

Recall Our Syllabus

- Basic concepts of generative models ✓
- **Generative models in statistical learning**
 - Naive Bayesian models, Hidden Markov models (HMM), Gaussian mixture models (GMM)
 - Probabilistic latent semantic analysis models (pLSA), latent Dirichlet allocation (LDA)
- Generative models in deep learning
 - Autoregressive models
 - Variational auto-encoders (VAEs)
 - Generative adversarial networks (GANs)
 - Flow-based models
 - Energy-based models (EBMs)
 - Score-based / diffusion generative models (DMs)
 - Large-language models (LLMs)
- Applications in different areas
 - Image processing: image generation
 - NLP: text generation
 - Speech recognition: audio generation
 - Video processing: video generation
 - Drug design & discovery: molecule generation



Statistical *vs.* Deep Generative Models

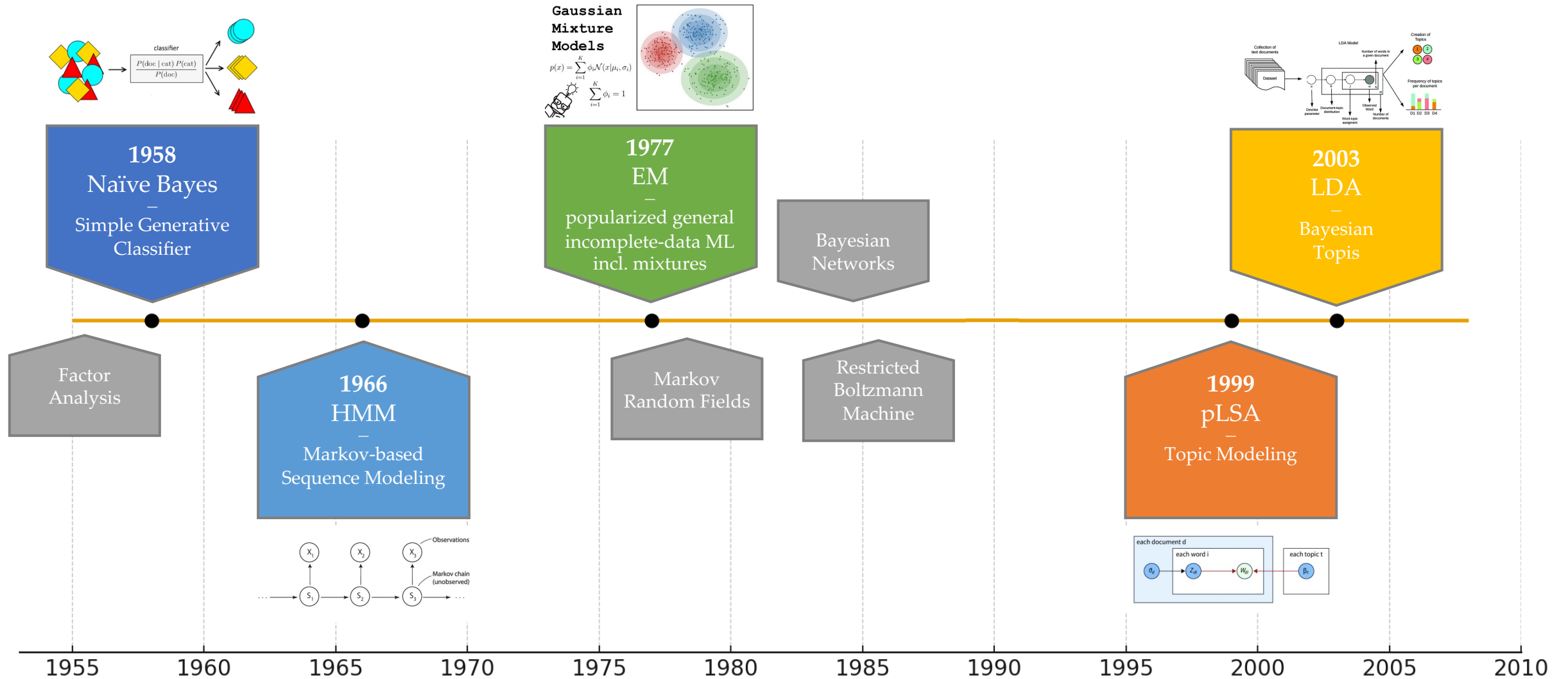
- Statistical generative models

- Rely on **strong distributional assumptions** (*e.g., Gaussian, Multinomial, Dirichlet*)
- Limited by the **chosen distributional forms** and **structures** (*e.g., graph*)
- representation; learning/estimation (closed form, MLE/MAP, EM) ; inference/sampling (exact, VI, MCMC, ancestral/iterative sampling)

- Deep generative models

- Make **weaker assumptions**, using DNNs to approximate complex distributions directly
- Highly expressive, capable of modeling **complex high-dimensional data** (*e.g., images, audio, text*)
- Learning (ELBO, denoising loss), Leverage **neural network-based inference** (*e.g., VAE, GAN, Flow, DM*)

Statistical Generative Models: A Timeline



Discrete *vs.* Continuous Data

- Assume we want to generate...
 - A piece of text *or* an image
 - The underlying statistical modeling are not the same
- Data can be discrete or continuous
 - The design of generative models must be chosen according to the data type



Generated by ChatGPT 5

Discrete *vs.* Continuous Data

- **Text data**

- Word counts in documents (*bag-of-words*)
- Email classification (*spam vs. ham*)

- **Categorical tabular features**

- Patient symptoms (*positive vs. negative*)
- Weather data (*sunny, cloudy, windy*)

- **Sequential data**

- Word sequences in NLP
- DNA sequences (*A/C/G/T symbols*)
- User click sequences (*modeling user behavior*)

- **Numerical tabular features**

- Height, weight, age
- Most medical test indicators (*blood pressure, glucose*)

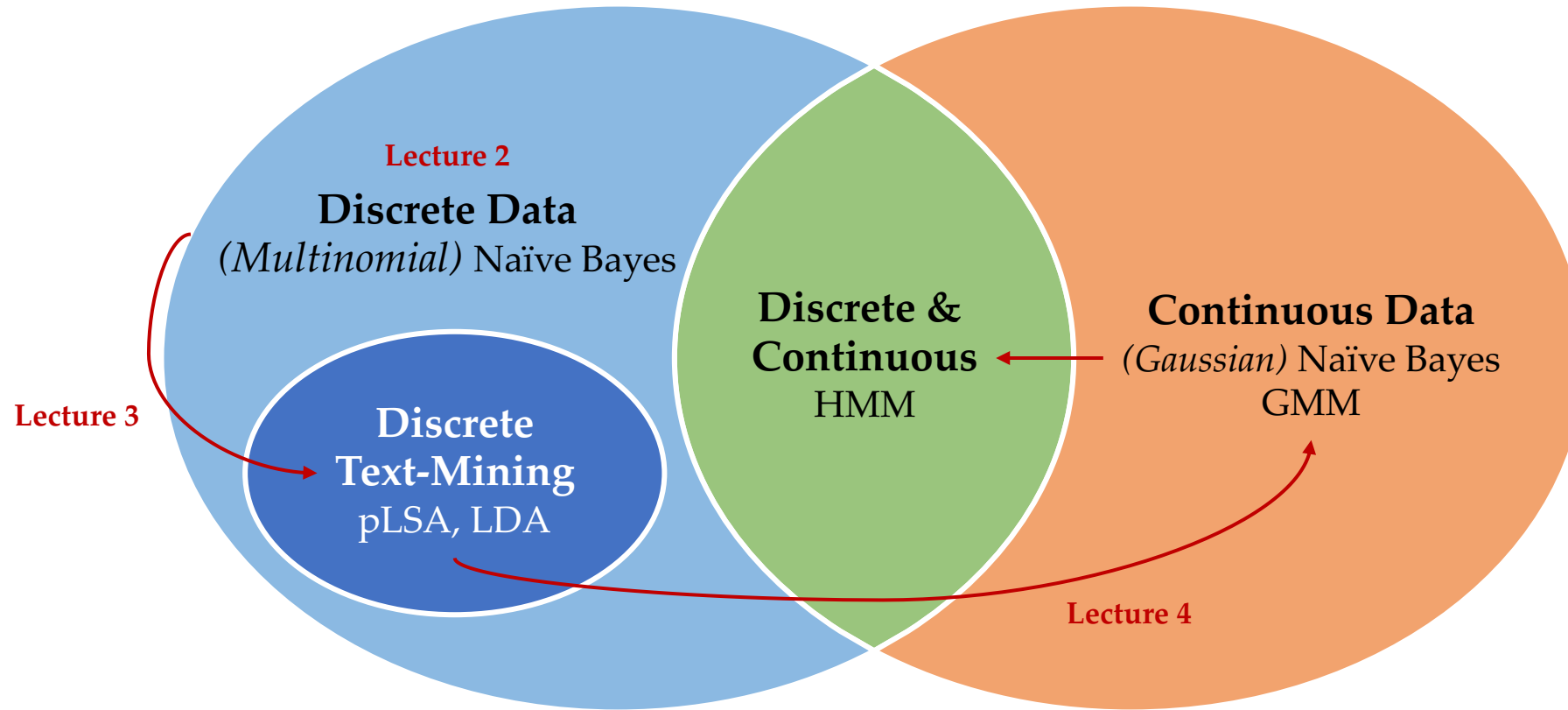
- **Image / speech features**

- Image pixel values
- Acoustic features of speech (*MFCC vectors*)

- **Financial / sensor data**

- Stock price time series
- IoT sensor data (*temperature, current, voltage*)

A Roadmap



Generative Models: Fundamentals and Applications

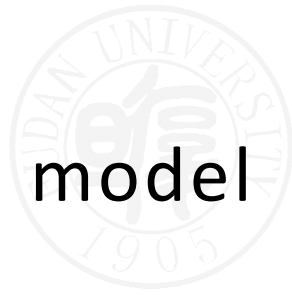
Lecture 2: Generative Models for Discrete Data

Shuigeng Zhou, Wenhui Liu
College of CSAI

September 21, 2026

Outline

- Generative vs. Discriminative: Revisit
- Bayesian concept learning
- The beta-binomial model
- The Dirichlet-multinomial model
- Naive Bayes classifiers



Generative vs. Discriminative: Revisit

- **Frequentist:** θ seen as fixed
 - e.g., a point estimation $\hat{\theta}$
- **Bayesian:** θ seen as learned
 - Given data D , $P(\theta|D) = P(D|\theta)P(\theta)/P(D) \propto P(D|\theta)P(\theta)$

Generative vs. Discriminative: Revisit

- Discriminative models are *often* frequentist, and generative models are *often* Bayesian
 - — but the two dimensions are orthogonal and can be mixed

	Frequentist	Bayesian
Discriminative	Logistic Regression, SVM, Neural Nets	Bayesian Logistic Regression, Bayesian Neural Nets
Generative	GMM (with MLE), HMM (with MLE)	Naive Bayes , Bayesian GMM, Variational Bayes models

Generative vs. Discriminative: Revisit

- Goal of discriminative model
 - Learn the **conditional distribution** $P(Y|X; \theta)$
 - θ often seen as fixed, hence simplifies to $P(Y|X)$



Generative vs. Discriminative: Revisit

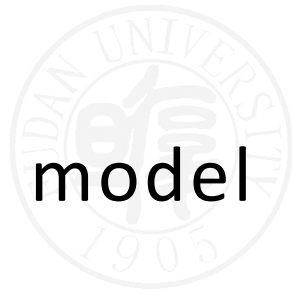
- Goal of generative model
 - Learn the **joint distribution** $P(X, Y; \theta)$
- Generative classifier

$$\begin{aligned} P(Y = c|x, \theta) &= \frac{P(Y=c, x | \theta)}{P(x|\theta)} \\ &= \frac{P(Y=c | \theta)P(x|Y=c, \theta)}{\sum_{c' \in \mathcal{C}} P(x, Y=c' | \theta)} \\ &= \frac{\text{class-prior distribution} \quad \boxed{P(Y=c | \theta)} \quad \boxed{P(x|Y=c, \theta)} \quad \text{class-conditional distribution}}{\sum_{c' \in \mathcal{C}} P(Y=c' | \theta) P(x|Y=c', \theta)} \end{aligned}$$

- Key: class-conditional distribution $P(x|Y = c, \theta)$

Outline

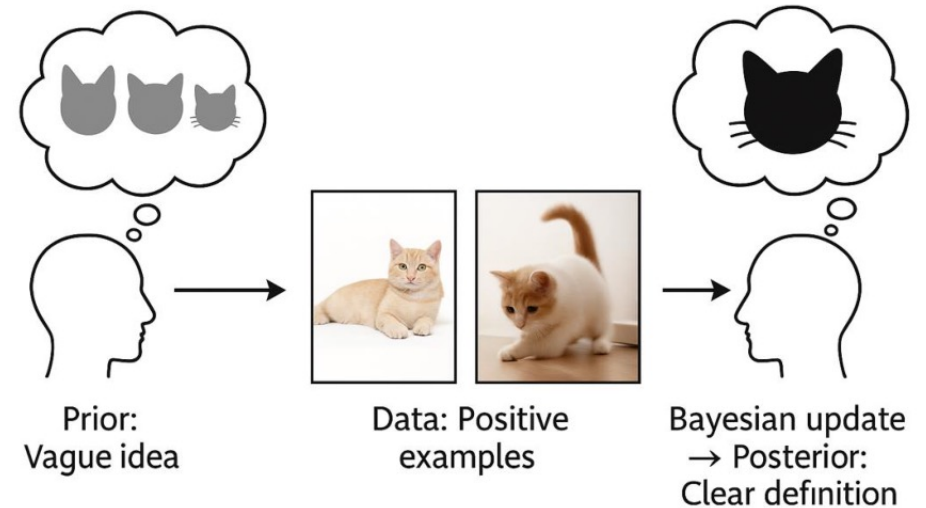
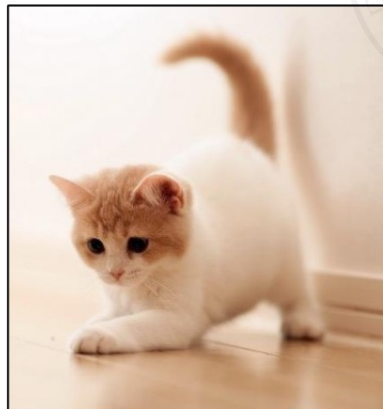
- Generative vs. Discriminative: Revisit
- Bayesian concept learning
- The beta-binomial model
- The Dirichlet-multinomial model
- Naive Bayes classifiers



Bayesian concept learning

- Concept learning
 - Provide **only the positive** examples
 - Learn the meaning of the example

This is a cat



Bayesian concept learning

- Concept learning

- Be equivalent to binary classification

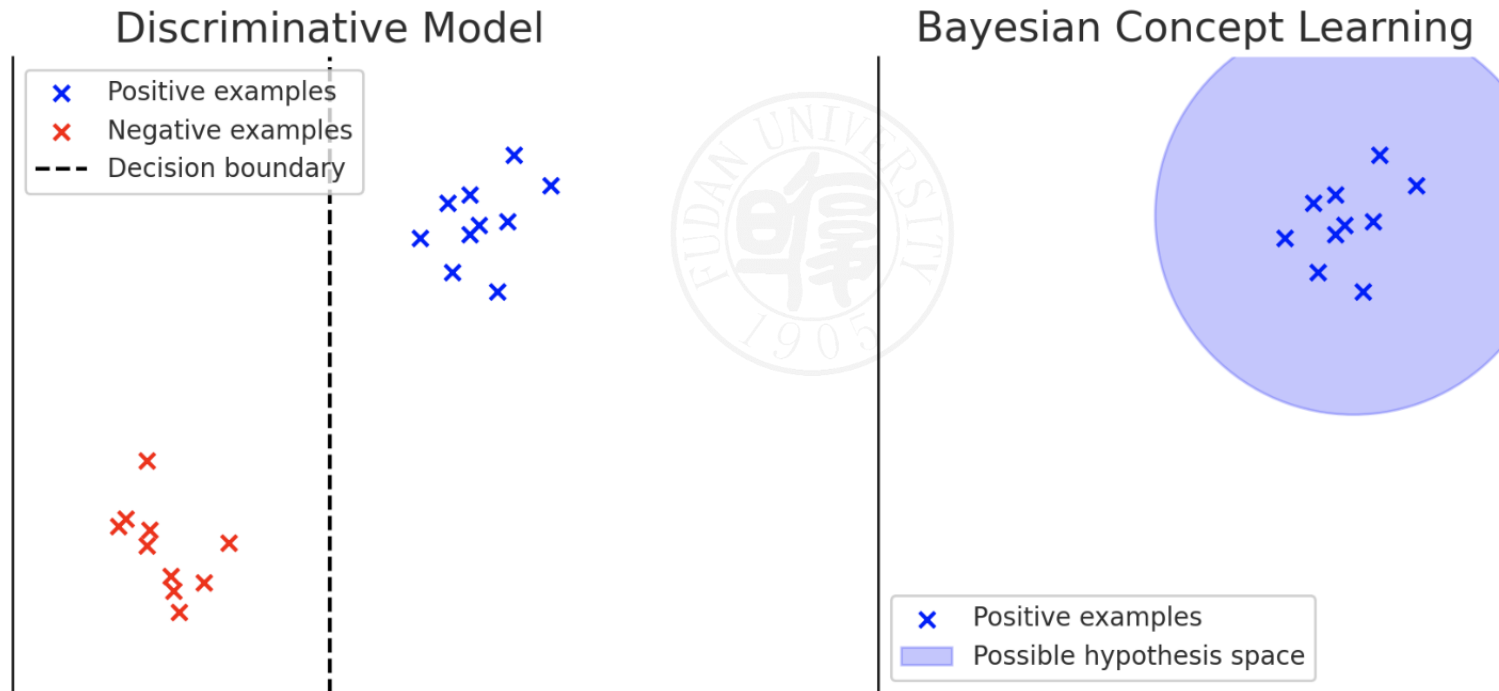
- The goal is to learn the **indicator function f** , determining which elements are in the set C .

$$f(x) = \begin{cases} 1, & x \in C \\ 0, & x \notin C \end{cases}$$

- **Difference** : binary classification provides both the positive and negative examples, while concept learning provides **only the positive examples**

Bayesian concept learning

■ Bayesian concept learning vs. binary classification



Example: number game

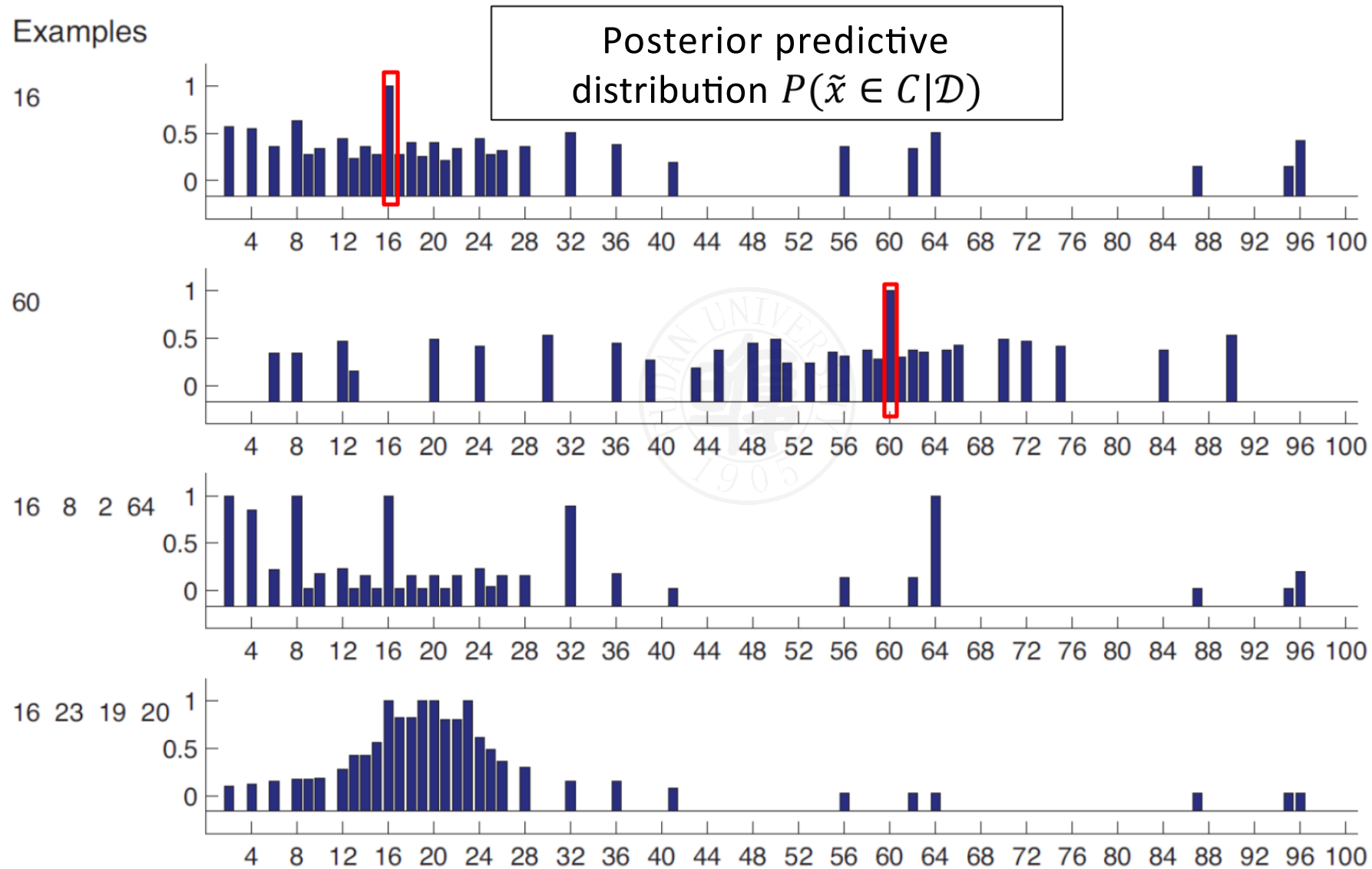
- Given some **unknown** arithmetical concepts $\mathcal{C}$, such as “prime number” or “a number between 1 and 10”
- Given positive examples $\mathcal{D} = \{x_1, x_2, \dots, x_N\} \subseteq \mathcal{C}$
- **Question**: the new sample $\tilde{x} \in \mathcal{C}$?
 - Learn a basic rule to determine whether the data follows the unknown arithmetical concepts $\mathcal{C}$

Example: number game

- Suppose $\mathcal{D}$ and the test set are from $[1,100]$
- Concept set $\mathcal{C}$ is **not clear**
- Four experiments
 - Experiment 1: $\mathcal{D} = \{16\}$
 - Experiment 2: $\mathcal{D} = \{60\}$
 - Experiment 3: $\mathcal{D} = \{16,8,2,64\}$
 - Experiment 4: $\mathcal{D} = \{16,23,19,20\}$



Example: number game



Bayesian concept learning

- Posterior predictive distribution $p(\tilde{x} \in \mathcal{C}|\mathcal{D})$
 - Describe the probability that $\tilde{x} \in \mathcal{C}$ given the data $\mathcal{D}$ for any $\tilde{x} \in \{1, 2, \dots, 100\}$
- Question: How to explain the behavior and emulate it in a machine?

Bayesian concept learning

■ Hypothesis space $\mathcal{H}$:

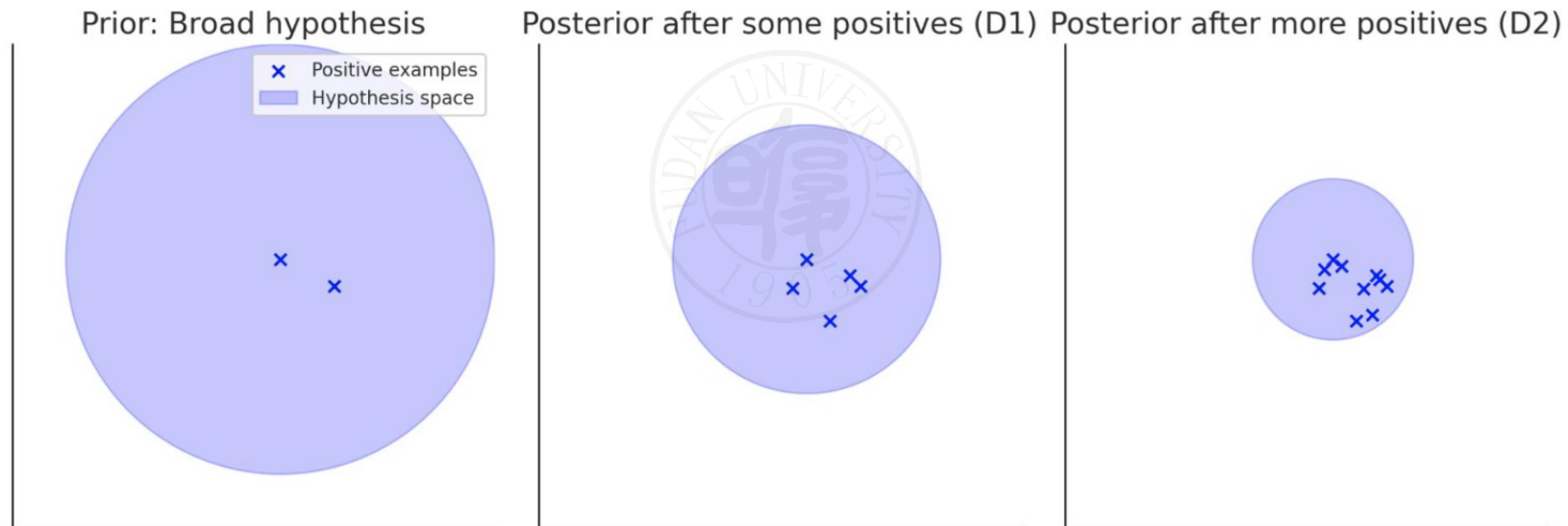
- Include all possible models or rules for the concept set
- Eg: odd numbers, even numbers, powers of two, all numbers ending in j (for $0 \leq j \leq 9$)

■ Version space

- The subset of $\mathcal{H}$ that is consistent with the data $\mathcal{D}$
- As we see more examples, the version space shrinks and we become increasingly certain about
 - Eg. $\mathcal{D} = \{16\} \rightarrow \mathcal{D} = \{16, 8, 2, 64\}$

Bayesian concept learning

■ The shrinking of version space



Likelihood

- Given $\mathcal{D} = \{16, 8, 2, 64\}$, which hypothesis is more possible?
 - $h_{two} = \text{"powers of two"}$
 - $h_{even} = \text{"even number"}$
- **Extension** of a concept
 - The set of numbers that belong to the concept
 - the extension of h_{two} is $\{2, 4, 8, 16, 32, 64\}$
 - the extension of h_{even} is $\{2, 4, 6, \dots, 98, 100\}$



Likelihood

- Strong sampling assumption
 - Assume N examples are sampled **uniformly** at **random** from the extension of a concept.
- **Likelihood** $p\{\mathcal{D}|h\}$
 - the probability of independently sampling N items (**with replacement**) from h that happen to constitute $\mathcal{D}$

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

Likelihood

- Compute the probability $P\{\mathcal{D}|h\}$?

- Let $\mathcal{D} = \{16\}$

- $p(\mathcal{D}|h_{two}) = 1/6$
- $p(\mathcal{D}|h_{even}) = 1/50$

- Let $\mathcal{D} = \{16, 8, 2, 64\}$

- $p(\mathcal{D}|h_{two}) = (1/6)^4 = 7.7 \times 10^{-4}$
- $p(\mathcal{D}|h_{even}) = (1/50)^4 = 1.6 \times 10^{-7}$
- **likelihood ratio** : $\frac{p(\mathcal{D}|h_{two})}{p(\mathcal{D}|h_{even})} = 4822.5$



Likelihood

- Size principle (Occam's razor)
 - The model favors the simplest (smallest) hypothesis consistent with the data

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

- Among all hypotheses consistent with $\mathcal{D}$, the fewer data a hypothesis h covers, the higher its likelihood.

Prior

- Given $\mathcal{D} = \{16, 8, 2, 64\}$, which hypothesis seems “conceptually unnatural” ?
 - $h_{two} = \text{"powers of two"}$
 - $h_{two} = \{2, 4, 8, 16, \textcolor{red}{32}, 64\}$
 - $p(\mathcal{D}|h_{two}) = (1/6)^4$
 - $h'_{two} = \text{"powers of two except } \textcolor{red}{32}\text{"}$
 - $h'_{two} = \{2, 4, 8, 16, 64\}$
 - $p(\mathcal{D}|h'_{two}) = (1/5)^4$
 - The likelihood of h'_{two} is higher than that of h_{two}

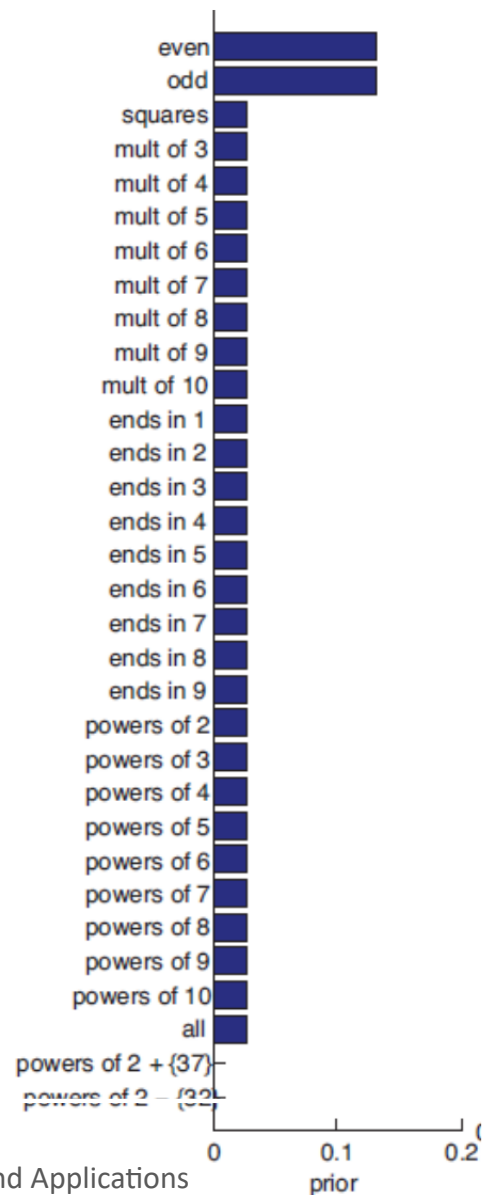
Prior

- Prior
 - Assign **low prior probability** to unnatural concepts
- Eg: given $\mathcal{D}=\{1200,1500,900,1400\}$, classify 400 and 1183 ?
 - The numbers are from some arithmetic rule
 - 400 is likely but 1183 is unlikely
 - The numbers are examples of healthy cholesterol levels
 - 400 is unlikely but 1183 is likely

Example: what prior to use

■ Hypothesis space

- Even numbers
- Odd numbers
- Squares
- Multiples of j ($3 \leq j \leq 10$)
- Ends in j ($1 \leq j \leq 9$)
- Powers of j ($2 \leq j \leq 10$)
- All
- **Powers of 2, plus 37**
- **Powers of 2, except 32**



Posterior

- posterior = normalization (likelihood times prior)

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}, h')} = \frac{p(h)\mathbb{I}(\mathcal{D} \in h)/|h|^N}{\sum_{h' \in \mathcal{H}} p(h')\mathbb{I}(\mathcal{D} \in h')/|h'|^N}$$

- Prior vs. Posterior
 - Prior: **Without** training data, quantify the possibility of each hypothesis
 - Posterior: **Given** training data, quantify the possibility of each hypothesis

Example

- Given $\mathcal{D}=\{16\}$
 - h ="powers of 2"

- Prior

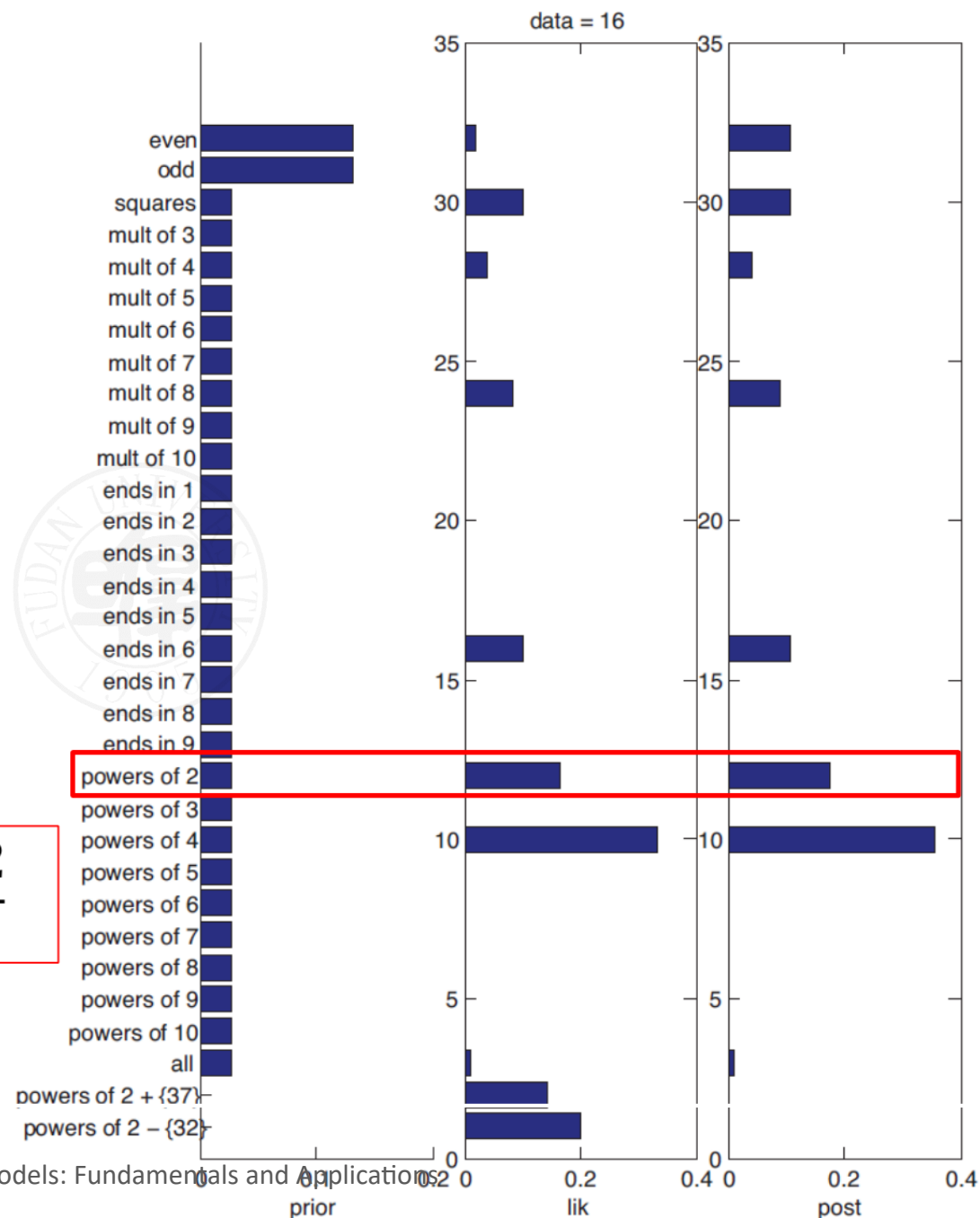
$$p(h) = 0.025$$

- Likelihood

$$p(\mathcal{D}|h) = 1/6$$

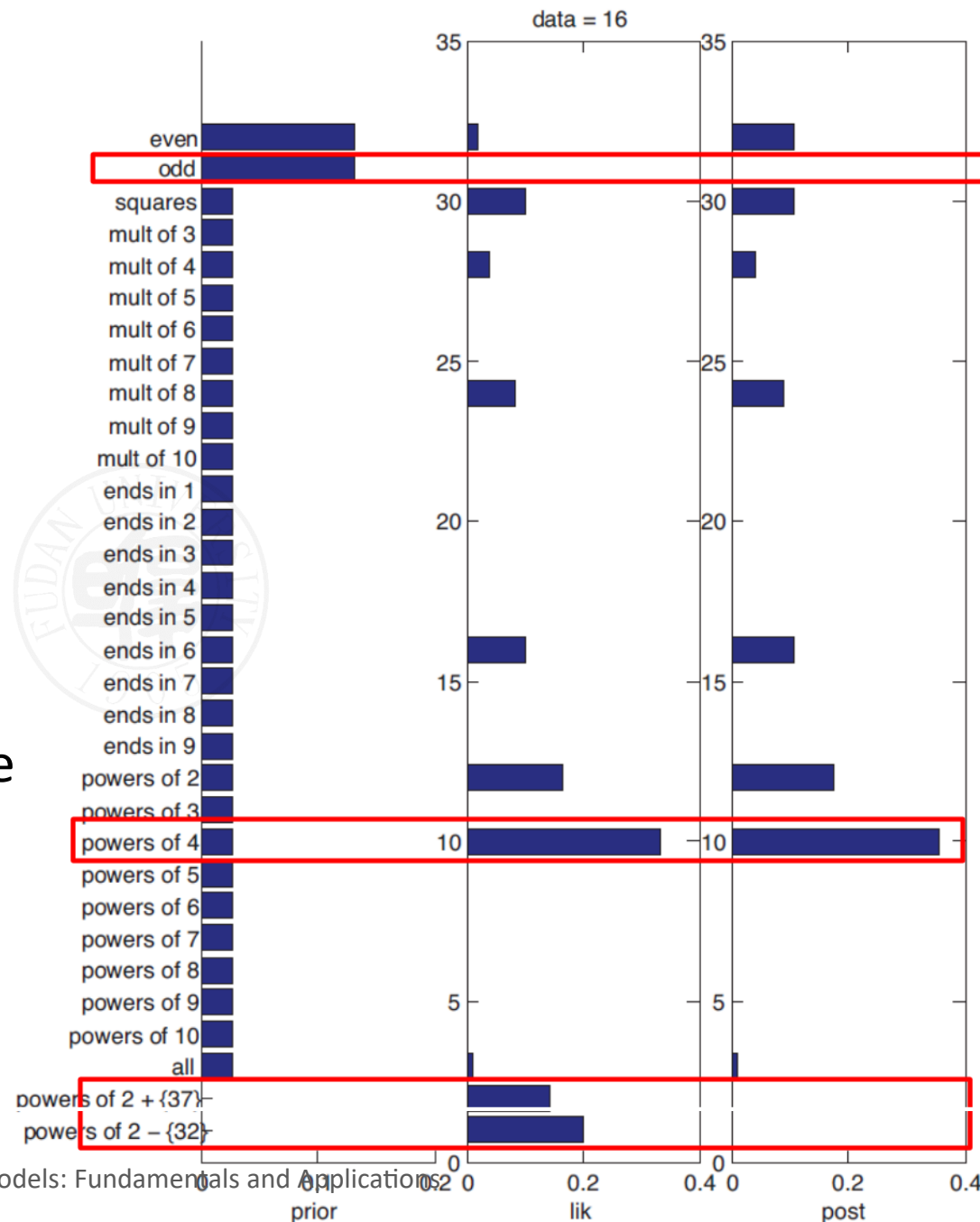
- Posterior

$$p(h|\mathcal{D}) = \frac{0.0042}{\text{sum}}$$



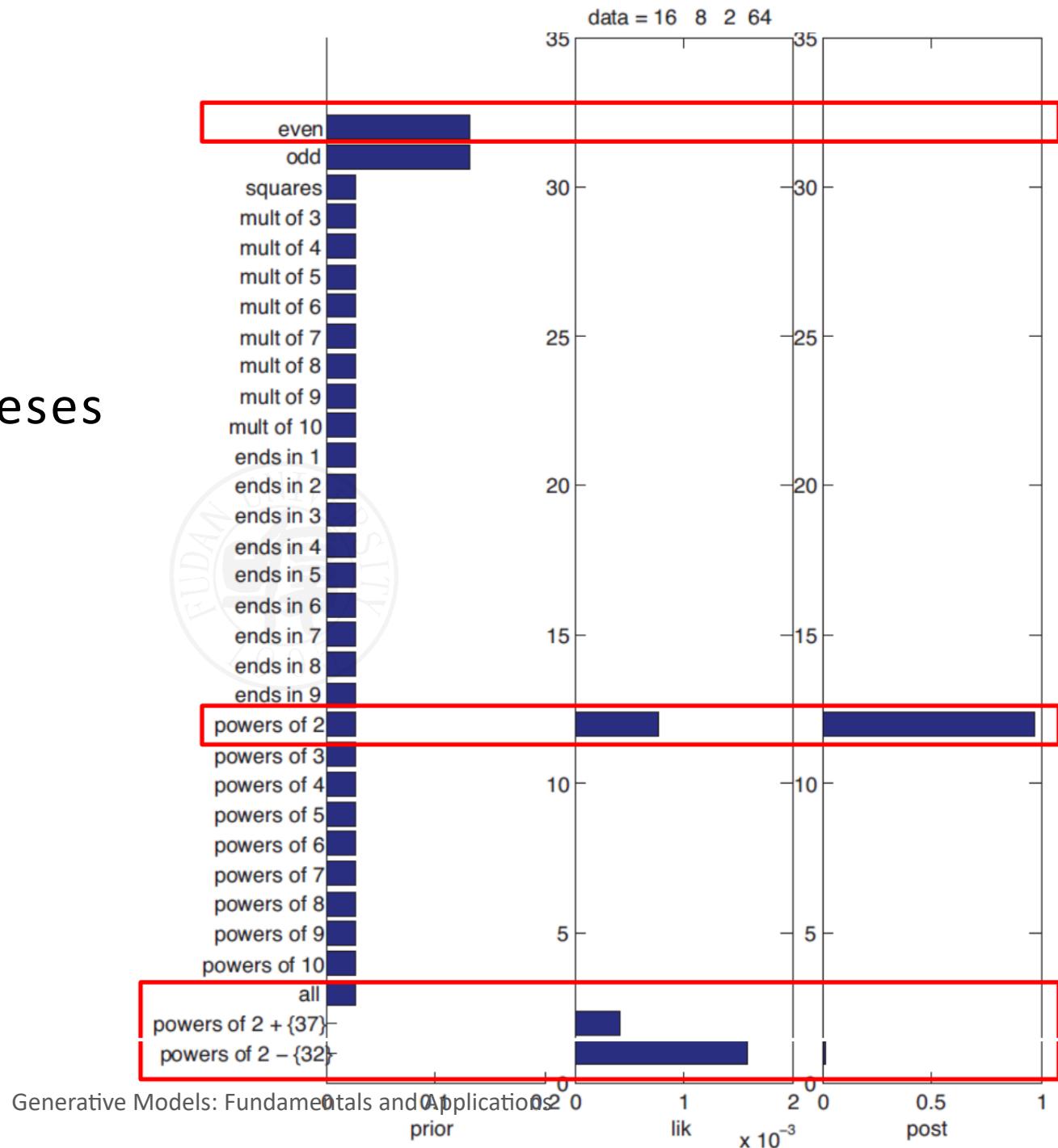
Example

- From the results, we learn
 - The hypothesis with the highest posterior is "**powers of 4**"
 - Prior and likelihood jointly determine the posterior
 - Even if the prior is large, if the likelihood is too small, the posterior can still be small
 - And vice versa



Example

- Given $\mathcal{D}=\{16,8,2,64\}$
- Five possible hypotheses



Example

- Given $\mathcal{D}=\{16,8,2,64\}$

- h ="powers of 2"

- Prior

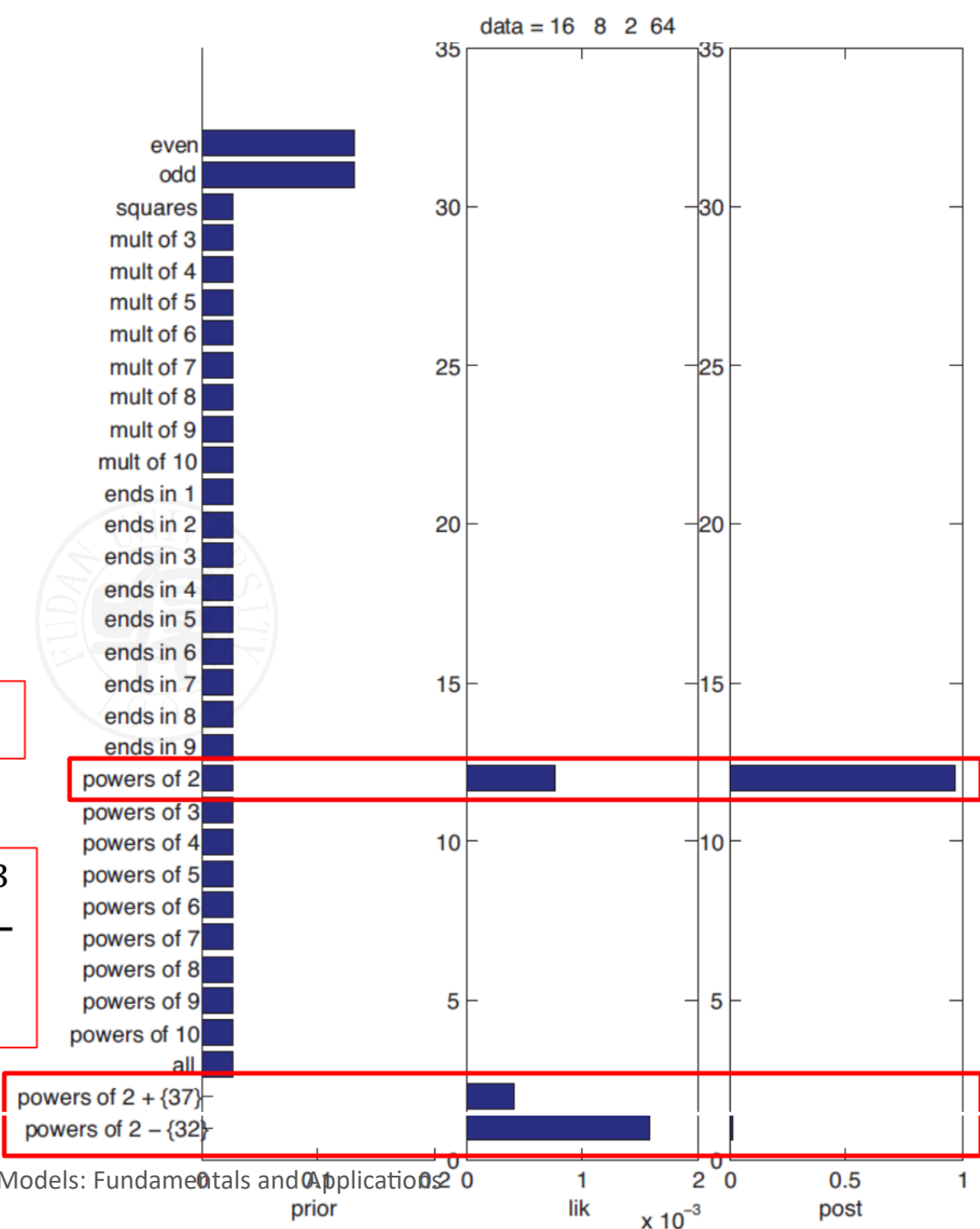
$$p(h) = 0.025$$

- Likelihood

$$p(\mathcal{D}|h) = 0.77 \times 10^{-3}$$

- Posterior

$$p(h|\mathcal{D}) = \frac{0.019 \times 10^{-3}}{\text{sum}} \approx 1$$



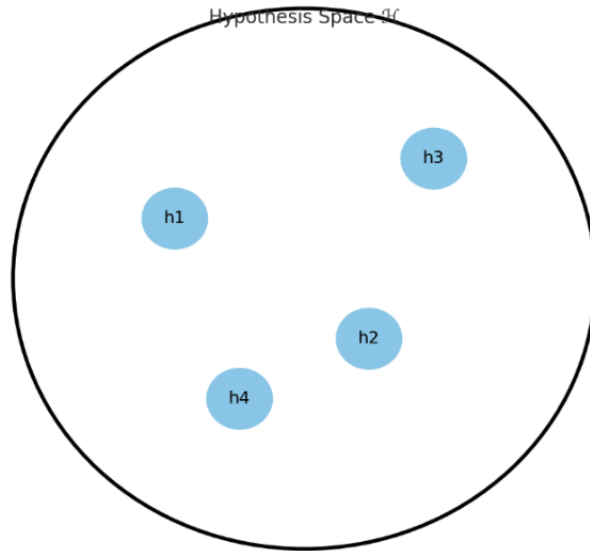
Example

■ Result:

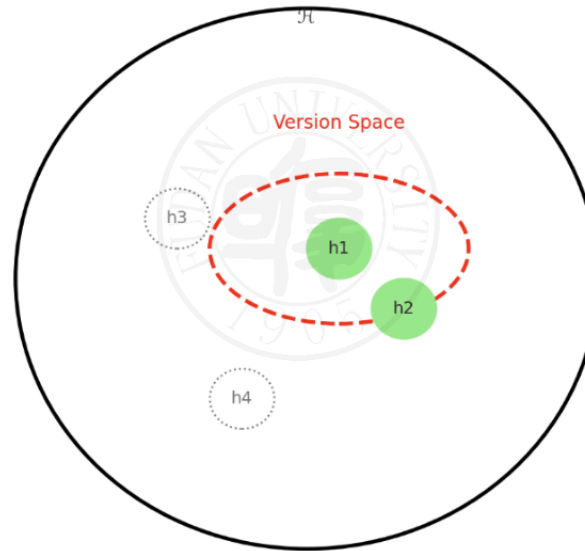
- Assigning small priors to unnatural hypotheses **helps avoid overfitting the data**
- When $|\mathcal{D}|$ is sufficiently large, $p(h \mid \mathcal{D})$ will peak at a single hypothesis (reach its maximum)
- Taken as the **optimal hypothesis**
- Take-away:
 - *Prior prevents overfitting, and more data changes likelihood which sharpens the posterior to the true hypothesis.*

Example

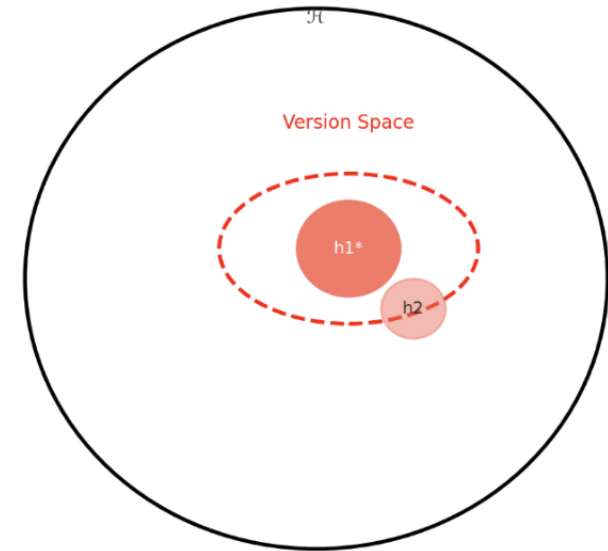
Step 1: Hypothesis space



Step 2: Version space after data



Step 3: Posterior update



Summary of Concepts

- **Hypothesis space:** the set of all possible hypotheses (concepts)
- **Hypothesis:** a specific member or concept of the hypothesis space
- **Version space:** the subset of hypotheses consistent with observed data
- **Extension:** the set of samples that satisfy a given hypothesis
- **Prior:** initial belief or bias over hypotheses before seeing data
- **Likelihood:** probability of observing the data given a hypothesis (sampling from its extension)
- **Posterior:** updated belief about a hypothesis after observing data
- **Optimal hypothesis:** the hypothesis with the highest posterior probability

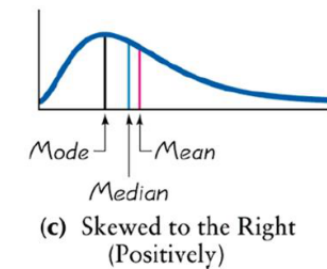
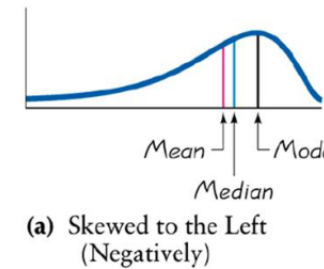
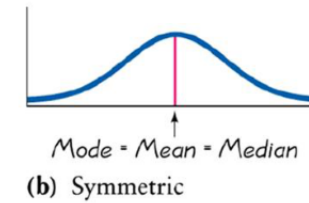
Posterior

- **MAP** (Maximum *a Posteriori*) estimate

$$\max_{h \in \mathcal{H}} p(h|\mathcal{D})$$

- It simply finds the posterior mode

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



Posterior

■ MAP estimate

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}, h')}$$



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

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



□ Note:

- The log-likelihood grows linearly with the number of elements N .
- When N is sufficiently large, MAP is dominated by the log-likelihood and becomes almost independent of the log-prior.

Posterior

- **MLE** (Max-likelihood Estimation)

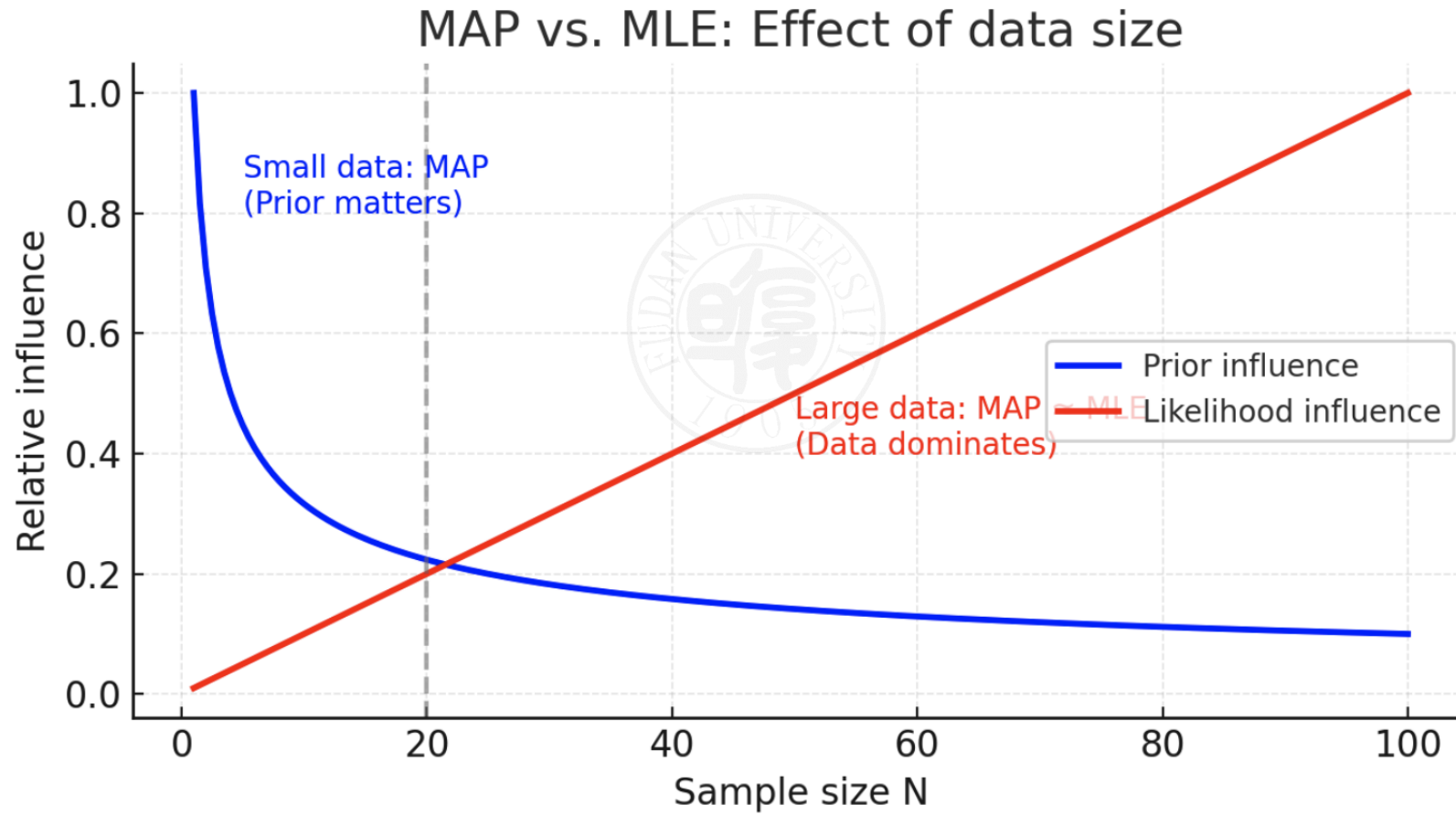
- Choose $\hat{\theta}$ that maximize the likelihood of the observed data

$$\hat{\theta}_{MLE} = \arg \max_{\theta} p(\mathcal{D} \mid \theta)$$

- For MAP, if we have enough data, we see that **the data overwhelms the prior**. In this case, the MAP estimate converges towards the MLE

$$\hat{\theta}_{MAP} = \arg \max_{\theta} p(\mathcal{D} \mid \theta)p(\theta)$$

Posterior



Posterior

- If the true hypothesis is in the hypothesis space
 - The MAP estimate will **converge upon** this hypothesis.
- If our hypothesis class is not rich enough to represent the “truth”
 - Converge on the hypothesis that is **as close as possible** to the truth

Posterior predictive distribution

- Posterior predictive distribution

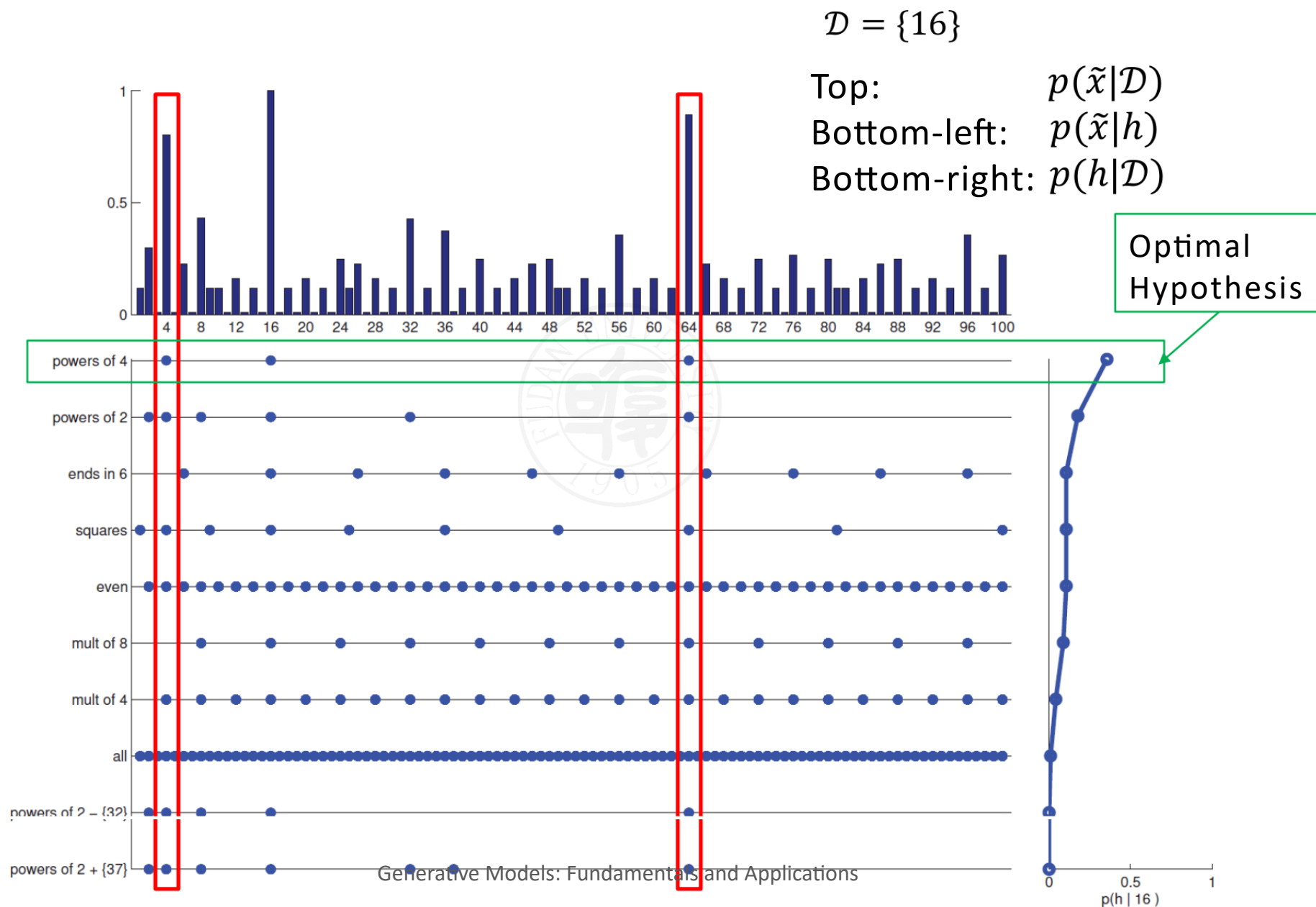
$$p(\tilde{x}|\mathcal{D}) = \sum_{h \in \mathcal{H}} p(h|\mathcal{D})p(\tilde{x}|h, \mathcal{D}) = \sum_{h \in \mathcal{H}} p(h|\mathcal{D})p(\tilde{x}|h)$$

- Consider $p(h|\mathcal{D})$ as a kind of weight

$$\sum_{h \in \mathcal{H}} p(h|\mathcal{D}) = 1$$

- a weighted average of the predictions of each individual hypothesis
 - **Bayes model averaging** (BMA) method

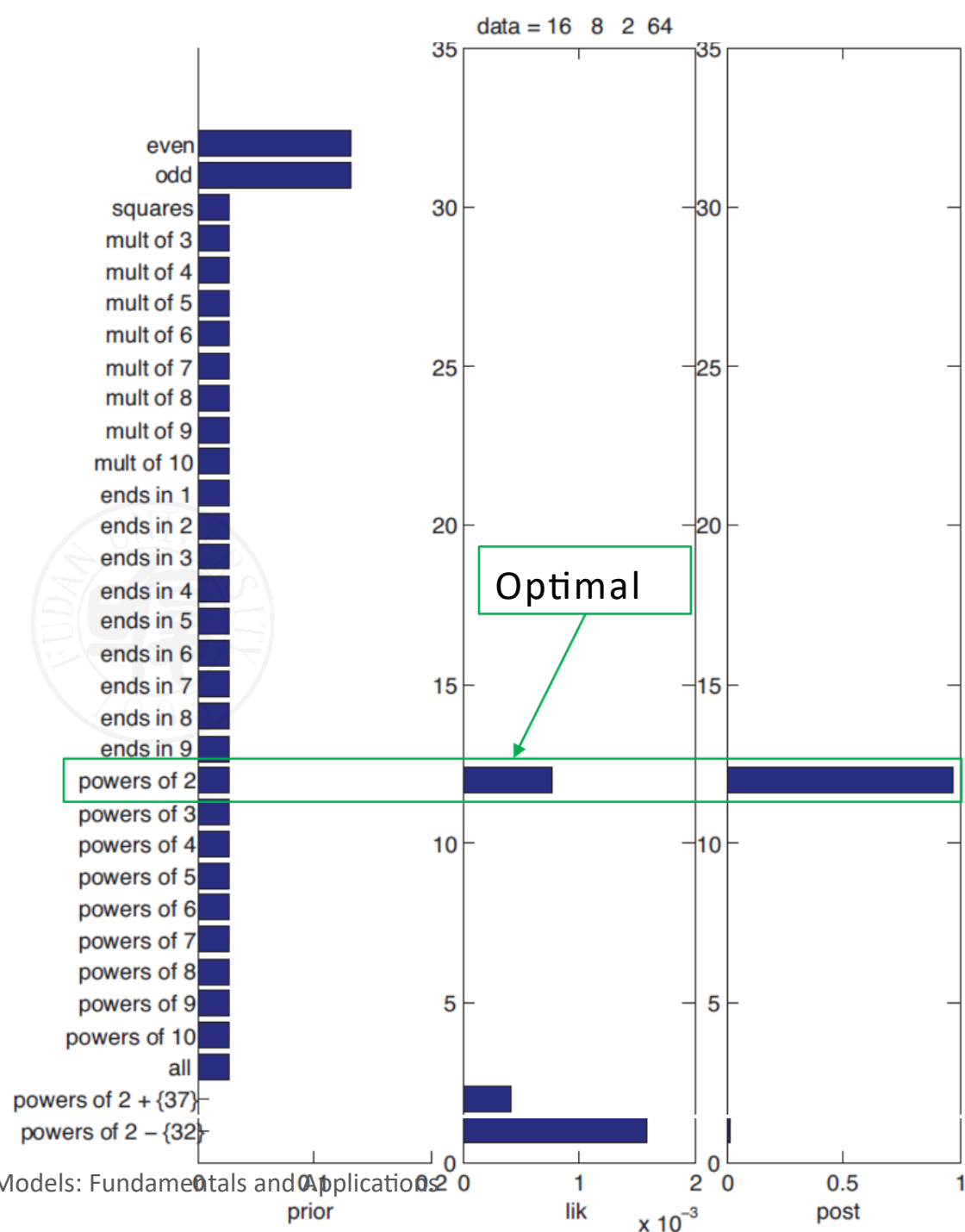
Example



Example

- Given $\mathcal{D}=\{16,8,2,64\}$
 - The posterior for the hypothesis “powers of 2” is close to 1

$$\begin{aligned}
 p(\tilde{x}|\mathcal{D}) &= \sum_{\tilde{h} \in \mathcal{H}} p(h|\mathcal{D})p(\tilde{x}|h) \\
 &= \sum_{\tilde{h} \in \mathcal{H}} \delta_{\hat{h}^{MAP}}(h) p(\tilde{x}|h) \\
 &\approx p(\tilde{x}|\hat{h}^{MAP})
 \end{aligned}$$



Posterior predictive distribution

- Plug-in approximation
 - If we have “figured things out”, posterior becomes a delta function centered at the MAP estimate

$$p(\tilde{x}|\mathcal{D}) = \int_{\overline{h \in \mathcal{H}}} p(h|\mathcal{D})p(\tilde{x}|h)$$
$$\approx p(\tilde{x}|\hat{h}^{MAP})$$

Posterior predictive distribution

- BMA vs. plug-in (MAP)
 - Sample space by MAP: from narrow to wide
 - For a single hypothesis
 - As training data increases, the sample space generated by the MAP-selected hypothesis expands
 - **Example**
 - $D = \{16\}$:the optimal hypothesis is "powers of 4", generated samples $\{4, 16, 64\}$
 - $D = \{16, 8, 2, 64\}$:the optimal hypothesis is "powers of 2", generated samples $\{2, 4, 8, 16, 32, 64\}$

Posterior predictive distribution

- BMA vs. plug-in (MAP)
 - Sample space by BMA: from wide to narrow
 - For the entire hypothesis space
 - As training data increases, some hypotheses are assigned weight 0, so the effective sample space shrinks
 - **Example**
 - $D = \{16\}$: many hypotheses are consistent with the data
 - $D = \{16, 8, 2, 64\}$: fewer hypotheses remain consistent with the data, inconsistent ones have weight 0

A more complex prior

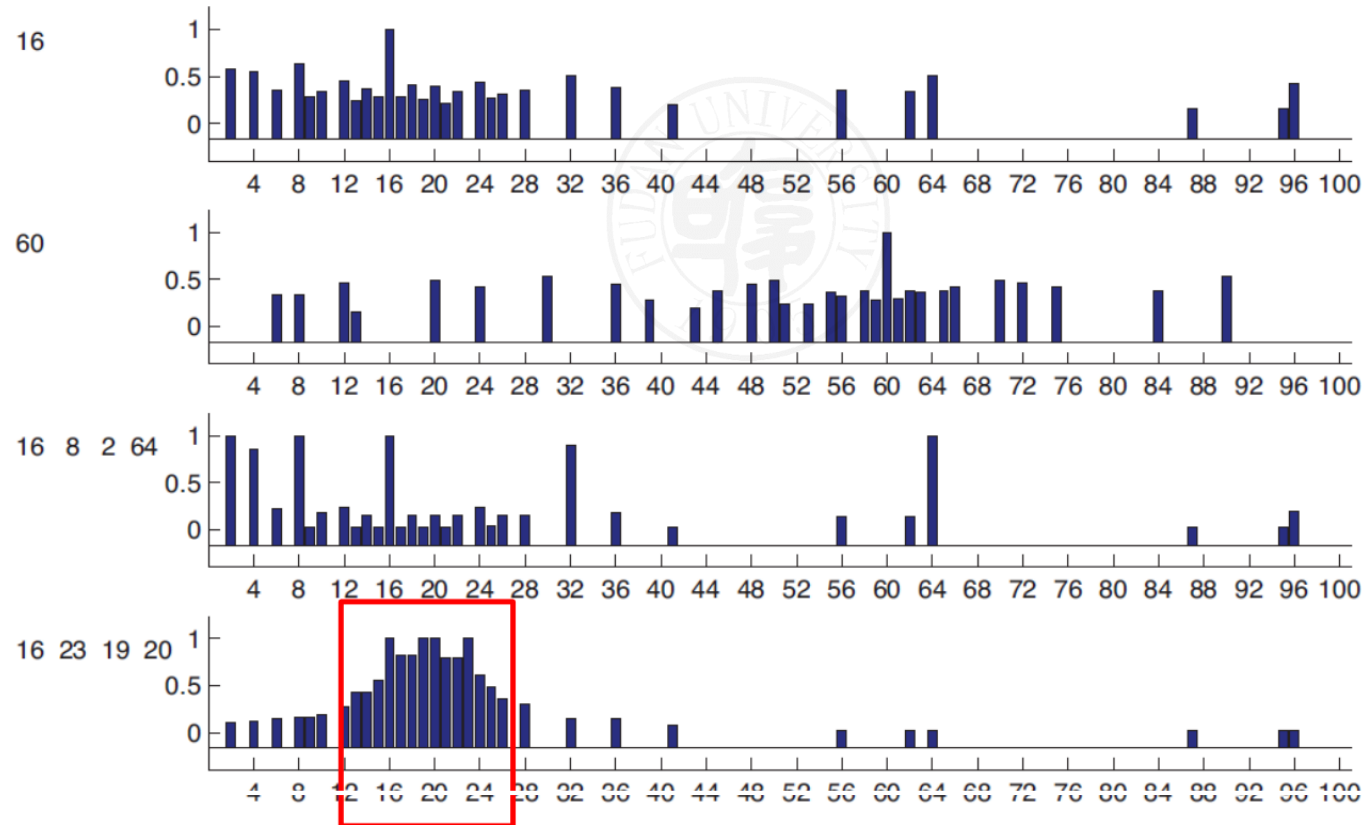
- Mixture of several priors
 - E.g., in number game
 - p_{rules} : arithmetical concepts
 - $p_{interval}$: intervals between n and m

$$p(h) = \pi_0 p_{rules}(h) + (1 - \pi_0) p_{interval}(h)$$

A more complex prior

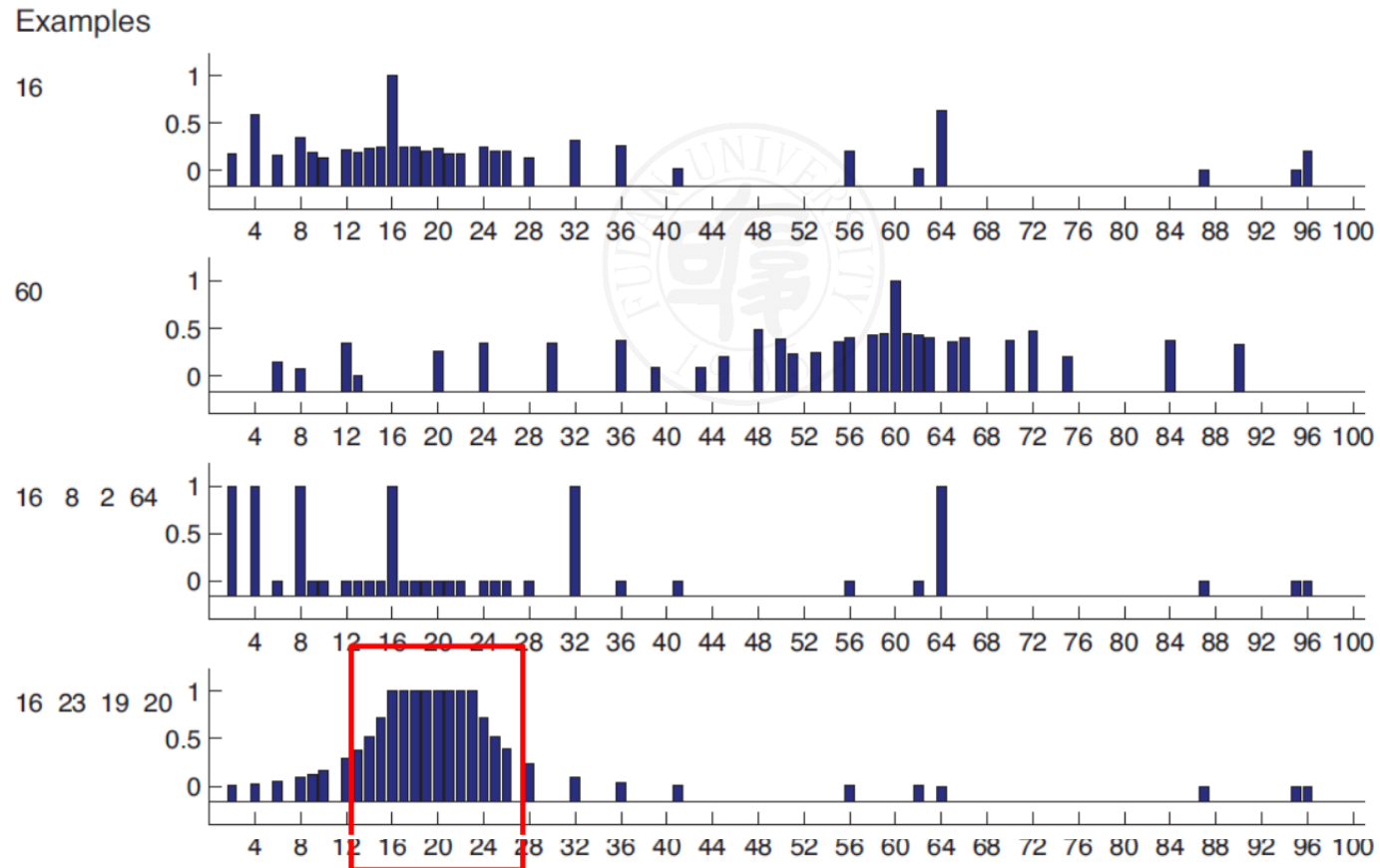
■ Back to number game

Examples



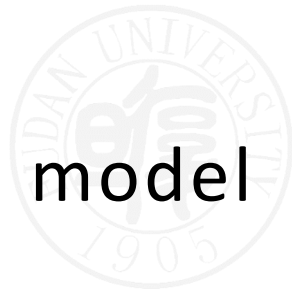
A more complex prior

- Arithmetical concepts + intervals between n and m



Outline

- Generative vs. Discriminative: Revisit
- Bayesian concept learning
- The beta-binomial model
- The Dirichlet-multinomial model
- Naive Bayes classifiers



The beta-binomial model

■ Continuous case

- Unknown parameters are **continuous**
- Hypothesis space is (some subset) of $\mathbb{R}^K$
 - K is the number of parameters
- Replace sums with integrals

- Posterior

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}, h')}$$

- Posterior Predictive Distribution

$$p(\tilde{x}|\mathcal{D}) = \sum_{h \in \mathcal{H}} p(h|\mathcal{D})p(\tilde{x}|h, \mathcal{D}) = \sum_{h \in \mathcal{H}} p(h|\mathcal{D})p(\tilde{x}|h)$$

The beta-binomial model

- Example: coin toss
 - Toss coins N times
 - N_1 times head up, and $N_0 = N - N_1$ times tail up
 - Suppose $x_i \sim \text{Ber}(\theta), i = 1, \dots, N$
 - $x_i = 1$: head up
 - $x_i = 0$: tail up
 - $\theta \in [0,1]$: rate parameter (probability of heads)

Likelihood

- Suppose the data $x_1, x_2, \dots, x_N$ are i.i.d.

$$p(\mathcal{D}|\theta) = \theta^{N_1} (1 - \theta)^{N_0}$$

- **Sufficient statistic**: it captures all the information in the data about the parameter θ

$$N_1 = \sum_{i=1}^N \mathbb{I}(x_i = 1)$$

$$N_0 = \sum_{i=1}^N \mathbb{I}(x_i = 0)$$

The beta-binomial model

- Suppose N_1 is a random variable
 - $N_1 \sim \text{Bin}(N, \theta)$

$$\text{Bin}(N_1|N, \theta) = \binom{N}{N_1} \theta^{N_1} (1 - \theta)^{N - N_1}$$

- The likelihood for the binomial sampling model is the same as the likelihood for the Bernoulli model

Prior

- The prior has the same form as the likelihood

- Likelihood $p(\mathcal{D}|\theta) = \theta^{N_1} (1 - \theta)^{N_0}$

- Prior, assume in the form of $p(\theta) \propto \theta^{\gamma_1} (1 - \theta)^{\gamma_2}$

- Hence, posterior $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta) = \theta^{N_1} (1 - \theta)^{N_0} \theta^{\gamma_1} (1 - \theta)^{\gamma_2} = \theta^{N_1+\gamma_1} (1 - \theta)^{N_0+\gamma_2}$

- Conjugate prior

- the prior and the posterior have the same form

$$\text{Beta}(\theta|a, b) \propto \theta^{a-1} (1 - \theta)^{b-1}$$

Posterior

- Multiply the likelihood by the beta prior

$$p(\theta|\mathcal{D}) \propto \text{Bin}(N_1|N_1 + N_0, \theta)\text{Beta}(\theta|a, b)$$

$$\propto \text{Beta}(\theta|N_1 + a, N_0 + b)$$

Example

- (a) With a relatively weak prior, the posterior is very close to the likelihood, as the data overwhelms the prior.
- (b) With a stronger prior, the posterior becomes a compromise between the prior and the likelihood.

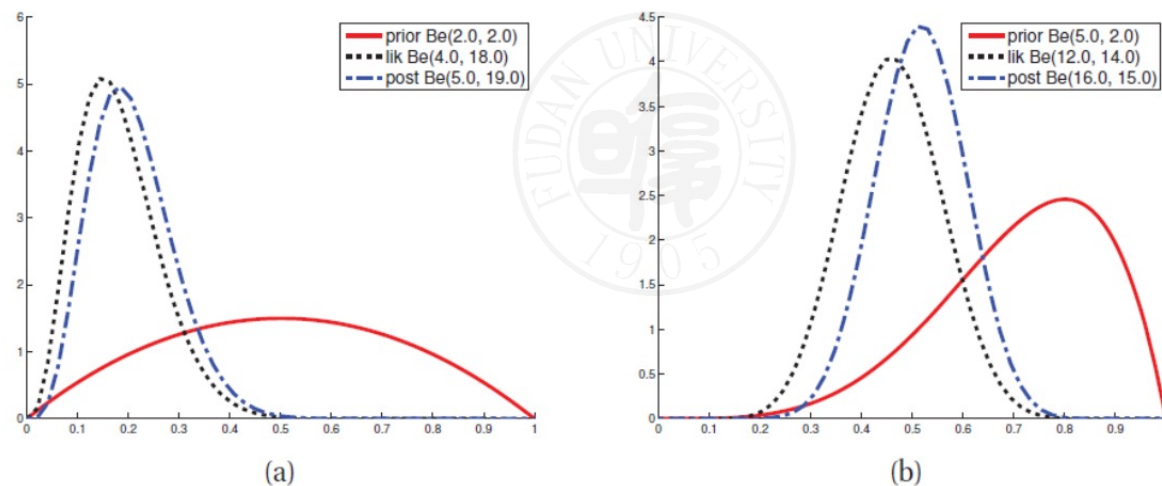


Figure 3.6 (a) Updating a $\text{Beta}(2, 2)$ prior with a Binomial likelihood with sufficient statistics $N_1 = 3, N_0 = 17$ to yield a $\text{Beta}(5, 19)$ posterior. (b) Updating a $\text{Beta}(5, 2)$ prior with a Binomial likelihood with sufficient statistics $N_1 = 11, N_0 = 13$ to yield a $\text{Beta}(16, 15)$ posterior. Figure generated by `binomialBetaPosteriorDemo`.

Posterior

- MAP estimate

$$\text{Beta}(a, b): \text{mode} = \frac{a-1}{a+b-2}$$

$$\max_{\theta} p(\theta|\mathcal{D}) \propto \text{Beta}(\theta|N_1 + a, N_0 + b)$$



$$\hat{\theta}_{MAP} = \frac{a + N_1 - 1}{a + N_1 + b + N_0 - 2} = \frac{a + N_1 - 1}{a + b + N - 2}$$

- MLE

- If the prior is uniformly distributed ($a=b=1$)

$$\hat{\theta}_{MLE} = \frac{N_1}{N}$$

Posterior

■ Posterior mean

$$\text{Beta}(a, b): \text{mean} = \frac{a}{a+b}$$

$$\bar{\theta} = \frac{a + N_1}{a + N_1 + b + N_0} = \frac{a + N_1}{a + b + N}$$

- Let $\alpha_0 = a + b$, and let the prior mean $m_1 = a/\alpha_0$

$$\mathbb{E}[\theta|\mathcal{D}] = \frac{\alpha_0 m_1 + N_1}{N + \alpha_0} = \frac{\alpha_0}{N + \alpha_0} m_1 + \frac{N}{N + \alpha_0} \frac{N_1}{N} = \lambda m_1 + (1 - \lambda) \hat{\theta}_{MLE}$$

where $\lambda = \frac{\alpha_0}{N + \alpha_0}$

- **Result:** the posterior mean is **convex combination** of the prior mean and the MLE

Posterior

- Equivalent **sample size** of the prior $\alpha_0 = a + b$

$$\text{Beta}(\theta|a, b) \propto \theta^{a-1}(1 - \theta)^{b-1}$$

- Equivalent **sample size** of the posterior $N + \alpha_0$

$$p(\theta|\mathcal{D}) \propto \text{Beta}(\theta|N_1 + a, N_0 + b)$$

- Ratio of the prior to posterior equivalent sample size λ
 - The smaller λ , the closer the **posterior mean** is to the MLE
 - The smaller λ , the closer the **posterior mode** is to the MLE

Posterior

■ Posterior variance

$$\text{Beta}(a, b): \text{var} = \frac{ab}{(a+b)^2(a+b+1)}$$

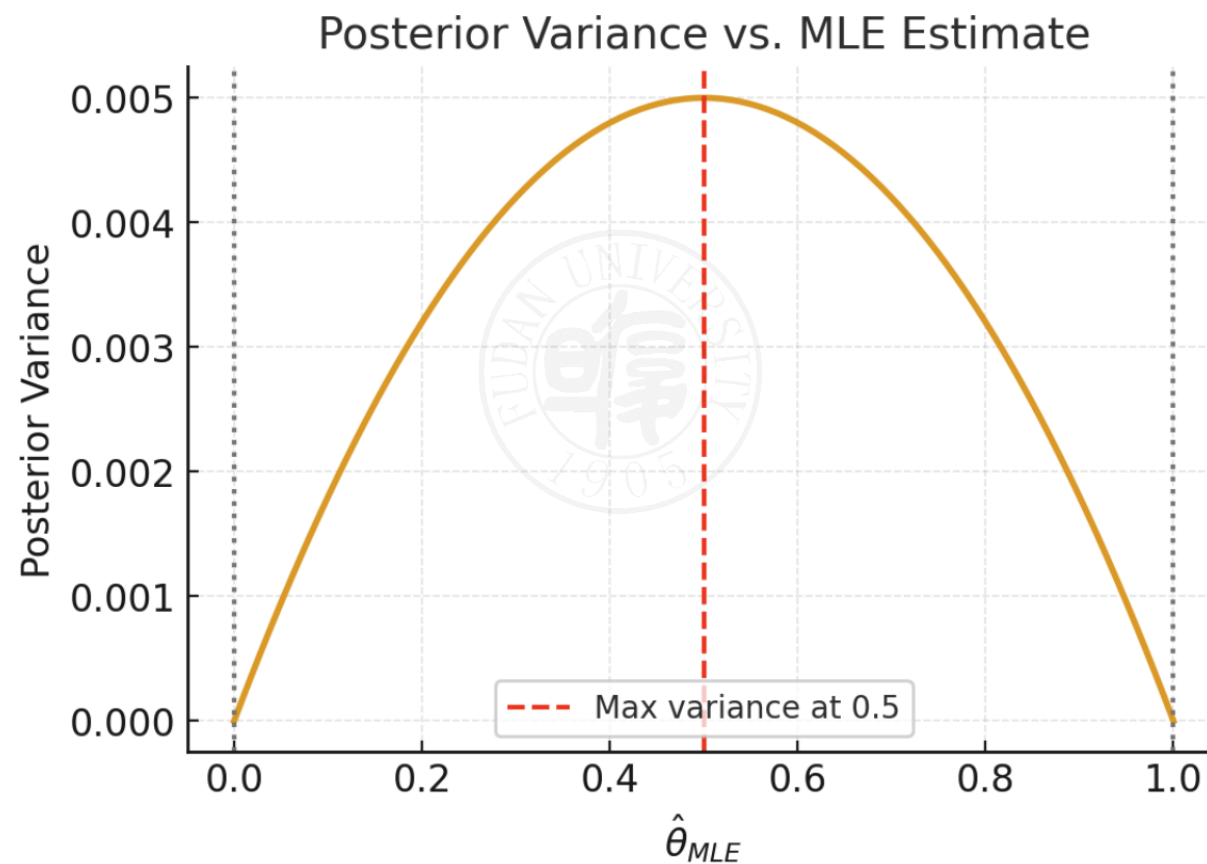
$$\text{var} [\theta | \mathcal{D}] = \frac{(a + N_1)(b + N_0)}{(a + N_1 + b + N_0)^2(a + N_1 + b + N_0 + 1)}$$

□ If $N \gg a, b$

$$\text{Var}[\theta | \mathcal{D}] \approx \frac{N_1 N_0}{N^2 N} = \frac{\hat{\theta}_{MLE}(1 - \hat{\theta}_{MLE})}{N}$$

- The variance is maximized when $\hat{\theta}_{MLE} = 0.5$
- The variance is minimized when $\hat{\theta}_{MLE} = 0, 1$

Posterior



Posterior predictive distribution

- Generate one data

BMA

$$p(\tilde{x} = 1 \mid \mathcal{D}) = \int_0^1 p(\tilde{x} = 1 \mid \theta) p(\theta \mid \mathcal{D}) d\theta$$

$$= \int_0^1 \theta \text{Beta}(\theta \mid N_1 + a, N_0 + b) d\theta$$

$$= \mathbb{E}[\theta \mid \mathcal{D}]$$

Posterior mean

$$= \frac{N_1 + a}{N + a + b}$$

Posterior predictive distribution

- Overfitting and the black swan paradox
 - Zero count problem or the sparse data problem
 - Suppose the sample size is very small,
 - E.g., $N = 3, N_1 = 0$
 - Suppose we use plug-in the MLE
 - $\hat{\theta}_{MLE} = \frac{0}{3} = 0$
 - Result: heads are impossible.

Posterior predictive distribution

- Overfitting and the black swan paradox
 - Bayes theorem: Laplace's rule of succession
 - Suppose the prior is uniform on $[0,1]$

$$p(\tilde{x} = 1|\mathcal{D}) = \frac{a + N_1}{a + b + N} = \frac{N_1 + 1}{N + 2} = \frac{N_1 + 1}{N_1 + 1 + N_0 + 1}$$

- Add-one smoothing: Beta(1, 1)

Posterior predictive distribution

- Generate multiple data: **beta-binomial distribution**
 - Predict the number of heads, x , in M future trials

$$\begin{aligned} p(x|\mathcal{D}, M) &= \int_0^1 \text{Bin}(x|M, \theta) \text{Beta}(\theta|N_1 + a, N_0 + b) d\theta \\ &= \binom{M}{x} \frac{1}{B(N_1 + a, N_0 + b)} \int_0^1 \theta^x (1 - \theta)^{M-x} \theta^{N_1+a-1} (1 - \theta)^{N_0+b-1} d\theta \\ &= \binom{M}{x} \frac{B(x + N_1 + a, M - x + N_0 + b)}{B(N_1 + a, N_0 + b)} \\ &\triangleq Bb(x|N_1 + a, N_0 + b, M) \end{aligned}$$

Posterior predictive distribution

- beta-binomial distribution

- Mean

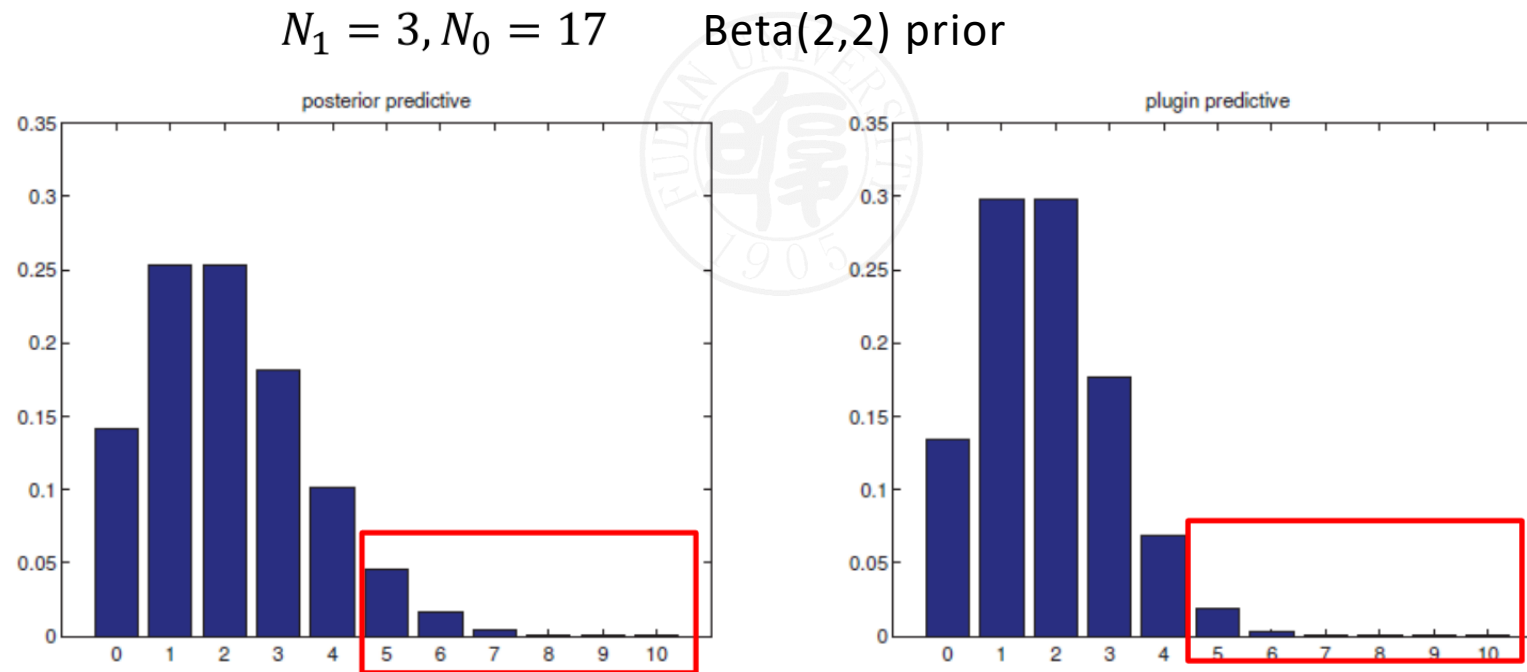
$$\mathbb{E}[x|\mathcal{D}] = M \frac{N_1 + a}{N + a + b}$$

- Variance

$$\text{Var}[x|\mathcal{D}] = \frac{M(N_1 + a)(N_0 + b)}{(N + a + b)^2} \frac{N + a + b + M}{N + a + b + 1}$$

Experiment

- Bayesian prediction has longer tails, spreading its probability mass more widely, and is therefore less prone to overfitting and black swan type paradoxes



Outline

- Generative vs. Discriminative: Revisit
- Bayesian concept learning
- The beta-binomial model
- The Dirichlet-multinomial model
- Naive Bayes classifiers



The Dirichlet-multinomial model

- Example:

- Infer the probability that a dice with K sides comes up as face k
- Given $\mathcal{D} = \{x_1, x_2, \dots, x_N | x_i \in \{1, 2, \dots, K\}\}$

- Likelihood

where

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^K \theta_k^{N_k}$$

$$N_k = \sum_{i=1}^N \mathbb{I}(x_i = k)$$

The Dirichlet-multinomial model

- Prior: conjugate prior

$$\text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K \theta_k^{\alpha_k-1} \mathbb{I}(\boldsymbol{\theta} \in S_K)$$

- Posterior

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &\propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &= \prod_{k=1}^K \theta_k^{N_k} \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K \theta_k^{\alpha_k-1} \mathbb{I}(\boldsymbol{\theta} \in S_K) \\ &\propto \prod_{k=1}^K \theta_k^{N_k+\alpha_k-1} \\ &\propto \text{Dir}(\boldsymbol{\theta}|N_1 + \alpha_1, N_2 + \alpha_2, \dots, N_K + \alpha_K) \end{aligned}$$

The Dirichlet-multinomial model

- MAP estimate

$$\hat{\theta}_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}$$

where

$$\alpha_0 \triangleq \sum_{k=1}^K \alpha_k$$

- MLE:

$$\hat{\theta}_k = N_k / N$$

The Dirichlet-multinomial model

■ Posterior predictive

$$\begin{aligned} p(X = j|\mathcal{D}) &= \int p(X = j|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \\ &= \int p(X = j|\theta_j) \left[\int p(\boldsymbol{\theta}_{-j}, \theta_j|\mathcal{D})d\boldsymbol{\theta}_{-j} \right] d\theta_j \\ &= \int \theta_j p(\theta_j|\mathcal{D})d\theta_j = \mathbb{E}[\theta_j|\mathcal{D}] = \frac{\alpha_j + N_j}{\sum_k (\alpha_k + N_k)} = \frac{\alpha_j + N_j}{\alpha_0 + N} \end{aligned}$$

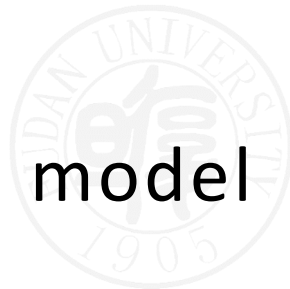
where $\boldsymbol{\theta}_{-j}$ are all the components of $\boldsymbol{\theta}$ except θ_j

Beta-Binomial vs. Dirichlet-Multinomial

	Beta-Binomial Model	Dirichlet-Multinomial Model
Task scenario	Binary classification (coin toss: head/tail)	Multi-class classification (dice roll: 1,2,...,K)
Parameters	$\theta \in [0,1]$:probability of “head”	$\theta = (\theta_1, \dots, \theta_K)$,class probabilities, $\sum_k \theta_k = 1$
Prior	Beta distribution: $\text{Beta}(\theta a, b)$	Dirichlet distribution: $\text{Dir}(\theta \alpha_1, \dots, \alpha_K)$
Likelihood	Binomial: $\text{Bin}(N_1 N, \theta)$	Multinomial: $\text{Mult}(N_1, \dots, N_K N, \theta)$
Conjugacy	Beta is the conjugate prior of Binomial	Dirichlet is the conjugate prior of Multinomial
Posterior	$\theta D \sim \text{Beta}(a + N_1, b + N_0)$	$\theta D \sim \text{Dir}(\alpha_1 + N_1, \dots, \alpha_K + N_K)$
Posterior predictive	$p(\tilde{x} = 1 D) = \frac{a + N_1}{a + b + N}$	$p(\tilde{x} = j D) = \frac{\alpha_j + N_j}{\alpha_0 + N}$
Intuition	“Pseudo-counts”: $a - 1$ heads, $b - 1$ tails	“Pseudo-counts”: $\alpha_k - 1$ for each class, total $\alpha_0 - K$

Outline

- Generative vs. Discriminative: Revisit
- Bayesian concept learning
- The beta-binomial model
- The Dirichlet-multinomial model
- Naive Bayes classifiers



Naive Bayes classifiers

- Classify vectors of **discrete-valued** features to C classes

$$\mathbf{x} = (x_1, x_2, \dots, x_D)^T \in \{1, 2, \dots, K\}^D$$

- K is the number of values for each feature
- D is the number of features
- Compute the probability $P(Y = c | \mathbf{x}, \theta)$

Naive Bayes classifiers

- Generative model, revisit
 - Key: learn class-conditional density $P(\mathbf{x}|Y = c, \theta)$

$$\begin{aligned} P(Y = c|\mathbf{x}, \theta) &= \frac{P(\mathbf{x}, Y=c | \theta)}{P(\mathbf{x}|\theta)} \\ &= \frac{P(\mathbf{x}, Y=c | \theta)}{\sum_{c' \in \mathcal{C}} P(\mathbf{x}, Y=c' | \theta)} \\ &= \frac{P(Y=c | \theta) \boxed{P(\mathbf{x}|Y=c, \theta)}}{\sum_{c' \in \mathcal{C}} P(Y=c' | \theta) P(\mathbf{x}|Y=c', \theta)} \end{aligned}$$

class-conditional
distribution

Naive Bayes classifiers

- Assumption: the features are **conditionally independent** given the class label

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^D p(x_j|y = c, \boldsymbol{\theta}_{jc})$$

- Even if the naive Bayes assumption is not true, it often results in classifiers that work well
 - The model is quite simple (it only has $O(CD)$ parameters, for C classes and D features)

Naive Bayes classifiers

- In the **case of binary** features $x_j \in \{0,1\}$
 - Multivariate Bernoulli naive Bayes
 - Use the Bernoulli distribution

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^D \text{Ber}(x_j|\mu_{jc})$$

- μ_{jc} is the probability that feature j occurs in class c

Categorical Naive Bayes

- In the **case of categorical** features $x_j \in \{1, \dots, K\}$
 - Categorical Naive Bayes
 - Use the categorical distribution

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^D \text{Cat}(x_j|\mu_{jc})$$

- μ_{jc} is a histogram over the K possible values for x_j in class c.

Naive Bayes classifiers

- In the case of **real-valued** features:
 - Gaussian Naive Bayes
 - Use the Gaussian distribution

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^D \mathcal{N}(x_j | \mu_{jc}, \sigma_{jc}^2)$$

- μ_{jc} is the mean of feature j in objects of class c
- σ_{jc}^2 is the variance of feature j in objects of class c

Model fitting

- How to “train” a naive Bayes classifier
 - MLE
 - MAP estimate
 - Bayesian estimation
 - Compute the full posterior $p(\theta|\mathcal{D})$



MLE for NBC

- The probability for a single data $(\mathbf{x}_i, y_i)$

$$\begin{aligned} p(\mathbf{x}_i, y_i | \boldsymbol{\theta}) &= p(y_i | \boldsymbol{\pi}) \prod_j p(x_{ij} | y_i, \theta_j) \\ &= \prod_c \pi_c^{\mathbb{I}(y_i=c)} \prod_j \prod_c p(x_{ij} | \theta_{jc})^{\mathbb{I}(y_i=c)} \end{aligned}$$

- $\boldsymbol{\pi}$: the parameters for class prior

MLE for NBC

- The probability for data set $\mathcal{D}$

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^N \left(\sum_c \pi_c^{\mathbb{I}(y_i=c)} \prod_j p(x_{ij}|\theta_{jc})^{\mathbb{I}(y_i=c)} \right)$$

- The log-likelihood for data set $\mathcal{D}$

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{c=1}^C N_c \log \pi_c + \sum_{j=1}^D \sum_{c=1}^C \sum_{i:y_i=c} \log p(x_{ij}|\boldsymbol{\theta}_{jc})$$

MLE for NBC

- The log-likelihood for data set $\mathcal{D}$
 - The categorical class prior

$$\hat{\pi}_c = \frac{N_c}{N}$$

- Suppose all features are binary

$$\hat{\theta}_{jc} = \frac{N_{jc}}{N_c}$$

Bayesian naive Bayes

- Trouble for MLE: overfitting
- Solution: Bayesian estimation

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{P(\boldsymbol{\theta})P(\mathcal{D}|\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}'} P(\boldsymbol{\theta}')P(\mathcal{D}|\boldsymbol{\theta}') d\boldsymbol{\theta}'}$$

Bayesian naive Bayes

■ Prior

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_{j=1}^D \prod_{c=1}^C p(\theta_{jc})$$

- $\boldsymbol{\pi}$: Dir($\boldsymbol{\alpha}$) prior
- θ_{jc} : Beta(β_0, β_1) prior

■ Special case: $\boldsymbol{\alpha} = 1, \boldsymbol{\beta} = 1$

- Add-one or **Laplace smoothing**

Bayesian naive Bayes

■ Posterior

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\theta)p(\mathcal{D}|\theta) \\ &= \left\{ p(\pi) \prod_{j=1}^D \prod_{c=1}^C p(\theta_{jc}) \right\} \prod_{i=1}^N \left\{ \left[\prod_{c=1}^C \pi_c^{\mathbb{I}(y_i=c)} \right] \left[\prod_{c=1}^C \prod_{j=1}^D \text{Ber}(x_{ij}|\theta_{jc})^{\mathbb{I}(y_i=c)} \right] \right\} \\ &= \left\{ p(\pi) \prod_{i=1}^N \prod_{c=1}^C \pi_c^{\mathbb{I}(y_i=c)} \right\} \prod_{j=1}^D \prod_{c=1}^C \left\{ p(\theta_{jc}) \prod_{i=1}^N \text{Ber}(x_{ij}|\theta_{jc})^{\mathbb{I}(y_i=c)} \right\} \\ &= \left\{ \text{Dir}(\alpha) \prod_{c=1}^C \pi_c^{N_c} \right\} \prod_{j=1}^D \prod_{c=1}^C \left\{ \text{Beta}(\beta_0, \beta_1) (1 - \theta_{jc})^{N_c - N_{jc}} \theta_{jc}^{N_{jc}} \right\} \\ &\propto \text{Dir}(N_1 + \alpha_1, N_2 + \alpha_2, \dots, N_C + \alpha_C) \prod_{j=1}^D \prod_{c=1}^C \text{Beta}(N_c - N_{jc} + \beta_0, N_{jc} + \beta_1) \end{aligned}$$

Bayesian naive Bayes

■ Posterior

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= p(\boldsymbol{\pi}|\mathcal{D}) \prod_{j=1}^D \prod_{c=1}^C p(\theta_{jc}|\mathcal{D}) \\ p(\boldsymbol{\pi}|\mathcal{D}) &= \text{Dir}(N_1 + \alpha_1, \dots, N_C + \alpha_C) \\ p(\theta_{jc}|\mathcal{D}) &= \text{Beta}((N_c - N_{jc}) + \beta_0, N_{jc} + \beta_1) \end{aligned}$$

□ Posterior mean

$$\bar{\theta}_{jc} = \frac{N_{jc} + \beta_1}{N_c + \beta_0 + \beta_1}$$

$$\bar{\pi}_c = \frac{N_c + \alpha_c}{N + \alpha_0}$$

Bayesian naive Bayes

■ MAP estimate

$$\hat{\pi}_c = \frac{N_c + \alpha_c - 1}{N + \alpha_0 - C}, \quad \hat{\theta}_{jc} = \frac{N_{jc} + \beta_1 - 1}{N_c + \beta_0 + \beta_1 - 2}$$

- The MAP estimate can be directly computed by combining **prior counts** with **empirical counts**.
- Very simple to implement, with low computational complexity, and easy to extend.

Using the model for prediction

- Predict the label y for new data $\mathbf{x}$

$$\begin{aligned} p(y = c | x, \mathcal{D}) &\propto p(y = c | \mathcal{D}) p(x | y = c, \mathcal{D}) \\ &= p(y = c | \mathcal{D}) \prod_{j=1}^D p(x_j | y = c, \mathcal{D}) \end{aligned}$$

- Bayesian procedure

$$\begin{aligned} p(y = c | x, \mathcal{D}) &\propto \left[\int_{\pi} p(y = c | \pi) p(\pi | \mathcal{D}) d\pi \right] \prod_{j=1}^D \left[\int_{\theta_{jc}} p(x_j | y = c, \theta_{jc}) p(\theta_{jc} | \mathcal{D}) d\theta_{jc} \right] \\ &= \left[\int_{\pi} \text{Cat}(y = c | \pi) p(\pi | \mathcal{D}) d\pi \right] \prod_{j=1}^D \left[\int_{\theta_{jc}} \text{Ber}(x_j | y = c, \theta_{jc}) p(\theta_{jc} | \mathcal{D}) d\theta_{jc} \right] \end{aligned}$$

Using the model for prediction

- Rewrite the prediction probability

$$p(y = c|x, \mathcal{D}) \propto \left[\int_{\pi} \pi_c p(\pi|\mathcal{D}) d\pi \right] \prod_{j=1}^D \left[\int_{\theta_{jc}} \theta_{jc}^{\mathbb{I}(x_j=1)} (1 - \theta_{jc})^{\mathbb{I}(x_j=0)} p(\theta_{jc}|\mathcal{D}) d\theta_{jc} \right]$$

π_c 关于分布 $p(\pi|\mathcal{D})$ 的期望

$\theta_{jc}^{\mathbb{I}(x_j=1)} (1 - \theta_{jc})^{\mathbb{I}(x_j=0)}$ 关于分布 $p(\theta_{jc}|\mathcal{D})$ 的期望

Using the model for prediction

- Plug in the posterior mean parameters

$$p(y = c | \mathbf{x}, \mathcal{D}) \propto \bar{\pi}_c \prod_{j=1}^D (\bar{\theta}_{jc})^{\mathbb{I}(x_j=1)} (1 - \bar{\theta}_{jc})^{\mathbb{I}(x_j=0)}$$

$$\bar{\pi}_c = \frac{N_c + \alpha_c}{N + \alpha_0}, \quad \bar{\theta}_{jc} = \frac{N_{jc} + \beta_1}{N_c + \beta_0 + \beta_1}$$

Using the model for prediction

- Plug in the point estimate: MLE or MAP

$$\begin{aligned} p(y = c|x, \mathcal{D}) &\propto p(y = c|\mathcal{D}) \prod_{j=1}^D p(x_j|y = c, \mathcal{D}) \\ &= p(y = c|\hat{\pi}) \prod_{j=1}^D p(x_j|y = c, \hat{\theta}_{jc}) \\ &= \hat{\pi}_c \prod_{j=1}^D \hat{\theta}_{jc}^{\mathbb{I}(x_j=1)} (1 - \hat{\theta}_{jc})^{\mathbb{I}(x_j=0)}, \end{aligned}$$

$$\hat{\pi}_c = \frac{N_c}{N}$$

$$\hat{\theta}_{jc} = \frac{N_{jc}}{N_c}$$

Using the model for prediction

■ Classifier

$$\begin{aligned} y = f^*(x) &= \arg \max_{c=1,2,\dots,C} p(y = c | X = x, \mathcal{D}) \\ &= \arg \max_{c=1,2,\dots,C} \tilde{\pi}_c \prod_{j=1}^D \tilde{\theta}_{jc}^{\mathbb{I}(x_j=1)} (1 - \hat{\theta}_{jc})^{\mathbb{I}(x_j=0)} \end{aligned}$$

Example

- 给定训练数据如下表所示，其中， $X^{(1)} \in \{1,2,3\}$, $x^{(2)} \in \{S,M,L\}$ 为特征， $Y \in \{-1,1\}$ 为类标记。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$x^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

试根据给定训练数据学习一个朴素贝叶斯分类器并确定 $x = (2, S)^T$ 的类标记。

Example

■ MLE

□ 计算 class prior

$$N_1 = 9$$

$$N_{-1} = 6$$

$$N = 15$$

$$\hat{\pi}_{-1} = \frac{N_{-1}}{N} = \frac{6}{15}$$

$$\hat{\pi}_1 = \frac{N_1}{N} = \frac{9}{15}$$

□ 计算 $\hat{\theta}_{k,j,c} = p(X^{(j)} = k | Y = c)$

$$\hat{\theta}_{1,1,1} = \frac{N_{1,1,1}}{N_1} = \frac{2}{9}$$

$$\hat{\theta}_{2,1,1} = \frac{N_{2,1,1}}{N_1} = \frac{3}{9}$$

$$\hat{\theta}_{3,1,1} = \frac{N_{3,1,1}}{N_1} = \frac{4}{9}$$

$$\hat{\theta}_{S,2,1} = \frac{N_{S,2,1}}{N_1} = \frac{1}{9}$$

$$\hat{\theta}_{M,2,1} = \frac{N_{M,2,1}}{N_1} = \frac{4}{9}$$

$$\hat{\theta}_{L,2,1} = \frac{N_{L,2,1}}{N_1} = \frac{4}{9}$$

Example

□ 计算 $\hat{\theta}_{k,j,c}$

$$\begin{aligned}\hat{\theta}_{1,1,-1} &= \frac{N_{1,1,-1}}{N_{-1}} = \frac{3}{6} & \hat{\theta}_{S,2,-1} &= \frac{N_{S,2,-1}}{N_{-1}} = \frac{3}{6} \\ \hat{\theta}_{2,1,-1} &= \frac{N_{2,1,-1}}{N_{-1}} = \frac{2}{6} & \hat{\theta}_{M,2,-1} &= \frac{N_{M,2,-1}}{N_{-1}} = \frac{2}{6} \\ \hat{\theta}_{3,1,-1} &= \frac{N_{3,1,-1}}{N_{-1}} = \frac{1}{6} & \hat{\theta}_{L,2,-1} &= \frac{N_{L,2,-1}}{N_{-1}} = \frac{1}{6}\end{aligned}$$

Example

- 预测 $p(Y = c | \mathbf{x}, \mathcal{D})$

$$\begin{aligned} p(Y = 1 | (2, S)^T, \mathcal{D}) &\propto \hat{\pi}_1(\hat{\theta}_{2,1,1})^{\mathbb{I}(X^{(1)}=2)} (\hat{\theta}_{S,2,1})^{\mathbb{I}(X^{(2)}=S)} \\ &= \frac{9}{15} \times \frac{3}{9} \times \frac{1}{9} = \frac{1}{45} \end{aligned}$$

$$\begin{aligned} p(Y = -1 | (2, S)^T, \mathcal{D}) &\propto \hat{\pi}_{-1}(\hat{\theta}_{2,1,-1})^{\mathbb{I}(X^{(1)}=2)} (\hat{\theta}_{S,2,-1})^{\mathbb{I}(X^{(2)}=S)} \\ &= \frac{6}{15} \times \frac{2}{6} \times \frac{3}{6} = \frac{1}{15} \end{aligned}$$

所以数据 $(2, S)^T$ 预测的标签为 -1

Example

- Bayes estimation

- 计算 class prior

$$N_1 = 9$$

$$N_{-1} = 6$$

$$N = 15$$

$$\bar{\pi}_{-1} = \frac{N_{-1} + 1}{N + 2} = \frac{7}{17}$$

$$\bar{\pi}_1 = \frac{N_1 + 1}{N + 2} = \frac{10}{17}$$

Example

■ Bayes estimation

□ 计算 $\bar{\theta}_{k,j,c} = p(X^{(j)} = k | Y = c)$

$$\bar{\theta}_{1,1,1} = \frac{N_{1,1,1} + 1}{N_1 + 3} = \frac{3}{12}$$

$$\bar{\theta}_{2,1,1} = \frac{N_{2,1,1} + 1}{N_1 + 3} = \frac{4}{12}$$

$$\bar{\theta}_{3,1,1} = \frac{N_{3,1,1} + 1}{N_1 + 3} = \frac{5}{12}$$

$$\bar{\theta}_{S,2,1} = \frac{N_{S,2,1} + 1}{N_1 + 3} = \frac{2}{12}$$

$$\bar{\theta}_{M,2,1} = \frac{N_{M,2,1} + 1}{N_1 + 3} = \frac{5}{12}$$

$$\bar{\theta}_{L,2,1} = \frac{N_{L,2,1} + 1}{N_1 + 3} = \frac{5}{12}$$

Example

□ 计算 $\bar{\theta}_{k,j,c}$

$$\begin{aligned}\bar{\theta}_{1,1,-1} &= \frac{N_{1,1,-1} + 1}{N_{-1} + 3} = \frac{4}{9} & \bar{\theta}_{S,2,-1} &= \frac{N_{S,2,-1} + 1}{N_{-1} + 3} = \frac{4}{9} \\ \bar{\theta}_{2,1,-1} &= \frac{N_{2,1,-1} + 1}{N_{-1} + 3} = \frac{3}{9} & \bar{\theta}_{M,2,-1} &= \frac{N_{M,2,-1} + 1}{N_{-1} + 3} = \frac{3}{9} \\ \bar{\theta}_{3,1,-1} &= \frac{N_{3,1,-1} + 1}{N_{-1} + 3} = \frac{2}{9} & \bar{\theta}_{L,2,-1} &= \frac{N_{L,2,-1} + 1}{N_{-1} + 3} = \frac{2}{9}\end{aligned}$$

Example

- 预测 $p(Y = c | \mathbf{x}, \mathcal{D})$

$$\begin{aligned} p(Y = 1 | (2, S)^T, \mathcal{D}) &\propto \bar{\pi}_1(\bar{\theta}_{2,1,1})^{\mathbb{I}(X^{(1)}=2)} (\bar{\theta}_{S,2,1})^{\mathbb{I}(X^{(2)}=S)} \\ &= \frac{10}{17} \times \frac{4}{12} \times \frac{2}{12} = 0.0327 \end{aligned}$$

$$\begin{aligned} p(Y = -1 | (2, S)^T, \mathcal{D}) &\propto \bar{\pi}_{-1}(\bar{\theta}_{2,1,-1})^{\mathbb{I}(X^{(1)}=2)} (\bar{\theta}_{S,2,-1})^{\mathbb{I}(X^{(2)}=S)} \\ &= \frac{7}{17} \times \frac{3}{9} \times \frac{4}{9} = 0.0610 \end{aligned}$$

所以数据 $(2, S)^T$ 预测的标签为 -1

Summary

■ Bayesian concept learning

- In discrete hypothesis spaces, update hypothesis posterior using Bayes' rule
- MAP estimation: select the hypothesis with the highest posterior probability

■ Parametric Bayesian estimation

- **Beta-Binomial model:** extend parameter estimation to the binomial case
- **Dirichlet-Multinomial model:** further extend to the multinomial case
- Provides a unified framework from discrete to continuous distributions

Summary

■ Naive Bayes classifiers

- **Generative modeling:** estimating class-conditional distribution $p(x | y)$
- To simplify, conditional independence assumption

■ Parameter estimation in Naive Bayes

- **Point estimates:** MLE / MAP — simple and efficient
- **Bayesian Naive Bayes:** keep full posterior distribution over parameters
 - In prediction, often approximated by **posterior mean** or **MAP** for tractability

Thanks!



Questions?