Probability & Bayesian Learning

Hanwool Jeong

hwjeong@kw.ac.kr

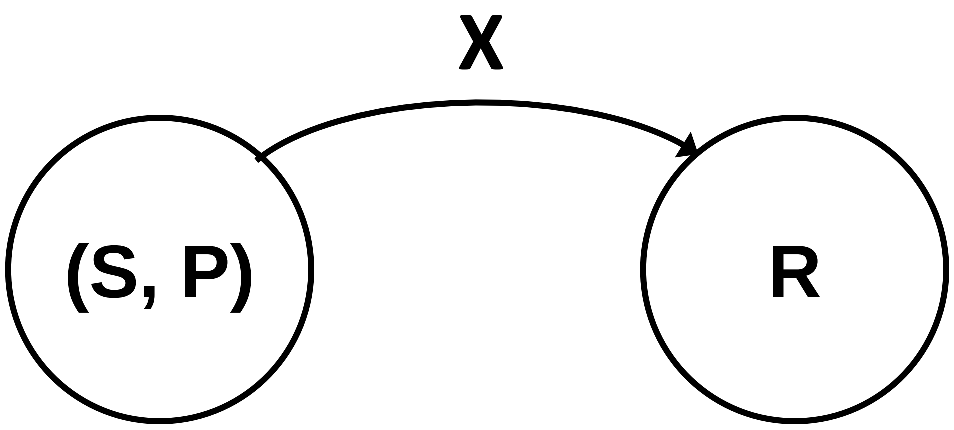
Revisit:
Why We Need Probability & Statistics

- Basically, we want “the machine” can perform delicate jobs.
- Real world data is “uncertain” and “ambigous”
- Handling exception case (or outliers)
- Example

Random Variable

Though we only examine briefly, understanding probability is extremely essential for an engineer!

- A random variable **X** is a ***function*** mapping a probability space (**S**, **P**) into the real line **R**.



Discrete Random Variable (RV)

- Consider a set $\mathcal{X}$ which is a finite or countable infinite set.
- With **discrete random variable** X, the probability of the event that X = x is denoted by p(X = x), or shortly p(x), where $x \in \mathcal{X}$.
- Here p() is called a probability mass function (**PMF**).
- PMF is an example of **probability distribution** which is a function that represents probability that a random variable have a certain value.

Continuous RV

- Suppose X is some uncertain continuous quantity.
- The probability that X lies in any interval a ≤ X ≤ b can be computed as follows.
  - Define the events A = (X ≤ a), B = (X ≤ b) and W = (a < X ≤ b).
  - p(B) = p(A) + p(W) ➔ p(W) = p(B)-p(A)
  - Define the function F(q) = p(X ≤ q). This is called **the cumulative distribution function** (CDF) of X. Thus,

$$p(a < X \leq b) = F(b) - F(a)$$

  - Define $f(x) = \frac{d}{dx}F(x)$ as the **probability density function** or **pdf.**

$$P(a < X \leq b) = \int_a^b f(x)dx$$

- f(x) > 0 for all x, and the density should be integrated to 1

Gaussian (Normal) Distribution

- The most widely used distribution in statistics and ML:

$$\mathcal{N}(x|\mu, \sigma^2) \quad \triangleq \quad \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- $\mu = E[X]$: mean (and mode), $\sigma^2 = \mathrm{var}[X]$: variance
- $X \sim \mathcal{N}(\mu, \sigma^2)$ means $p(X=x) = \mathcal{N}(x|\mu, \sigma^2)$
- The CDF of Gaussian is

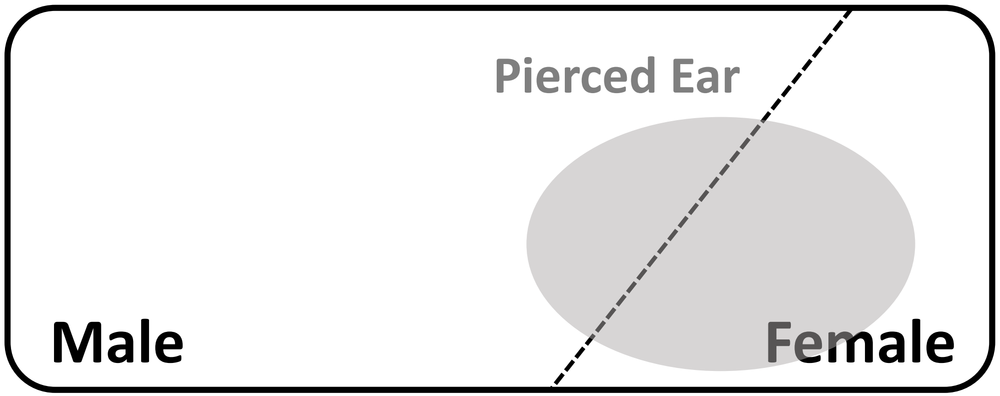
$$\Phi(x; \mu, \sigma^2) \triangleq \int_{-\infty}^{x} \mathcal{N}(z|\mu, \sigma^2)dz$$

Joint Probability

- The probability of joint event A and B is

$$p(A, B) = p(A \wedge B)$$

- Consider an example:
  - A is gender in KW, like Male or Female
  - B is to have pierced ear



Conditional Probability

- We define the conditional probability of event A, given that event B is true, as follows:

$$p(A|B) = \frac{p(A, B)}{p(B)} \text{ if } p(B) > 0$$

- How do we apply this probability to ML?

Bayes Rule

- Bayes rule or Bayes theorem:

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')}$$

Medical Diagnosis Example

- Suppose there is a test which will be positive with probability of 0.8 if one has a cancer, and you had this test.
- If the test is positive for you, what is the probability you have a cancer?
- Base rate fallacy is ignoring the prior

Wrong thought

	Cancer	**No Cancer**
Positive		
Negative		

Example 2

- Should you work hard to get A+? **vs.** If you work hard, can you get A+?
- After examining my grade, I conclude that

	Got an A+	Got something else
Working Hard	18	2
Not working hard	2	8

- How about this?

	Got an A+	Got something else
Working Hard	8	2
Not working hard	12	8

	Got an A+	Got something else
Working Hard	10	10
Not working hard	2	8

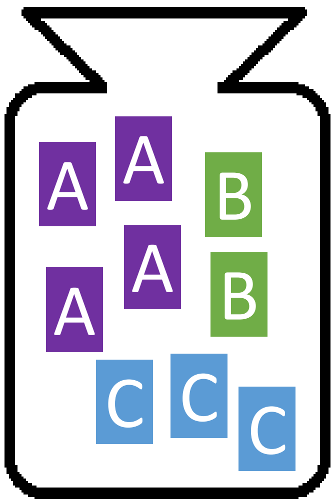
Revisit the Question

- If you work hard, can you get A+?

	Got an A+	Got something else
Working Hard	18	2
Not working hard	2	8

Ocean View Room Example
Let’s Relate it to Machine Learning!

- Suppose that you pick a hotel ticket randomly from the draw. Then you will visit hotel you picked among A, B, and C.
- The probability they give you ocean view room are: 75%, 25%, and 50% for Hotel A, B, and C respectively.
- Please find which hotel you visit, given that you have ocean view room.



Hotel Ticket Draw

Hotel A	Hotel B	Hotel C
Ocean view portion 75%	Ocean view portion 25%	Ocean view portion 50%

Ocean View Room Example (cont’d)

- **Note : the probability that you have the ocean view is not important at all**

Generative Classifier Example

- Generative classifier specifies how to generate the data using p(**x**|y = c) and the class prior p(y = c) as

$$\underbrace{p(y = c \mid \boldsymbol{x})}_{\textbf{Posterior}} = \frac{p(y=c,\boldsymbol{x})}{p(x)} = \frac{\overbrace{p(\boldsymbol{x}|y=c)}^{\textbf{Likelihood}}\overbrace{p(y=c)}^{\textbf{Prior}}}{p(x)}$$

- And we can determine the class by

$$\hat{c} = \underset{c'}{\operatorname{argmax}}\ P(y = c'|x)$$

- This approach is called Maximum a posterior (MAP)

Application to Classification

- Revisiting the Iris flower classification
- Class y ∈ {setosa, versicolor, virginica}, x is 4D feature

Feature Extraction

$$\mathbf{x} = \begin{bmatrix} 8.1 \\ 2.8 \\ 5.0 \\ 1.1 \end{bmatrix}$$

Posterior Estimation

P(setosa|**x**) = 0.15
P(versicolor|**x**) = 0.75
P(virginica|**x**) = 0.10

argmax

versicolor

Bayesian Concept Learning

- Let's deep dive into MAP by looking into the meanings of each component in MAP.

$$\underbrace{p(y = c \mid \boldsymbol{x})}_{\text{Posterior}} = \frac{p(y=c,\boldsymbol{x})}{p(x)} = \frac{p(\boldsymbol{x}|y=c)p(y=c)}{p(x)} \propto \overbrace{p(\boldsymbol{x}|y = c)}^{\text{Likelihood}}\overbrace{p(y = c)}^{\text{Prior}}$$

- Let's include parameter vector θ that characterizes the model.

$$\underbrace{p(y = c \mid \boldsymbol{x}, \boldsymbol{\theta})}_{\text{Posterior}} \propto \underbrace{p(\boldsymbol{x}|y = c, \boldsymbol{\theta})}_{\text{Likelihood}} \times \underbrace{p(y = c|\boldsymbol{\theta})}_{\text{Prior}}$$

Number Game Example

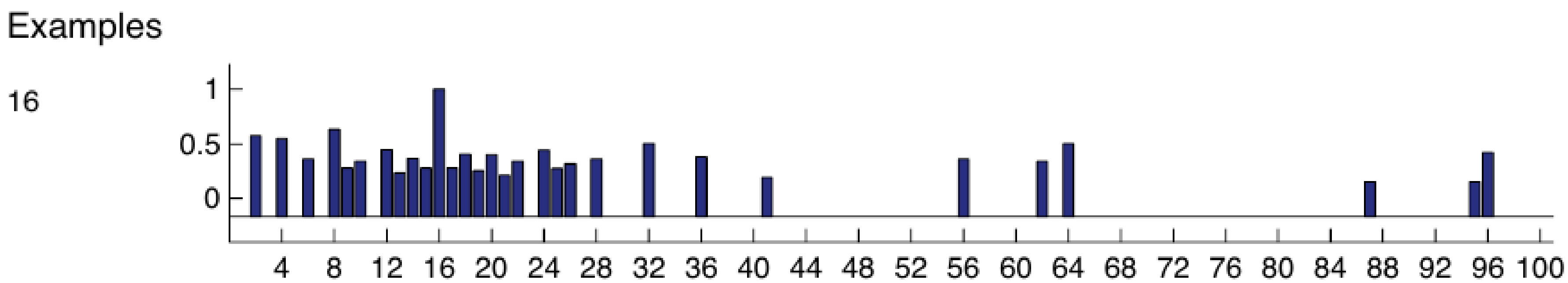
- Number game is introduced in Josh Tenenbaum's PhD thesis (Tenenbaum 1999), which proceeds as follows.
- I choose some simple arithmetical concept C, such as "prime number" or "a number between 1 and 10".
- I then give you a series of randomly chosen positive examples $D = \{x_1, \ldots, x_N\}$ drawn from C, and ask you whether some new test case $\tilde{x}$ belongs to C, i.e., I ask you to classify $\tilde{x}$.
- Suppose, for simplicity, that all numbers are integers between 1 and 100. Now I tell you "16" is a positive example of the concept. What other numbers do you think are positive?
  - 17? 6? 32?
  - How about 99?

Posterior Predictive Distribution

- When I say 16,
  - 17, 6, 32 is more likely than 99, etc.
- We should represent this “degree of being likely” with

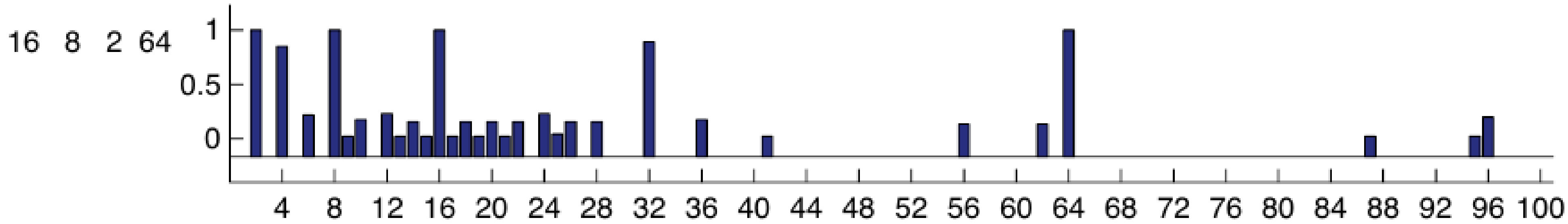
  p($\tilde{x}/\mathcal{D}$) = probability that $\tilde{x} \in$C, given the data $\mathcal{D}$

  which is called **posterior predictive distribution.**
- For $\mathcal{D}$={16}, the following posterior predictive distribution is obtained by people prediction

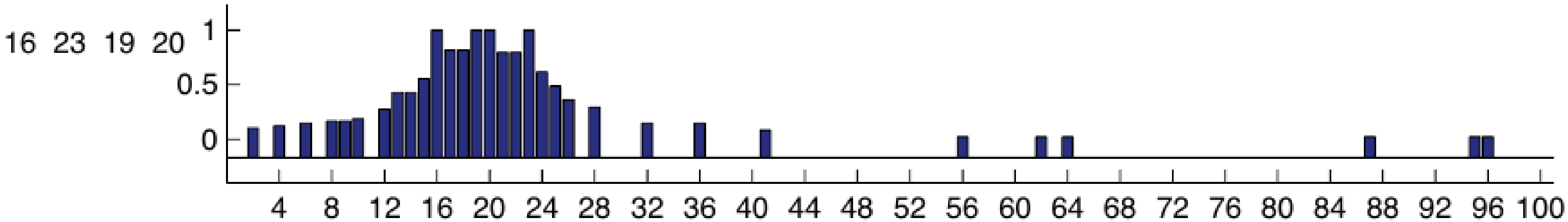


How About with More Data?

- Now I give 8, 2, and 64 are also positive examples, which means $\mathcal{D}$={2,8,16,64}
- You may say the hidden concept is "power of 2" ➔ **induction**
- Posterior predictive distribution is changed into



- For $\mathcal{D}$={16,23,19,20}, generalization gradient is different



We Should Emulate in Machine For

- For induction, we define hypothesis space of concepts, $\mathcal{H}$, such as: odd numbers, even numbers, all numbers between 1 and 100, powers of two, all numbers ending in j.
  - The subset of $\mathcal{H}$ consistent with the $\mathcal{D}$ is **the version space**.
- After seeing $\mathcal{D}$ = {16}, there are many consistent rules; how do you combine them to predict if $\tilde{x} \in$ C?
- After seeing $\mathcal{D}$ = {16, 8, 2, 64}, why did you choose the rule “powers of 2” and not, say, “all even numbers”, or “powers of 2 except for 32”?
- We will see Bayesian explanation for above.

Formalizing "Likelihood"

- After seeing $\mathcal{D}$ = {16, 8, 2, 64}, we will more likely to choose h_{two} = "power of 2", rather than, h_{even} = "even numbers"
- We should explain why and **formalize** this.
- The key is to avoid **suspicious coincidences**:

  ➔ If the true concept was even numbers, how come we only saw numbers that happened to be powers of two?

Quantifying Likelihood

- With the strong sampling assumption, the probability of independently sampling N items (w/ replacement) from *h* is

$$p(\mathcal{D}|h) = \left[\frac{1}{\text{size}(h)}\right]^N = \left[\frac{1}{|h|}\right]^N$$

- This embodies that "the model favors simplest (smallest) hypothesis consistent with $\mathcal{D}$." ➔ **Occam's razor**
- When $\mathcal{D}$ = {16}, $p(\mathcal{D}|h_{two}) = 1/6$ and $p(\mathcal{D}|h_{even}) = 1/50$
  ➔ The likelihood of h_{two} is higher than h_{even}
- When $\mathcal{D}$ = {16,8,2,64}, $p(\mathcal{D}|h_{two}) = (1/6)^4 = 7.7\times10^{-4}$ and $p(\mathcal{D}|h_{even}) = (1/50)^4 = 1.6\times10^{-7}$ ➔ 5000:1 likelihood ratio
- This quantifies the degree of suspicious coincidence

Revisiting Posterior Estimation

- The likelihood in the following corresponds to p($\mathcal{D}$|h).

$$\underbrace{p(y = c \mid \boldsymbol{x}, \boldsymbol{\theta})}_{\textbf{Posterior}} \propto \underbrace{p(\boldsymbol{x} \mid y = c, \boldsymbol{\theta})}_{\textbf{Likelihood}} \times \underbrace{p(y = c \mid \boldsymbol{\theta})}_{\textbf{Prior}}$$

- As our goal is to derive $\underbrace{\mathrm{p}(\mathrm{h}|\mathcal{D})}_{\text{Posterior}} \propto \underbrace{\mathrm{p}(\mathcal{D}|\mathrm{h})}_{\text{Likelihood}} \times \underbrace{\mathrm{p}(\mathrm{h})}_{\text{Prior}}$

- What we also need to examine prior term, p(h). What is the meaning of this?

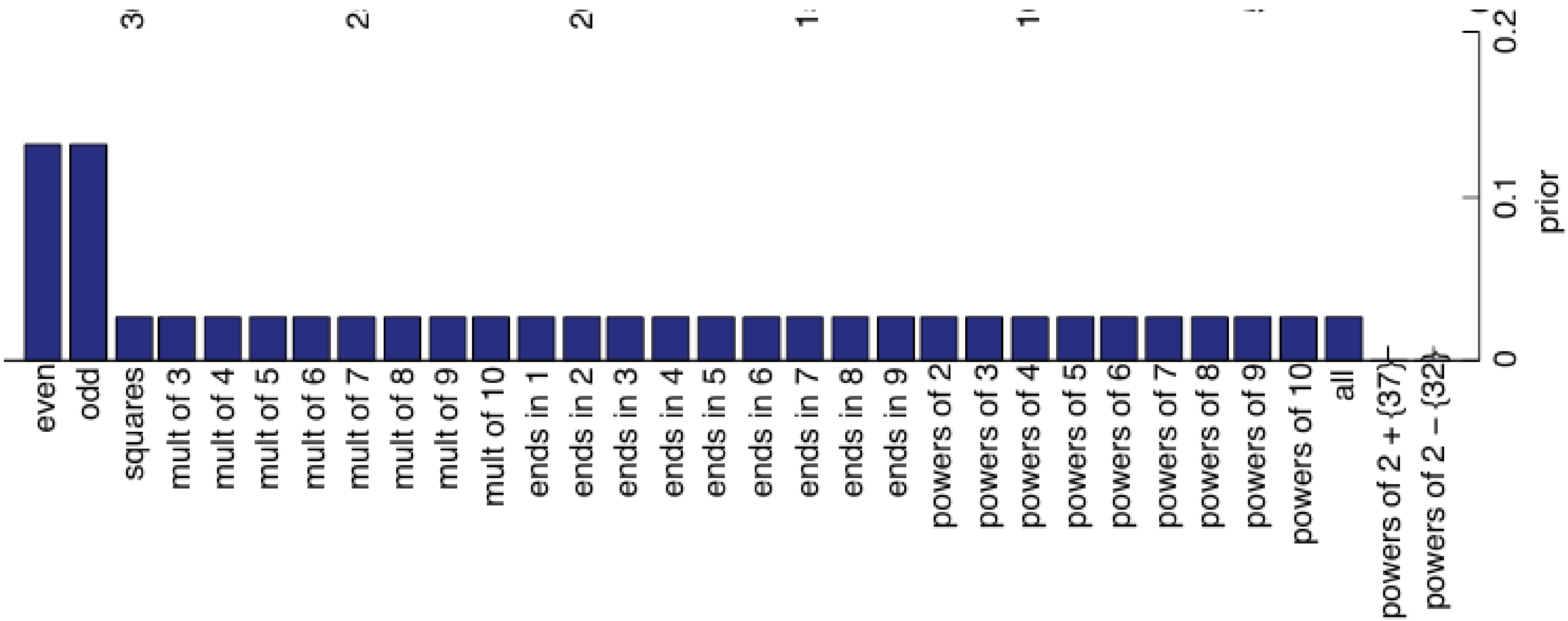
Necessity of Prior

- The likelihood is higher for h' = "powers of 2 except 32" than h_{two} = "powers of 2"
- h ="powers of two except 32" seems "conceptually unnatural"
  ➔ We should reflect this as p(h) preventing **overfitting**
- p(h) is subjective thus making Bayesian reasoning unreliable
- However, p(h) is useful because it reflects the background knowledge about data
- Ex) with $\mathcal{D}$ = {1200, 1500, 900, 1400} ➔ 400 vs. 1183.
  - Background 1) : the data are picked based on arithmetic rule.
  - Background 2) : the data are human cholesterol level.

  ➔ Different background for data determines p(h) and significantly enhances the efficiency of ML.

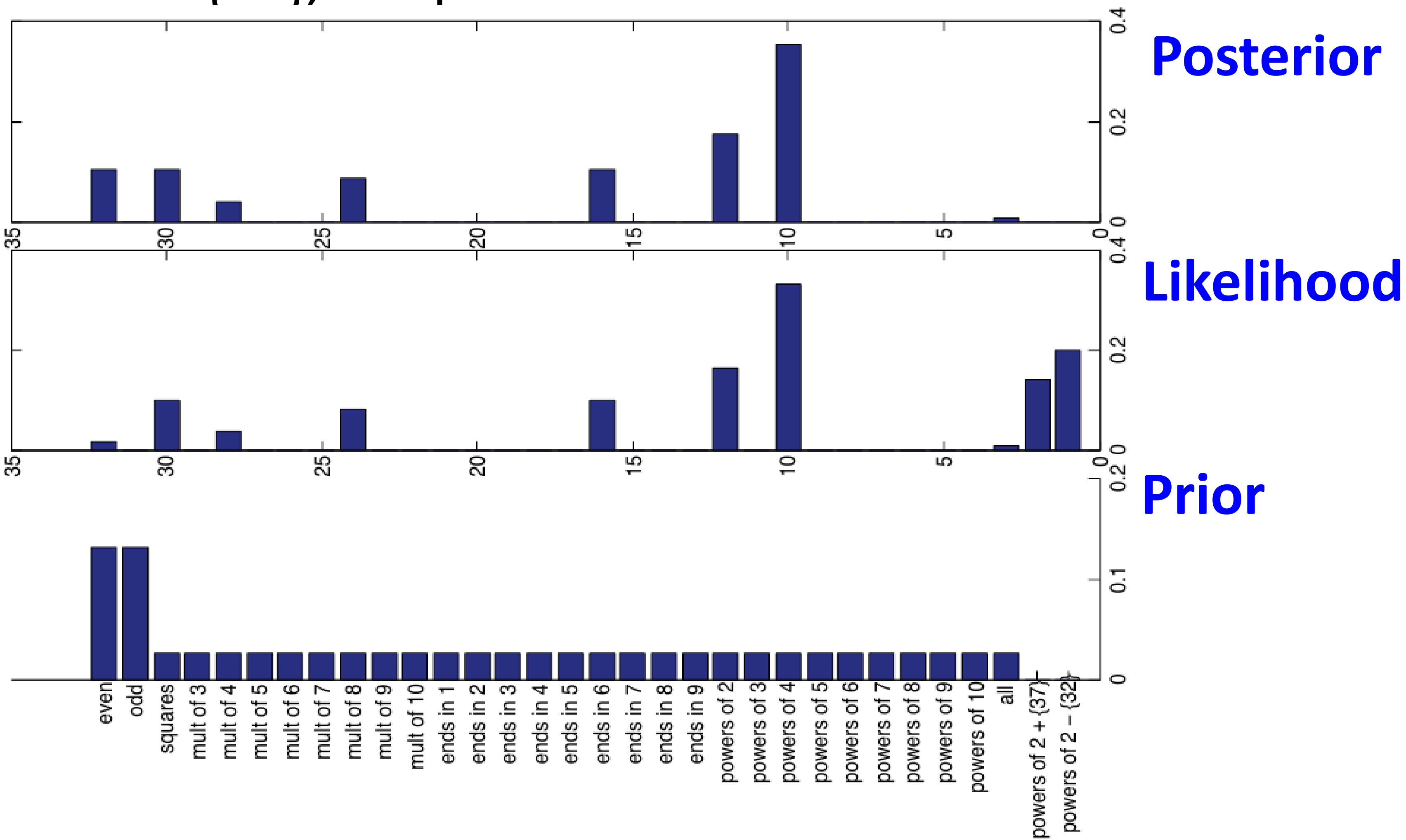
Prior Example for Number Game

- The “unnatural” concepts of “powers of 2, plus 37” and “powers of 2, except 32” have very low prior.



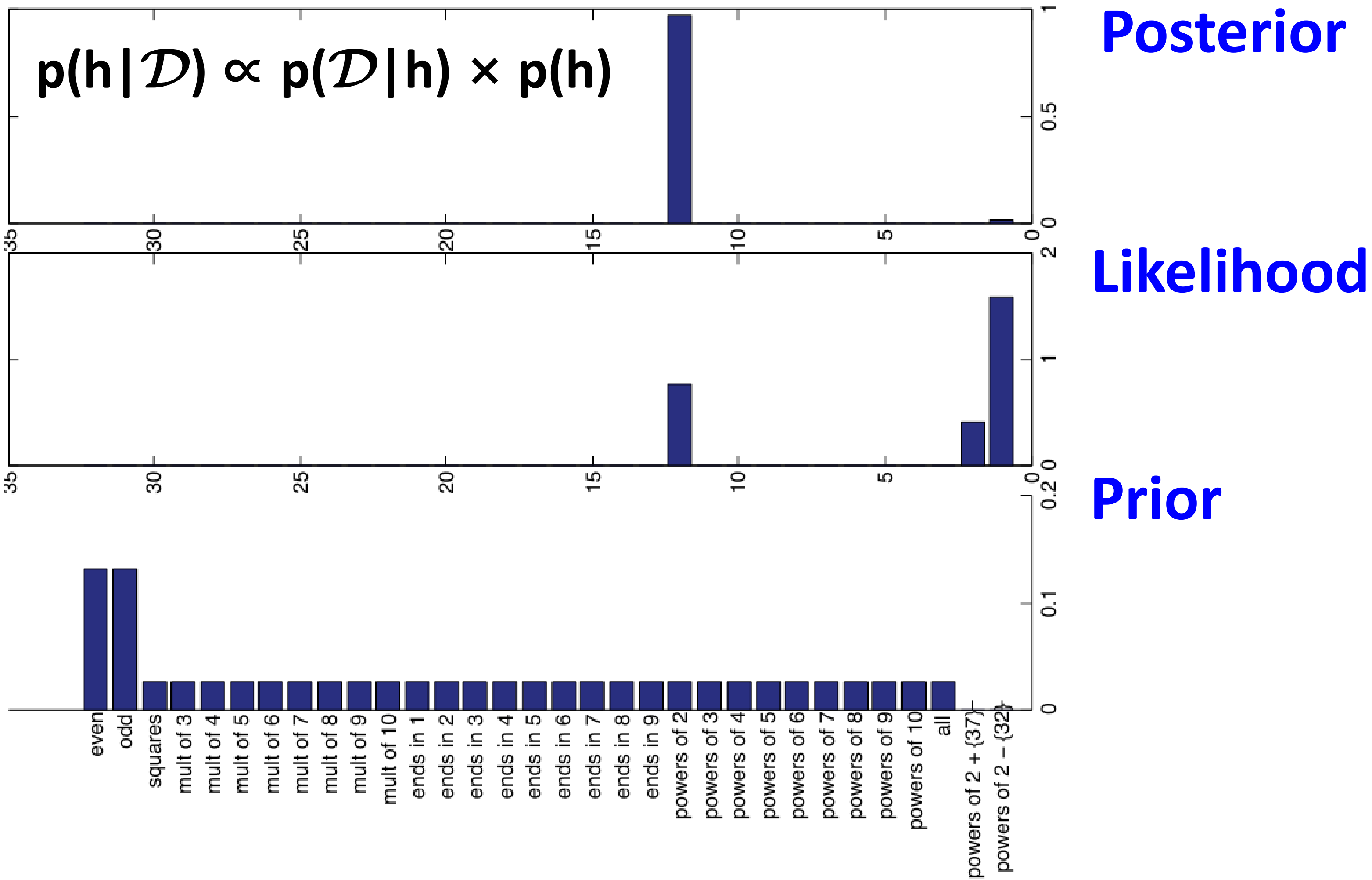
Finally, Posterior

- Revisiting $p(h|\mathcal{D}) \propto p(\mathcal{D}|h) \times p(h)$
  Posterior Likelihood Prior
- When $\mathcal{D}$ = {16}, the posterior is derived as follows



Posterior when $\mathcal{D}$ = {16, 8, 2, 64}

- Having enough data, the posterior becomes peaked on a single concept, namely MAP esimate.



MAP Estimate vs. MLE

- MAP Esimate means

$$\hat{h}^{MAP} = \text{argmax}_h \, p(h|\mathcal{D})$$

- which can be written as

$$\hat{h}^{MAP} = \underset{h}{\text{argmax}} \, p(\mathcal{D}|h)p(h) = \underset{h}{\text{argmax}} \, [\log p(\mathcal{D}|h) + \log p(h)]$$

- Since the likelihood term depends exponentially on N , and the prior stays constant, as we get more and more data, the MAP estimate converges towards the maximum likelihood estimate or MLE:

$$\hat{h}^{mle} \triangleq \underset{h}{\text{argmax}} \, p(\mathcal{D}|h) = \underset{h}{\text{argmax}} \log p(\mathcal{D}|h)$$

- If we have enough data, data overwhelms the prior.

MAP vs. MLE

- $p(h|\mathcal{D}) \propto p(\mathcal{D}|h) \times p(h)$
- If p(h) is constant over various h's, MAP is equivalent to MLE