Generative models for discrete data

Xiaosong Jia

07/17/2018

Outline

- Bayesian concept learning
- The beta-binomial model
- The Dirichlet-multinomial model
- Naive Bayes classifier

Introduction

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(y = c|\boldsymbol{\theta})p(\mathbf{x}|y = c, \boldsymbol{\theta})}{\sum_{c'} p(y = c'|\boldsymbol{\theta})p(\mathbf{x}|y = c', \boldsymbol{\theta})}$$

- This is called a **generative classifier**, since it specifies how to generate the data using the **class conditional density** p(**x**|y = c) and the class prior p(y = c).

- Posterior = $\frac{prior * likelihood}{evidence}$

Bayesian concept learning

- Goal: to learn the indicator function f, which just defines which elements are in the set C.
- Example: **the number game**

1. choose some simple arithmetical concept C

2. give a series of randomly chosen positive examples D = $\{x_1, . . . , x_N\}$ drawn from C

3. ask whether some new test case x' belongs to C ?
eg: p(x'|D)? which is the probability that $\mathrm{x}' \in$ C given the data D for any $\mathrm{x}' \in \{1, . . . , 100\}$

Likelihood

- Strong sampling assumption:

Examples are sampled uniformly at random from the extension of a concept.

(eg: the extension of h_{even} is {2, 4, 6, . . . , 98, 100})

$$p(\mathcal{D}|h) = \left[\frac{1}{\text{size}(h)}\right]^N = \left[\frac{1}{|h|}\right]^N$$ N is the size of D (**Occam's razor)**

eg: for D={16} p(D|$h_{PowerOfTwo}$) = $\frac{1}{6}$ p(D|h_{even}) $= \frac{1}{50}$

Prior

- Suppose D = {16, 8, 2, 64}. Given this data, the concept $\mathbf{h}'$ = “powers of two except 32” is more likely than h =“powers of two”
- conceptually unnatural！
- We can capture such intuition by assigning low prior probability to unnatural concepts. （subjective）

Posterior

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

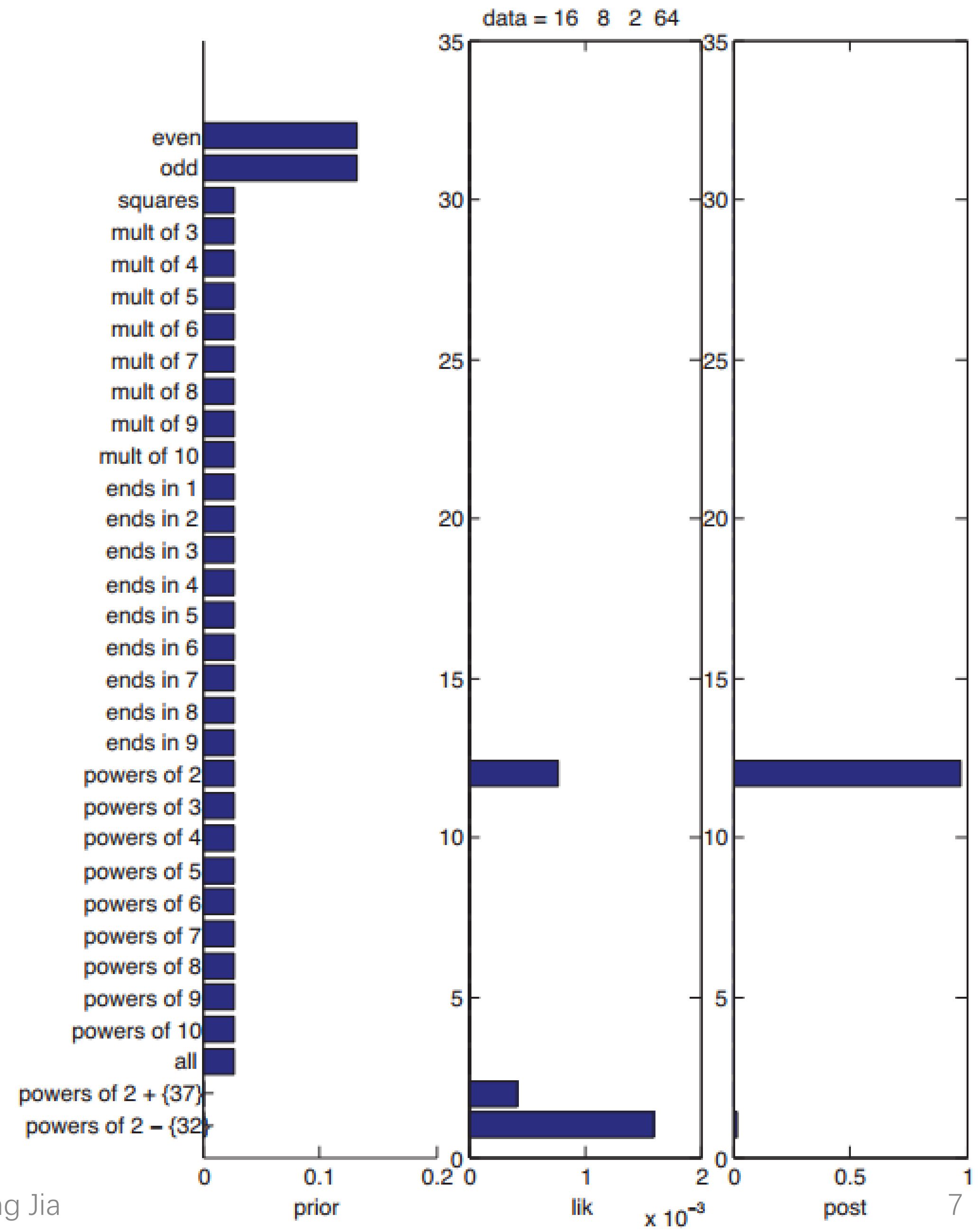
- When we have enough data, the posterior p(h|D) becomes peaked on a single concept, namely the MAP(Maximum A Posteriori Estimation) estimate.

$$p(h|\mathcal{D}) \rightarrow \delta_{\hat{h}^{MAP}}(h)$$

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

- 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 MLE(Maximum Likelihood Estimation)

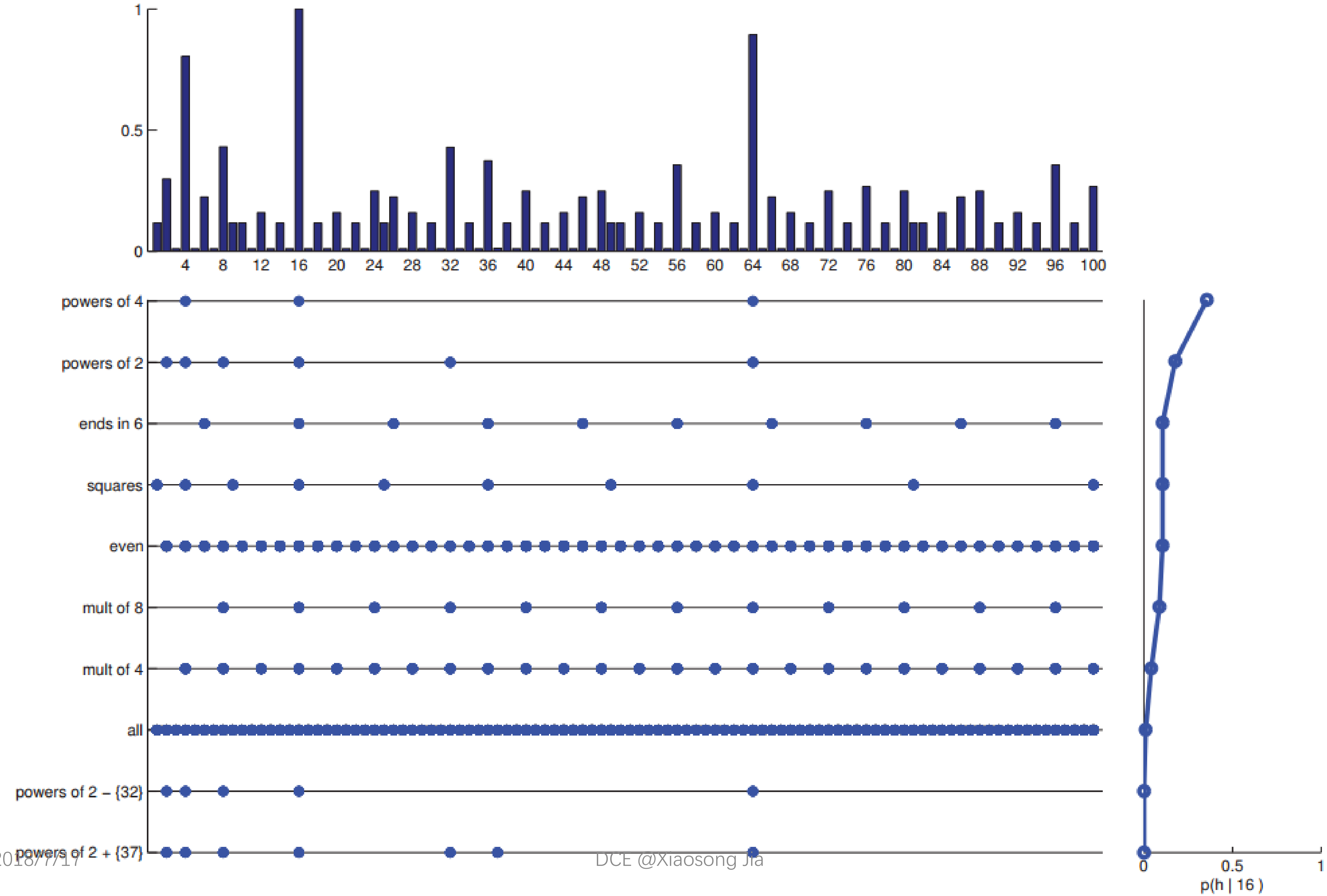
$$\hat{h}^{mle} \triangleq \underset{h}{\operatorname{argmax}}\, p(\mathcal{D}|h) = \underset{h}{\operatorname{argmax}} \log p(\mathcal{D}|h)$$



Posterior predictive distribution

- **Bayes model averaging**

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



1
0.5
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
64
68
72
76
80
84
88
92
96
100
powers of 4
powers of 2
ends in 6
squares
even
mult of 8
mult of 4
all
powers of 2 − {32}
powers of 2 + {37}
0
0.5
1
p(h | 16)

The beta-binomial model

- the unknown parameters are continuous
- Example:

the problem of inferring the probability that a coin shows up heads, given a series of observed coin tosses.

Likelihood

- Suppose $X_i \sim$ Ber(θ), where X_i = 1 represents "heads", X_i = 0 represents "tails", and θ ∈ [0, 1] is the rate parameter (probability of heads).

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

Prior

- To make the math easier, it would convenient if the prior had the same form as the likelihood.

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

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

- When a=1 and b=1, it's called a uniform prior.

Posterior

- p(θ|D) ∝ Bin(N1|θ, N0 + N1)Beta(θ|a, b) ∝ Beta(θ|N1 + a, N0 + b)
- When the prior and the posterior have the same form, we say that the prior is a **conjugate prior** for the corresponding likelihood.
- MAP MLE

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

- Mean

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

- The strength of the prior, also known as the **effective sample size** of the prior, is the sum of the pseudo counts a + b;
- Variance

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

Posterior predictive distribution

$$
\begin{aligned}
p(\tilde{x} = 1|\mathcal{D}) &= \int_0^1 p(x = 1|\theta)p(\theta|\mathcal{D})d\theta \\
&= \int_0^1 \theta \ \mathrm{Beta}(\theta|a, b)d\theta = \mathbb{E}\left[\theta|\mathcal{D}\right]
\end{aligned}
$$

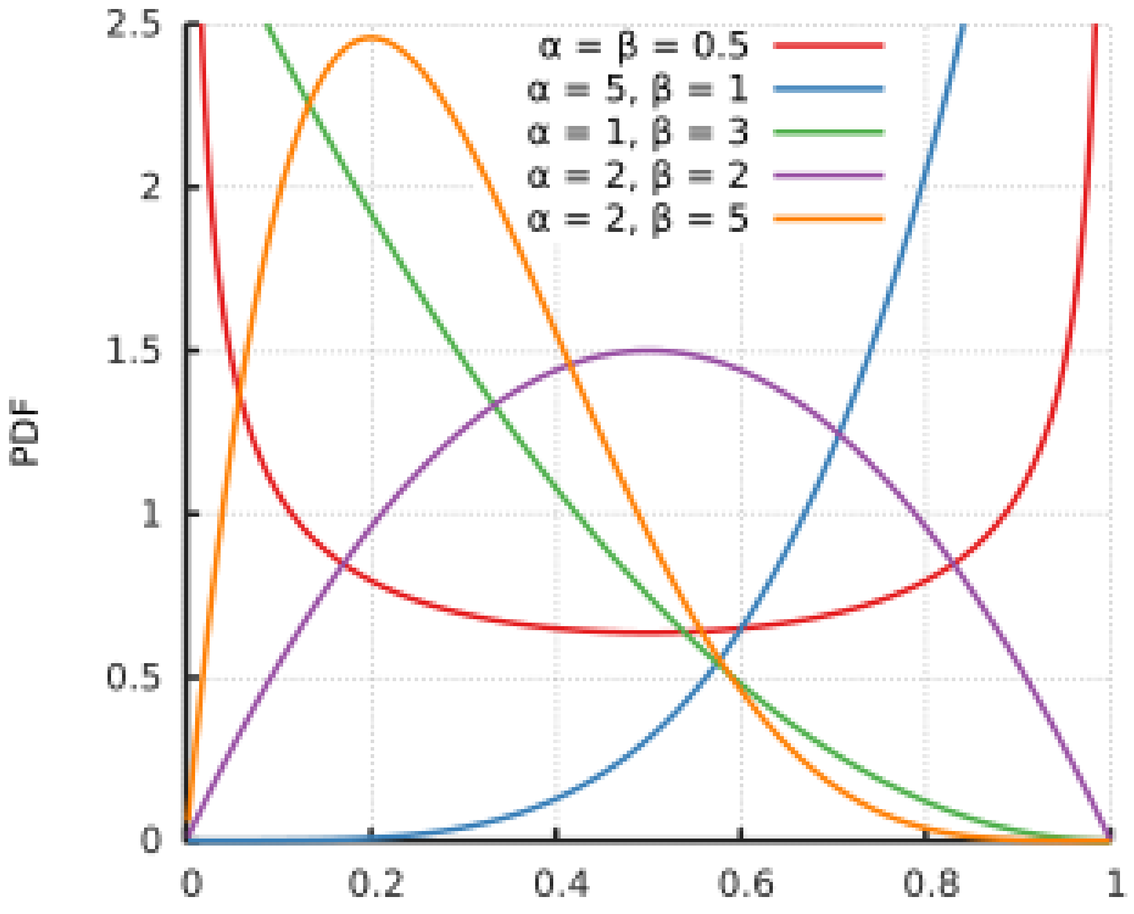
- **Predicting the outcome of multiple future trials**
  Suppose now we were interested in predicting the number of heads, x, in M future trials

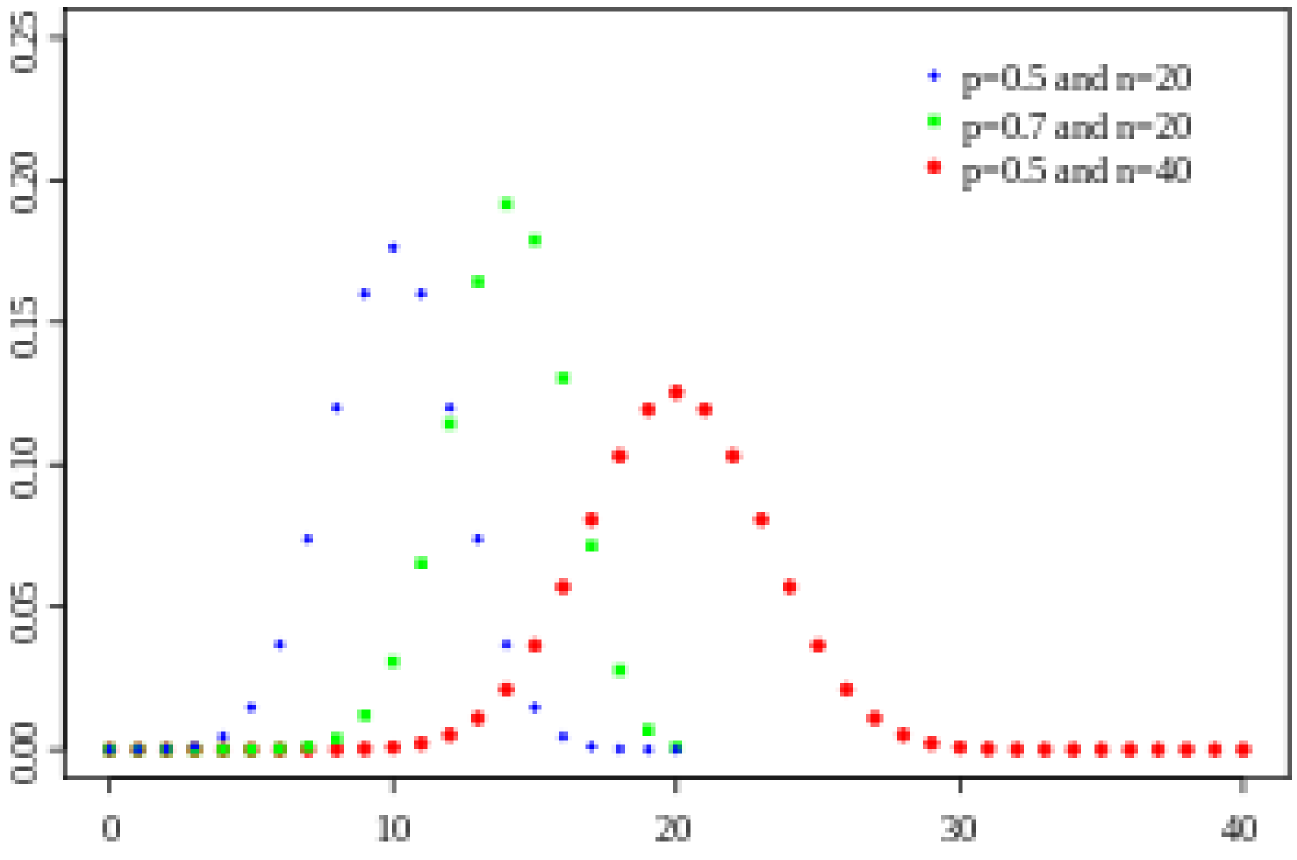
beta-binomial distribution

$$
Bb(x|a, b, M) \triangleq \binom{M}{x} \frac{B(x + a, M - x + b)}{B(a, b)}
$$

$$
\mathbb{E}[x] = M\frac{a}{a + b}, \ \mathrm{var}[x] = \frac{Mab}{(a + b)^2} \frac{(a + b + M)}{a + b + 1}
$$

Probability Density Graph





Overfitting and the black swan paradox

- black swan paradox
- **zero count problem** or the **sparse data problem**
- Laplace’s rule of succession(uniform prior)

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

The Dirichlet-multinomial model

- Infer the probability that a dice with *K* sides comes up as face *k*.
- Likelihood

Suppose we observe N dice rolls, D = $\{x_1, \ldots, x_N\}$, where $x_i \in \{1, \ldots, K\}$.

If we assume the data is iid, the likelihood has the form

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

- Prior(conjugate: Dirichlet distribution)

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

Posterior and Posterior predictive

- Posterior

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

- MAP

MLE

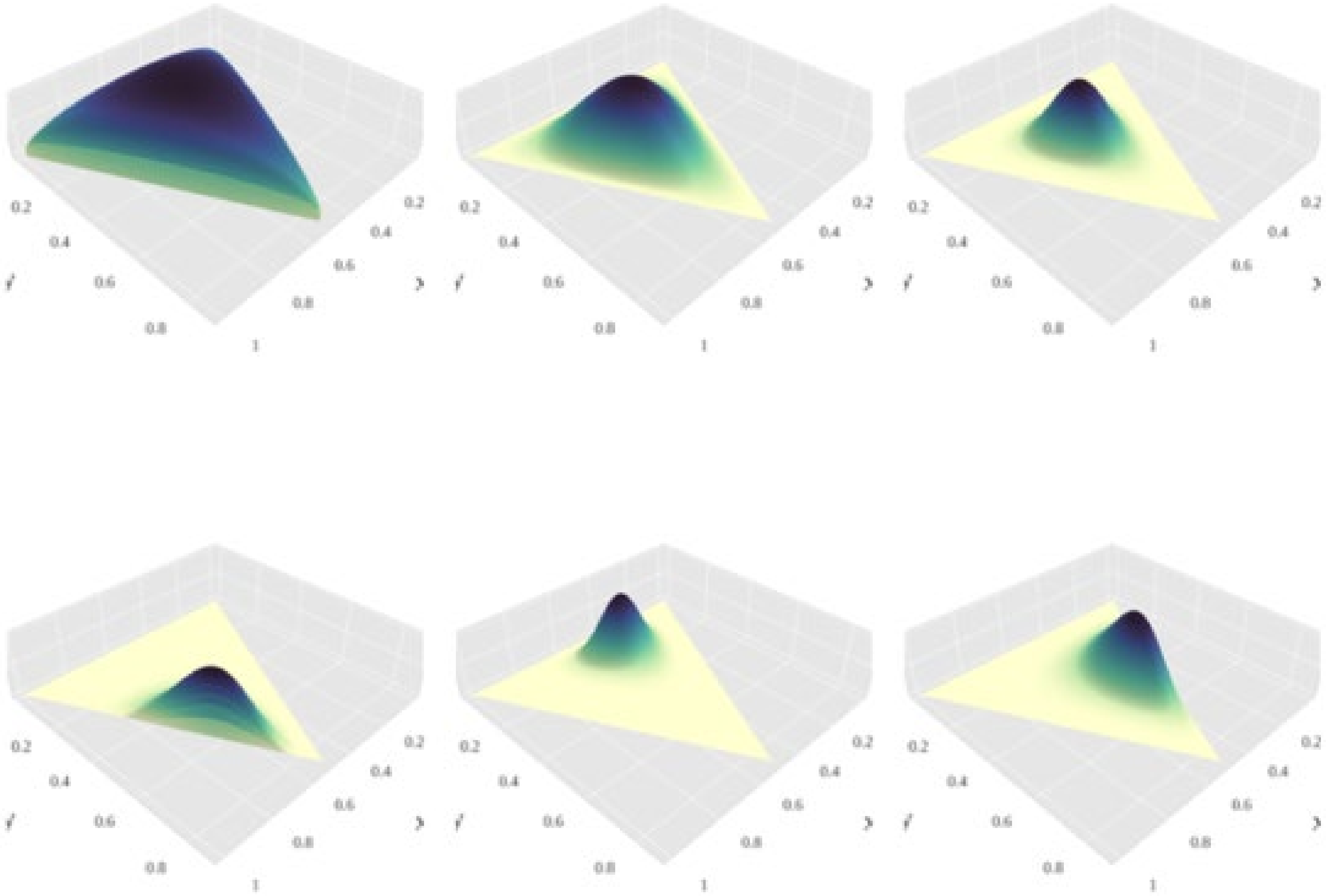
$$\hat{\theta}_{MAP} = \frac{(N_k + \alpha_k - 1)}{N + \alpha_0 - K} \qquad \hat{\theta}_{MLE} = \frac{N_k}{N}$$

- 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 $\alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$ is the equivalent sample size of the prior.

Probability Density Graph



Naive Bayes classifiers

- Goal: classify vectors of discrete-valued features, $\mathbf{x} \in \{1, \ldots, K\}^D$, where K is the number of values for each feature, and D is the number of features.
- Assumption:

the features are **conditionally independent** given the class label.

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

Naive Bayes classifiers

- In the case of real-valued features, we can use the Gaussian distribution: $p(\mathbf{x}|y=c,\boldsymbol{\theta}) = \prod_{j=1}^{D} \mathcal{N}(x_j|\mu_{jc},\sigma_{jc}^2)$, where μ_{jc} is the mean of feature j in objects of class c, and σ_{jc}^2 is its variance.
- In the case of binary features, $x_j \in \{0,1\}$, we can use the Bernoulli distribution: $p(\mathbf{x}|y=c,\boldsymbol{\theta}) = \prod_{j=1}^{D} \mathrm{Ber}(x_j|\mu_{jc})$, where μ_{jc} is the probability that feature j occurs in class c. This is sometimes called the **multivariate Bernoulli naive Bayes** model. We will see an application of this below.

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

Model fitting

- How to train?
- The probability for a single data(Using assumption)

$$p(\mathbf{x}_i, y_i|\boldsymbol{\theta}) = p(y_i|\boldsymbol{\pi}) \prod_j p(x_{ij}|\boldsymbol{\theta}_j)$$

- the log-likelihood

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

Method: let us suppose all features are binary

$$\hat{\pi}_c = \frac{N_c}{N} \qquad \hat{\theta}_{jc} = \frac{N_{jc}}{N_c}$$

Example

- Data Set

帅？	性格好？	身高？	上进？	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
不帅	不好	矮	不上进	不嫁
帅	好	高	不上进	嫁
不帅	好	高	上进	嫁
帅	好	高	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

Example

- 现在给我们的问题是，如果一对男女朋友，男生向女生求婚，男生的四个特点分别是不帅，性格不好，身高矮，不上进，请你判断一下女生是嫁还是不嫁？
- 转换为数学语言
- p(嫁|(不帅、性格不好、身高矮、不上进))与p(不嫁|(不帅、性格不好、身高矮、不上进))哪个大？

Example

- 朴素贝叶斯公式

$$p\left(嫁\,|\,不帅、性格不好、身高矮、不上进\right)=\frac{p\left(不帅、性格不好、身高矮、不上进|嫁\right)*\mathrm{p}(嫁)}{\mathrm{p}(不帅、性格不好、身高矮、不上进)}$$

$$=\frac{\mathrm{p}(不帅\,|\,嫁)*\mathrm{p}(性格不好|嫁)*\mathrm{p}(身高矮|嫁)*\mathrm{p}(不上进|嫁)*\mathrm{p}(嫁)}{\mathrm{p}(不帅)*\mathrm{p}(性格不好)*\mathrm{p}(身高矮)*\mathrm{p}(不上进)}$$

p(嫁) = 6/12（总样本数） = 1/2

p(不嫁) = 6/12（总样本数） = 1/2

Example

帅？	性格好？	身高？	上进？	嫁与否
不帅	好	高	上进	嫁
不帅	好	中	上进	嫁
不帅	不好	高	上进	嫁

p(不帅|嫁) = 3/6 = 1/2

……

Example

帅？	性格好？	身高？	上进？	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
帅	不好	矮	上进	不嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	好	中	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

p（不帅） = 4/12 = 1/3

……

Example

$$p\left(\text{嫁}\,|\,\text{不帅、性格不好、身高矮、不上进}\right)=\frac{p\left(\text{不帅、性格不好、身高矮、不上进}|\text{嫁}\right)*\mathrm{p}(}{\mathrm{p}(\text{不帅、性格不好、身高矮、不上进})}$$

$$=\frac{\mathrm{p}(\text{不帅}\,|\,\text{嫁})*\mathrm{p}(\text{性格不好}|\text{嫁})*\mathrm{p}(\text{身高矮}|\text{嫁})*\mathrm{p}(\text{不上进}|\text{嫁})*\mathrm{p}(\text{嫁})}{\mathrm{p}(\text{不帅})*\mathrm{p}(\text{性格不好})*\mathrm{p}(\text{身高矮})*\mathrm{p}(\text{不上进})}$$

= (1/2*1/6*1/6*1/6*1/2)/(1/3*1/3*7/12*1/3)

同理可得 p (不嫁|不帅、性格不好、身高矮、不上进)

Bayesian naive Bayes

- Above method： it can **overfit**.
- Solution： use a factored prior

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

We will use a Dir(α) prior for π and a Beta(β_0, β_1) prior for each θjc. Often we just take α = 1 and β = 1, corresponding to add-one or Laplace smoothing

Bayesian naive Bayes

- Posterior

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

- Using the model for prediction

$$
\begin{aligned}
p(y = c|\mathbf{x}, \mathcal{D}) &\propto \overline{\pi}_c \prod_{j=1}^{D} (\overline{\theta}_{jc})^{\mathbb{I}(x_j=1)} (1 - \overline{\theta}_{jc})^{\mathbb{I}(x_j=0)} \\
\overline{\theta}_{jk} &= \frac{N_{jc} + \beta_1}{N_c + \beta_0 + \beta_1} \\
\overline{\pi}_c &= \frac{N_c + \alpha_c}{N + \alpha_0}
\end{aligned}
$$

where $\alpha_0 = \sum_c \alpha_c$.

Document classification

- Goal: classify text documents into different categories
- let x_i be a vector of counts for document i, so $x_{ij} \in \{0, 1, . . . , N_i\}$, where N_i is the number of terms in document i (so $\sum_{j=1}^{D} x_{ij} = N_i$).
- For the class conditional densities, we can use a multinomial distribution

$$p(\mathbf{x}_i|y_i = c, \boldsymbol{\theta}) = \mathrm{Mu}(\mathbf{x}_i|N_i, \boldsymbol{\theta}_c) = \frac{N_i!}{\prod_{j=1}^{D} x_{ij}!} \prod_{j=1}^{D} \theta_{jc}^{x_{ij}}$$

Here θ_{jc} is the probability of generating word j in documents of class c; these parameters satisfythe constraint that $\sum_{j=1}^{D} \theta_{jc}$ =1 for each class c.

Quiz

- Calculate p (不嫁|不帅、性格不好、身高矮、不上进) with naïve Bayesian Classifier