Basics of Statistical Estimation

Doug Downey, Northwestern EECS 395/495, Spring 2016

(several illustrations from P. Domingos, University of Washington CSE)

Bayes' Rule

- $P(A \mid B) = P(B \mid A)\, P(A) / P(B)$
- Example:

P(symptom| disease) = 0.95, P(symptom| $\neg$disease) = 0.05
P(disease) = 0.0001

$$P(\text{disease} \mid \text{symptom}) = \frac{P(\text{symptom} \mid \text{disease}) * P(\text{disease})}{P(\text{symptom})}$$

$$= \frac{0.95*0.0001}{0.95*0.0001+0.05*0.9999} = \mathbf{0.002}$$

Bayes' Rule

- $P(A \mid B) = P(B \mid A)\, P(A) / P(B)$
- Also:
  - $P(A \mid B, C) = P(B \mid A, C)\, P(A \mid C) / P(B \mid C)$
- More generally:
  - $P(\boldsymbol{A} \mid \boldsymbol{B}) = P(\boldsymbol{B} \mid \boldsymbol{A})\, P(\boldsymbol{A}) / P(\boldsymbol{B})$
  - (Boldface indicates vectors of variables)

Bayes' Rule

- Why is Bayes' Rule so important?
  - Often, we want to deduce **P(*Hidden state* | *Data*)**
    - E.g., Hidden state = disease, Data = symptoms
  - and the simplest way to express that is in terms of "causes" of the model: **P(*Data* | *Model*)**
    - E.g., how common is a symptom, with or without a given disease
  - times a prior belief about the model, **P(*Model*)**
    - E.g., probability of a disease

Terms for Bayes

- P(*Model* | *Data*) = P(*Data* | *Model*) P(*Model*) / P(*Data*)
- P(*Model*) : **Prior**
- P(*Data* | *Model*) : **Likelihood**
- P(*Model* | *Data*) : **Posterior**

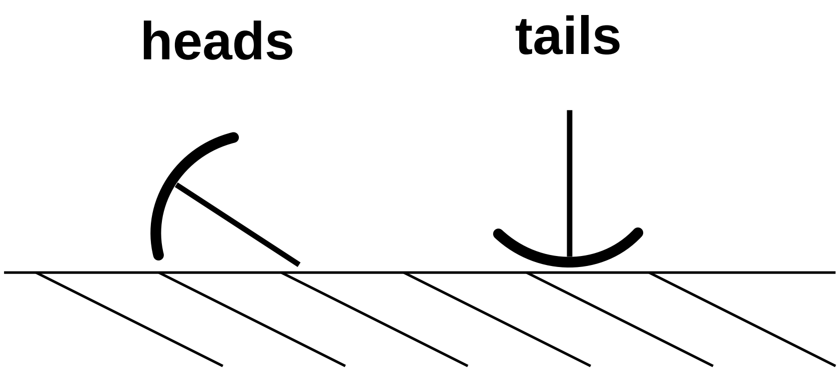
Probabilistic Models

- Joint Distribution can answer queries
  - P(**symptoms**, disease) can be used to predict whether person has disease based on symptoms

- But:
  - **Where do the probabilities come from (learning)?**
  - How do we represent a joint compactly using conditional independencies? (representation – graphical models)

Learning Probabilities:Classical Approach

Simplest case: Flipping a thumbtack



<u>**True probability**</u> θ **is unknown**

Given: flips generated independently with the same $\boldsymbol{\theta}$**,**
(a.k.a. **I**ndependent and **i**dentically **d**istributed data - iid),
Estimate: θ

Estimating Probabilities

- Three Methods:
  - Maximum Likelihood Estimation (ML)
  - Bayesian Estimation
  - Maximum A posteriori Estimation (MAP)

Maximum Likelihood Principle

$$\boxed{\text{Choose the parameters that maximize the probability of the observed data}}$$

Think/Pair/Share

If Data={h heads and t tails}, what parameter θ maximizes the probability of Data?

|Think |

Start End

Think/Pair/Share

If Data={h heads and t tails}, what parameter θ maximizes the probability of Data?

Pair

Start

End

Think/Pair/Share

If Data={*h* heads and *t* tails}, what parameter θ maximizes the probability of Data?

P(Data | θ) = θ^h * (1-θ)^t
hhthtth…
d/dθ : h*θ^(h-1) * (1-θ)……

Share

arg(subθ)max F(θ) =
arg(subθ)max logF(θ)

Find θ maximizing h*lgθ + t*lg(1-θ)
d/dθ (logP(Data | θ)) = h/θ - t/(1-θ) = 0

maximizing f(x) is the same as
maximizing any monotonic f(x)

h/θ = t/(1-θ)
(1-θ)*h = tθ
h - hθ = tθ

h = tθ+hθ = θ(t+h)

h/(h+t) = θ
#h / (#h + #t)

Maximum Likelihood Estimation

$$p(\text{heads} \mid \theta) = \theta$$

$$p(\text{tails} \mid \theta) = (1 - \theta)$$

$$p(hhth\ldots ttth \mid \theta) = \theta^{\#h}(1 - \theta)^{\#t}$$

(Number of heads is binomial distribution)

Computing the ML Estimate

- Use log-likelihood
- Differentiate with respect to parameter(s)
- Equate to zero and solve
- Solution:

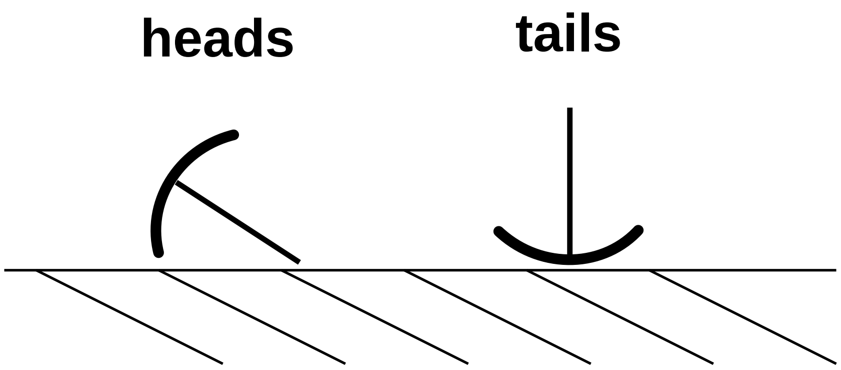
$$\theta = \frac{\#h}{\#h + \#t}$$

Sufficient Statistics

$$p(hhth...ttth \mid \theta) = \theta^{\#h}(1-\theta)^{\#t}$$

(#h,#t) are sufficient statistics

Bayesian Estimation

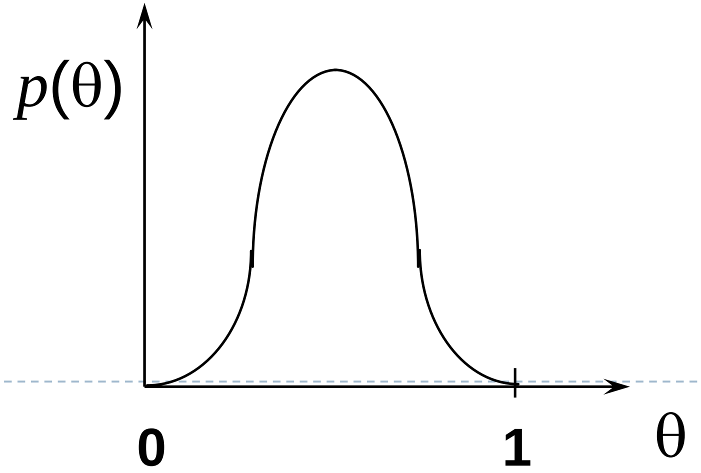


if flipped once and tails,

then θ = 0/(1+0) = 0

<u>True probability</u> θ is unknown

<u>Bayesian probability density for</u> θ



Use of Bayes' Theorem

prior

likelihood

posterior

$$p(\theta \mid \text{heads}) = \frac{p(\theta)\,p(\text{heads} \mid \theta)}{\int p(\theta')\,p(\text{heads} \mid \theta')\; d\theta'}$$

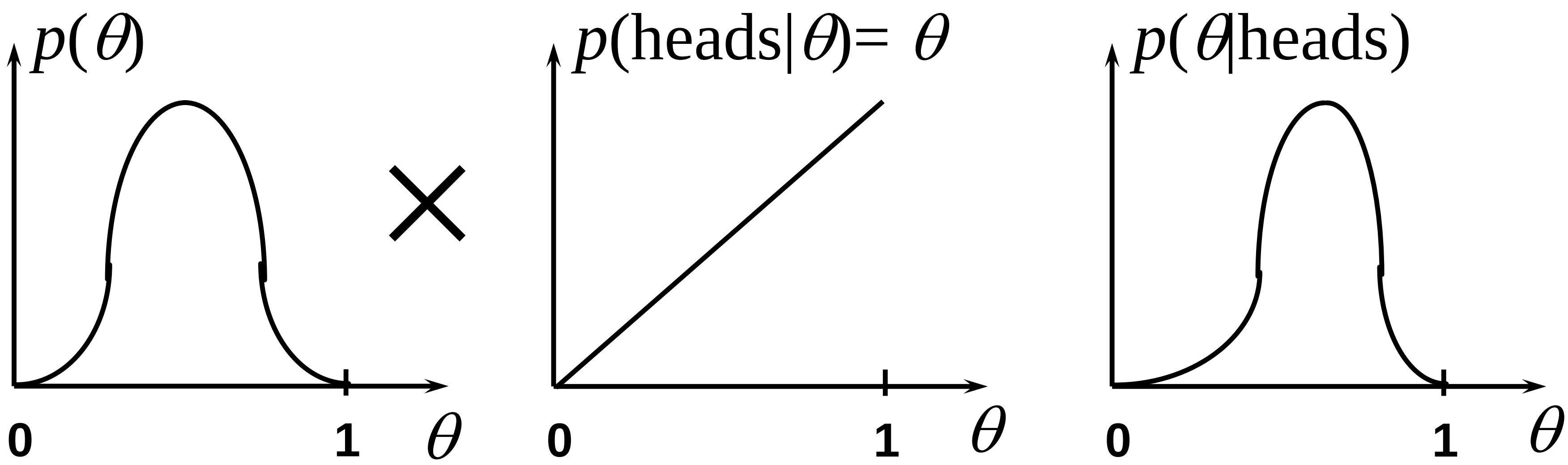
just means probability of heads ->

"proportional" -> $\propto \; p(\theta)\,p(\text{heads} \mid \theta)$

probability of a probability is how bayesian estimation works

Example: Observation of “Heads"

less variance where we think θ is



prior likelihood posterior

Probability of Heads on Next Toss

$$
\begin{aligned}
p(n+1\text{th toss is } h \mid \mathbf{d}) &= \int p(X_{N+1} = h \mid \theta)\, p(\theta \mid \mathbf{d})\, d\theta \\
&= \int \theta \;\; p(\theta \mid \mathbf{d}) \;\; d\theta \\
&= E_{p(\theta \mid \mathbf{d})}(\theta)
\end{aligned}
$$

Q: but isn't the probability independent of the previous toss?

not just that they're independent,
they're independent and identically distributed
("conditionally independent given their parameters")

MAP Estimation

- Approximation:
  - Instead of averaging over all parameter values
  - Consider only the **most probable value** (i.e., value with highest posterior probability)
- Usually a very good approximation, and much simpler
- MAP value ≠ Expected value
- MAP → ML for infinite data (as long as prior ≠ 0 everywhere)

Prior Distributions for θ

- Direct assessment
- Parametric distributions
  - Conjugate distributions (for convenience)

Conjugate Family of Distributions

Beta distribution:

$$p(\theta) = \mathrm{Beta}(\alpha_h, \alpha_t) \propto \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1}$$

$$\alpha_h, \alpha_t > 0$$

Resulting posterior distribution:

$$p(\theta \mid h \text{ heads}, t \text{ tails}) \propto \theta^{\#h + \alpha_h - 1}(1-\theta)^{\#t + \alpha_t - 1}$$

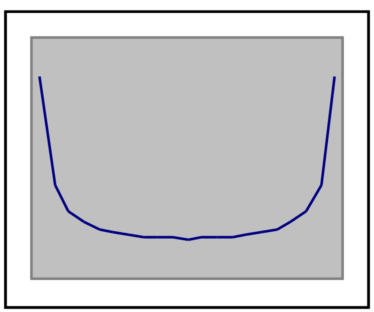
Estimates Compared

- Prior prediction: $$E(\theta) = \frac{\alpha_h}{\alpha_h + \alpha_t}$$
- Bayesian posterior prediction $$E(\theta) = \frac{\#h + \alpha_h}{\#h + \alpha_h + \#t + \alpha_t}$$
- MAP estimate: $$\theta = \frac{\#h + \alpha_h - 1}{\#h + \alpha_h - 1 + \#t + \alpha_t - 1}$$
- ML estimate: $$\theta = \frac{\#h}{\#h + \#t}$$

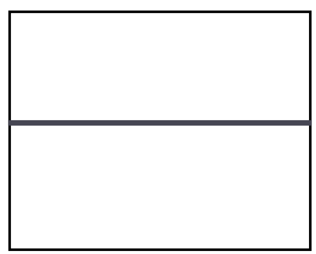
Intuition

- The hyperparameters α_h and α_t can be thought of as imaginary counts from our prior experience, starting from "pure ignorance"
- Equivalent sample size = $\alpha_h + \alpha_t$
  - (“equivalent” in terms of effect on *Bayesian* estimate)
- The larger the equivalent sample size, the more confident we are about the true probability

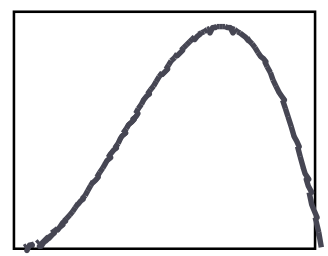
Beta Distributions



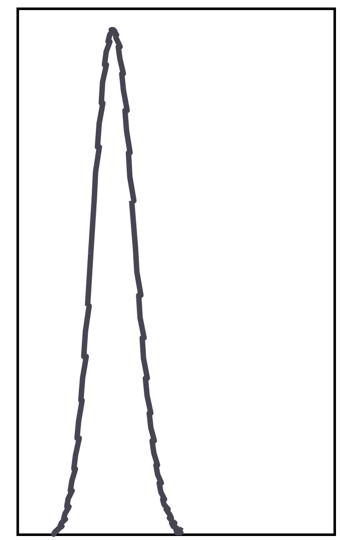
Beta(0.5, 0.5)



Beta(1, 1)



Beta(3, 2)



Beta(19, 39)

Assessment of a Beta Distribution

Method 1: Equivalent sample

- assess α_h and α_t

- assess $\alpha_h+\alpha_t$ and $\alpha_h/(\alpha_h+\alpha_t)$

Method 2: Imagined future samples

$$p(\text{heads}) = 0.2 \text{ and } p(\text{heads} \mid 3 \text{ heads}) = 0.5 \Rightarrow \alpha_h = 1, \alpha_t = 4$$

$$\text{check: } 0.2 = \frac{1}{1+4}, \quad 0.5 = \frac{1+3}{1+3+4}$$

Generalization to *m* Outcomes (Multinomial Distribution)

Dirichlet distribution:

$$p(\theta_1,\ldots,\theta_m) = \mathrm{Dirichlet}(\alpha_1,\ldots,\alpha_m) \propto \prod_{i=1}^{m} \theta_i^{\alpha_i - 1}$$

$$\sum_{i=1}^{m} \theta_i = 1 \qquad \alpha_i > 0$$

Properties:

$$E(\theta_i) = \frac{\alpha_i}{\sum_{i=1}^{m} \alpha_i}$$

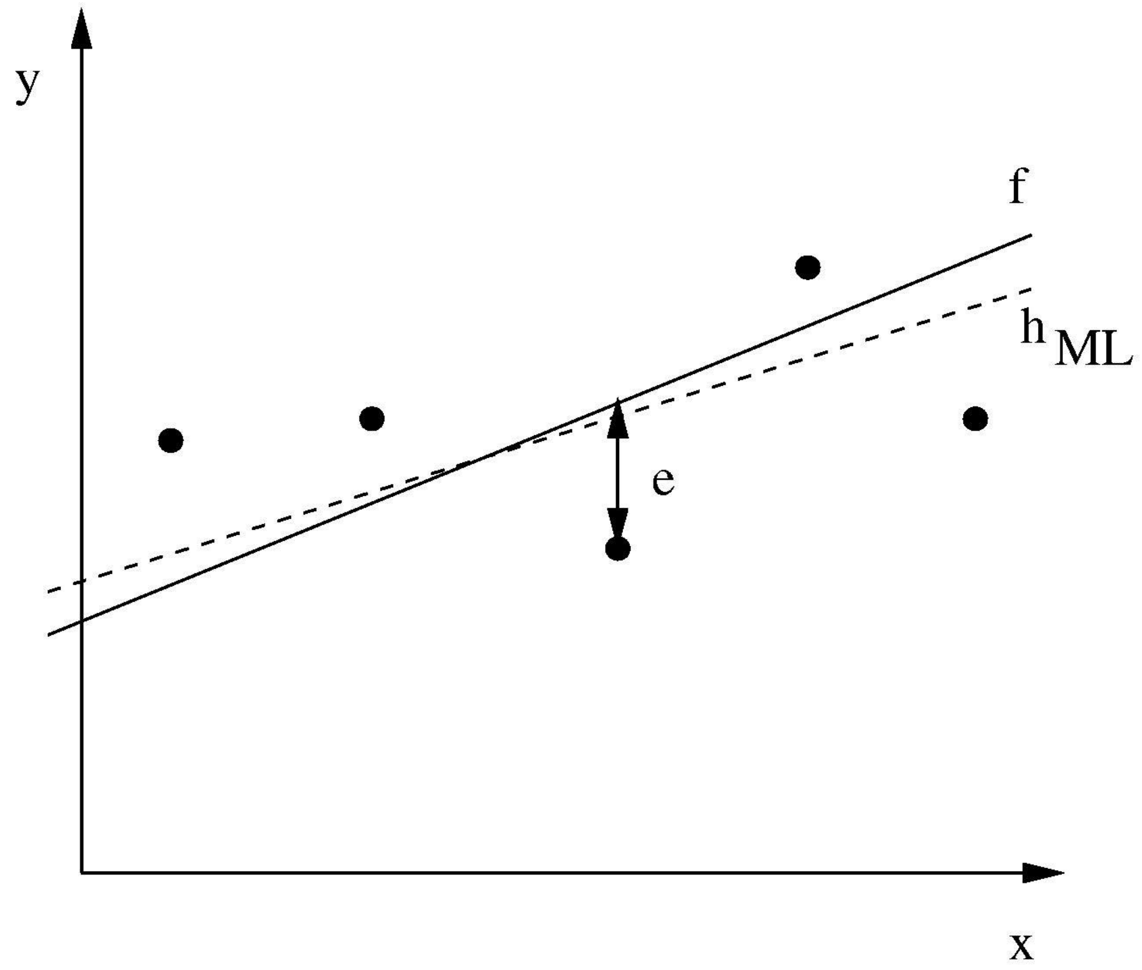
$$p(\theta \mid N_1,\ldots,N_m) \propto \prod_{i=1}^{m} \theta_i^{\alpha_i + N_i - 1}$$

Other Distributions

Likelihoods from the exponential family

- Binomial
- Multinomial
- Poisson
- Gamma
- Normal

Learning a Real-Valued Function



Consider any real-valued target function f

Training examples $\langle x_i, d_i \rangle$, where d_i is noisy training value

- $d_i = f(x_i) + e_i$
- e_i is random variable (noise) drawn independently for each x_i according to some Gaussian distribution with mean=0

Then the maximum likelihood hypothesis h_{ML} is the one that minimizes the sum of squared errors:

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

Maximum likelihood hypothesis:

$$
\begin{aligned}
h_{ML} &= \underset{h \in H}{\operatorname{argmax}}\, p(D|h) = \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} p(d_i|h) \\
&= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2}
\end{aligned}
$$

Maximize natural log of this instead . . .

$$
\begin{aligned}
h_{ML} &= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\
&= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\
&= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\left(d_i - h(x_i)\right)^2 \\
&= \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{m} \left(d_i - h(x_i)\right)^2
\end{aligned}
$$

why we like squared error as a regression