : D) = \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta$$

4

Maximum Likelihood Estimation

MLE Principle:

Choose parameters that maximize the likelihood function

- This is one of the most commonly used estimators in statistics
- Intuitively appealing

Example: MLE in Binomial Data

- It can be shown that the MLE for the probability of heads is given by

$$\hat{\theta} = \frac{N_H}{N_H + N_T}$$

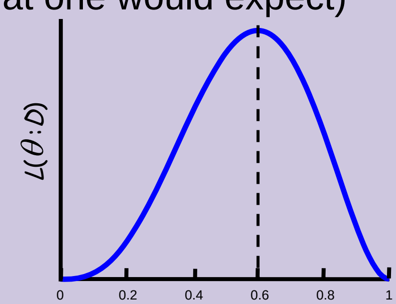
We prove this after the next slide

(which coincides with what one would expect)

Example:

$(N_H, N_T) = (3,2)$

MLE estimate is $3/5 = 0.6$



From Binomial to Multinomial

- For example, suppose X can have the values $1,2,\ldots,K$
- We want to learn the parameters $\theta_1, \theta_2, \ldots, \theta_K$

Observations:

- $N_1, N_2, \ldots, N_K$ - the number of times each outcome is observed

Likelihood function: $$L(\Theta : D) = \prod_{k=1}^{K} \theta_k^{N_k}$$

MLE: $$\hat{\theta}_k = \frac{N_k}{\sum_\ell N_\ell}$$

We prove this on next several slides

7

MLE for Multinomial

Theorem: For the multinomial distribution, the MLE for the probability P(x=k) is given by

$$\hat{\theta}_k = \frac{N_k}{\sum_\ell N_\ell}$$

Proof: The likelihood function is $$L(\Theta : D) = \prod_{k=1}^{K} \theta_k^{N_k}$$

To maximize it, it is equivalent to maximize the log-likelihood

$$LL(\theta_1, \theta_2, \ldots, \theta_K) = \ln L = \sum_\ell N_\ell \ln \theta_\ell$$

But we must impose the constraints

$\sum_\ell \theta_\ell = 1$ and $\theta_\ell \geq 0 \quad \forall \ell$

8

MLE for Multinomial (cont.)

We use the method of Lagrange multipliers.

Since there is one constraint equation, we introduce one Lagrange multiplier λ

We want to find $\theta_1, \theta_2, \ldots, \theta_K$ and λ so that the Lagrangian function

$$G(\theta_1, \ldots, \theta_K; \lambda) = LL(\theta_1, \ldots, \theta_K) - \lambda \left(\sum_{\ell} \theta_{\ell} - 1 \right)$$

attains a maximum as the θ_k values vary (and a minimum as λ varies).

9

MLE for Multinomial (cont.)

- Take partial derivatives:

$$\frac{\partial G}{\partial \theta_k} = \frac{N_k}{\theta_k} - \lambda \quad \forall k, \qquad \frac{\partial G}{\partial \lambda} = 1 - \sum_{\ell} \theta_{\ell}$$

- Equate to zero and rearrange:

$$\theta_k = \frac{N_k}{\lambda} \quad \forall k, \qquad \sum_{\ell} \theta_{\ell} = 1$$

- Thus $\theta_k \propto N_k \ \forall k.$

10

MLE for Multinomial (cont.)

- Normalizing so the probabilities sum to 1 yields

$$\theta_k = \frac{N_k}{\sum_\ell N_\ell} \quad \forall k.$$

- To see that this is a maximum as the θ_k values vary, it's sufficient to observe that the second partial derivatives of G satisfy

$$\frac{\partial^2 G}{\partial\theta_i \partial\theta_j} = 0 \quad \forall i \neq j, \qquad \frac{\partial^2 G}{\partial\theta_i^2} = -\frac{N_i}{\theta_i^2} < 0 \quad \forall i$$

11

Is MLE all we need?

- Suppose that after 10 observations,
  - ML estimates ***P(H) = 0.7*** for the thumbtack
  - Would you bet on heads for the next toss?
- Suppose now that after 10 observations,
  - ML estimates ***P(H) = 0.7*** for a coin
  - Would you place the same bet?

12

Bayesian Inference

Frequentist Approach:

- Assumes there is an unknown but fixed parameter θ
- Estimates θ with some confidence
- Prediction by using the estimated parameter value

Bayesian Approach:

- Represents uncertainty about the unknown parameter
- Uses probability to quantify this uncertainty:
  - Unknown parameters as **random variables**
- Prediction follows from the rules of probability:
  - Expectation over the unknown parameters

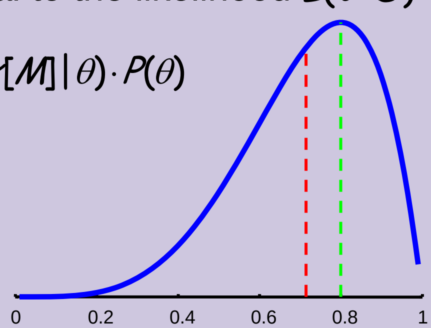
13

Example: Binomial Data Revisited

- Prior: uniform for θ in [0,1]
  - $P(\theta) = 1$
- Then $P(\theta \mid D)$ is proportional to the likelihood $L(\theta : D)$

$$P(\theta \mid x[1],\ldots x[M]) \propto P(x[1],\ldots x[M] \mid \theta) \cdot P(\theta)$$

$(N_H, N_T) = (4,1)$



- MLE for $P(X = H)$ is 4/5 = 0.8
- Bayesian prediction is

$$P(x[M+1] = H \mid D) = \int \theta \cdot P(\theta \mid D) d\theta = \frac{5}{7} = 0.7142\ldots$$

14

Bayesian Inference and MLE

- In our example, MLE and Bayesian prediction differ
- But…

  If: prior is well-behaved (i.e., does not assign 0 density to any "feasible" parameter value)

  Then: both MLE and Bayesian prediction converge to the same value as the number of training data increases

15

Dirichlet Priors

- Recall that the likelihood function is

$$L(\Theta : D) = \prod_{k=1}^{K} \theta_k^{N_k}$$

- A **Dirichlet** prior with hyperparameters $\alpha_1,\ldots,\alpha_K$ is defined as

$$P(\Theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \text{ for legal } \theta_1,\ldots,\theta_K$$

Then the posterior has the same form, with

hyperparameters $\alpha_1+N_1,\ldots,\alpha_K+N_K$

$$P(\Theta \mid D) \propto P(\Theta)P(D \mid \Theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{k=1}^{K} \theta_k^{N_k} = \prod_{k=1}^{K} \theta_k^{\alpha_k + N_k - 1}$$

16

Dirichlet Priors (cont.)

- We can compute the prediction on a new event in closed form:
- If $P(\Theta)$ is Dirichlet with hyperparameters $\alpha_1,\ldots,\alpha_K$ then

$$P(X[1]=k)=\int \theta_k \cdot P(\Theta)d\Theta = \frac{\alpha_k}{\sum_\ell \alpha_\ell}$$

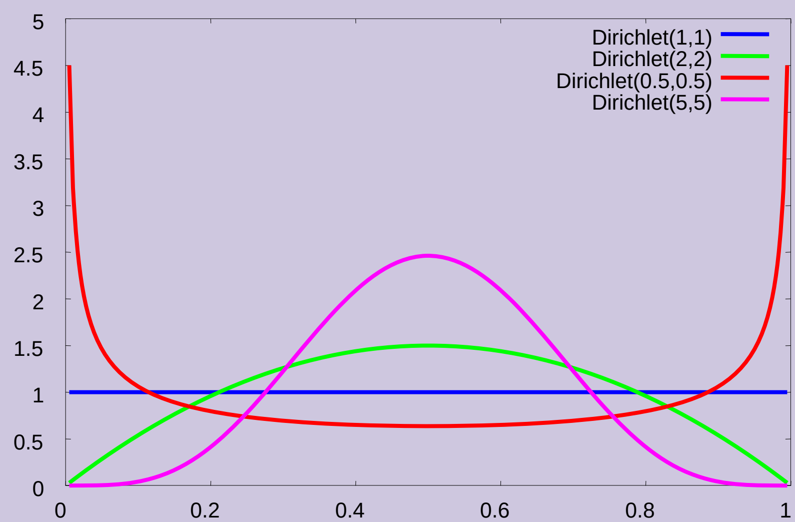
We won't prove this

- Since the posterior is also Dirichlet, we get

$$P(X[M+1]=k \mid D)=\int \theta_k \cdot P(\Theta \mid D)d\Theta = \frac{\alpha_k + N_k}{\sum_\ell (\alpha_\ell + N_\ell)}$$

17

Dirichlet Priors -- Example



18

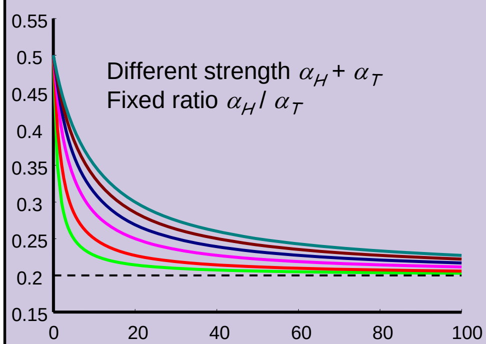
Prior Knowledge

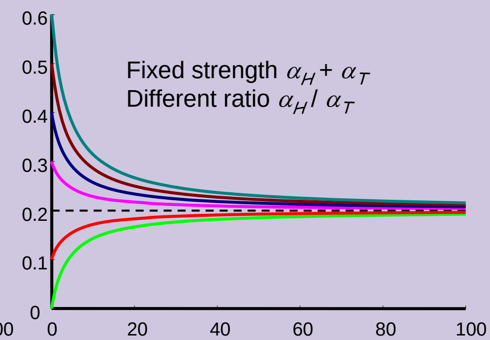
- The hyperparameters $\alpha_1,\ldots,\alpha_K$ can be thought of as "imaginary" counts from our prior experience
- Equivalent sample size = $\alpha_1+\ldots+\alpha_K$
- The larger the **equivalent sample size** the more confident we are in our prior

19

Effect of Priors

Prediction of $P(X=H)$ after seeing data with $N_H = 0.25 \cdot N_T$ for different sample sizes

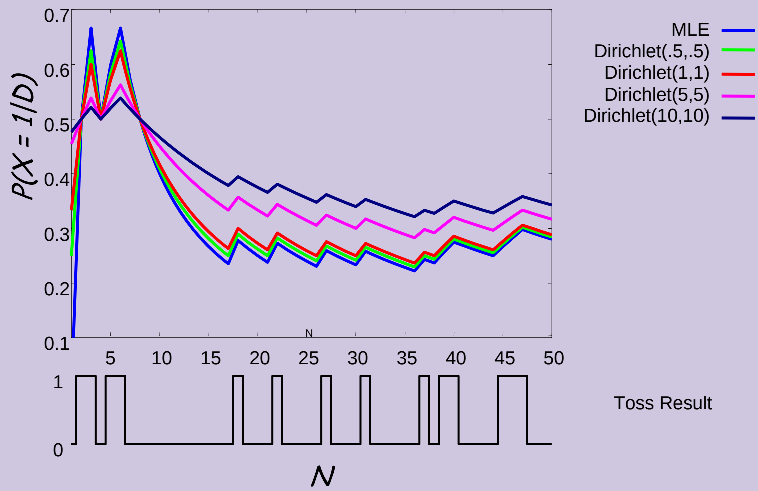




20

Effect of Priors (cont.)

- In real data, Bayesian estimates are less sensitive to noise in the data



21

One reason to prefer Bayesian method

- If any value fails to occur in the training data, MLE for the corresponding probability will be zero
- But even with uniform prior, Bayesian estimate for this same probability will be non-zero
- Probability estimates of zero can have very bad effects on just about any learning algorithm
  - Only want zero probability estimates when non-occurrence of an event is justified by prior belief

22