Probability Estimation

Machine Learning 10-601B

Seyoung Kim

Many of these slides are derived from Tom Mitchell, William Cohen, Eric Xing. Thanks!

Overview

- Joint probability distribution
  - A functional mapping f: X->Y via probability distribution
- Probability estimation
  - Maximum likelihood estimation
  - Maximum a priori estimation

What does all this have to do with function approximation for f: X->Y?

Joint Probability Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Once you have the joint distribution, you can ask for the probability of any logical expression involving your attribute

Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor, Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

B is that people work less than 40.5

$$\boxed{P(A) = P(A \wedge B) + P(A \wedge \sim B)}$$

Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Learning and the Joint Distribution

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

Suppose we want to learn the function f: <G, H> → W

Equivalently, P(W | G, H)

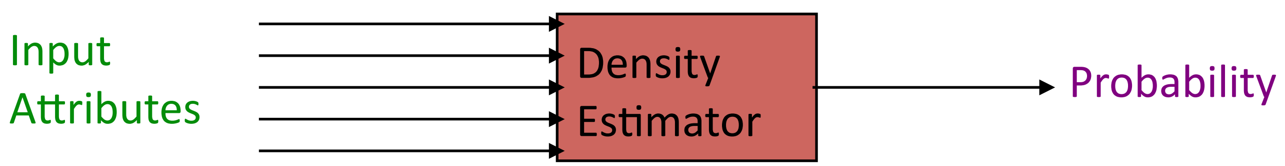
Solution: learn joint distribution from data, calculate P(W | G, H)

e.g., P(W=rich | G = female, H = 40.5-) =

[A. Moore]

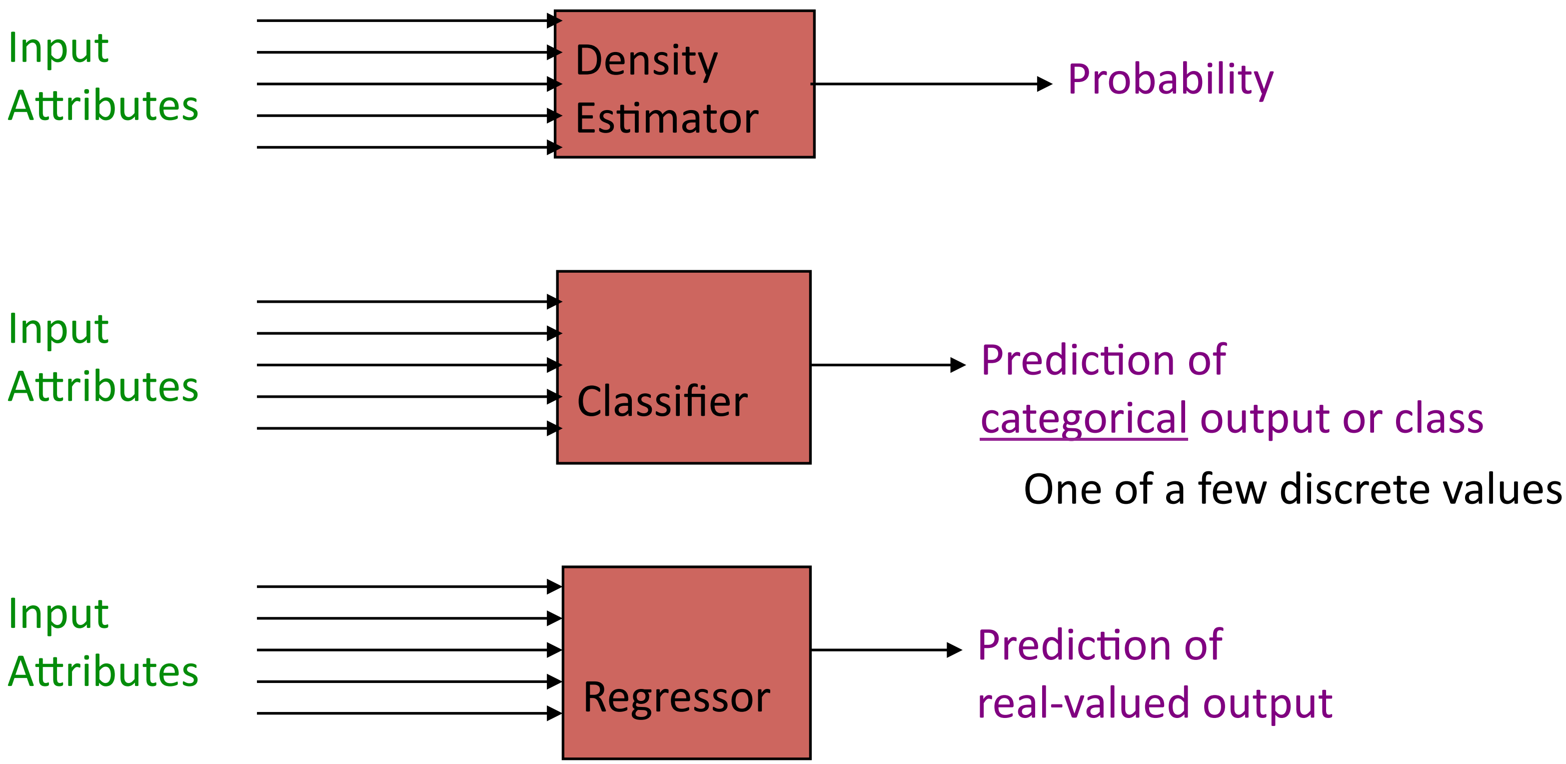
Density Estimation

- Our Joint Distribution learner is our first example of something called Density Estimation 密度估计
- A Density Estimator learns a mapping from a set of attributes values to a Probability

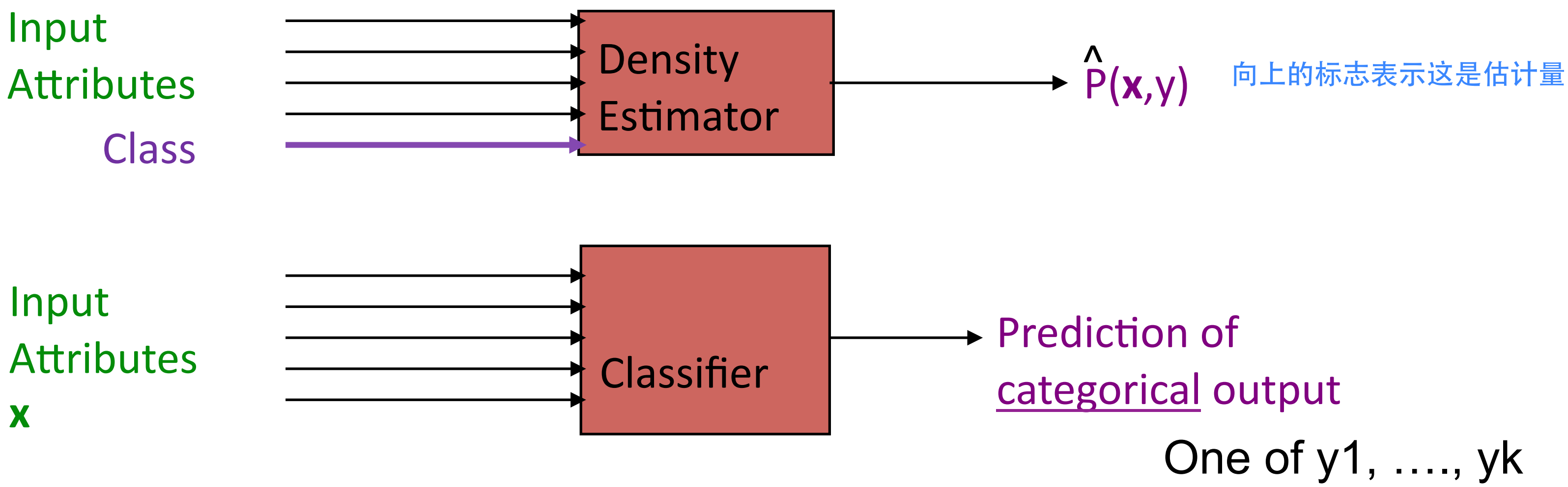


Density Estimation

- Compare it against the two other major kinds of models:



Density Estimation ➔ Classification



To classify **x**

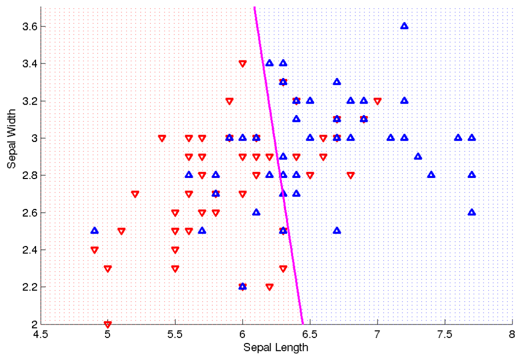
1. Use your estimator to compute $\hat{P}(\mathbf{x},y1)$, …., $\hat{P}(\mathbf{x},yk)$
2. Return the class y* with the highest predicted probability

Binary case:
predict POS if
$\hat{P}(\mathbf{x}, y_{pos})>0.5$

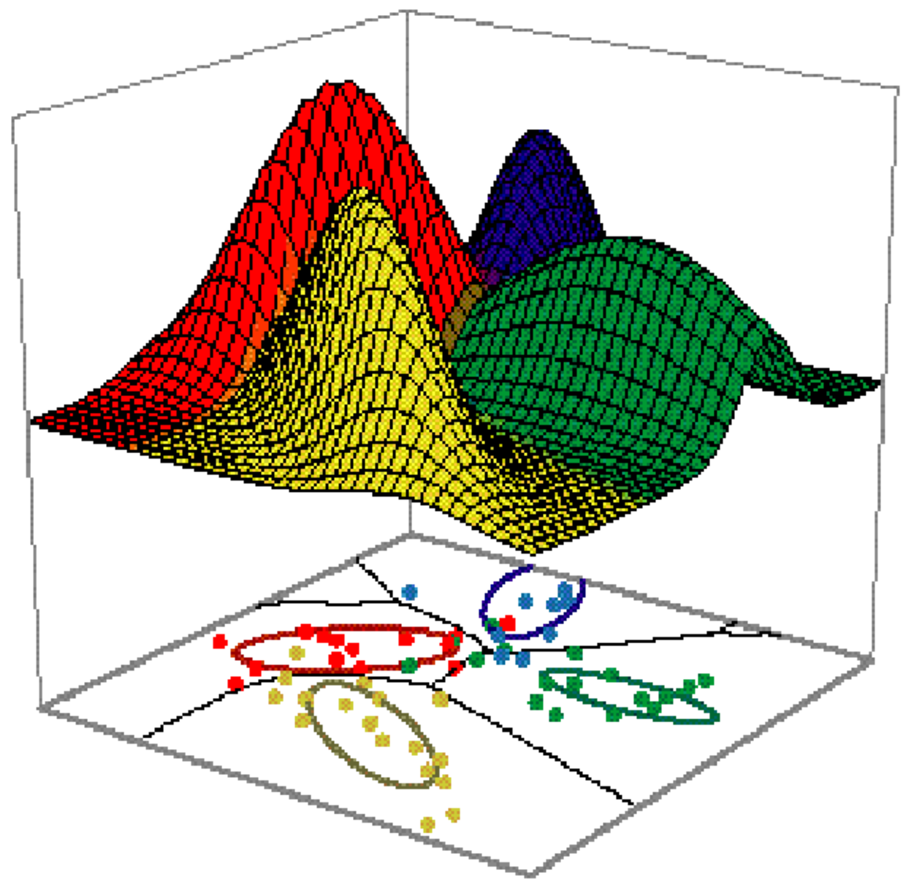
Ideally is correct with $\hat{P}(y^* \mid \mathbf{x}) = \hat{P}(\mathbf{x},y^*)/(\hat{P}(\mathbf{x},y1) + ... + \hat{P}(\mathbf{x},yk))$

Classification vs Density Estimation

Classification



Density Estimation



Classification vs density estimation

Modeling Uncertainty with Probabilities

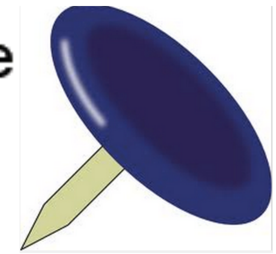
- Y is a Boolean-valued random variable if
  - Y denotes an event,
  - there is uncertainty as to whether Y occurs.
- More examples
  - Y = You wake up tomorrow with a headache
  - Y = The US president in 2023 will be male
  - Y = there is intelligent life elsewhere in our galaxy
  - Y = the 1,000,000,000,000th digit of π is 7
  - Y = I woke up today with a headache
- Define P(Y|X) as "the fraction of possible worlds in which Y is true, given X"

sounds like the solution to learning
F: X $\rightarrow$Y,
or P(Y | X).

Are we done?

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?



  - You say: Please flip it a few times:

  - You say: The probability is:
  - **He says: Why???**
  - You say: Because…

[C. Guestrin]

Thumbtack – Binomial Distribution

二项分布

- P(Heads) = θ, P(Tails) = 1-θ

D = { D1, D2, D3, D4, D5}

Flips produce data set D with α_H heads and α_T tails

- Flips are independent, identically distributed 1's and 0's (Bernoulli)
- α_H and α_T are counts that sum these outcomes (Binomial)

$$P(D|\theta) = P(\alpha_H, \alpha_T|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}\binom{\alpha_H + \alpha_T}{\alpha_H}$$

θ^3*(1-θ)^2=
θ^(αH)*(1-θ)^(αT)*组合（从总数中选择αH个head up）

[C. Guestrin]

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distributior $P(D|\theta) = P(\alpha_H, \alpha_T|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T} \binom{\alpha_H + \alpha_T}{\alpha_H}$
- Learning θ is an optimization problem
  - What's the objective function?

- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta) \end{aligned}$$

[C. Guestrin]

Maximum Likelihood Estimate for *Θ*

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \; \ln P(\mathcal{D} \mid \theta) = 0}$

[C. Guestrin]

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

$$\hat{\theta} = \arg\max_{\theta} \ \ln P(\mathcal{D} \mid \theta)$$
$$= \arg\max_{\theta} \ \ln \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

[C. Guestrin]

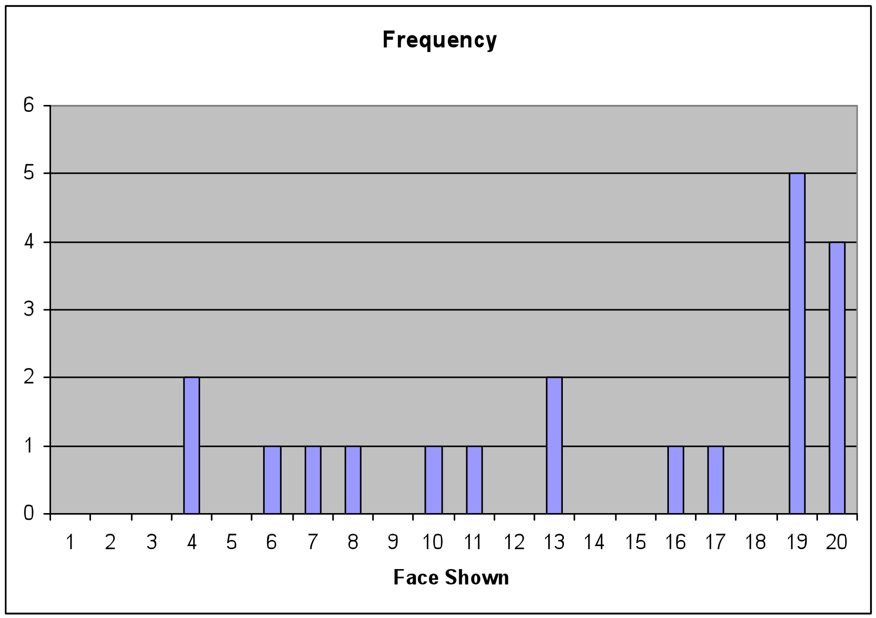
How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

[C. Guestrin]

Issues with MLE estimate

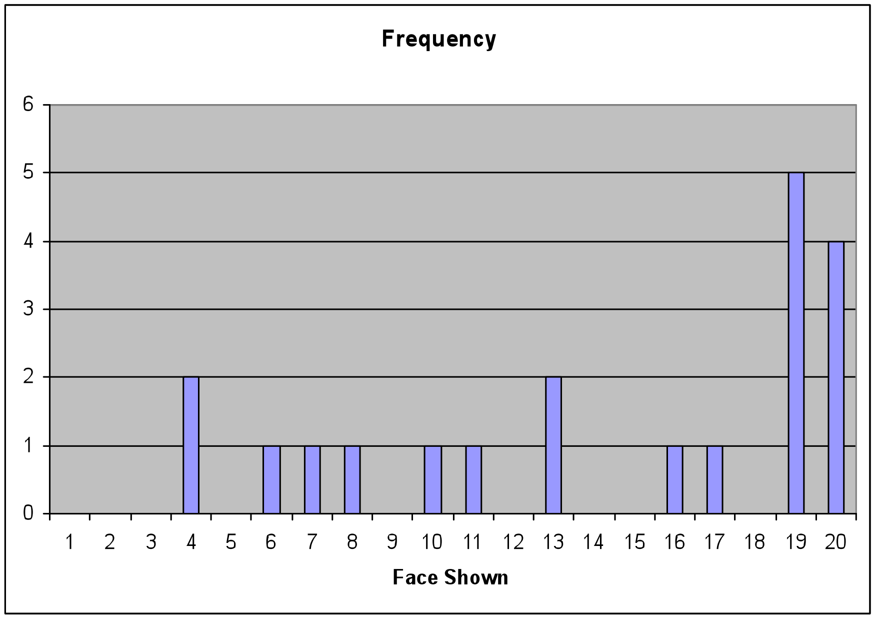
I bought a loaded 20-faced die (d20) on EBay…but it didn't come with any specs. How can I find out how it behaves?



1. Collect some data (20 rolls)
2. Estimate P(i)=CountOf(rolls of i)/CountOf(any roll)

Issues with MLE estimate

I bought a loaded d20 on EBay…but it didn't come with any specs. How can I find out how it behaves?



P(1)=0

P(2)=0

P(3)=0

P(4)=0.1

…

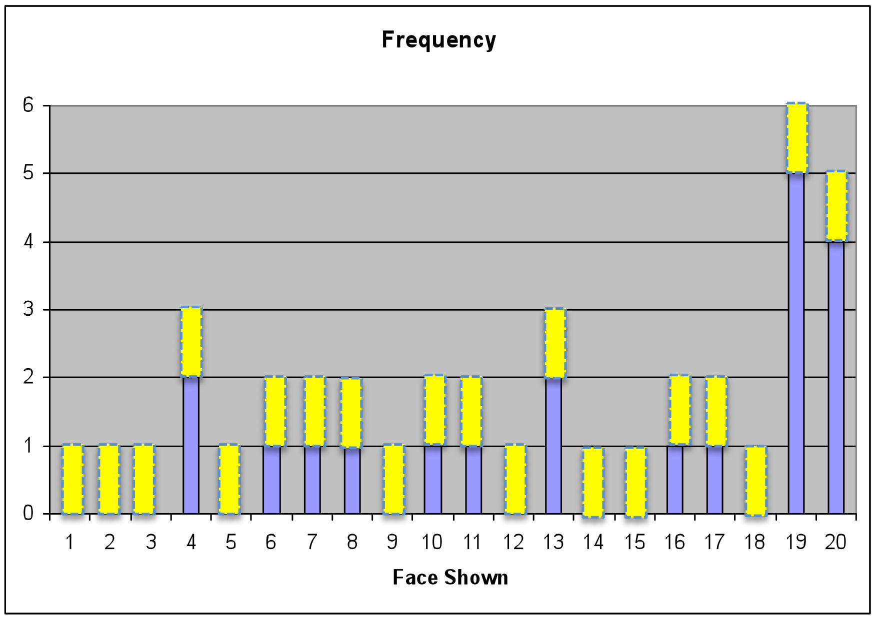
P(19)=0.25

P(20)=0.2

But: Do I really think it's *impossible* to roll a 1,2 or 3?

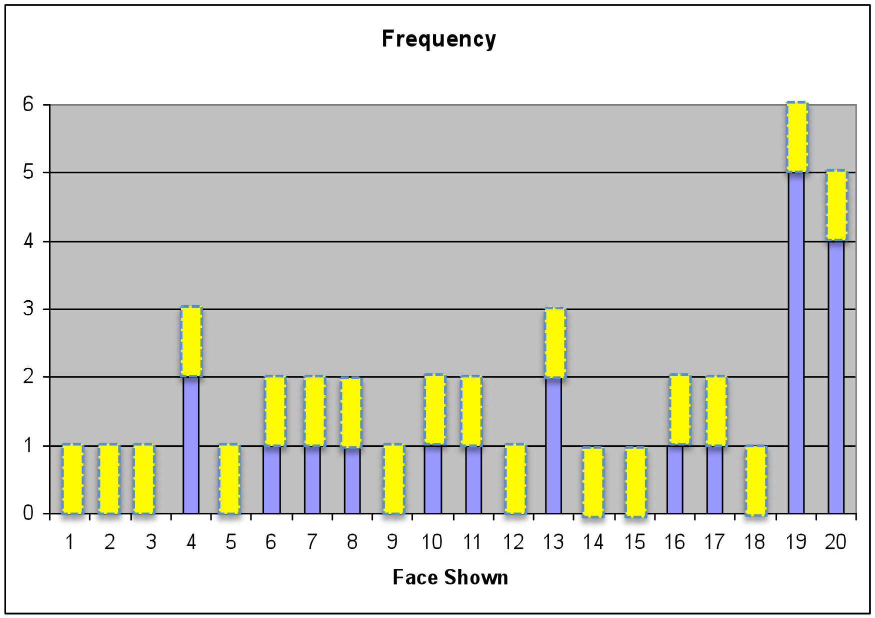
A better solution

I bought a loaded d20 on EBay…but it didn't come with any specs.
How can I find out how it behaves?



0. *Imagine* some data (20 rolls, each i shows up 1x)
1. Collect some data (20 rolls)
2. Estimate P(i)

A better solution?



P(1)=1/40

P(2)=1/40

P(3)=1/40

P(4)=(2+1)/40

…

P(19)=(5+1)/40

P(20)=(4+1)/40=1/8

MAP = maximum a posteriori estimate

$$\hat{P}(i) = \frac{CountOf(i)+1}{CountOf(ANY)+CountOf(IMAGINED)}$$

0.2 *vs.* 0.125 – really different! Maybe I should "imagine" less data?

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

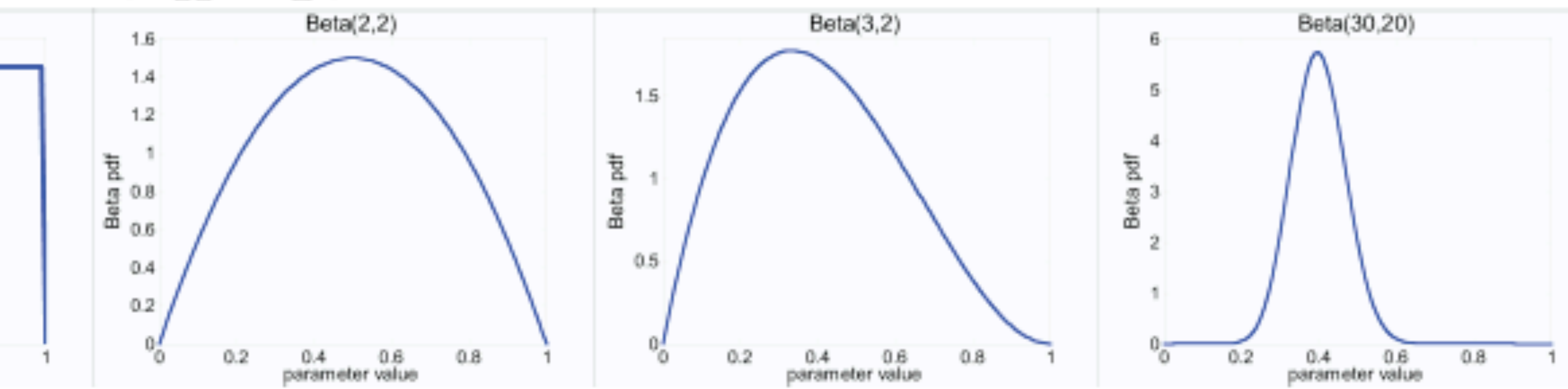
- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

[C. Guestrin]

Beta prior distribution – P(θ)

- $$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$



θ

[C. Guestrin]

Beta prior distribution – P(θ)

- $$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}\binom{\alpha_H + \alpha_T}{\alpha_H}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$

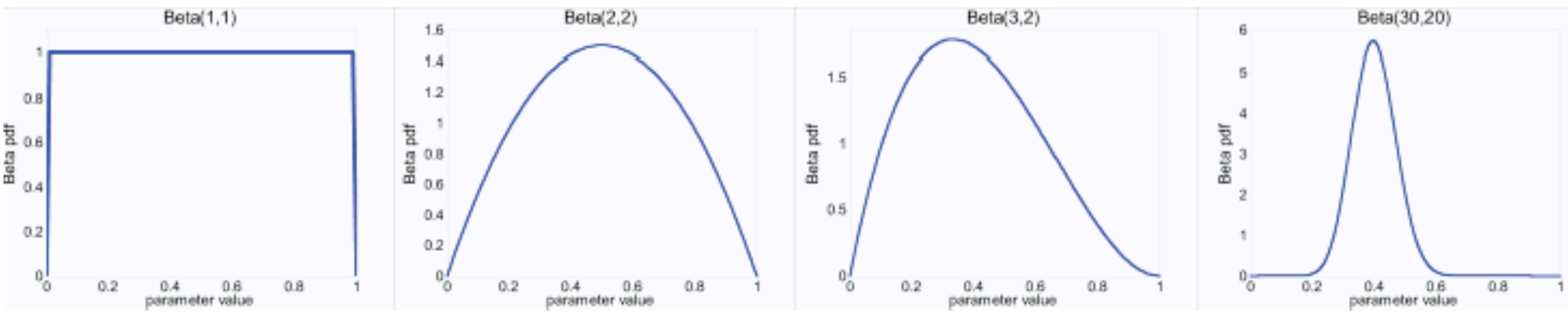
[C. Guestrin]

Posterior distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails
- Posterior distribution:

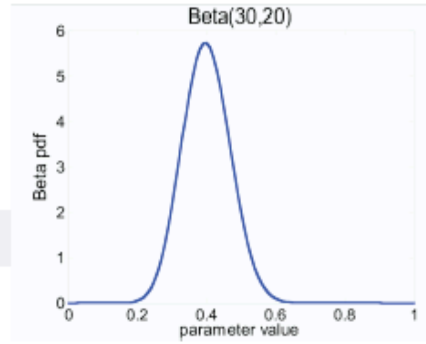
β是imageined data

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



[C. Guestrin]

MAP for Beta distribution



$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H, \beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T)$$

- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) =$$

$$\frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H - 1 + \alpha_T + \beta_T - 1}$$

- Beta prior equivalent to extra thumbtack flips
- As $N \to \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**

[C. Guestrin]

Conjugate priors

- P(θ) and P(θ|D) have the same form

Eg. 1 Coin flip problem

Likelihood is ~ Binomial

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Then posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

For Binomial, conjugate prior is Beta distribution.

[A. Singh]

Dirichlet distribution

- number of heads in N flips of a two-sided coin
  - follows a binomial distribution
  - Beta is a good prior (conjugate prior for binomial)
- what it' s not two-sided, but k-sided?
  - follows a *multinomial* distribution
  - *Dirichlet* distribution is the conjugate prior

$$P(\theta_1, \theta_2, ...\theta_K) = \frac{1}{B(\alpha)} \prod_{i}^{K} \theta_i^{(\alpha_1 - 1)}$$

Lejeune Dirichlet

Johann Peter Gustav Lejeune Dirichlet

Born	13 February 1805 Düren, French Empire
Died	5 May 1859 (aged 54) Göttingen, Hanover
Residence	Germany
Nationality	German
Fields	Mathematician
Institutions	University of Berlin University of Breslau University of Göttingen
Alma mater	University of Bonn
Doctoral advisor	Simeon Poisson Joseph Fourier
Doctoral students	Ferdinand Eisenstein Leopold Kronecker Rudolf Lipschitz Carl Wilhelm Borchardt
Known for	Dirichlet function Dirichlet eta function

Conjugate priors

- P(θ) and P(θ|D) have the same form

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2}\ldots\theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k}\theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

[A. Singh]

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximiz$\mathcal{D}$ probability of observed data

$$\widehat{\theta} \;=\; \arg\max_{\theta} \quad P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\widehat{\theta} \;=\; \arg\max_{\theta} \quad P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} \;=\; \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

Expected values

Given discrete random variable X, the expected value of X, written E[X] is

$$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$

We also can talk about the expected value of functions of X

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x) P(X = x)$$

Covariance

Given two discrete r.v.'s X and Y, we define the covariance of X and Y as

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

e.g., X=gender, Y=playsFootball

or X=gender, Y=leftHanded

Remember: $$E[X] = \sum_{x \in \mathcal{X}} xP(X = x)$$

You should know

- Density estimation and its relation to classification
- Estimating parameters from data
  - maximum likelihood estimates
  - maximum a posteriori estimates
  - distributions – binomial, Beta, Dirichlet, …
  - conjugate priors