



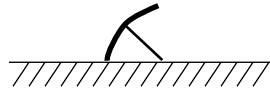
# MACHINE LEARNING

Vasant Honavar  
Artificial Intelligence Research Laboratory  
Department of Computer Science  
Bioinformatics and Computational Biology Program  
Center for Computational Intelligence, Learning, & Discovery  
Iowa State University  
honavar@cs.iastate.edu  
[www.cs.iastate.edu/~honavar/](http://www.cs.iastate.edu/~honavar/)  
[www.cild.iastate.edu/](http://www.cild.iastate.edu/)

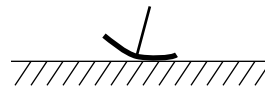
## Estimating probabilities from data (discrete case)

- Maximum likelihood estimation
- Bayesian estimation
- Maximum a posteriori estimation

## Example: Binomial Experiment



Head



Tail

- When tossed, the thumbtack can land in one of two positions: Head or Tail
- We denote by  $\theta$  the (unknown) probability  $P(H)$ .
- Estimation task
  - Given a sequence of toss samples  $x[1], x[2], \dots, x[M]$  we want to estimate the probabilities  $P(H) = \theta$  and  $P(T) = 1 - \theta$

## Statistical parameter fitting

Consider samples  $x[1], x[2], \dots, x[M]$  such that

- The set of values that  $X$  can take is known
  - Each is sampled from the same distribution
  - Each is sampled independently of the rest
- } i.i.d. samples

The task is to find a parameter  $\theta$  so that the data can be summarized by a probability  $P(x[j] | \theta)$ .

- The parameters depend on the given family of probability distributions: multinomial, Gaussian, Poisson, etc.
- We will focus first on **binomial** and then on **multinomial** distributions
- The main ideas generalize to other distribution families

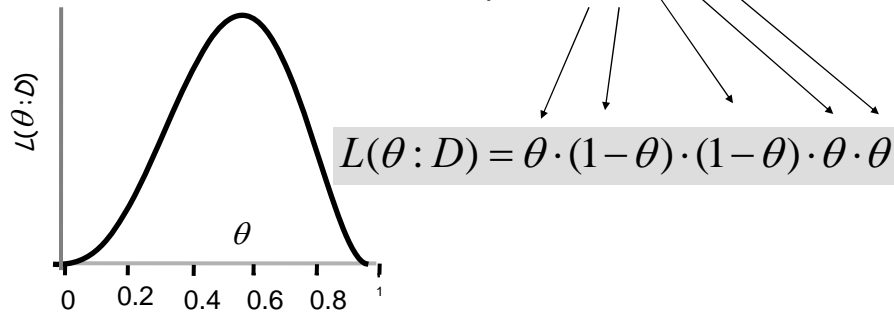
## The Likelihood Function

How good is a particular  $\theta$ ?

It depends on how likely it is to generate the observed data

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

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



Copyright Vasant Honavar, 2006.

## Likelihood function

- The likelihood function  $L(\theta : D)$  provides a measure of relative preferences for various values of the parameter  $\theta$  given a collection of observations  $D$  drawn from a distribution that is parameterized by fixed but unknown  $\theta$ .
- $L(\theta : D)$  is the *probability of the observed data  $D$  considered as a function of  $\theta$* .
- Suppose data  $D$  is 5 heads out of 8 tosses. What is the likelihood function assuming that the observations were generated by a binomial distribution with an unknown but fixed parameter  $\theta$ ?

$$\binom{8}{5} \theta^5 (1 - \theta)^3$$

Copyright Vasant Honavar, 2006.

## Sufficient Statistics

- To compute the likelihood in the thumbtack example we only require  $N_H$  and  $N_T$  (the number of heads and the number of tails)

$$L(\theta : D) = \theta^{N_H} \cdot (1 - \theta)^{N_T}$$

- $N_H$  and  $N_T$  are **sufficient statistics** for the parameter  $\theta$  that specifies the binomial distribution
- A **statistic** is simply a function of the data
- A **sufficient statistic**  $s$  for a parameter  $\theta$  is a function that summarizes from the data  $D$ , the relevant information  $s(D)$  needed to compute the likelihood  $L(\theta : D)$ .
- If  $s$  is a sufficient statistic for  $\theta$   
then  $L(\theta : D) = L(\theta : D')$

## Maximum Likelihood Estimation

- **Main Idea:** Learn parameters that maximize the likelihood function
- Maximum likelihood estimation is
- Intuitively appealing
- One of the most commonly used estimators in statistics
- Assumes that the parameter to be estimated is fixed, but unknown

## Example: MLE for Binomial Data

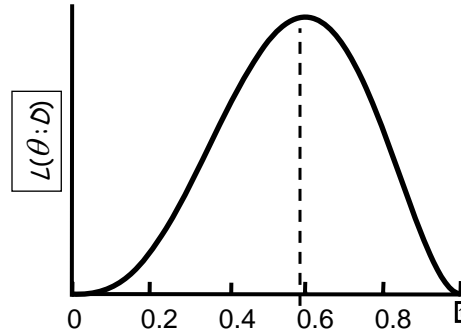
- Applying the MLE principle we get
- (Why?)

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

Example:

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

ML estimate is  $3/5 = 0.6$



Copyright Vasant Honavar, 2006.

## MLE for Binomial data

$$L(\theta; D) = \theta^{N_H} \cdot (1 - \theta)^{N_T}$$

$$\log L(\theta; D) = N_H \log \theta + N_T \log(1 - \theta)$$

The likelihood is positive for all legitimate values of  $\theta$

So maximizing the likelihood is equivalent to maximizing its logarithm i.e. log likelihood

$$\frac{\partial}{\partial \theta} \log L(\theta; D) = 0 \text{ at extrema of } L(\theta; D)$$

$$\frac{\partial}{\partial \theta} \log L(\theta; D) = \frac{N_H}{\theta} + \frac{N_T(-1)}{(1-\theta)} = 0$$

$$(N_H + N_T)\theta = N_H$$

$$\theta_{ML} = \frac{N_H}{(N_H + N_T)}$$

Note that the likelihood is indeed maximized at  $\theta = \theta_{ML}$  because in the neighborhood of  $\theta_{ML}$ , the value of the likelihood is smaller than it is at  $\theta = \theta_{ML}$

Copyright Vasant Honavar, 2006.

## Maximum and curvature of likelihood around the maximum

- At the maximum, the derivative of the log likelihood is zero
- At the maximum, the second derivative is negative.
- The curvature of the log likelihood is defined as

$$I(\theta) = -\frac{\partial}{\partial \theta^2} \log L(\theta : D)$$

- Large observed curvature  $I(\theta_{ML})$  at  $\theta = \theta_{ML}$
- is associated with a sharp peak, intuitively indicating less uncertainty about the maximum likelihood estimate
- $I(\theta_{ML})$  is called the Fisher information

## Maximum Likelihood Estimate

ML estimate can be shown to be

- Asymptotically unbiased  $\lim_{N \rightarrow \infty} E(\theta_{ML}) = \theta_{True}$
- Asymptotically consistent - converges to the true value as the number of examples approaches infinity

$$\lim_{N \rightarrow \infty} \Pr \{ \|\theta_{ML} - \theta_{True}\| \leq \varepsilon \} = 1$$

$$\lim_{N \rightarrow \infty} E(\|\theta_{ML} - \theta_{True}\|^2) = 0$$

- Asymptotically efficient – achieves the lowest variance that any estimate can achieve for a training set of a certain size (satisfies the Cramer-Rao bound)

## Maximum Likelihood Estimate

- ML estimate can be shown to be representationally invariant – If  $\theta_{ML}$  is an ML estimate of  $\theta$ , and  $g(\theta)$  is a function of  $\theta$ , then  $g(\theta_{ML})$  is an ML estimate of  $g(\theta)$
- When the number of samples is large, the probability distribution of  $\theta_{ML}$  has *Gaussian distribution with mean  $\theta_{True}$*  (the actual value of the parameter) – a consequence of the central limit theorem – a random variable which is a sum of a large number of random variables has a Gaussian distribution – ML estimate is related to the sum of random variables
- We can use the likelihood ratio to reject the null hypothesis corresponding to  $\theta = \theta_0$  as unsupported by data if the ratio of the likelihoods evaluated at  $\theta_0$  and at  $\theta_{ML}$  is *small*. (The ratio can be calibrated when the likelihood function is approximately quadratic)

## Naïve Bayes Classifier

- We can define the likelihood for a Naïve Bayesian Classifier
- Let  $\Theta_j$  be the class conditional probabilities for class  $j$
- Let  $L_j$  be the corresponding likelihood
- $L_j$  factorizes

$$\begin{aligned}
 L_j(\Theta : D) &= \prod_p P(x_1[p], \dots, x_n[p] : \Theta_j) \\
 &= \prod_p \prod_i P(x_i[p] : \Theta_{ij}) \\
 &= \prod_i \prod_p P(x_i[p] : \Theta_{ij}) \\
 &= \prod_i L_{ij}(\Theta_{ij} : D)
 \end{aligned}$$

i.i.d. samples

Independence factorization

- Each  $\Theta_{ij}$  specifies a binomial distribution associated with class  $j$  for  $i$ th attribute

## Naïve Bayes Classifier

- Decomposition  $\rightarrow$  Independent Estimation Problems
- If the parameters for each family are decoupled via independence, then they can be estimated independently of each other

## From Binomial to Multinomial

- Suppose a random variable  $X$  can take the values  $1, 2, \dots, K$
- We want to learn the parameters  $\theta_1, \theta_2, \dots, \theta_K$
- Sufficient statistics:  $N_1, N_2, \dots, N_K$  - the number of times each outcome is observed
- Likelihood function

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

- ML estimate

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

## MLE estimates for Naive Bayes Classifiers

- When we assume that  $P(X_i | C)$  is multinomial, we get the decomposition:

$$\begin{aligned} L_i(\Theta_i : D) &= \prod_m P(x_i[m] | c[m] : \Theta_i) \\ &= \prod_{c_j} \prod_{x_i} P(x_i | c_j : \Theta_i)^{N(x_i, c_j)} = \prod_{c_j} \prod_{x_i} \theta_{x_i | c_j}^{N(x_i, c_j)} \end{aligned}$$

- For each class we get an independent multinomial estimation problem
- The MLE is

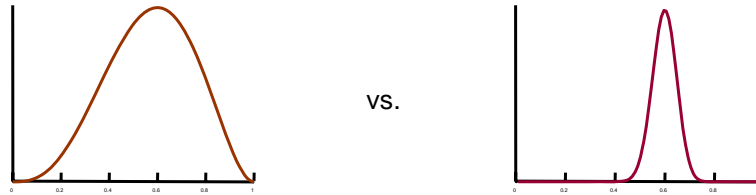
$$\hat{\theta}_{x_i | c_j} = \frac{N(x_i, c_j)}{N(c_j)}$$

## Summary of Maximum Likelihood estimation

- Define a likelihood function which is a measure of how likely it is that the observed data were generated from a probability distribution with a particular choice of parameters
- Select the parameters that maximize the likelihood
- In simple cases, ML estimate has a closed form solution
- In other cases, ML estimation may require numerical optimization
- **Problem with ML estimate** – assigns zero probability to unobserved values – can lead to difficulties when estimating from small samples
- **Question** – How would Naïve Bayes classifier behave if some of the class conditional probability estimates are zero?

## Bayesian Estimation

- MLE commits to a specific value of the unknown parameter (s)
- MLE is the same in both cases shown



Of course, in general, one cannot summarize a function by a single number!

Intuitively, the confidence in the estimates should be different

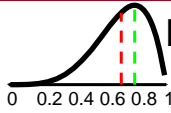
## Bayesian Estimation

Maximum Likelihood approach is Frequentist at its core

- Assumes there is an unknown but fixed parameter  $\theta$
- Estimates  $\theta$  with some confidence
- Prediction of probabilities 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



## Example: Binomial Data Revisited

- Suppose that we choose a uniform prior  $p(\theta) = 1$  for  $\theta$  in  $[0,1]$
- $P(\theta | D)$  is proportional to the likelihood  $L(\theta : D)$

$$p(\theta | D) = \frac{p(D | \theta)p(\theta)}{\int_0^1 p(D | \theta)p(\theta)d\theta}$$

In this case,  $p(D | \theta) = \theta^4(1-\theta)^1$  and  $\forall \theta \in [0,1], p(\theta) = \frac{1}{1-0} = 1$

$$\int_0^1 p(D | \theta)p(\theta) = \int_0^1 (\theta^4 - \theta^5) d\theta = \left[ \frac{\theta^5}{5} - \frac{\theta^6}{6} \right]_0^1 = \frac{1}{30}$$

$$p(\theta | D) = 30\theta^4(1-\theta)$$

$$P(X[m+1] = H | D) = \int_0^1 p(\theta | D)\theta d\theta = 30 \int_0^1 \theta^4(1-\theta)\theta d\theta = 30 \left[ \frac{\theta^6}{6} - \frac{\theta^7}{7} \right]_0^1 = \frac{5}{7} = 0.7142$$

## Example: Binomial Data Revisited

$(NH, NT) = (4, 1)$

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

Bayesian estimate is

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

In this example, MLE and Bayesian prediction differ

It can be proved that

If the prior is well-behaved – i.e. does not assign 0 density to any *feasible* parameter value

Then both MLE and Bayesian estimate converge to the same value in the limit

Both *almost surely* converge to the underlying distribution  $P(X)$

But the ML and Bayesian approaches behave differently when the number of samples is small

## All relative frequencies are not equi-probable

- In practice we might want to express priors that allow us to express our beliefs regarding the parameter to be estimated
- For example, we might want a prior that assigns a higher probability to parameter values that describe a fair coin than it does to an unfair coin
- The beta distribution allows us to capture such prior beliefs

## Beta distribution

Gamma Function:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

The integral converges if and only if  $x > 0$ .

If  $x$  is an integer that is greater than 0, it can be shown

that  $\Gamma(x) = (x-1)!$  So  $\frac{\Gamma(x+1)}{\Gamma(x)} = x$

The beta density function with parameters  $a, b, N = a + b$ , where  $a, b$  are real numbers  $> 0$ ,  $beta(\theta; a, b)$  is:

$$p(\theta) = \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} \text{ where } 0 \leq \theta \leq 1$$

## Beta distribution

If  $a, b$  are real numbers  $> 0$ , then

$$\int_0^1 \theta^a (1-\theta)^b d\theta = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$$

If  $\theta$  has distribution given by  $beta(\theta; a, b)$ , then  $E(\theta) = \frac{a}{N}$ .

Let  $D = \{X[1], \dots, X[M]\}$  be

a sequence of iid samples from a binomial distribution;

Let  $N_H = s$ ;  $N_T = t$ ; and  $p(\theta) = beta(\theta; a, b)$

Then we can show that  $p(\theta|D) = beta(\theta; a + s, b + t)$

Update of the parameter with a beta prior based on data yields a beta posterior

## Conjugate Families

- The property that the posterior distribution follows the same parametric form as the prior distribution is called **conjugacy**
- Conjugate families are useful because:
  - For many distributions we can represent them with hyper parameters
  - They allow for sequential update to obtain the posterior
  - In many cases we have closed-form solution for prediction
- Beta prior is a **conjugate family** for the binomial likelihood

## Bayesian prediction

prior :  $beta(\theta; a, b)$ Data :  $D = \{X[1], \dots, X[M]\}$ posterior :  $p(\theta | D) = beta(\theta; a + N_H, b + N_T)$ prediction :  $P(X[M+1] = H | D) = \frac{a + N_H}{N + M} = \frac{(a + N_H)}{(a + b) + (N_H + N_T)}$ 

## Dirichlet Priors

- Recall that the likelihood function is  $L(\Theta : D) = \prod_{k=1}^K \theta_k^{N_k}$
- A **Dirichlet** prior with hyperparameters  $\alpha_1, \dots, \alpha_K$  is defined as

$$P(\Theta) = \frac{\Gamma(N)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}; \quad 0 \leq \theta_k \leq 1; \quad \sum_{k=1}^K \theta_k = 1$$

where  $\Theta = (\theta_1, \dots, \theta_K)$ 

- Then the posterior has the same form, with hyperparameters  $\alpha_1 + N_1, \dots, \alpha_K + N_K$

$$P(\Theta | D) \propto P(\Theta)P(D | \Theta)$$

$$\propto \prod_{k=1}^K \theta_k^{\alpha_k - 1} \prod_{k=1}^K \theta_k^{N_k} = \prod_{k=1}^K \theta_k^{\alpha_k + N_k - 1}$$

## Dirichlet Priors

- Dirichlet priors enable closed form prediction based on multinomial samples:

– If  $P(\Theta)$  is Dirichlet with hyperparameters  $\alpha_1, \dots, \alpha_K$  then

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

- Since the posterior is also Dirichlet, we get

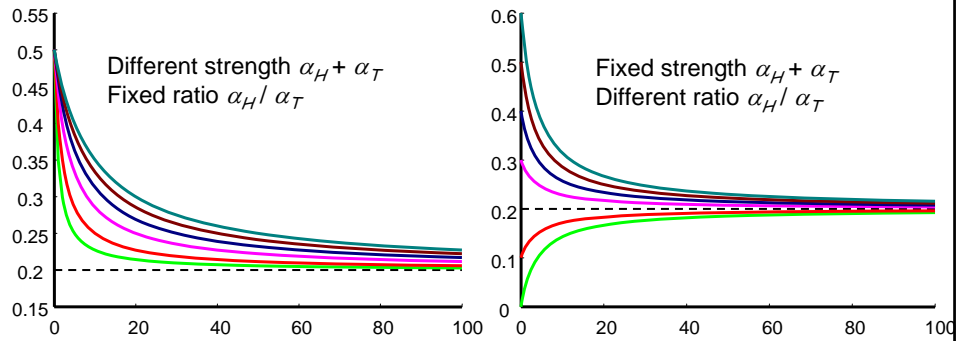
$$P(X[M+1] = k | D) = \int \theta_k \cdot P(\Theta | D) d\Theta = \frac{\alpha_k + N_k}{\sum_{\ell} (\alpha_{\ell} + N_{\ell})}$$

## Intuition behind priors

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

## Effect of Priors

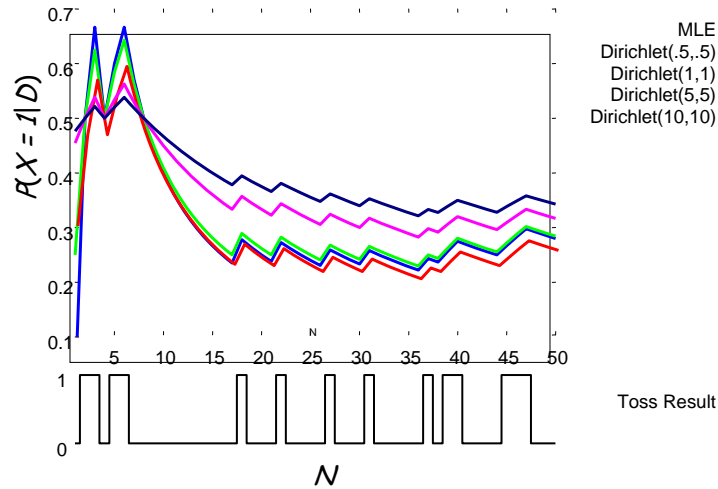
Prediction of  $P(X=H)$  after seeing data with  $N_H = 0.25 \cdot N_T$   
for different sample sizes



Copyright Vasant Honavar, 2006.

## Effect of Priors

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



Copyright Vasant Honavar, 2006.

## Conjugate Families

- The property that the posterior distribution follows the same parametric form as the prior distribution is called **conjugacy**
  - Dirichlet prior is a **conjugate family** for the multinomial likelihood
- Conjugate families are useful because:
  - For many distributions we can represent them with hyperparameters
  - They allow for sequential update within the same representation
  - In many cases we have closed-form solution for prediction

## Bayesian Estimation

$$\begin{aligned}
 P(x[M+1] | x[1], \dots, x[M]) \\
 &= \int P(x[M+1] | \theta, x[1], \dots, x[M]) P(\theta | x[1], \dots, x[M]) d\theta \\
 &= \int P(x[M+1] | \theta) P(\theta | x[1], \dots, x[M]) d\theta
 \end{aligned}$$

where

$$P(\theta | x[1], \dots, x[M]) = \frac{P(x[1], \dots, x[M] | \theta) P(\theta)}{P(x[1], \dots, x[M])}$$

Likelihood

Prior

Posterior

Probability of data

## Summary of Bayesian estimation

- Treat the unknown parameters as random variables
- Assume a prior distribution for the unknown parameters
- Update the distribution of the parameters based on data
- Use Bayes rule to make prediction

## Maximum a posteriori (MAP) estimates – Reconciling ML and Bayesian approaches

$$P(\Theta|D) = \frac{P(D|\Theta)P(\Theta)}{P(D)}$$
$$\Theta_{MAP} = \arg \max_{\Theta} P(\Theta|D)$$
$$= \arg \max_{\Theta} P(D|\Theta)P(\Theta)$$
$$= \arg \max_{\Theta} P(\Theta)L(\Theta : D)$$

## Maximum a posteriori (MAP) estimates – Reconciling ML and Bayesian approaches

$$\Theta_{MAP} = \arg \max_{\Theta} P(\Theta)L(\Theta : D)$$

Like in Bayesian estimation, we treat the unknown parameters as random variables

But we estimate a single value for the parameter – the maximum a posteriori estimate that corresponds to the most probable value of the parameter given the data for a given choice of the prior

## Back to Naïve Bayes Classifier

$$\hat{P}(X_i = a_{i_k} | \omega_j) = 0 \rightarrow \hat{P}(\omega_j) \prod_l \hat{P}(X_l = a_{l_k} | \omega_j) = 0$$

If one of the attribute values has estimated class conditional probability of 0, it dominates all other attribute values

When we have few examples, this is more likely

Solution – use priors e.g., assume each value to be equally likely unless data indicates otherwise