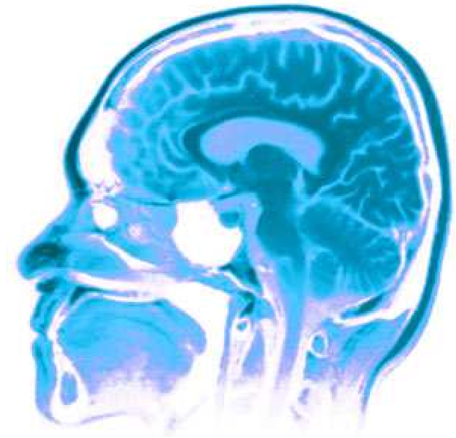
CPSC340



Dirichlet and Categorical variables: Naïve Bayes classifier

Nando de Freitas
November, 2012
University of British Columbia

Outline of the lecture

This lecture introduces the Dirichlet and categorical distributions, as well as the Naïve Bayes classifier. The goal is for you to:

- ❑ Learn categorical distributions.
- ❑ Derive the Dirichlet posterior from the Dirichlet prior and categorical likelihood.
- ❑ Understand how a classifier for text is set up.
- ❑ Understand the Naïve Bayes classifier for text classification.

Revision: Beta-Bernoulli

Suppose $X_i \sim \text{Ber}(\theta)$, so $X_i \in \{0, 1\}$. We know that the likelihood has the form

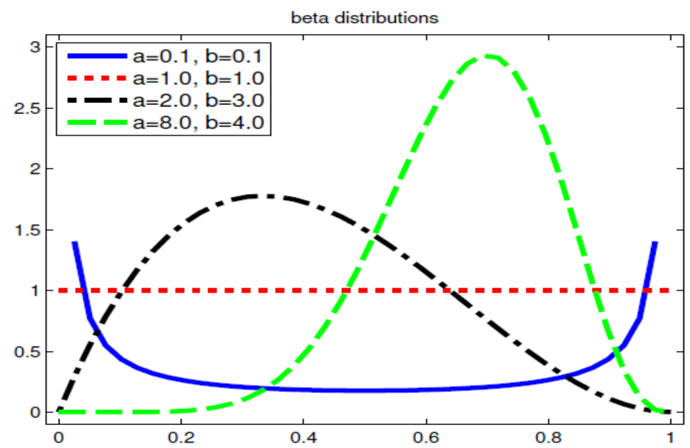
$$\boxed{p(\underset{\sim\, X_{1:N} = \text{data}}{\mathcal{D}}|\theta) = \theta^{N_1}(1-\theta)^{N_0}}$$

where we have $N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1)$ heads and $N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0)$ tails.

The **beta prior** has pdf:

$0 \leq \theta \leq 1$

$$\text{Beta}(\theta|\alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$$



Revision: Beta-Bernoulli

If we multiply the Bernoulli likelihood by the beta prior we get

$$
\begin{aligned}
p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta) \\
&\propto \overbrace{[\theta^{N_1}(1-\theta)^{N_2}]}^{\text{Bernoulli}}\overbrace{[\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}]}^{\text{Beta}} \\
&= \theta^{N_1+\alpha_1-1}(1-\theta)^{N_2+\alpha_2-1} \\
&\propto \mathrm{Beta}(\theta|N_1+\alpha_1, N_2+\alpha_2)
\end{aligned}
$$

We see that the posterior has the same functional form (beta) as the prior (beta), since it is $\boxed{\text{conjugate.}}$

Categorical distribution

The multivariate version of the Bernoulli distribution is the ***Categorical distribution*** *(an instance of the* ***multinomial distribution****).*

$\theta_K = 1 - \sum_{j=1}^{K-1} \theta_j$

We are given $\mathbf{n}$ *data points,* $\mathbf{x}_{1:n} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$*. Each point* $\mathbf{x}_i$ *indicates one of* $\mathbf{K}$ *values. For example if* $K=3$ *,then the possible vectors are (****100****), (****010****) and (****001****).*

$$x_i = \begin{cases} 1 & 100 \\ 2 & 010 \\ 3 & 001 \end{cases}$$

The likelihood of the data is then:

$P(x_i=1|\theta) = \theta_1$

$P(x_i=2|\theta) = \theta_2$

$\vdots$

$$\boldsymbol{p}(\mathbf{x}_i \,|\, \boldsymbol{\theta}) = \boldsymbol{Cat}\,(\mathbf{x}_i \,|\, \boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_{ij}=1)}$$

$$\boldsymbol{p}(\mathbf{x}_{1:n} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_{ij}=1)}$$

$x_i = 001$, $x_{i1} = 0$, $x_{i3} = 1$

$P(x_i=3|\theta) = \theta_1^0 \theta_2^0 \theta_3^1$

$= \theta_3$

$\theta_1 + \theta_2 + \theta_3 = 1$

Dirichlet distribution

The conjugate prior is the **Dirichlet distribution** which is the natural generalization of the beta distribution to multiple dimensions.

The pdf is defined as follows:

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

Beta

$$\propto \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1}$$

$$= \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 2}$$

defined on the **probability simplex**, i.e., the set of vectors such that $0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{K} \theta_k = 1$.

In addition, $B(\alpha_1, \ldots, \alpha_K)$ is the natural generalization of the beta function to K variables:

$$B(\boldsymbol{\alpha}) := \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$$

where $\alpha_0 := \sum_{k=1}^{K} \alpha_k$.

~~Beta~~ Dir Die

$$\theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \theta_3^{\alpha_3 - 1} \theta_4^{\alpha_4 - 1} \theta_5^{\alpha_5 - 1}$$

$$(1 - \theta_1 - \theta_2 - \theta_3 - \theta_4 - \theta_5)^{\alpha_6 - 1}$$

Dirichlet-categorical model

$$P(\theta \mid x_{1:n}) \propto P(x_{1:n} \mid \theta)\, P(\theta)$$

$$\propto \underbrace{\prod_{i=1}^{n} \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_{ij}=1)}} \; \underbrace{\prod_{j=1}^{K} \theta_j^{\alpha_j - 1}}$$

$$N_j = \sum_{i=1}^{n} \mathbb{I}(x_{ij}=1)$$

$$= \prod_{j=1}^{K} \theta_j^{N_j} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1}$$

For the die

N_5 is the # times you saw a 5.

$$= \prod_{j=1}^{K} \theta_j^{(N_j + \alpha_j) - 1}$$

$$\alpha'_j = N_j + \alpha_j$$

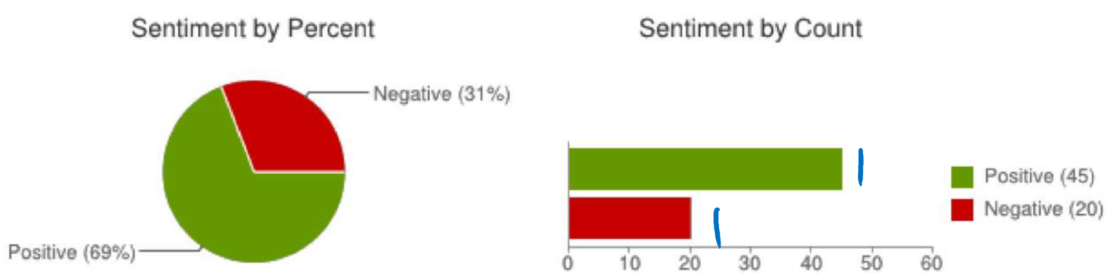
Posterior is Dirichlet $\boxed{}$

Text classification example

Sentiment140 Tweet 381 Like 173 85

obama | English | Search | Save this search

Sentiment analysis for obama



Tweets about: obama

lillien1984: RT @Barack**Obama**: President **Obama**: "You know that I know what real change looks like because you've seen me fight for it."
Posted 22 seconds ago

a_girl_irl: Romney endorsed by cool H'wood celebs: Kid Rock, Chuck Woolery, hot chick from Clueless, that's literally it, everyone else likes **Obama**
Posted 22 seconds ago

AmericanWoman8: RT @RBReich: If **Obama** wins, will radical right see it as a repudiation and become more reasonable, or as a provocation and grow even more extreme?
Posted 22 seconds ago

y is used to indicate ***C*** classes. E.g., the classes could be *positive*, *negative* and *neutral*. That is, ***C*=3**.

The input ***x*** *in this example is a vector of* ***d*** *zeros with ones indicating which words occur in the tweet.*

$$\begin{array}{cccccc} & \text{cut} & \text{rat} & \text{sat} & \text{hi} & \text{omg} \\ x_i = & 1 & 0 & 1 & 0 & 1 \\ & j=1 & j=2 & & & j=d \end{array}$$

← dictionary tokens

$d = 5$

Nandodf: Omg ~~the~~ cat sat!

Naïve Bayes classifier $Y \in \{1, 2, \ldots, C\}$

We are interested in the posterior distribution of y *given the model parameters* θ *and* π *and the inputs (features)* x*:*

$$\underline{P(y_i \mid x_i, \theta, \pi)} = \underline{p(y_i \mid \pi)}\ \underline{p(x_i \mid y_i, \theta)} \,/\, p(x_i \mid \theta, \pi)$$

$$= \frac{P(Y_i \mid \pi)\, P(x_i \mid Y_i, \theta)}{\sum_{Y_i = c} P(Y_i = c \mid \pi)\, P(x_i \mid Y_i = c, \theta)}$$

$$\boxed{P(Y_i \mid \pi) = \prod_{c=1}^{C} \pi_c^{\mathbb{I}(Y_i = c)}} \leftrightarrow \boxed{P(Y_i = c \mid \pi) = \pi_c}$$

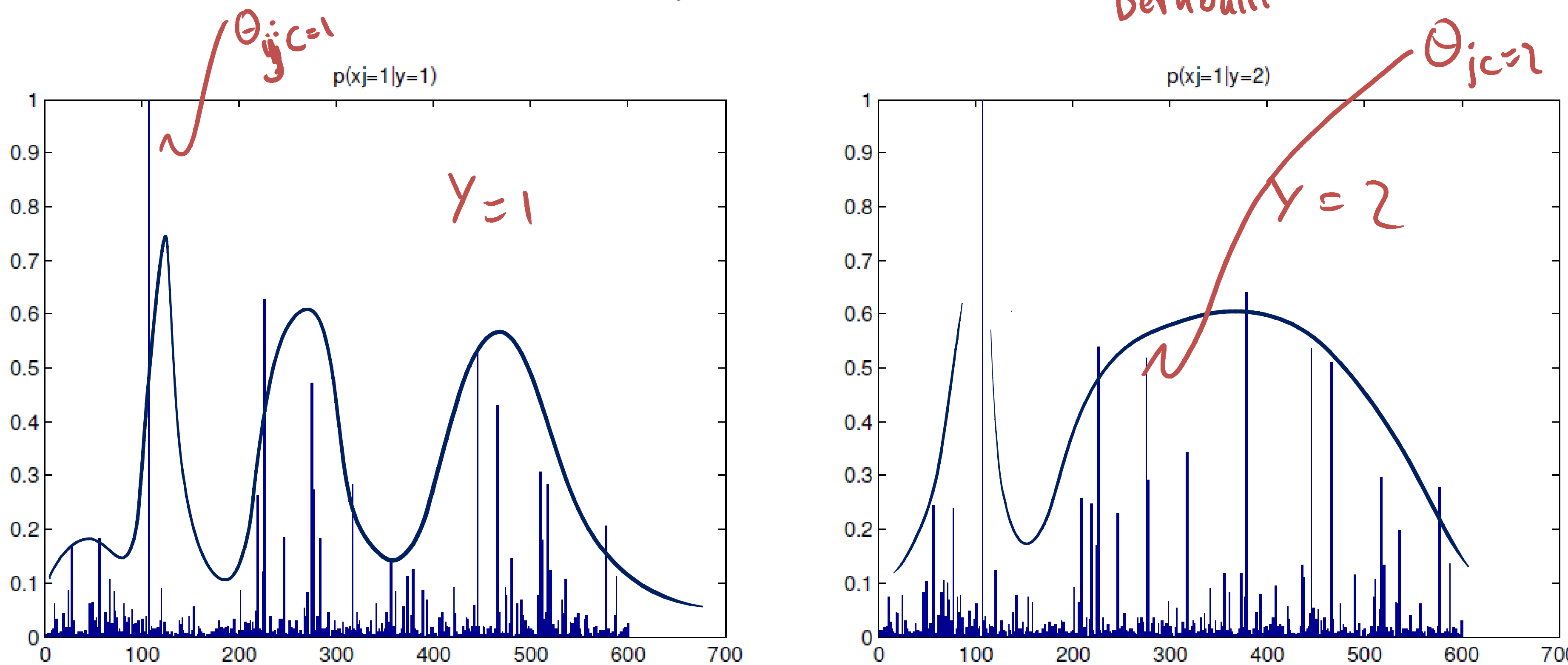
$C = 2$ $c = 1$ positive

$c = 2$ negative

Naïve Bayes classifier

Assume the features are conditionally independent given the class label. That is,

$$\underbrace{P(x_i \mid \theta, y_i = c)}_{\text{ith tweet}} = \prod_{j=1}^{d} \underbrace{P(x_{ij} \mid \theta, y_i = c)}_{\text{Bernoulli}}$$



Naïve Bayes classifier with binary features x

$$x_i = [1\ 0\ 0\ 0\ 1\ 0\ 0\ 1]$$

$$j=1 \rightarrow j=d=8$$

P(y | x, θ, π) α p(y|π) p(x | y, θ)

$$y_i = \begin{cases} 1 & c=1 \\ 2 & c=2 \end{cases}$$

$$p(\mathbf{y}|\boldsymbol{\pi},\boldsymbol{\theta},\mathbf{x}) \propto \prod_{i=1}^{n}\prod_{c=1}^{C}\left(\pi_c^{\mathbb{I}_c(y_i)}\prod_{j=1}^{d}\theta_{jc}^{\mathbb{I}_c(y_i)\mathbb{I}_1(x_{ij})}(1-\theta_{jc})^{\mathbb{I}_c(y_i)\mathbb{I}_0(x_{ij})}\right)$$

$$P(y|\pi) = \prod_{i=1}^{n} \underbrace{\prod_{c=1}^{C} \pi_c^{\mathbb{I}_c(y_i)}} \longleftrightarrow P(y_i = c|\pi) = \pi_c$$

$$P(x_{ij}|y_i=c,\theta) = \theta_{jc}^{\mathbb{I}_c(y_i)\mathbb{I}_1(x_{ij})}(1-\theta_{jc})^{\mathbb{I}_c(y_i)\mathbb{I}_0(x_{ij})}$$

$$\theta_{jc} = P(x_{ij}=1|y_i=c,\theta)$$

MLE for Naïve Bayes classifier with binary features x

Note: $\mathbb{I}_c(y_i)$ is the same as $\mathbb{I}(y_i = c)$

$$P(y \mid x, \theta, \pi) \; \alpha \; p(y \mid \pi) \, p(x \mid y, \theta)$$

$$p(\mathbf{y} \mid \boldsymbol{\pi}, \boldsymbol{\theta}, \mathrm{x}) \propto \prod_{i=1}^{n} \prod_{c=1}^{C} \left(\pi_c^{\mathbb{I}_c(y_i)} \prod_{j=1}^{d} \theta_{jc}^{\mathbb{I}_c(y_i)\mathbb{I}_1(x_{ij})} (1-\theta_{jc})^{\mathbb{I}_c(y_i)\mathbb{I}_0(x_{ij})} \right)$$

n is the number of data points

$\underline{N_c} = \sum_{i=1}^{n} \mathbb{I}(y_i = c)$ is the number of data of class c.

$\underline{N_{jc}} = \sum_{i=1}^{n} \mathbb{I}_c(y_i)\mathbb{I}_1(x_{ij})$ # times $x_{ij}=1$ given we are in class c.

$$\hat{\pi}_c = N_c / n$$

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

Predicting the class of new data

Given a new data point (say tweet) $\underline{x}^*$, the class prediction is:

training data

$$P(\underline{y}=c \mid x^*, D) \propto \underline{\hat{\pi}}_c \prod_{j=1}^{d} \underline{\hat{\theta}}_{jc}^{\,\mathbb{I}(x_j^*=1)} \left(1-\hat{\theta}_{jc}\right)^{\mathbb{I}(x_j^*=0)}$$

Do this for all classes and then normalize so that

$$\sum_{c'} P(y=c' \mid x^*, D) = 1$$

Naïve Bayes classifier with binary features

$N_c = 0$, $N_{jc} = 0$;

for $i = 1 : n$ **do**

  $c = y_i$ // Class label of i'th example

  $N_c := N_c + 1$;

  for $j = 1 : d$ **do**

    if $x_{ij} = 1$ **then**

      $N_{jc} := N_{jc} + 1$

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

Log-sum-exp trick

$$\begin{aligned} \log p(y = c|\mathbf{x}) &= b_c - \log\left[\sum_{c'=1}^{C} e^{b_{c'}}\right] \\ b_c &:= \log p(\mathbf{x}|y=c) + \log p(y=c) \end{aligned}$$

$$\log[\sum_{c'} e^{b_{c'}}] = \log \sum_{c'} p(y=c', \mathbf{x}) = \log p(\mathbf{x})$$

Log sum exp function

$$\log \sum_c e^{b_c} = \log\left[(\sum_c e^{b_c - B})e^B\right] = \left[\log(\sum_c e^{b_c - B})\right] + B \quad \text{where } B = \max_c b_c.$$

For example,

$$\log(e^{-120} + e^{-121}) = \log\left(e^{-120}(e^0 + e^{-1})\right) = \log(e^0 + e^{-1}) - 120$$

NBC prediction with log-sum-exp trick

for $i = 1 : n$ **do**
  for $c = 1 : C$ **do**
    $L_{ic} = \log \hat{\pi}_c$;
    for $j = 1 : d$ **do**
      if $x_{ij} = 1$ **then** $L_{ic} := L_{ic} + \log \hat{\theta}_{jc}$ **else** $L_{ic} := L_{ic} + \log(1 - \hat{\theta}_{jc})$
  $p_{ic} = \exp(L_{ic} - \text{logsumexp}(L_{i,:}))$;
  $\hat{y}_i = \text{argmax}_c \, p_{ic}$;

MLE

$2^3 2^4 = 2^{3+4}$

$$\prod_i \prod_c \pi_c^{\mathbb{I}_c(y_i)} = \prod_c \pi_c^{\sum_{i=1}^{n} \mathbb{I}_c(y_i)}$$

P(y | x, θ, π) α p(y | π) p(x | y, θ)

$$p(\mathbf{y}|\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{x}) \propto \prod_{i=1}^{n} \prod_{c=1}^{C} \left(\pi_c^{\mathbb{I}_c(y_i)} \prod_{j=1}^{d} \theta_{jc}^{\mathbb{I}_c(y_i)\mathbb{I}_1(x_{ij})} (1-\theta_{jc})^{\mathbb{I}_c(y_i)\mathbb{I}_0(x_{ij})} \right)$$

$$= \prod_{c=1}^{C} \pi_c^{N_c} \prod_{j=1}^{d} \theta_{jc}^{N_{jc}} (1-\theta_{jc})^{N_c - N_{jc}}$$

$$\mathcal{L}(\theta, \pi) = \underline{\left(\sum_{c=1}^{C} N_c \log \pi_c \right)} + \underline{\left(\sum_{c=1}^{C} \sum_{j=1}^{d} N_{jc} \log \theta_{jc} + (N_c - N_{jc}) \log(1-\theta_{jc}) \right)}$$

MLE for π

$$\ell(\pi,\lambda) = \left(\sum_{c=1}^{C} N_c \log \pi_c\right) + \lambda\left[1 - \sum_{c=1}^{C} \pi_c\right]$$

$$\frac{\partial \ell(\pi,\lambda)}{\partial \lambda} = 0 + \left[1 - \sum_{c=1}^{C} \pi_c\right] \xrightarrow{=0} \boxed{\sum_{c=1}^{C} \pi_c = 1} \text{ — } 1$$

$$\frac{\partial \ell(\pi,\lambda)}{\partial \pi_c} = N_c \frac{1}{\pi_c} + \lambda(-1) \xrightarrow{=0} \frac{N_c}{\pi_c} = \lambda$$

$$\sum_{c=1}^{C} N_c = \sum_{c=1}^{C} \pi_c \lambda$$

$$n = \lambda \left[\sum_{c=1}^{C} \pi_c\right] = \lambda\, 1$$

$$\lambda = n$$

$$N_c = \lambda \pi_c$$

$$\therefore \pi_c = N_c / n$$

MLE for π

MLE for θ

$$\ell(\theta) = \sum_{j=1}^{d} \sum_{c=1}^{C} N_{jc} \log \theta_{jc} + (N_c - N_{jc}) \log(1 - \theta_{jc})$$

$$\frac{\partial \ell(\theta)}{\partial \theta_{jc}} = N_{jc} \frac{1}{\theta_{jc}} + (N_c - N_{jc}) \frac{-1}{1 - \theta_{jc}}$$

$$= \left[N_{jc} (1 - \theta_{jc}) + (N_{jc} - N_c) \theta_{jc} \right] / \theta_{jc} (1 - \theta_{jc})$$

Equating to zero:

$$N_{jc} - \cancel{N_{jc}} \theta_{jc} + \cancel{N_{jc}} \theta_{jc} - N_c \theta_{jc} = 0$$

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

Bayesian analysis

Likelihood:

$$p(\mathbf{y}|\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{x}) \propto \prod_{i=1}^{n} \prod_{c=1}^{C} \left(\pi_c^{\mathbb{I}_c(y_i)} \prod_{j=1}^{d} \theta_{jc}^{\mathbb{I}_c(y_i)\mathbb{I}_1(x_{ij})} (1-\theta_{jc})^{\mathbb{I}_c(y_i)\mathbb{I}_0(x_{ij})} \right)$$

$$= \prod_{c=1}^{C} \left(\pi_c^{N_c} \prod_{j=1}^{d} \theta_{jc}^{N_{jc}} (1-\theta_{jc})^{N_c - N_{jc}} \right)$$

Prior:

$$P(\pi) \propto \prod_{c=1}^{C} \pi_c^{\alpha_c - 1} \qquad \text{Dirichlet}$$

$$P(\theta_{jc}) = \theta_{jc}^{\beta_1 - 1} (1-\theta_{jc})^{\beta_2 - 1} \qquad d \times C \text{ beta priors}$$

$$P(\pi, \theta) = \prod_{c=1}^{C} \left(\pi_c^{\alpha_c - 1} \prod_{j=1}^{d} \theta_{jc}^{\beta_1 - 1} (1-\theta_{jc})^{\beta_2 - 1} \right)$$

Bayesian analysis

$$P(\theta,\pi \mid x,y) \propto \prod_{c=1}^{C} \pi_c^{\overbrace{N_c+\alpha_c-1}^{\alpha_c'}} \prod_{j=1}^{d} \theta_{jc}^{\overbrace{\beta_1+N_{jc}-1}^{\beta_1'}} (1-\theta_{jc})^{\overbrace{\beta_2+N_c-N_{jc}-1}^{\beta_2'}}$$

Dirichlet !

$$\text{Post. mean}(\pi) = \frac{N_c+\alpha_c}{n+\sum_{c'=1}^{C}\alpha_{c'}}$$

$$\text{Post. mean}(\theta_{jc}) = \frac{N_{jc}+\beta_1}{N_c+\beta_1+\beta_2}$$

Next lecture

In the next lecture, we learn about another very popular classifier: logistic regression. This classifier will be a building block for neural networks.