CS 8501 Advanced Topics in Machine Learning

Lecture 02: Generative Modeling

Yangfeng Ji
Information and Language Processing Lab
Department of Computer Science
University of Virginia
https://yangfengji.net/

Generative Modeling

$$p(y \mid x; \theta) \propto p(y; \theta) \cdot p(x \mid y; \theta)$$

- θ represents all the parameters in this model
- $p(x \mid y; \theta)$: likelihood
- $p(y; \theta)$: prior

Concept Learning

- Pick a concept
- Give some examples of this concept
- Ask someone whether a new example belongs to this concept

Concept Learning in Generative Modeling

Given

- y: concept
- $\mathcal{D} = \{x_i\}$: observations

Learning

$$p(y \mid \mathcal{D}; \theta) \propto p(y; \theta) \cdot p(\mathcal{D} \mid y; \theta)$$

Number Game

- Assume all observed numbers are randomly drawn from $\{1, \ldots, 100\}$ with equal chance
- Hypothesis space $\mathcal{H}$: a set of hypotheses
- Version space $\mathcal{V}$: the subset of $\mathcal{H}$ that is consistent with $\mathcal{D}$. For example, if $\mathcal{D} = \{2\}$
  - h_{two} = "powers of two"
  - h_{even} = "even numbers"

Likelihood

The likelihood function of each hypothesis is

$$p(\mathcal{D} \mid h) = \left(\frac{1}{|h|}\right)^N$$

- $|h|$: the size of numbers can be explained by this hypothesis
  - For example, $|h_{\text{two}}| = 6$
- N: the size of observation $\mathcal{D}$

Example

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

- With hypothesis h_{two}

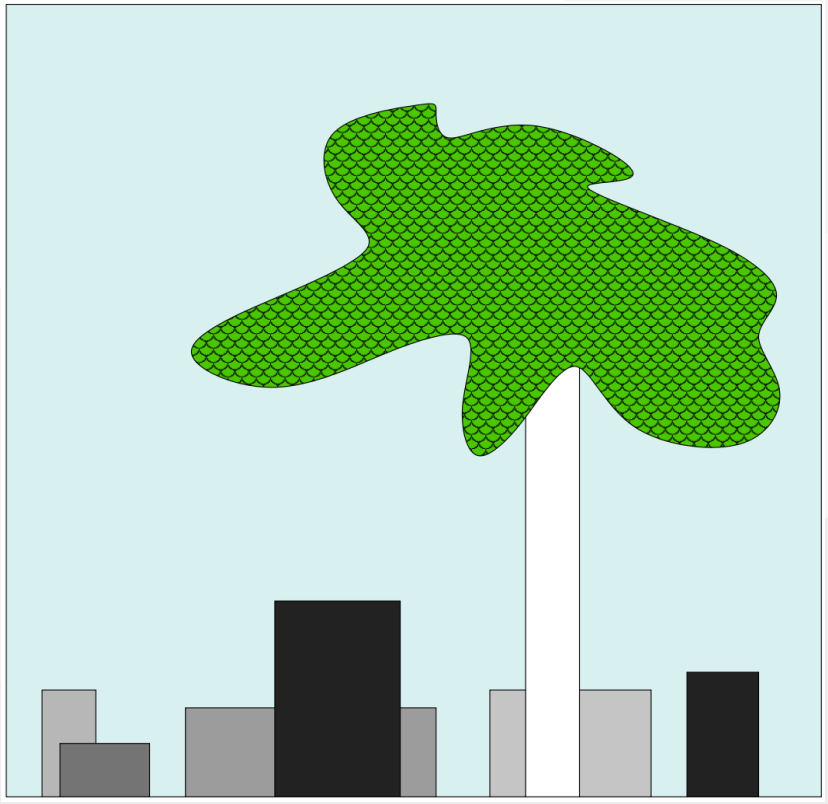
$$p(\mathcal{D} \mid h_{\text{two}}) = \left(\frac{1}{6}\right)^4$$

- With hypothesis h_{even}

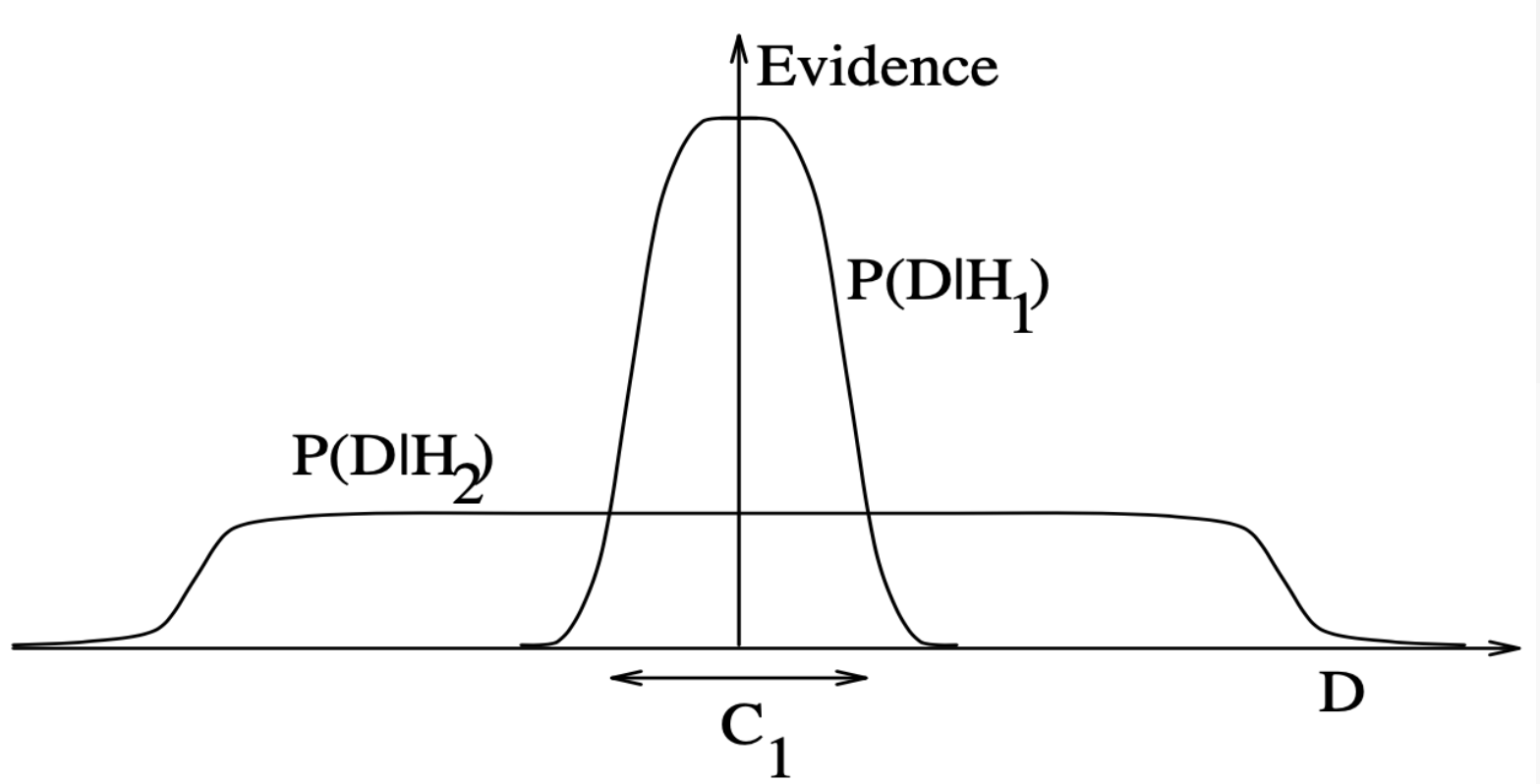
$$p(\mathcal{D} \mid h_{\text{even}}) = \left(\frac{1}{50}\right)^4$$

Occam's Razor

- [Mackay, 2006]: "Accept the simplest explanation that fits the data"
- For example, how many boxes in the following image?



Model Selection



- William of Ockham: "Entities are not to be multiplied without necessity"

Implication of Occam's Razor

There are many implications of Occam's Razor in machine learning research. For example,

- You should always select the simpliest models that can solve the problem
- However, it also means you should understand your research problem (or your data)

Example (Cont.)

Given $\mathcal{D} = \{16, 8, 2, 64\}$, now let's compare two similar hypotheses

- h_{two} = "power of two"

$$p(\mathcal{D} \mid h_{\text{two}}) = \left(\frac{1}{6}\right)^6$$

- h_{another} = "power of two except 4 and 32"

$$p(\mathcal{D} \mid h_{\text{another}}) = \left(\frac{1}{4}\right)^4$$

- The limitation of will be addressed by the Bayesian version of Occam's razor

About Maximizing the Likelihood

If the goal is solely about maximizing the likelihood function, then we may pick the hypothesis that explains the current data **too well**.

- This is overfitting
- The simple explanation applies to any other learning scenarios

Prior

Follow the previous discussion, and consider the two hypotheses:

- h: "powers of two"
- h': "powers of two except 4 and 32"
- With the previous discussion on likelihood functions, h' is more likely
- However, in practice, hypotheses like h' is more complicated to implement
    - Or "conceptually unnatural", as discussed in the textbook

Subjectivity

In Bayesian modeling

- Prior is the mechanism by which background knowledge can be brought to bear on a problem
  - It is also the key of "rapid learning" (e.g., learning with small sample sizes)
- The choice of prior sometimes is subjective
  - The subjectivity is a controversial issue in Bayesian modeling

Posterior

The posterior distribution of a hypothesis given $\mathcal{D}$ is

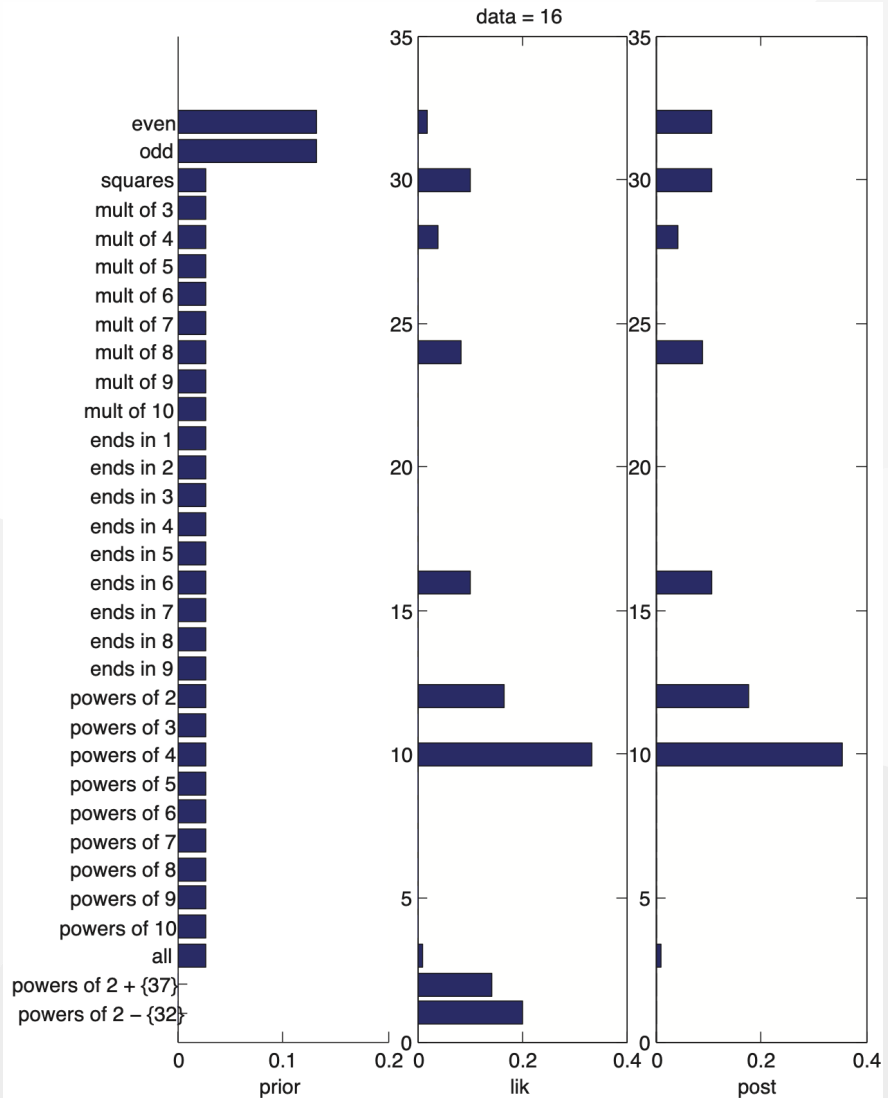
$$p(h \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid h)p(h)}{p(\mathcal{D})}$$

where

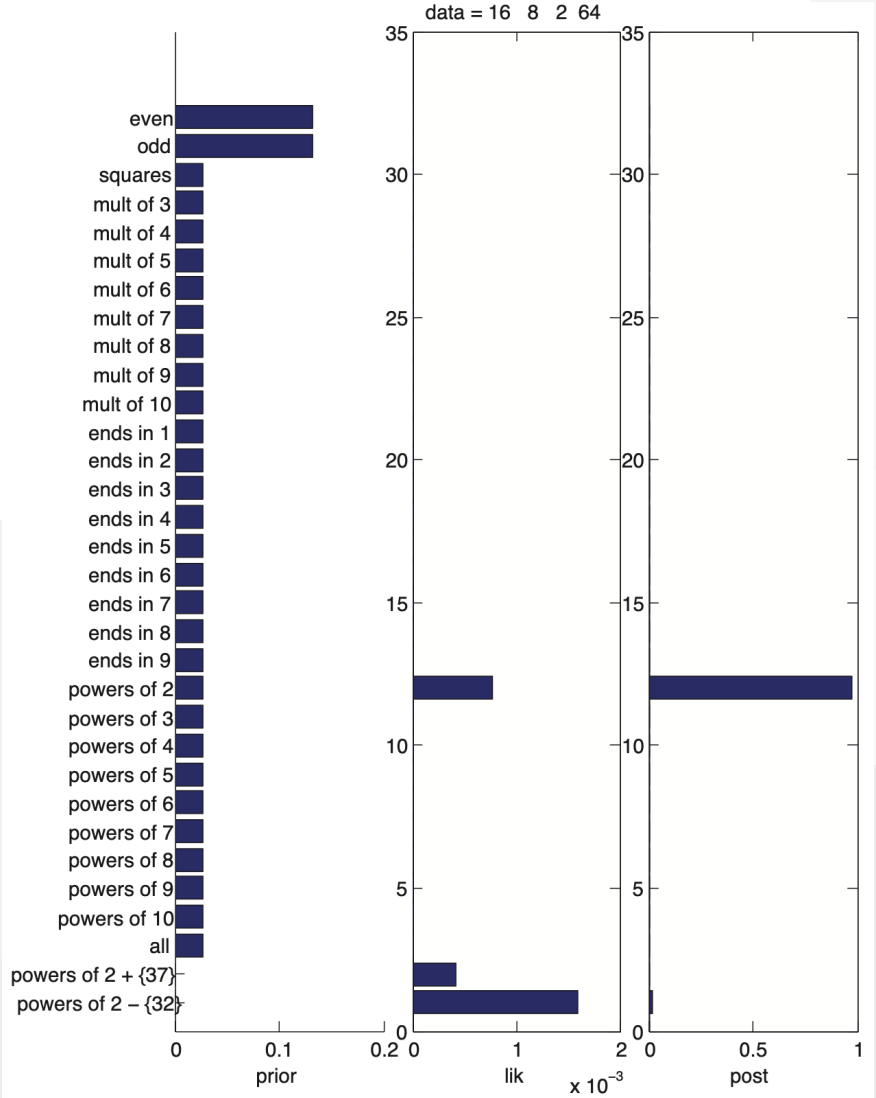
- $p(\mathcal{D}) = \sum_{h'} p(\mathcal{D} \mid h')p(h')$ is the major challenge in Bayesian inference
- If we need a single hypothesis from the posterior distribution, we can use the MAP estimate

$$\hat{h}_{\text{MAP}} = \text{argmax}_h p(h \mid \mathcal{D}) = \text{argmax}_h p(\mathcal{D} \mid h)p(h)$$

Example


data = 16
even
odd
squares
mult of 3
mult of 4
mult of 5
mult of 6
mult of 7
mult of 8
mult of 9
mult of 10
ends in 1
ends in 2
ends in 3
ends in 4
ends in 5
ends in 6
ends in 7
ends in 8
ends in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
all
powers of 2 + {37}
powers of 2 – {32}
0
5
10
15
20
25
30
35
0
0.1
0.2
prior
0
0.2
0.4
lik
0
0.2
0.4
post

Example (Cont.)



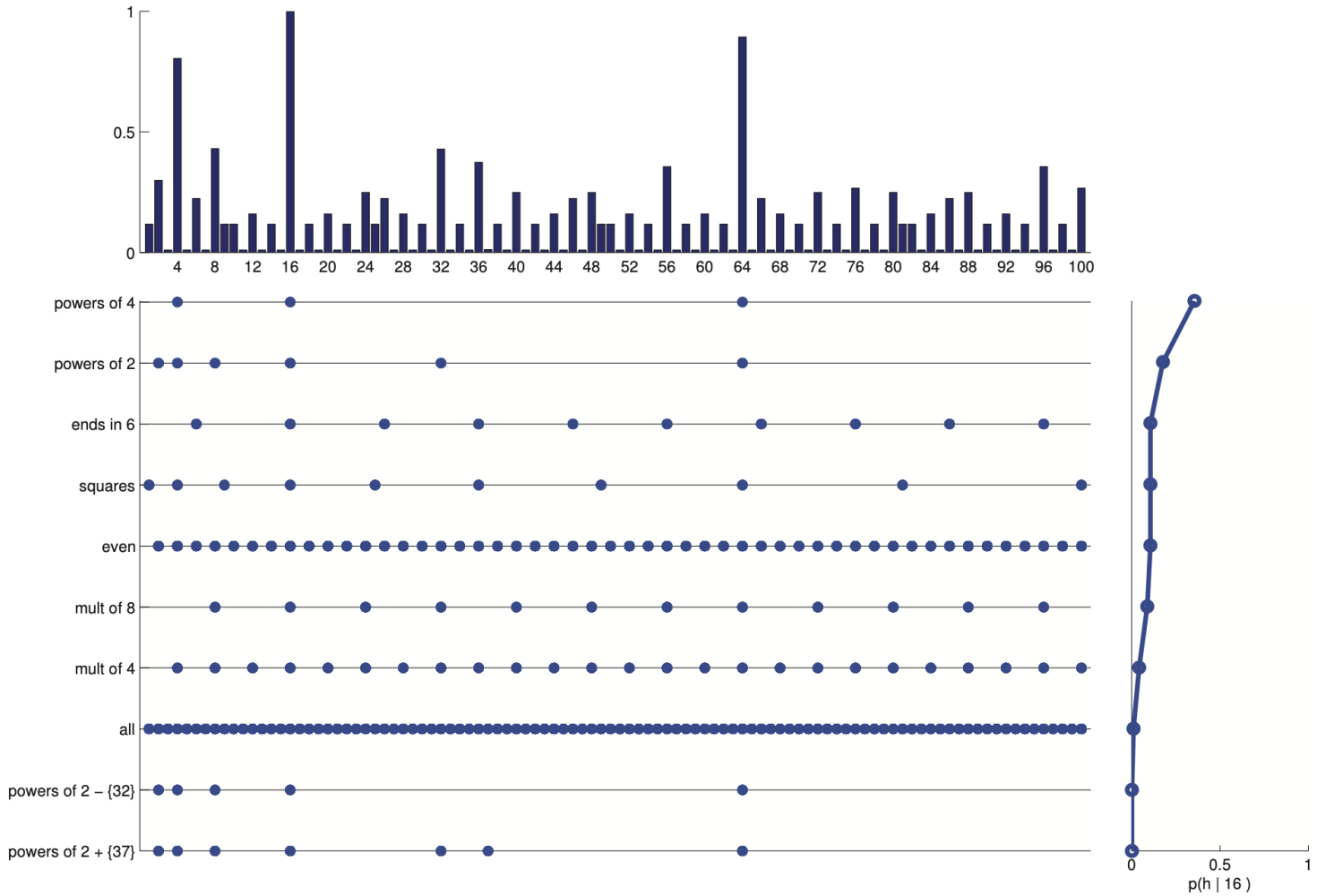
Posterior Predictive Distribution

Provide a way to prodict the next number

$$p(\tilde{x} \in C \mid \mathcal{D}) = \sum_h p(y = 1 \mid \tilde{x}, h) p(h \mid \mathcal{D})$$

Instead of considering one single hypothesis h for the prediction, it *averages* the possibility of all hypotheses.

Example of Number Prediction



The Beta-Binomial model

Binomial Distribution

Let $X_i \sim \text{Bernoulli}(\theta)$, where $p(X_i = 1) = \theta$ and $p(X_i = 0) = 1 - \theta$, then the probability of $\sum_{i=1}^{n} X_i = k$ is

$$p(k \mid n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

- Example: tossing a coin n times, the probability of getting the head k times

Prior

A popular prior distribution for θ is the Beta distribution

$$p(\theta; \gamma_1, \gamma_2) \propto \theta^{\gamma_1 - 1}(1 - \theta)^{\gamma_2 - 1}$$

where γ_1 and γ_2 are the parameters for the prior.

- Regarding θ, these two are called **hyper-parameters**

A Formal Definition

$$p(\theta;\gamma_1,\gamma_2) = B(\gamma_1,\gamma_2)\theta^{\gamma_1-1}(1-\theta)^{\gamma_2-1}$$

where $B(\gamma_1,\gamma_2)$ is the Beta function, which is also the normalization consistent.

Mean

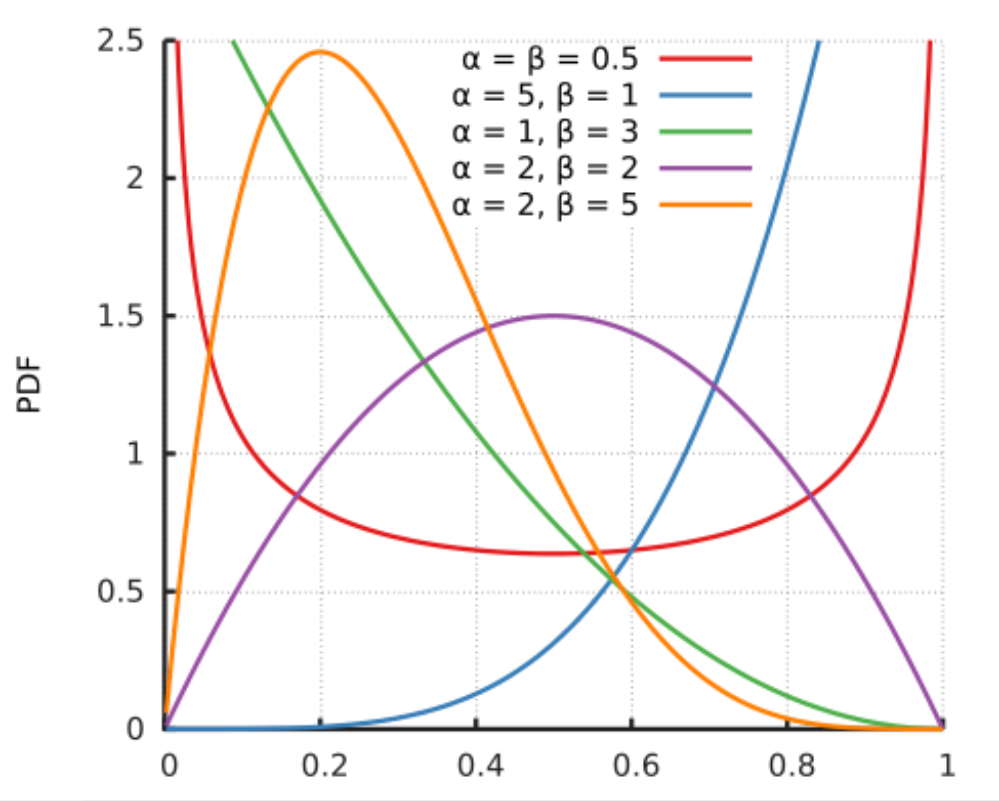
$$E(\theta) = \int_\theta \theta p(\theta;\gamma_1,\gamma_2) = \frac{\gamma_1}{\gamma_1+\gamma_2}$$

Mode

$$\hat{\theta} = \arg\max_\theta p(\theta;\gamma_1,\gamma_2) = \frac{\gamma_1-1}{\gamma_1+\gamma_2-2}$$

Beta Distribution

With different γ_1 and γ_2 (α and β in the following plot)



Posterior

The posterior distribution of θ

$$p(\theta \mid n, k; \gamma_1, \gamma_2) = \frac{p(\theta; \gamma_1, \gamma_2) p(k \mid n, \theta)}{p(k \mid n; \gamma_1, \gamma_2)}$$

where

$$p(k \mid n; \gamma_1, \gamma_2) = \int_\theta p(k, \theta \mid n; \gamma_1, \gamma_2) d\theta$$

Posterior (II)

Without the denominator, we have

$$p(\theta \mid n, k; \gamma_1, \gamma_2) \propto \theta^k (1-\theta)^{n-k} \cdot \theta^{\gamma_1 - 1} (1-\theta)^{\gamma_2 - 1}$$

or

$$p(\theta \mid n, k; \gamma_1, \gamma_2) \propto \theta^{k+\gamma_1 - 1} (1-\theta)^{n-k+\gamma_2 - 1}$$

- Beta distribution is the conjugate prior, because the posterior has the same form as the prior

The Dirichlet-Multinomial Model

Distributions

Multinomial distribution

$$p(x;\theta) = \frac{n!}{x_1! \cdots x_K!} \theta_1^{x_1} \cdots \theta_k^{x_K}$$

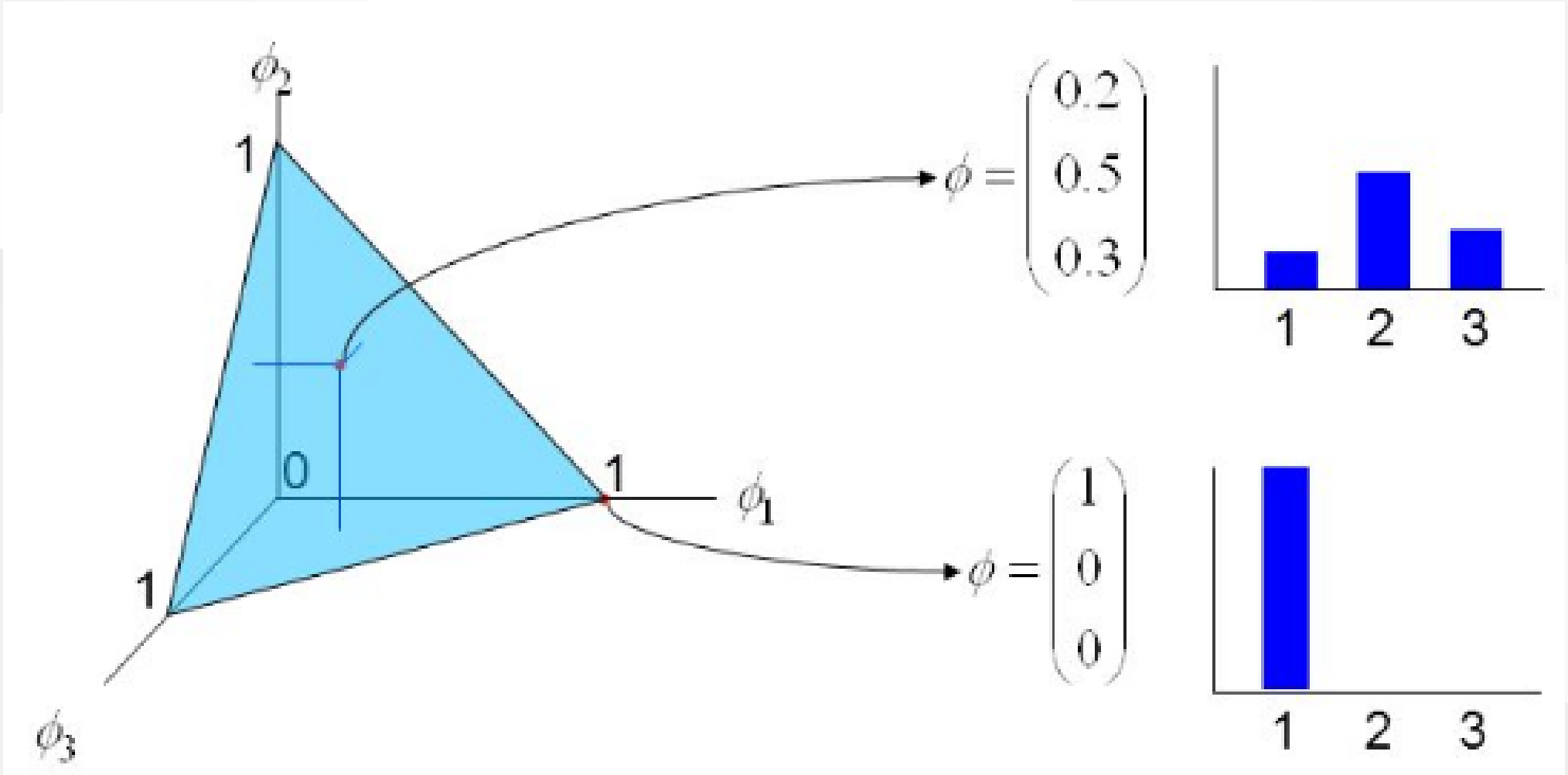
where x and θ are both K-dimensional vectors, and $\sum_{k=1}^{K} \theta_k = 1$

Dirichlet distribution

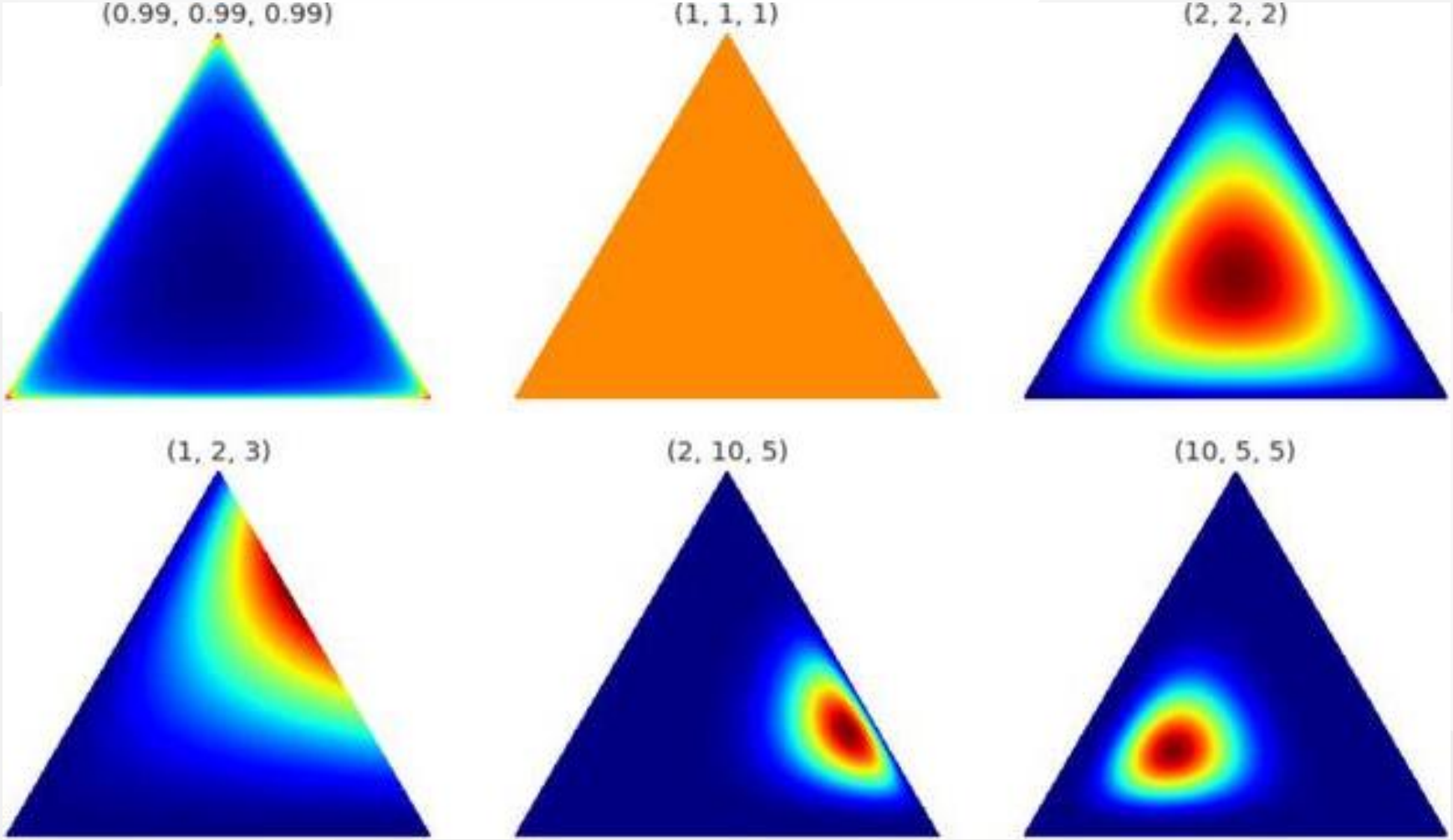
$$p(\theta;\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

where α is a K-dimensional vector too, with $\alpha_k > 0$

Dirichlet Distribution



Dirichlet Distribution (II)



Naive Bayes Classifiers

Problem Setup

Consider a classification problem, with $x \in \mathbb{R}^N$ as input and $y \in \{0, 1\}$ as output

- Focus on the posterior distribution of y, instead of the model parameter θ

$$p(y \mid x; \theta)$$

- We can also give θ a prior, which is called Bayesian naive Bayes classifier (section 3.5.1.2)

Likelihood

Given $\{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^K$ is a K-dimensional vector.

Navie Bayes assume that different dimensions in input are independent from each other

$$p(x_i \mid y_i; \theta_{x|y}) = \prod_{j=1}^{K} p(x_{i,j} \mid y_i; \theta_{x|y,j})$$

- This is a **naive** assumption
- Choose $p(x_{i,j} \mid y_i, \theta_{x|y,j})$ based on your inputs. For example,
  - Continuous variables: Gaussian
  - Discrete variables: Binomial or Multinomial

Likelihood (II)

The overall likelihood given the training set $\{(x_i, y_i)\}_{i=1}^{N}$ is defined as

$$\mathrm{lik}(\theta_{x|y}) = \prod_{i=1}^{N} \prod_{j=1}^{K} p(x_{i,j} \mid y_i; \theta_{x|y,j})$$

Prior

$$p(y; \theta_y)$$

Choices of the distribution:

- Uniform distribution
- Bernoulli distribution with the parameter estimated from data

MAP Estimate

$$\log p(y;\theta_y) + \sum_{i=1}^{N}\sum_{j=1}^{K} \log p(x_{i,j} \mid y_i;\theta_{x|y,j})$$

Re-arrange it:

$$\log p(y;\theta_y) + \sum_{j=1}^{K}\{\sum_{i=1}^{N} \log p(x_{i,j} \mid y_i;\theta_{x|y,j})\}$$

We need to solve $K+1$ one-dimensional problems

Thank You!