Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012

Prof. Erik Sudderth

Lecture 3:

Bayesian Learning, MAP & ML Estimation

Classification: Naïve Bayes

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Bayes Rule (Bayes Theorem)

$\theta \longrightarrow$ unknown parameters (many possible models)

$\mathcal{D} \longrightarrow$ observed data available for learning

$p(\theta) \longrightarrow$ prior distribution (domain knowledge)

$p(\mathcal{D} \mid \theta) \longrightarrow$ likelihood function (measurement model)

$p(\theta \mid \mathcal{D}) \longrightarrow$ posterior distribution (learned information)

$$p(\theta, \mathcal{D}) = p(\theta)p(\mathcal{D} \mid \theta) = p(\mathcal{D})p(\theta \mid \mathcal{D})$$

$$p(\theta \mid \mathcal{D}) = \frac{p(\theta, \mathcal{D})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\sum_{\theta' \in \Theta} p(\mathcal{D} \mid \theta')p(\theta')}$$

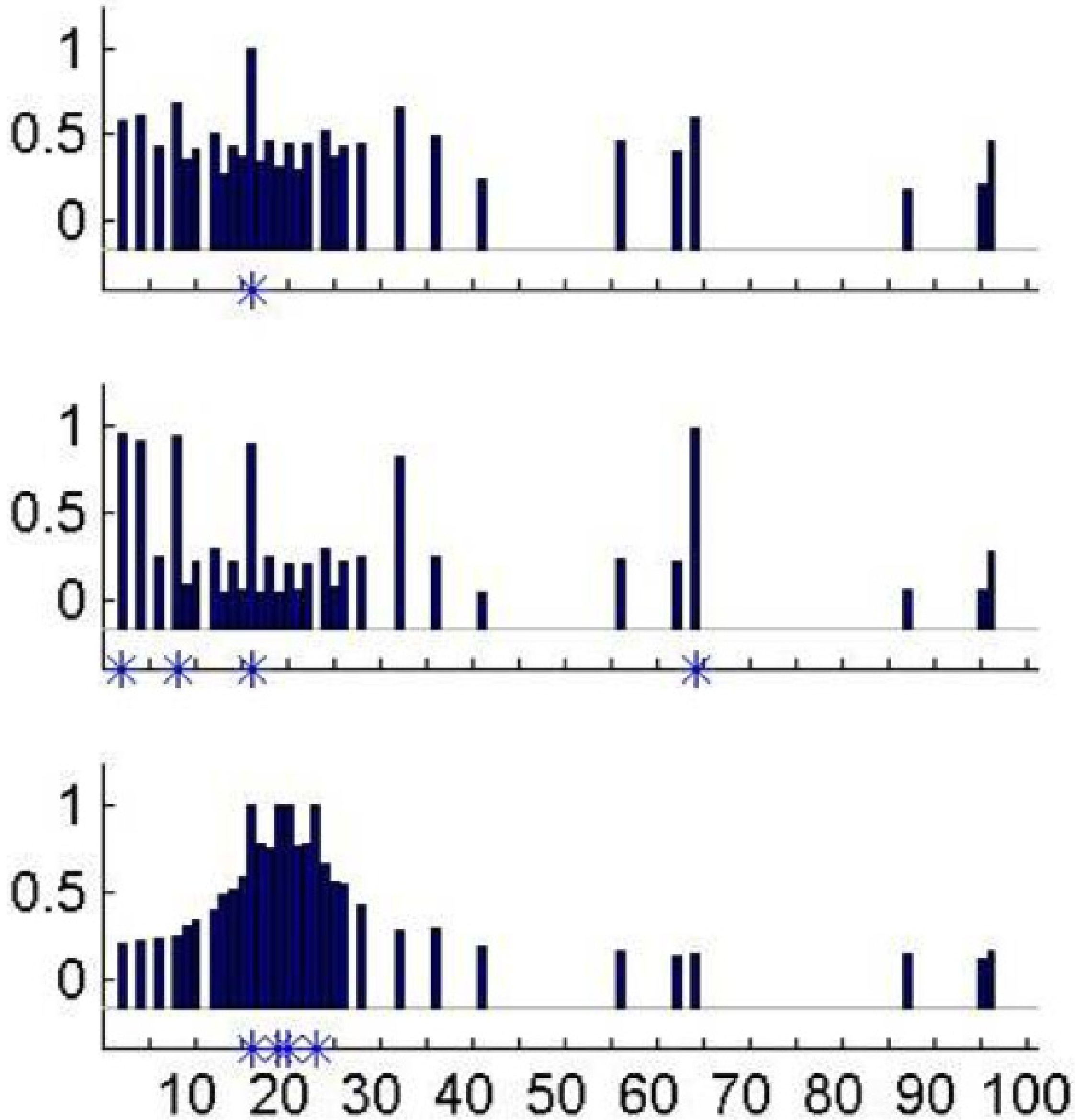
$$\propto p(\mathcal{D} \mid \theta)p(\theta)$$

The Number Game

- I am thinking of some arithmetical concept, such as:
  - Prime numbers
  - Numbers between 1 and 10
  - Even numbers
  - …
- I give you a series of randomly chosen *positive* examples from the chosen class
- Question: Are other test digits also in the class?

Tenenbaum 1999

Predictions of 20 Humans



D = {16}

D = {16,8,2,64}

D = {16,23,19,20}

A Bayesian Model

Likelihood:

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

- Assume examples are sampled uniformly at random from all numbers that are consistent with the hypothesis
- Size principle: Favors smallest consistent hypotheses

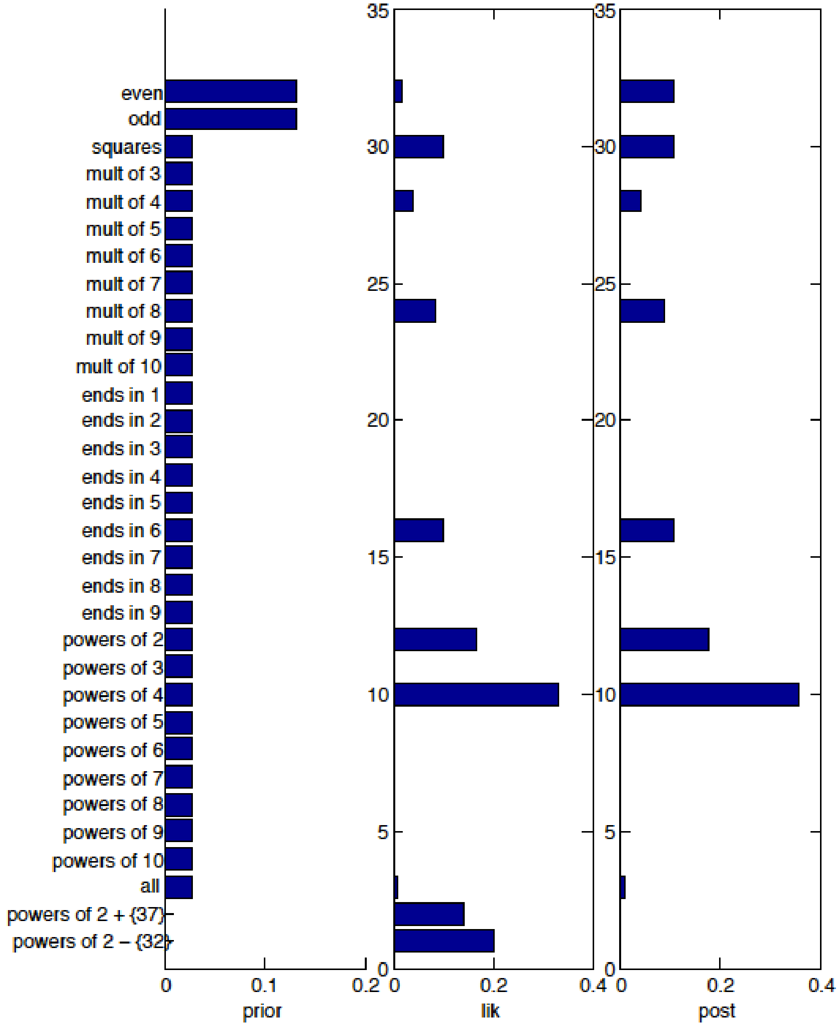
Prior:

- Based on prior experience, some hypotheses are more probable ("natural") than others
  - Powers of 2
  - Powers of 2 except 32, plus 37
- Subjectivity: May depend on observer's experience

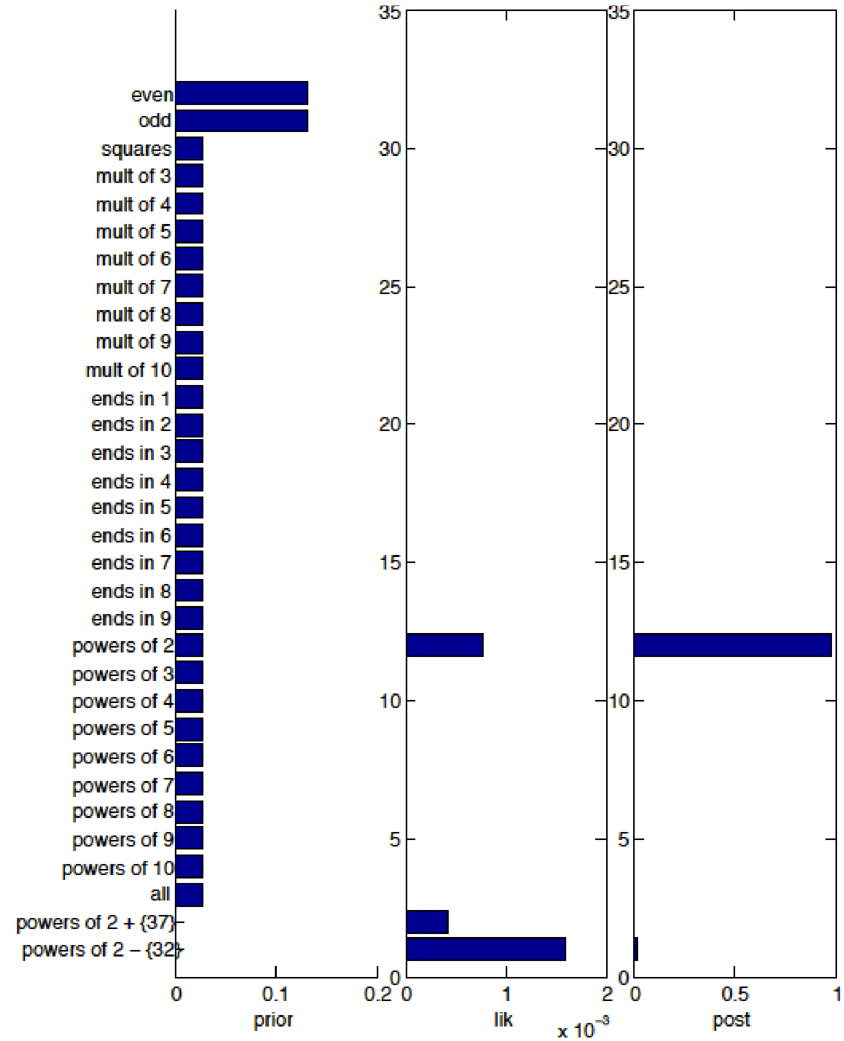
D = {5, 34, 2, 89, 1, 13}?

Posterior Distributions

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



D = {16}



D = {16,8,2,64}

Posterior Estimation

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

- As the amount of data becomes large, weak conditions on the hypothesis space and measurement process imply that

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

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

- This is a *maximum a posteriori (MAP)* estimate:

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

- With a large amount of data, and/or an (almost) uniform prior, we approach the *maximum likelihood (ML)* estimate:

$$h^{mle} \triangleq \operatorname*{argmax}_h p(\mathcal{D}|h) = \operatorname*{argmax}_h \log p(\mathcal{D}|h)$$

- More theory to come later…

Posterior Predictions

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

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

- Suppose we want to predict the next number that will be revealed to us. One option is to use the MAP estimate:

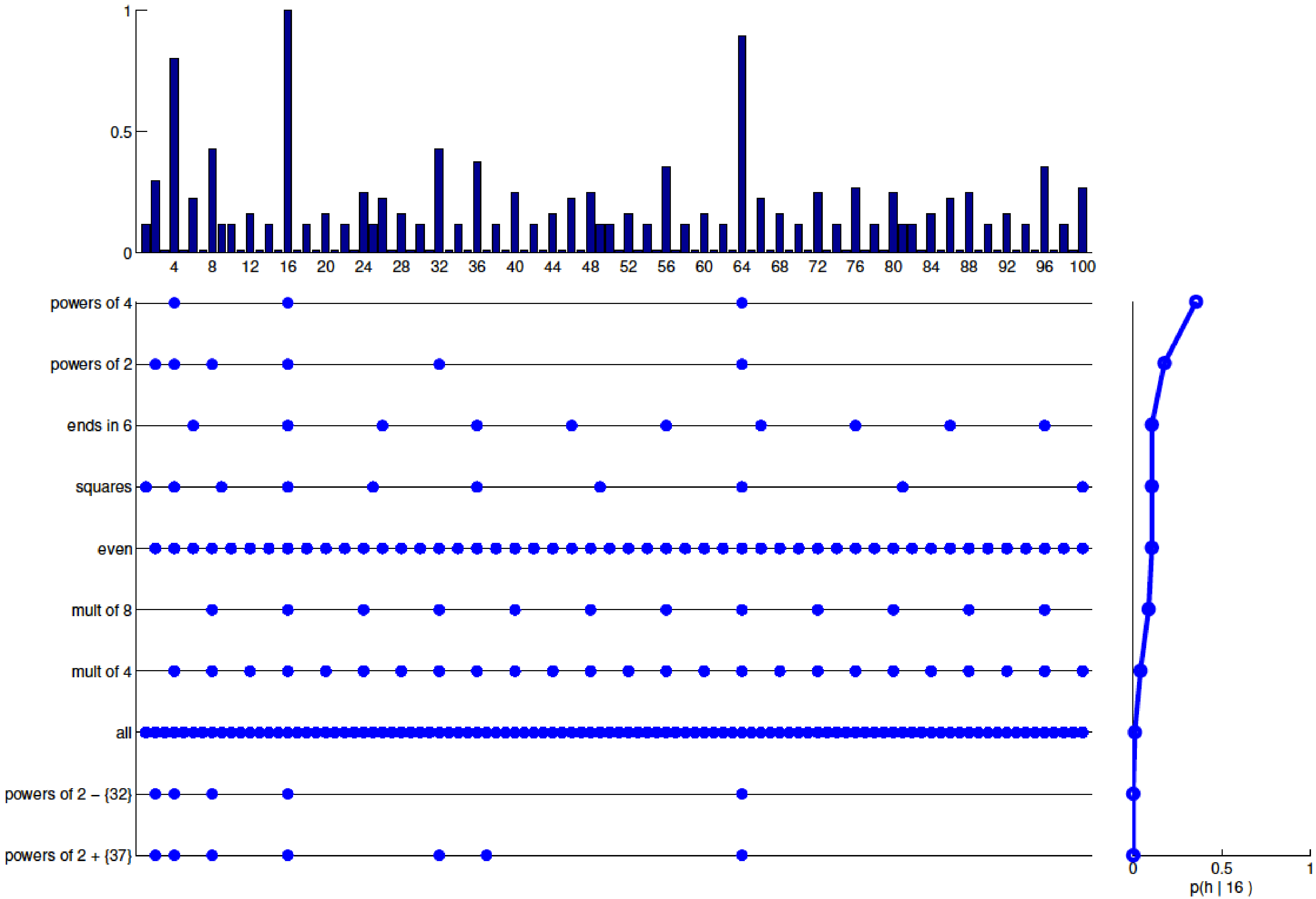
$$p(\tilde{x} \mid \mathcal{D}) \approx p(\tilde{x} \mid \hat{h}) \qquad \tilde{x} \in \{1, 2, 3, \ldots\}$$

- But if we correctly apply Bayes rule to the specified model, we instead obtain the *posterior predictive distribution*:

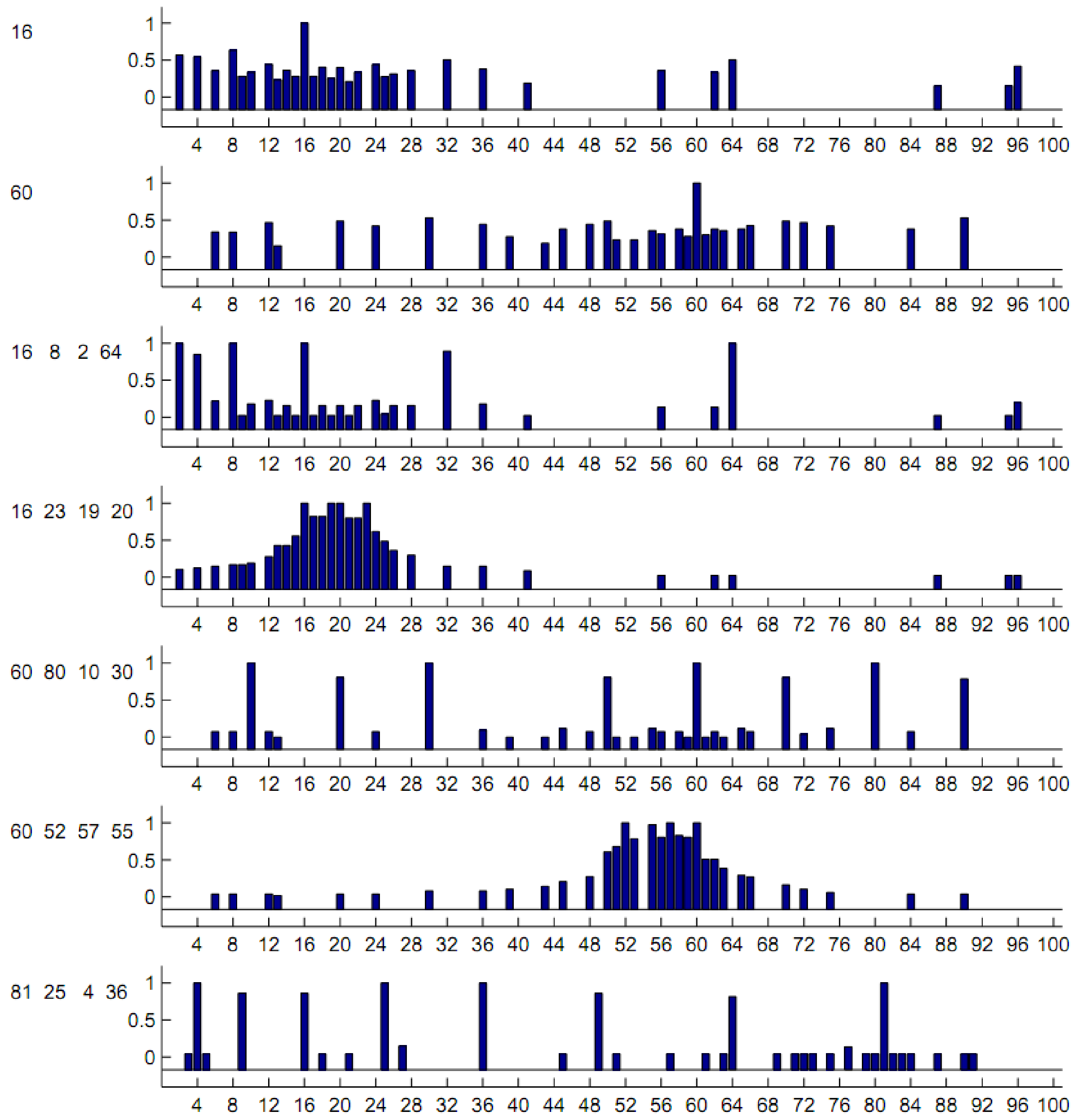
$$p(\tilde{x} \mid \mathcal{D}) = \sum_{h \in \mathcal{H}} p(\tilde{x} \mid h) p(h \mid \mathcal{D})$$

- This is sometimes called *Bayesian model averaging*.

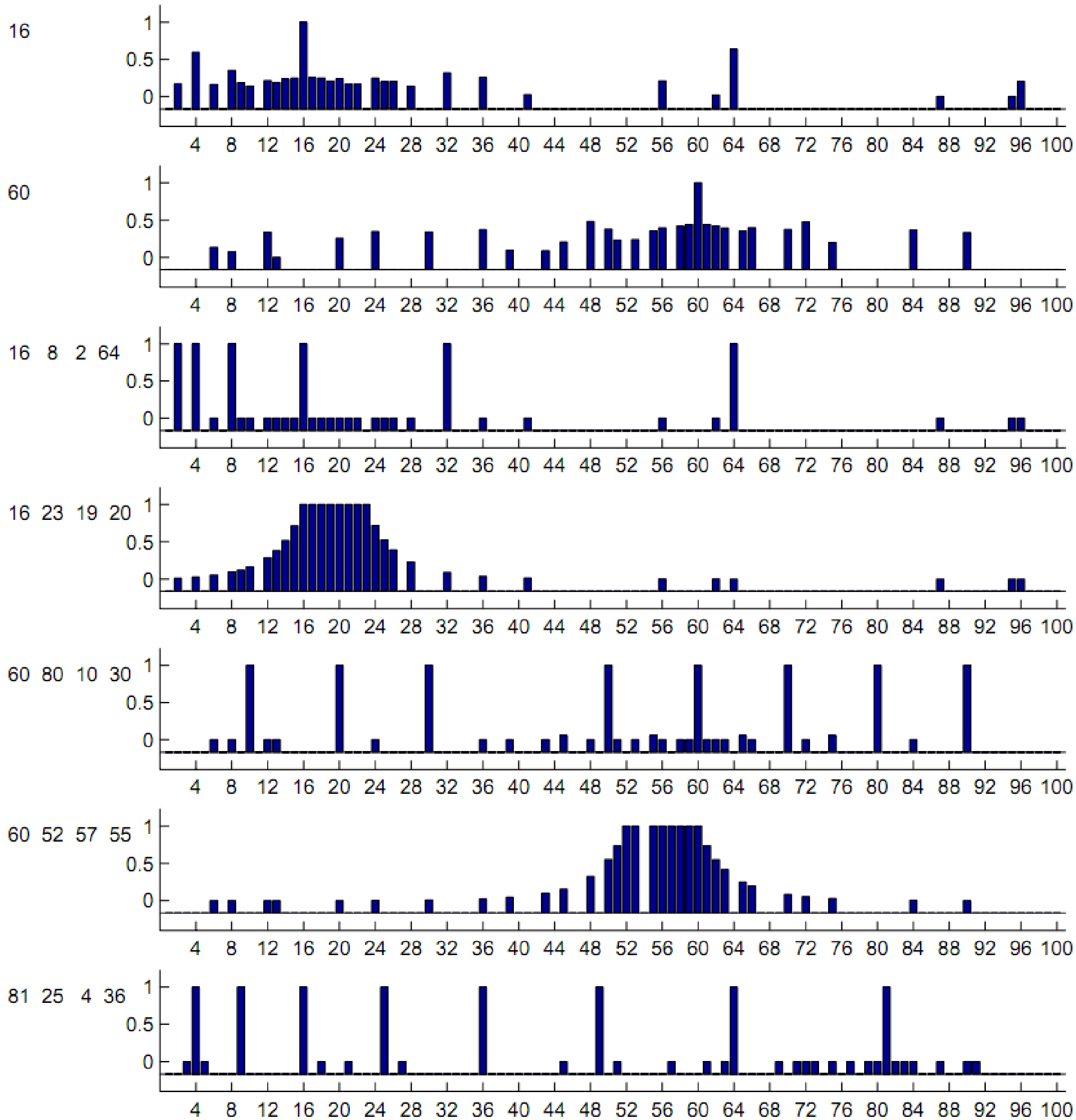
Posterior Predictive Distribution



Experiment: Human Judgments



Experiment: Bayesian Predictions



Machine Learning Problems

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Generative Classifiers

$y \longrightarrow$ class label in *{1,…,C}*, observed in training

$x \in \mathcal{X} \longrightarrow$ observed features to be used for classification

$\theta \longrightarrow$ parameters indexing family of models

$$p(y, x \mid \theta) = p(y \mid \theta) p(x \mid y, \theta)$$

prior distribution *likelihood function*

- Compute class *posterior distribution* via Bayes rule:

$$p(y = c \mid x, \theta) = \frac{p(y = c \mid \theta) p(x \mid y = c, \theta)}{\sum_{c'=1}^{C} p(y = c' \mid \theta) p(x \mid y = c', \theta)}$$

- *Inference:* Find label distribution for some input example
- *Classification:* Make decision based on inferred distribution
- *Learning:* Estimate parameters θ from labeled training data

Specifying a Generative Model

$$p(y, x \mid \theta) = \underbrace{p(y \mid \theta)}_{\text{prior distribution}} \underbrace{p(x \mid y, \theta)}_{\text{likelihood function}}$$

- For generative classification, we take the prior to be some categorical (multinoulli) distribution:

$$p(y \mid \theta) = \mathrm{Cat}(y \mid \theta)$$

- The likelihood must be matched to the domain of the data
- Suppose *x* is a vector of *D* different *features*
- The simplest generative model assumes that these features are *conditionally independent*, given the class label:

$$p(x \mid y = c, \theta) = \prod_{j=1}^{D} p(x_j \mid y = c, \theta_{jc})$$

- This is a so-called *naïve Bayes model* for classification

Learning a Generative Model

$$p(\theta \mid y, x) \propto p(\theta, y, x) = p(\theta)p(y \mid \theta)p(x \mid y, \theta)$$

model *class* *features*

- Assume we have N training examples independently sampled from an unknown naïve Bayes model:

$y_i \longrightarrow$ observed class label for training example *i*

$x_{ij} \longrightarrow$ value of feature *j* for training example *i*

$$p(\theta \mid y, x) \propto p(\theta) \prod_{i=1}^{N} p(y_i \mid \theta) p(x_i \mid y_i, \theta)$$

$$\propto p(\theta) \prod_{i=1}^{N} p(y_i \mid \theta) \prod_{j=1}^{D} p(x_{ij} \mid y_i, \theta)$$

- Learning: ML estimate, MAP estimate, or posterior prediction

Naïve Bayes: ML Estimation

$$p(\mathbf{x}_i, y_i|\boldsymbol{\theta}) = p(y_i|\boldsymbol{\pi}) \prod_j p(x_{ij}|\boldsymbol{\theta}_j) = \prod_c \pi_c^{\mathbb{I}(y_i=c)} \prod_j \prod_c p(x_{ij}|\boldsymbol{\theta}_{jc})^{\mathbb{I}(y_i=c)}$$

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

$N_c \longrightarrow$ number of examples of training class *c*

- Even if we are doing discrete categorization based on discrete features, this is a *continuous* optimization problem!
- Bayesian reasoning about models also requires continuous probability distributions, even in this simple case
- Next week we will show that the ML class estimates are:

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

- For binary features, we have: $\hat{\theta}_{jc} = \frac{N_{jc}}{N_c}$ $\qquad x_j|y = c \sim \mathrm{Ber}(\theta_{jc})$