Royal Institute of Technology

DD2447 STAT. METH. IN CS FALL 2014

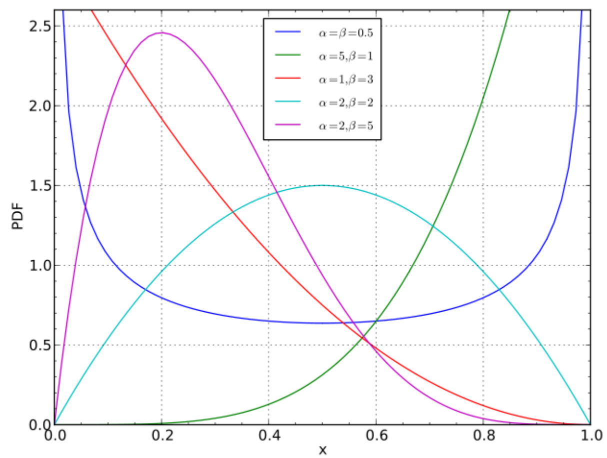
★ Lecture 3 - Bayesian

★ Chap. 3

BETA DISTRIBUTION

- PDF $\mathrm{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{(a-1)} x^{(b-1)}$
  - where $B(a,b) = \frac{\Gamma(a)\Gamma(a)}{\Gamma(a+b)}$
  - and $\Gamma(a) = (a-1)\Gamma(a-1)$
- In particular, for integer n

$$\Gamma(n) = (n-1)!$$



BETA DISTRIBUTION

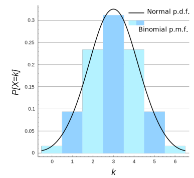
BIASED DICE - FIRST COIN

★ Assumption: we don't know the outcome probabilities

★ We need a prior over the outcome probabilities

★ First likelihood

- ★ N_1 heads and N_0 tails
- ★ p(head) = θ
- ★ sequence $p(D) = \theta^{N_1}(1-\theta)^{N_0}$ iid Bernoulli
- ★ counts $p(D) = \binom{N_1+N_0}{N_1}\theta^{N_1}(1-\theta)^{N_0}$ Binomial



BACK TO BAYES - PRIOR FOR BINOMIAL

- Beta distribution up to a constant $p(\theta|\gamma_1, \gamma_0) = \theta^{\gamma_1 - 1}(1-\theta)^{\gamma_0 - 1}$
- Posterior $$\begin{aligned} p(\theta|D) &\propto p(D|\theta)p(\theta|\gamma_1,\gamma_0) \\ &= \theta^{N_1}(1-\theta)^{N_0}\theta^{\gamma_1-1}(1-\theta)^{\gamma_0-1} \\ &= \theta^{N_1+\gamma_1-1}(1-\theta)^{N_0+\gamma_0-1} \\ &= \mathrm{Beta}(\theta|N_1+\gamma_1, N_0+\gamma_0) \end{aligned}$$
- Prior that gives posterior of the same sort is called conjunctive
- Beta is a conjunctive prior for Binomial

Uniform prior, i.e., $\gamma_1=\gamma_0=1$, gives

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

Black Swan “paradox”

LAPLACE’S RULE OF SUCCESSION

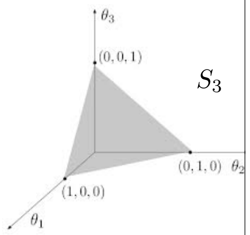
$D = \{x_1, \ldots, x_N\}$ where $x_i \in \{1, \ldots, K\}$

$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in S_K$ the K-dim. probability simplex, i.e., $\sum_{i \in [K]} \theta_k = 1$

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ hyperparameters

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

Likelihood $p(\boldsymbol{\theta}|D) \propto \prod_{k \in [K]} \theta_k^{N_k}$



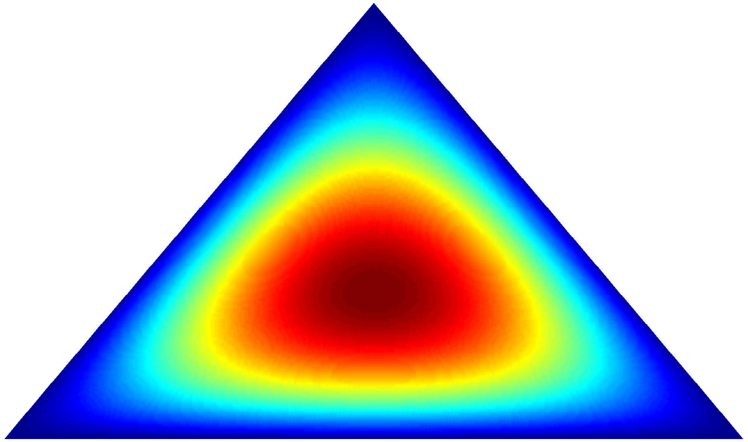
BACK TO THE DICE – DIRICHLET-MULTINOMIAL

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

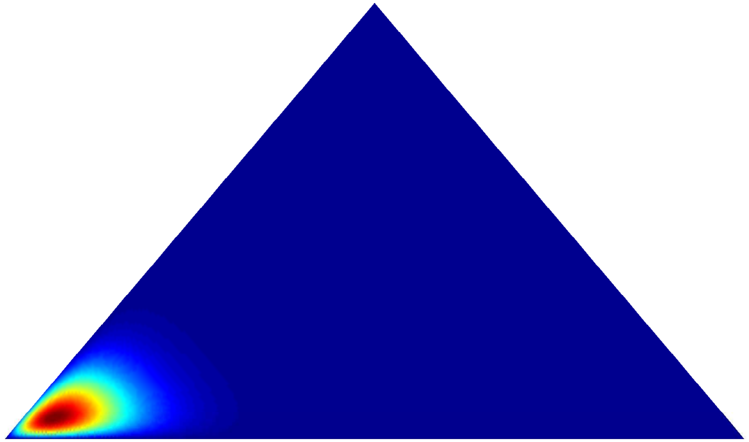
Likelihood $p(\boldsymbol{\theta}|D) \propto \prod_{k \in [K]} \theta_k^{N_k}$

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

POSTERIOR FOR DIRICHLET-MULTINOMIAL

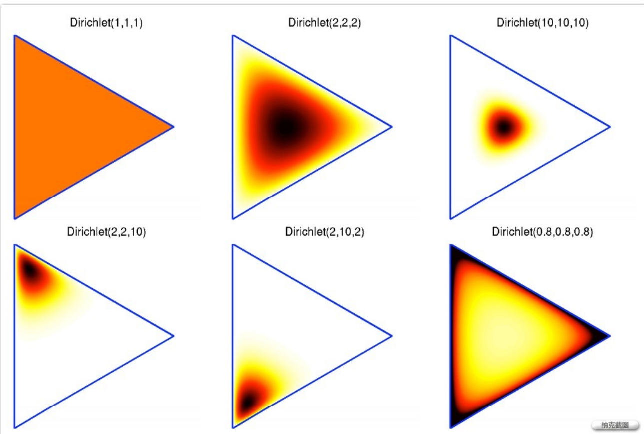


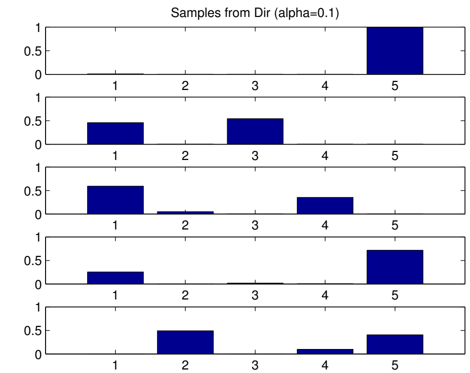
$\alpha = (2, 2, 2)$



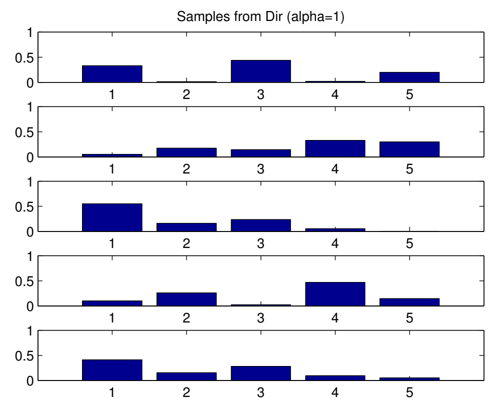
$\alpha = (20, 2, 2)$

ANOTHER COLOR CODING





SAMPLES $\alpha = (0.1, 0.1,0.1,0,1\ 0.1)$



SAMPLES $\alpha = (1,1,1,1,1)$

Probability that next is j, posterior to D

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

POSTERIOR PREDICTIVE CATEGORICAL-DIRICHLET

Text:

```
Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow
```

Vocabulary:

```
mary lamb little big fleece white black snow rain unk
1    2    3      4   5      6     7     8    9    10
```

Index occurrences:

1 10 3 2 3 2 3 2
1 10 3 2 10 5 10 6 8

Counts:

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

BAG OF WORDS

Counts:

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

Posterior predictive: $p(X = j|D) = E[\theta_j|D] = \dfrac{N_j + \alpha_j}{N + \alpha}$

BAG OF WORDS

Counts:

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

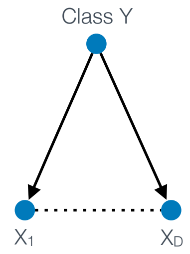
Posterior predictive: $p(X=j|D) = E[\theta_j|D] = \dfrac{N_j+\alpha_j}{N+\alpha} = \dfrac{N_j+1}{17+10}$

Posterior predictive:

If we set $\alpha_j = 1$, we get

$$(3/27, 5/27, 5/27, 1/27, 2/27, 2/27, 1/27, 2/27, 1/27, 5/27)$$

BAG OF WORDS



NAIVE BAYES CLASSIFIER – NBC

- ★ Distribution over class
- ★ Given class X_i:s independent

Datapoint $\boldsymbol{x} \in [K]^D$ Classes $[C]$

Class conditional independent $p(\boldsymbol{x}|y=c,\boldsymbol{\theta}) = \prod_{d=1}^{D} p(x_d|y=c,\boldsymbol{\theta}_{dc})$

$p(x_d|y=c,\boldsymbol{\theta}_{dc})$ is (now)

- Categorical, so $\boldsymbol{\theta}_{dc}$ probabilities of each outcome in [K]

but can also be

- Bernoulli, so $\boldsymbol{\theta}_{dc}$ probability of head
- Or x real valued and gaussian dist, so $\boldsymbol{\theta}_{dc}$ gives mean and variance

NAIVE BAYES CLASSIFIER – NBC

Data $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ Counts N, N_c, N_{dc}, N_{dck}

Likelihood

$$p(\boldsymbol{x}_n, y_n|\boldsymbol{\theta},\boldsymbol{\pi}) = p(y_n|\boldsymbol{\pi}) \prod_d p(\boldsymbol{x}_{nd}|\boldsymbol{\theta}_{y_n}) = \pi_{y_n} \prod_d \theta_{d y_n x_d}$$

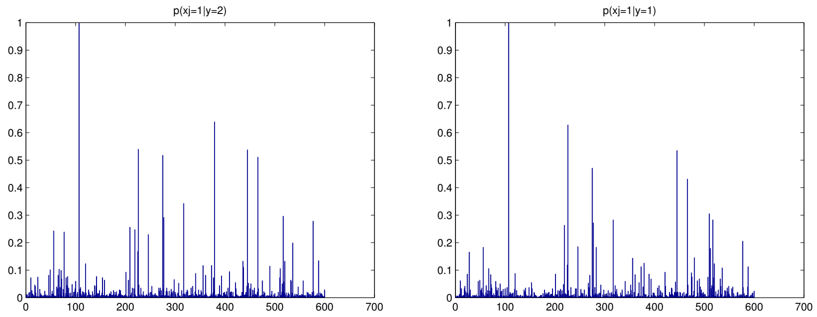
$$p(D|\boldsymbol{\pi},\boldsymbol{\theta}) = \prod_c \pi_c^{N_c} \prod_d \prod_c p(x=k|\boldsymbol{\theta}_{dc})^{N_{dck}}$$

Log-likelihood

$$\log p(D|\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_c N_c \log \pi_c + \sum_c \sum_d \left(\sum_k N_{dck} \log p(x=k|\boldsymbol{\theta}_{dc}) \right)$$

Optimized by $\hat{\pi}_c = N_c/N$ and $\hat{\theta}_{dck} = N_{dck}/N_{dc}$

TRAINING AN NBC



CLASS CONDITIONALS

BAYESIAN NAIVE BAYES CLASSIFIER

Prior (Dirichlet on all, perhaps add one)

$$p(\boldsymbol{\pi}, \boldsymbol{\theta}) = p(\boldsymbol{\alpha}) \prod_c \prod_d p(\boldsymbol{\theta}_{dc}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_c \prod_d \mathrm{Dir}(\boldsymbol{\theta}_{dc}|\boldsymbol{\beta})$$

(Recall) likelihood

$$p(D|\boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_c \pi_c^{N_c} \prod_d \prod_c \left(\prod_k p(x = k|\boldsymbol{\theta}_{dc})^{N_{dck}} \right)$$

Posterior

$$\begin{aligned} p(\boldsymbol{\pi}, \boldsymbol{\theta}|D) &= \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_c \pi_c^{N_c} \prod_c \prod_d \left(\mathrm{Dir}(\boldsymbol{\theta}_{dc}|\boldsymbol{\beta}) \prod_k p(x = k|\boldsymbol{\theta}_{dc})^{N_{dck}} \right) \\ &= \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{N} + \boldsymbol{\alpha}) \prod_c \prod_d \mathrm{Dir}(\boldsymbol{N}_{cd} + \boldsymbol{\beta}) \end{aligned}$$

What is the class, for unclassified **x**

$$p(y = c|\boldsymbol{x}, D) \propto p(y = c, \boldsymbol{x}|D)$$

Bayesan: integrate out the parameters

$$\begin{aligned} p(y = c, \boldsymbol{x}|D) &= \int_{\boldsymbol{\pi},\boldsymbol{\theta}} p(y = c, \boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\theta}|D) \mathrm{d}(\boldsymbol{\pi}, \boldsymbol{\theta}) \\ &= \int_{\boldsymbol{\pi},\boldsymbol{\theta}} p(y = c, \boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\theta}) p(\boldsymbol{\pi}, \boldsymbol{\theta}|D) \mathrm{d}(\boldsymbol{\pi}, \boldsymbol{\theta}) \\ &= \int_{\boldsymbol{\pi},\boldsymbol{\theta}} p(y = c|\boldsymbol{\pi}) p(\boldsymbol{\pi}|D) p(\boldsymbol{x}|y = c, \boldsymbol{\theta}) p(\boldsymbol{\theta}|D) \mathrm{d}(\boldsymbol{\pi}, \boldsymbol{\theta}) \\ &= \int_{\boldsymbol{\pi}} \mathrm{Cat}(y = c|\boldsymbol{\pi}) p(\boldsymbol{\pi}|D) \mathrm{d}\boldsymbol{\pi} \prod_d \int_{\boldsymbol{\theta}_{dc}} \mathrm{Cat}(x_d|y = c, \boldsymbol{\theta}_{dc}) p(\boldsymbol{\theta}|D) \mathrm{d}\boldsymbol{\theta}_{dc} \end{aligned}$$

PREDICTION – BASED ON DATA D

What is the class, for unclassified **x**

$$p(y = c|\boldsymbol{x}, D) \propto p(y = c, \boldsymbol{x}|D)$$

Bayesan: integrate out the parameters

$$p(y = c, \boldsymbol{x}|D) = \int_{\boldsymbol{\pi}} \mathrm{Cat}(y = c|\boldsymbol{\pi}) p(\boldsymbol{\pi}|D) \mathrm{d}\boldsymbol{\pi} \prod_d \int_{\boldsymbol{\theta}_{dc}} \mathrm{Cat}(x_d|y = c, \boldsymbol{\theta}_{dc}) p(\boldsymbol{\theta}|D) \mathrm{d}\boldsymbol{\theta}_{dc}$$

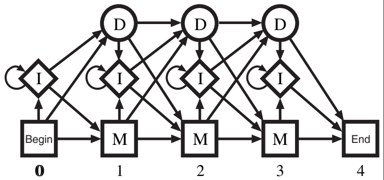
these are posterior means so

$$\int_{\boldsymbol{\pi}} \mathrm{Cat}(y = c|\boldsymbol{\pi}) p(\boldsymbol{\pi}|D) \mathrm{d}\boldsymbol{\pi} = \frac{N_c + \alpha_c}{N + \alpha_0} \qquad \alpha_0 := \sum_{i \geq 1} \alpha_i$$

$$\int_{\boldsymbol{\theta}_{dc}} \mathrm{Cat}(x_d|y = c, \boldsymbol{\theta}_{dc}) p(\boldsymbol{\theta}|D) \mathrm{d}\boldsymbol{\theta}_{dc} = \frac{N_{dc} + \alpha_c}{N_c + \beta_0} \qquad \beta_0 := \sum_{i \geq 1} \beta_i$$

PREDICTION – BASED ON DATA D

LOG-SUM-EXP TRICK



The end