FE540 금융공학 인공지능 및 기계학습

Generative Models

Kee-Eung Kim
Department of Computer Science
KAIST

Korea Advanced Institute of Science and Technology
한국과학기술원

Bayes Rule

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

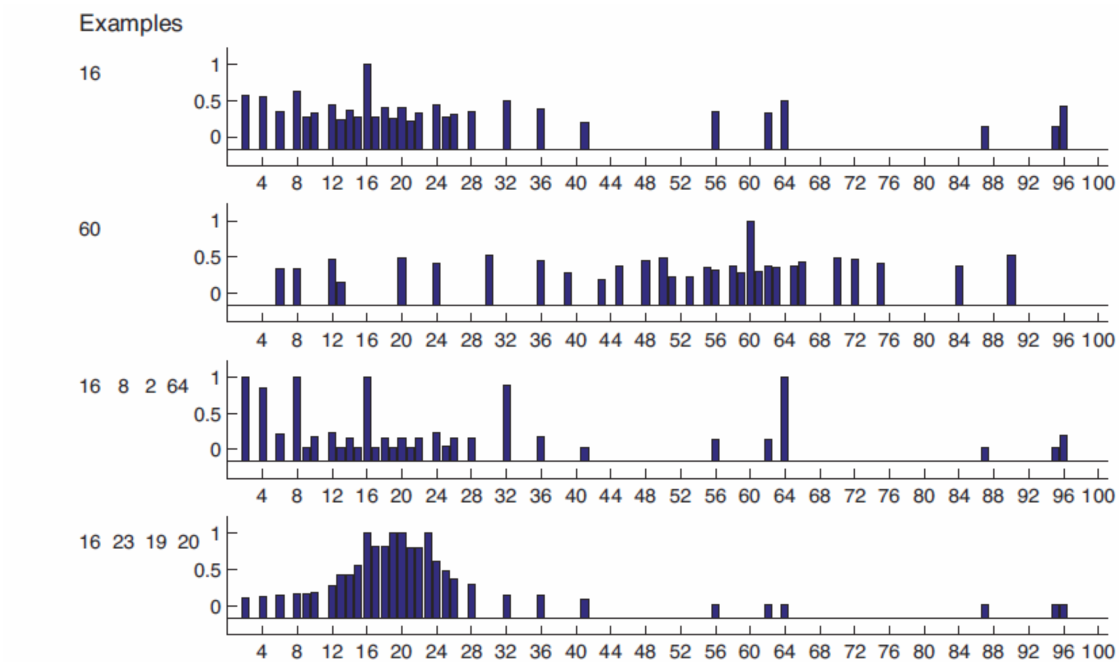
□ Example: medical diagnosis

- From a positive mammogram test result, what is the probability that a person has a breast cancer?
- Suppose sensitivity = 80%
  - Y = mammogram result, X = breast cancer
  - $p(Y=1|X=1)=0.8$
  - = 80% chance of breast cancer? (**base rate fallacy**)
- Two additional information
  - **Prior**: $p(X=1)=0.004$
  - **False positive** (i.e. false alarm) rate: $p(Y=1|X=0)=0.1$
- Correct answer: $p(X=1|Y=1)=0.031$

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Number Game

- Given a series of randomly chosen positive examples $\mathcal{D} = \{x_1, \ldots, x_N\}$ from some arithmetic concept, determine whether a new test case $\tilde{x}$ belongs to it.
  - e.g. “prime number” or “a number between 1 and 10”



$$p(\tilde{x}|\mathcal{D})$$

Figure 3.1 Empirical predictive distribution averaged over 8 humans in the number game. First two rows: after seeing $\mathcal{D} = \{16\}$ and $\mathcal{D} = \{60\}$. This illustrates diffuse similarity. Third row: after seeing $\mathcal{D} = \{16, 8, 2, 64\}$. This illustrates rule-like behavior (powers of 2). Bottom row: after seeing $\mathcal{D} = \{16, 23, 19, 20\}$. This illustrates focussed similarity (numbers near 20). Source: Figure 5.5 of (Tenenbaum 1999). Used with kind permission of Josh Tenenbaum.

Version Space

- Assume a hypothesis space of concepts, $\mathcal{H}$
  - "odd numbers", "even numbers", "all numbers ending in j", ...
- Version space = the set of all hypotheses that are consistent with the examples
  - The version space shrinks as more examples are given, i.e., we become increasingly certain about the concept
- After seeing $\mathcal{D} = \{16, 8, 2, 64\}$, what is your guess on the true concept?
  - Among *many* hypotheses in the version space, why this particular choice?
  - There is a Bayesian explanation of your choice...

Likelihood

- Suppose (strong sampling assumption):

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

  - N examples are assumed to be sampled from hypothesis h
  - Assuming all numbers are integers from 1...100,
    - $h_{\text{two}} = \{2, 4, 8, 16, 32, 64\} \quad \text{(powers of two)}$
    - $h_{\text{even}} = \{2, 4, 6, 8, 10, 12, \ldots, 100\} \quad \text{(even numbers)}$
    - $p(\mathcal{D} = \{16\}|h_{\text{even}}) =?, \quad p(\mathcal{D} = \{16\}|h_{\text{two}}) =?$
    - $p(\mathcal{D} = \{16, 8, 2, 64\}|h_{\text{even}}) =? \quad p(\mathcal{D}|h_{\text{two}}) =?$

Prior

- Prior: $p(h)$
  - encodes subjectivity, preference, or background knowledge

- Consider two hypotheses
  - $h = \text{"powers of two"}$
  - $h' = \text{"powers of two except 32"}$
  - conceptually natural vs. unnatural
  - low prior probability to unnatural concepts

- For Number Game, we use <u>*uniform*</u> prior over 30 arithmetic concepts

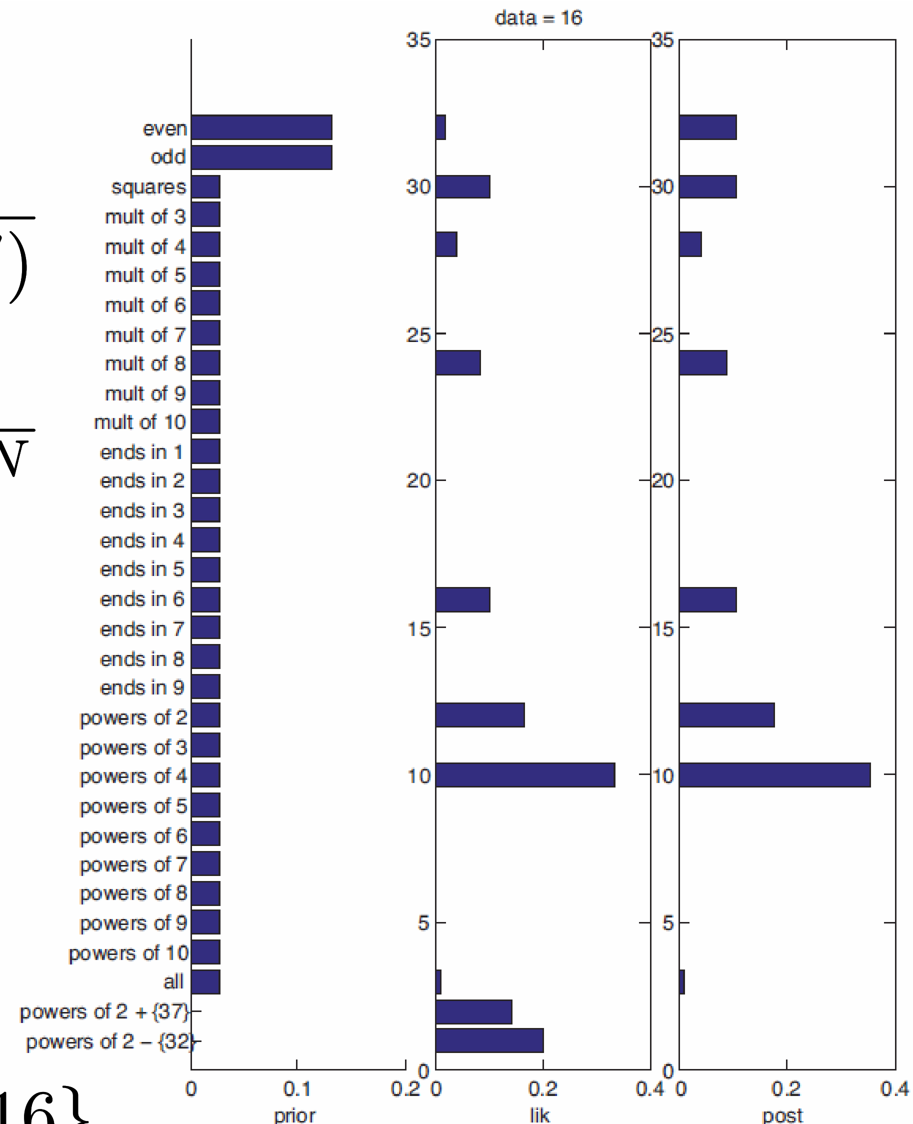
Posterior

- By Bayes rule,

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

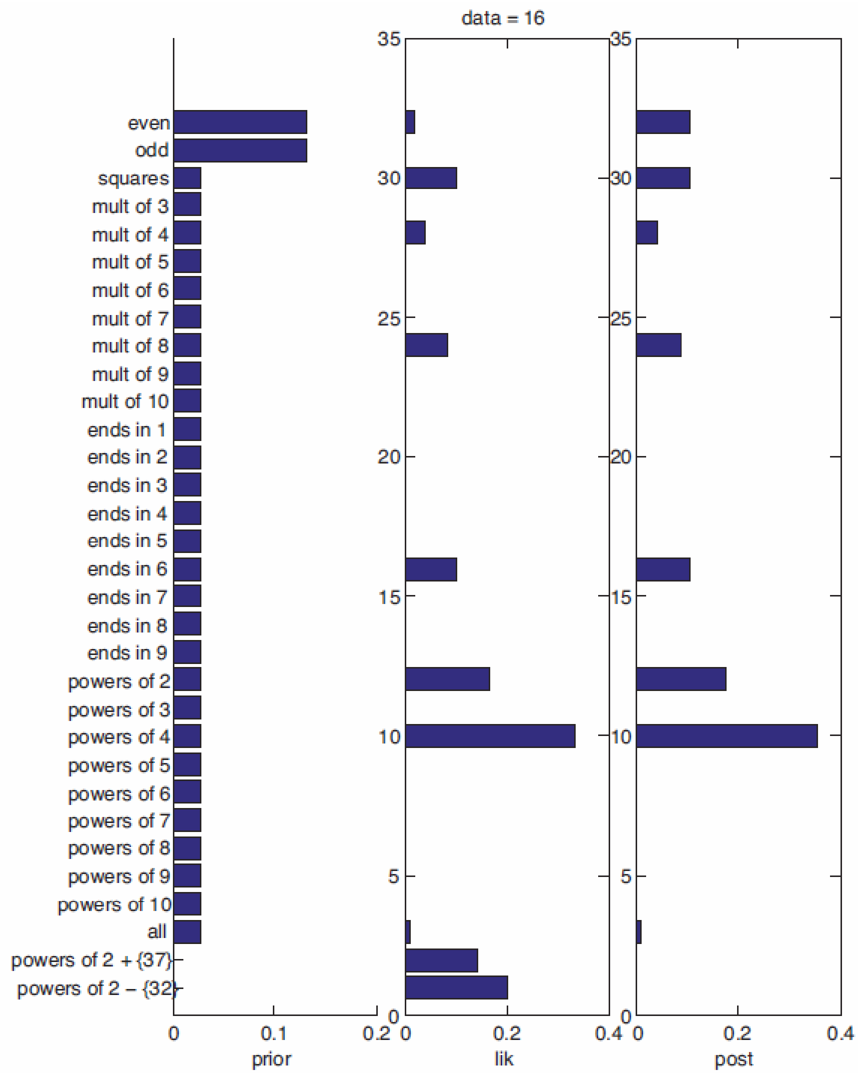
$$= \frac{p(h)\mathbb{I}(\mathcal{D} \in h)/|h|^N}{\sum_{h'} p(h')\mathbb{I}(\mathcal{D} \in h')/|h'|^N}$$

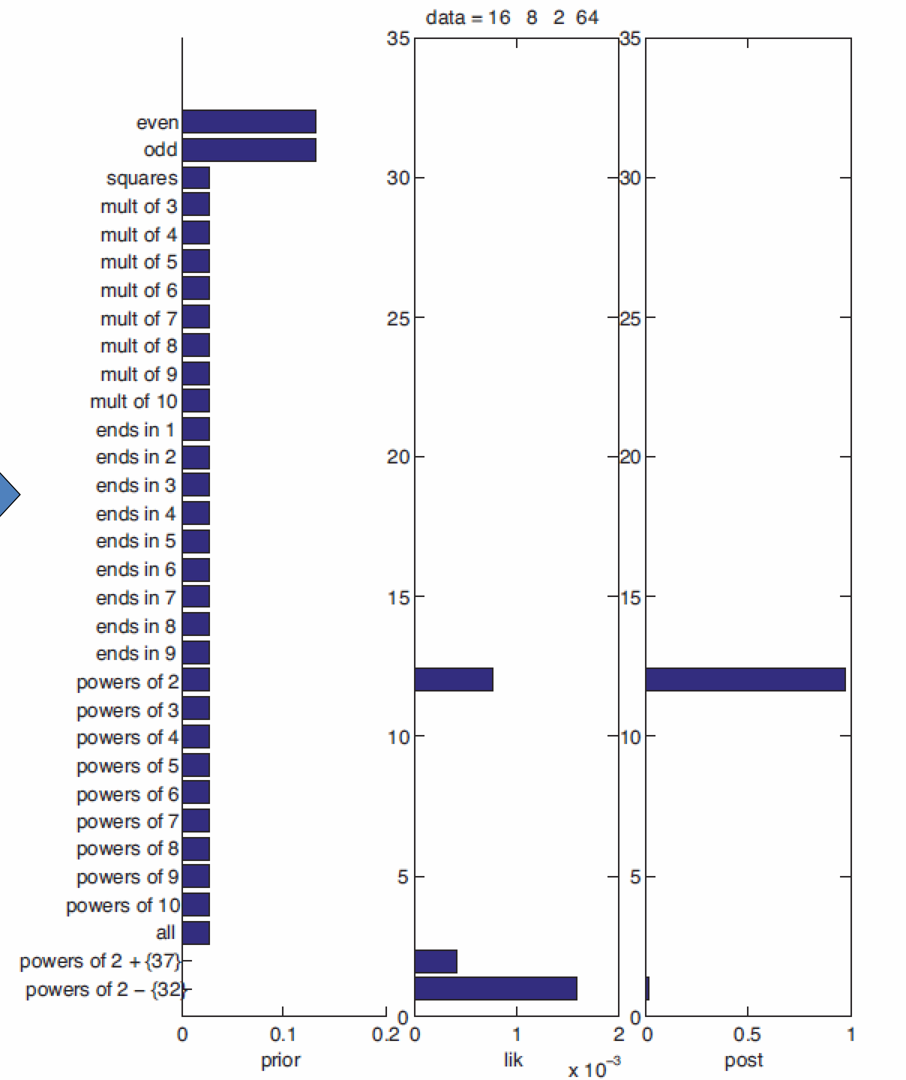
- Posterior = belief about the world



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

Posterior





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

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

Posterior Predictive Distribution

$$p(\tilde{x}|\mathcal{D}) = \sum_h p(\tilde{x}, h|\mathcal{D}) = \sum_h p(\tilde{x}|h, \mathcal{D}) p(h|\mathcal{D})$$
$$= \sum_h p(\tilde{x}|h) p(h|\mathcal{D})$$

□ also called **Bayesian model averaging (BMA)**

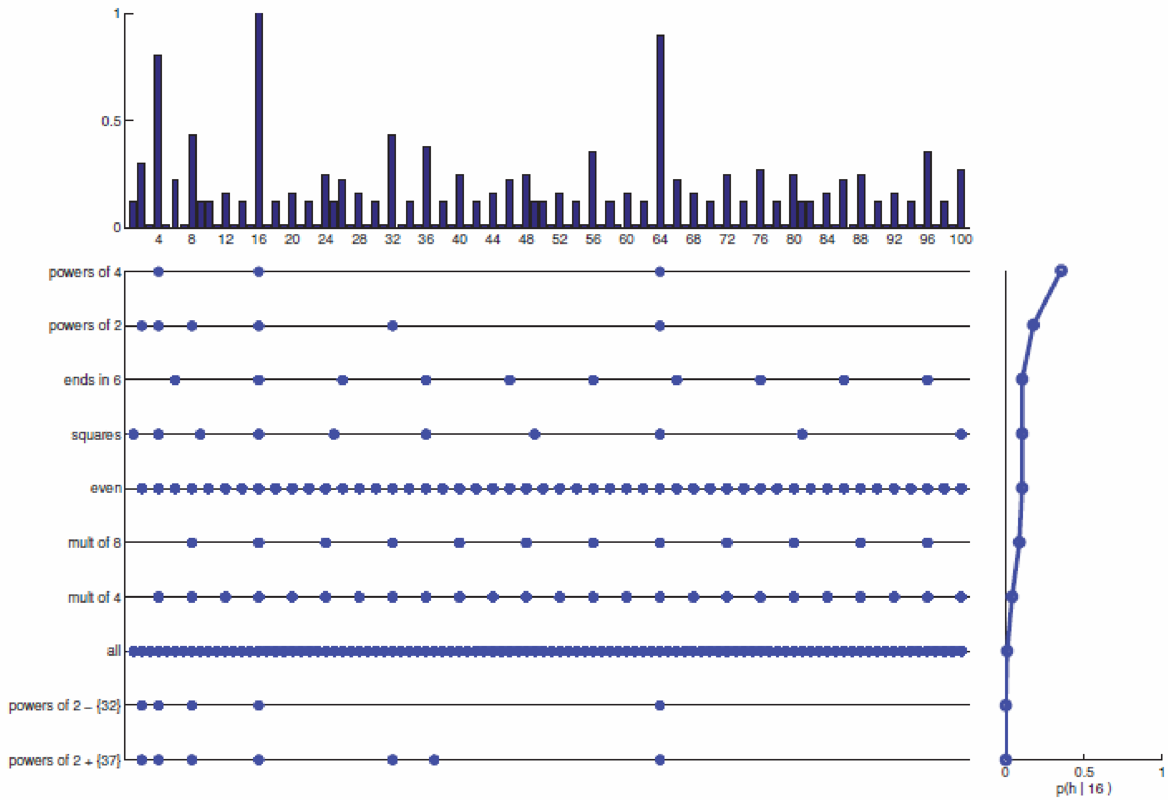


Figure 3.4 Posterior over hypotheses and the corresponding predictive distribution after seeing one example, $\mathcal{D} = \{16\}$. A dot means this number is consistent with this hypothesis. The graph $p(h|\mathcal{D})$ on the right is the weight given to hypothesis h. By taking a weighed sum of dots, we get $p(\tilde{x} \in C|\mathcal{D})$ (top).

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Posterior Predictive Distribution

$$p(\tilde{x}|\mathcal{D}) = \sum_h p(\tilde{x}|h)p(h|\mathcal{D})$$
$$\approx \sum_h p(\tilde{x}|h)\mathbb{I}(h = \hat{h}) = p(\tilde{x}|\hat{h})$$

□ Plug-in approximation

- Maximum-A-Posteriori (MAP) estimator

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

- Maximum Likelihood (ML) estimator

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

- Bayes estimator for continuous space of hypotheses

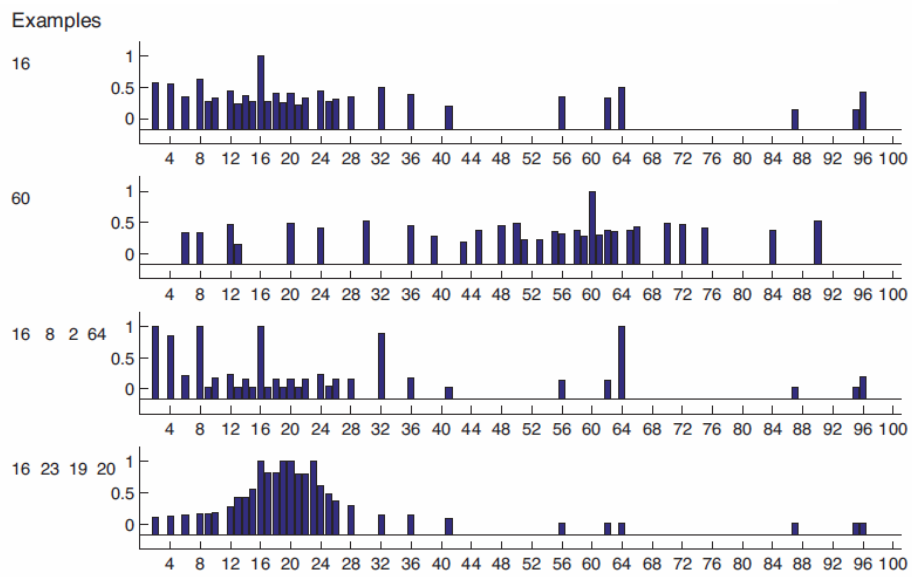
$$\hat{h}_{\text{BAYES}} = \int hp(h|\mathcal{D})dh$$

Posterior Predictive Distribution

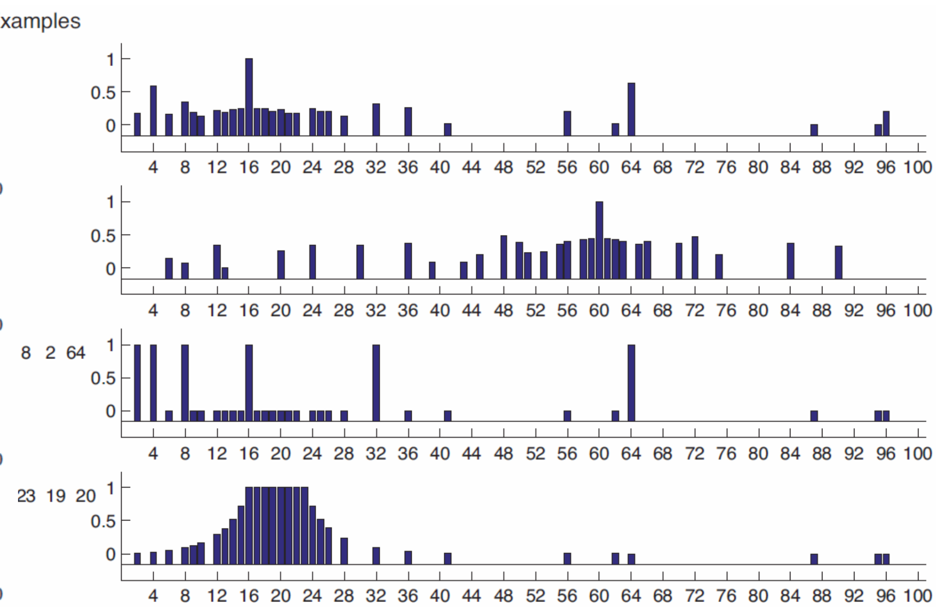
☐ Use a complex prior and fit to the human data

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

$$p_{\text{rule}}(h) = \frac{\mathbb{I}(h \in H_{\text{rules}})}{|H_{\text{rules}}|}, \quad p_{\text{interval}}(h) = \frac{\mathbb{I}(h \in H_{\text{interval}})}{|H_{\text{interval}}|}$$



human data



predictive distribution

Summary: Bayesian Concept Learning

□ Concept learning

- Train the learner to classify objects by showing a set of example objects

- Learn the unknown indicator function f such that

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is an example of concept } C \\ 0 & \text{otherwise} \end{cases}$$

- Binary classification: positive vs. negative examples

- Psychology: human can learn from *only* positive examples
  - We simulated this in the Number Game

Naive Bayes

Beta-Binomial Model

- Figure out the probability of a coin showing heads given a series of observed coin tosses
  - Hypothesis space is continuous!
  - Foundation for naive Bayes classifiers, Markov models, etc.
- Likelihood: two models with the same result
  - i-th outcome $X_i \sim \mathrm{Ber}(\theta)$ with 1=head, 0=tail
  - $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$
    - $p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$ $\quad \begin{cases} N_1 = \sum_{i=1}^N \mathbb{I}(x_i = 1) \\ N_0 = \sum_{i=1}^N \mathbb{I}(x_i = 0) \end{cases}$
  - N_1 and N_0 are sufficient statistics of the data (all we need to know to infer θ)
  - $\mathcal{D} = \{N_1, N_0\}$
    - $p(\mathcal{D}|\theta) = \mathrm{Bin}(N_1|\theta, N_1 + N_0) = \binom{N_1 + N_0}{N_1}\theta^{N_1}(1-\theta)^{N_0}$

Korea Advanced Institute of Science and Technology
한국과학기술원

Beta-Binomial Model

☐ Prior

- Beta distribution: conjugate prior for Bernoulli distribution
  $\mathrm{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$
- a and b are called **hyper-parameters**

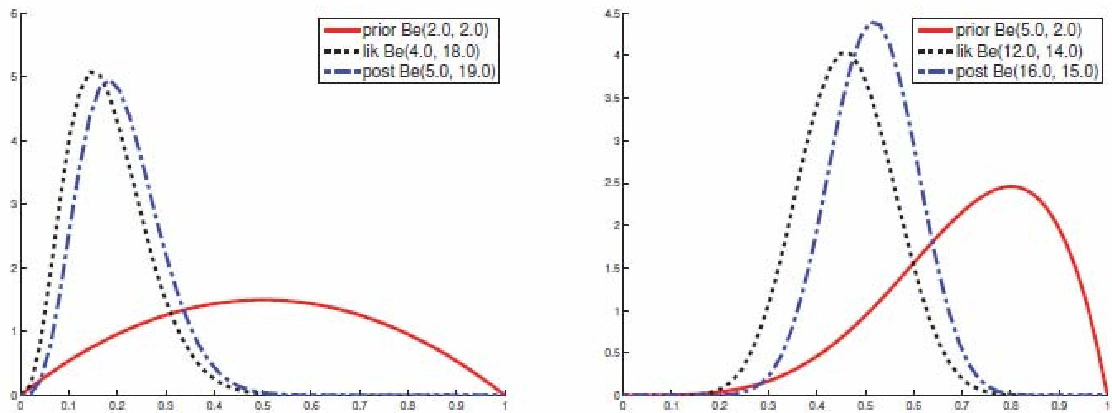
☐ Posterior

- $$p(\theta|\mathcal{D}) \propto \mathrm{Bin}(N_1|\theta, N_0+N_1)\mathrm{Beta}(\theta|a,b)$$
  $$= \mathrm{Beta}(\theta|N_1+a, N_0+b)$$
- hyper-parameters a and b are also called **pseudo-counts**
- **Sequential update** (**online learning**): let $\mathcal{D} = [\mathcal{D}'; \mathcal{D}'']$
  $$\begin{aligned} p(\theta|\mathcal{D}',\mathcal{D}'') &\propto p(\mathcal{D}''|\theta)p(\theta|\mathcal{D}') \\ &= \mathrm{Bin}(N_1''|\theta, N_1''+N_0'')\mathrm{Beta}(\theta|N_1'+a, N_0'+b) \\ &= \mathrm{Beta}(\theta|N_1'+N_1''+a, N_0'+N_0''+b) \\ &= \mathrm{Beta}(\theta|N_1+a, N_0+b) \end{aligned}$$

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Beta-Binomial Model

☐ Posterior distribution examples:



(a) (b)

☐ Posterior mode, mean, and variance

- $\hat{\theta}_{\text{MAP}} = \frac{a+N_1-1}{a+b+N-2}$, $\hat{\theta}_{\text{ML}} = \frac{N_1}{N}$ (uniform prior)
- posterior mean = ?
- $\text{Var}[\theta|\mathcal{D}] = \frac{(a+N_1)(b+N_0)}{(a+N_1+b+N_0)^2(a+N_1+b+N_0+1)} \approx \frac{N_1 N_0}{N N N} = \frac{\hat{\theta}(1-\hat{\theta})}{N}$
- $\sigma = \sqrt{\text{Var}[\theta|\mathcal{D}]} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{N}}$

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Beta-Binomial Model

- Posterior predictive distribution
  - $$p(\tilde{x}=1|\mathcal{D}) = \int_0^1 p(x=1|\theta)p(\theta|\mathcal{D})d\theta$$
    $$= \int_0^1 \theta \mathrm{Beta}(\theta|N_1+a, N_0+b)d\theta$$
    $$= E[\theta|\mathcal{D}] = \frac{a+N_1}{a+b+N}$$
  - i.e. $p(\tilde{x}|\mathcal{D}) = \mathrm{Ber}(\tilde{x}|E[\theta|\mathcal{D}])$
  - Coincides with **add-one smoothing** when uniform prior is used
    $$p(\tilde{x}=1|\mathcal{D}) = \frac{N_1+1}{N+2}$$

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Dirichlet-Multinomial Model

- Generalization of Beta-Binomial model
  - More than two outcomes, e.g. dice rolls $x_i \in \{1, \ldots, 6\}$

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

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

- Posterior:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &\propto p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= \prod_k \theta_k^{N_k} \theta_k^{\alpha_k - 1} = \prod_k \theta_k^{\alpha_k + N_k - 1} \\ &= \mathrm{Dir}(\boldsymbol{\theta}|\alpha_1 + N_1, \ldots, \alpha_K + N_K) \end{aligned}$$

Naive Bayes Classifiers (NBC)

- Classify vectors $\mathbf{x}$ into class $c \in \{1, \ldots, C\}$
- Assume **conditionally independent** features

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

- All binary features: $p(\mathbf{x}|Y = c, \boldsymbol{\Theta}) = \prod_{j=1}^{D} \mathrm{Ber}(x_j|\theta_{jc})$
- All categorical features: $p(\mathbf{x}|Y = c, \boldsymbol{\Theta}) = \prod_{j=1}^{D} \mathrm{Cat}(x_j|\theta_{jc})$
- All real-valued features: $p(\mathbf{x}|Y = c, \boldsymbol{\Theta}) = \prod_{j=1}^{D} \mathcal{N}(x_j|\mu_{jc}, \sigma_{jc}^2)$

- Various mix and match possible
  - e.g. student = [gender, weight, height] i.e. some features categorical, others real-valued

Training NBC

- Usually computing MLE or MAP estimate for $\boldsymbol{\Theta}$
- MLE:

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

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

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

- Suppose binary features: $x_j|y = c \sim \mathrm{Ber}(\theta_{jc})$

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

Bayesian NBC

- Prior

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

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha})$$

$$p(\theta_{jc}) = \mathrm{Beta}(\beta_1, \beta_0)$$

- Posterior

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

$$p(\boldsymbol{\pi} | \mathcal{D}) = \mathrm{Dir}(N_1 + \alpha_1, \ldots, N_C + \alpha_C)$$

$$p(\theta_{jc} | \mathcal{D}) = \mathrm{Beta}(N_{jc} + \beta_1, N_c - N_{jc} + \beta_0)$$

Prediction with Bayesian NBC

$$\begin{aligned}
p(y=c|\mathbf{x},\mathcal{D}) &= \int p(y=c|\mathbf{x},\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\mathcal{D})d\boldsymbol{\Theta} \\
&\propto \int p(y=c|\boldsymbol{\Theta})p(\mathbf{x}|y=c,\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\mathcal{D})d\boldsymbol{\Theta} \\
&= \left[\int \mathrm{Cat}(y=c|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathcal{D})d\boldsymbol{\pi}\right] \\
&\qquad \prod_j \left[\int \mathrm{Ber}(x_j|y=c,\theta_{jc})p(\theta_{jc}|\mathcal{D})d\theta_{jc}\right] \\
&= \bar{\pi}_c \prod_j (\bar{\theta}_{jc})^{\mathbb{I}(x_j=1)}(1-\bar{\theta}_{jc})^{\mathbb{I}(x_j=0)}
\end{aligned}$$

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

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

□ $\hat{\theta}_{\mathrm{MAP}}$? $\hat{\theta}_{\mathrm{ML}}$?

Summary: Generative Classifiers

- Use Bayes rule to classify feature vector $\mathbf{x}$ of any type

$$p(C=c|\vec{x}) = \frac{p(C=c)p(\vec{x}|C=c)}{\sum_{c'} p(C=c')p(\vec{x}|C=c')}$$

- **Class prior** $p(C)$
- **Class-conditional density** $p(\vec{x}|C)$
  - models how the data is *generated*

- vs. discriminative classifier
  - directly fit $p(C|\mathbf{x})$ from data