SPR day 4

Dirichlet-Categorical Joint Model for discrete data

Reading: Bishop PRML Sec 2.2
- Defines discrete and multinomial Dirichlet distributions

Outline:

(1) Categorical distribution
categorical vs. multinomial

(2) ML estimation for Categorical likelihoods

(3) Dirichlet distribution

(4) Dirichlet-Categorical joint model and its posterior

(5) MAP Estimation for Dir-Cat

From Binary to V-ary Discrete Distributions

Previously, we've modeled events with 2 possible outcomes

Now, consider V-possible outcomes, $2 \le V < \infty$
with finite V known in advance.

Example: Which letter in alphabet comes next?
Which team will win Super Bowl?

Binary aka Bernoulli

Sample Space: 0 or 1

Parameter: $\mu \in [0,1]$

PMF: $p(X=x|\mu) = \mu^x (1-\mu)^{1-x}$

Binomial

For results of N iid coin tosses, unordered

Sample Space: $[n_0, n_1]$ s.t. (n_0 = num. tails, n_1 = num. heads) integer valued $n_0 \ge 0, n_1 \ge 0$ $n_0 + n_1 = N$

Parameter: N number of trials

μ probability of heads $\mu \in [0,1]$

Categorical Discrete aka Multinoulli

Sample space: $\{1, 2, 3, \cdots V\}$

Parameter: $\boldsymbol{\mu} = [\mu_1, \mu_2 \cdots \mu_V]$, s.t. $\sum_{v=1}^{V} \mu_v = 1$, $\mu_v \ge 0 \; \forall v$

Space of valid vectors is the probability simplex denoted $\mu \in \Delta^V$

PMF: $p(X=x|\mu) = \prod_{v=1}^{V} \mu_v^{[x=v]} = \mu_x$

Multinomial

For results of N iid V-ary trials, unordered.

Sample Space: $[n_1, n_2, \cdots n_V]$ s.t. (n_1 = num 1s, n_2 = num 2s, n_V = num Vs) int valued $n_v \ge 0 \; \forall v$ $n_1 + \cdots n_V = N$

Parameter: N number of trials

$\mu \in \Delta^V$ probability vector

Notation for Categorical Likelihoods

Consider observing N words from vocabulary of size V.

Let random variables, $X_1, X_2, \cdots X_N$ indicate the words.

where $x_n = [0 \; 0 \; 0 \; 1 \; 0 \; 0 \; \cdots \; 0]$ indicator

$1 \; 2 \; 3 \; 4 \; 5 \; 6 \; \cdots \; V$ index

exactly one entry in x_n is "on" (=1), rest are "off" (=0).

We call this "one hot" encoding.

Suppose $V = 4$:

integer sample space	one hot encoding
1	1 0 0 0
2	0 1 0 0
3	0 0 1 0
4	0 0 0 1

Makes notation easier!

$$x_{nv} = \begin{cases} 1 & \text{if } n^{th} \text{ word is type } v \\ 0 & \text{otherwise} \end{cases}$$

Assume N words drawn independent & identically distributed from Categorical with parameter $\mu \in \Delta^V$

$$p(x_1 \cdots x_N \mid \mu) = \prod_{n=1}^{N} \text{CatPMF}(x_n \mid \mu) = \prod_{n=1}^{N} \prod_{v=1}^{V} \mu_v^{x_{nv}}$$

$$= \prod_{v=1}^{V} \mu_v^{\sum_{n=1}^{N} x_{nv}}$$

$\sum_{n=1}^{N} x_{nv}$ = number of times vocab symbol v appears in our dataset

$$= \prod_{v=1}^{V} \mu_v^{m_v(\mathbf{x})}$$

ML Estimation of Categorical Parameter

$$\hat{\mu} = \underset{\mu \in \Delta^V}{\operatorname{argmax}} \log p(x_1, \cdots x_N | \mu)$$

$$= \underset{\mu \in \Delta^V}{\arg\max} \sum_{v=1}^{V} m_v \log \mu_v$$

$$= \underset{\mu \in \mathbb{R}^V}{\arg\max} \sum_{v=1}^{V} m_v \log \mu_v$$

address with Lagrange multiplier → s.t. $1 - \sum_{v=1}^{V} \mu_v = 0$ sum to one

address with ignore, then check → $\mu_v \geq 0$ for all v all entries non-negative

Lagrange multiplier method

Step 1

↓

Step 2

$$\mathcal{L}(\mu, \lambda) = \sum_{v=1}^{V} m_v \log \mu_v + \lambda \left(1 - \sum_{v} \mu_v\right)$$

$\lambda \neq 0$ is multiplier

Step 3

$$\frac{\partial}{\partial \mu_1} \mathcal{L} = 0 \qquad \frac{m_1}{\mu_1} - \lambda = 0 \quad (1)$$

$$\frac{\partial}{\partial \mu_2} \mathcal{L} = 0 \quad \Rightarrow \quad \frac{m_2}{\mu_2} - \lambda = 0 \quad (2)$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{\partial}{\partial \mu_V} \mathcal{L} = 0 \qquad \frac{m_V}{\mu_V} - \lambda = 0 \quad (V)$$

$$\frac{\partial}{\partial \lambda} \mathcal{L} = 0 \qquad 1 - \sum_v \mu_v = 0 \quad (V+1)$$

Step 4: Solve!

Summing eq (1)-(V) we get

$$\frac{1}{\lambda} \sum_v m_v = \sum_v \mu_v$$

Plug into eq (V+1), simplify to

$$\frac{1}{\lambda} \sum_v m_v = 1 \Rightarrow \lambda = N = \sum_{v=1}^{V} m_v$$

total # of observed words

Now, returning to original eq (1)···(V),

$$\mu_1 = \frac{m_1}{N}$$

$$\vdots$$

$$\mu_V = \frac{m_V}{N}$$

done! ML estimate is

$$\mu^* = \left[\frac{m_1}{N}, \frac{m_2}{N}, \cdots \frac{m_V}{N} \right]$$

Check: Does this obey the $\mu_v \geq 0$ constraint we ignored?
Yes! $m_v \geq 0$ so $\frac{m_v}{N} \geq 0$. ✓

Dirichlet Distribution

Random Variable $\mu = [\mu_1, \mu_2, \cdots \mu_V]$

Sample Space $\mu \in \Delta^V$ set of V-dim vectors s.t. $\sum_v \mu_v = 1$, $\mu_v \geq 0$ for $v \in \{1 \cdots V\}$

Parameter $\alpha = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_V]$

$\alpha_1 \geq 0 \quad \alpha_2 \geq 0 \quad \cdots \quad \alpha_V \geq 0$

vector of positive entries (or at least non-negative)

PDF

$$\text{DirPDF}(\mu|\alpha) = \underbrace{\frac{\Gamma(\alpha_1 + \alpha_2 + \cdots \alpha_V)}{\prod_{v=1}^{V} \Gamma(\alpha_v)}}_{c(\alpha)} \cdot \underbrace{\prod_{v=1}^{V} \mu_v^{(\alpha_v - 1)}}_{f(\mu,\alpha)}$$

$c(\alpha)$: const wrt r.v. μ

$f(\mu,\alpha)$: function depends on μ

Can recognise that

$$\int_\mu \prod_{v=1}^{V} \mu_v^{\alpha_v - 1} \, d\mu = \frac{1}{c(\alpha)} = \frac{\prod_v \Gamma(\alpha_v)}{\Gamma(\sum_v \alpha_v)}$$

useful identity relating c and f

Dirichlet: Special Cases & Visuals

Special Case: When V=2, reduces to Beta distribution

Beta

Sample space: $\mu \in [0, 1]$

PDF: $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}$

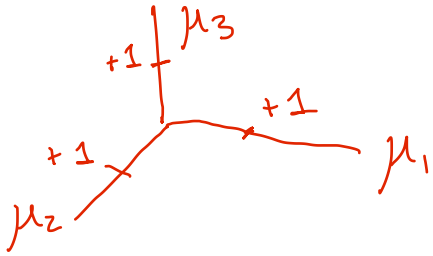
V=2 Dirichlet

$[\mu_1, \mu_2]$ s.t. $\mu_1 \geq 0$, $\mu_2 \geq 0$, $\mu_1 + \mu_2 = 1$

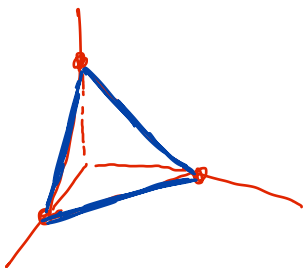
$\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \mu_1^{\alpha_1-1} \mu_2^{\alpha_2-1}$

Constraints imply $\mu_2 = 1 - \mu_1$.
In general, V dim. Dirichlet has V-1 free dimensions due to sum-to-one constraint.

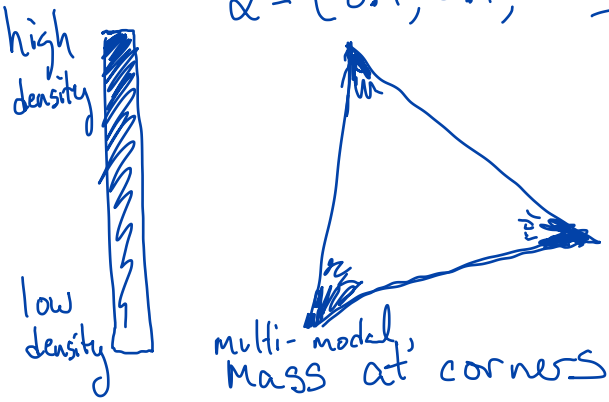
Visualization



set of points in Δ^3 is triangle with vertices (1,0,0), (0,1,0), (0,0,1)

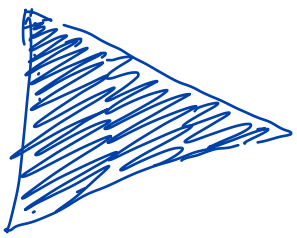


$\alpha = [0.1, 0.1, 0.1]$



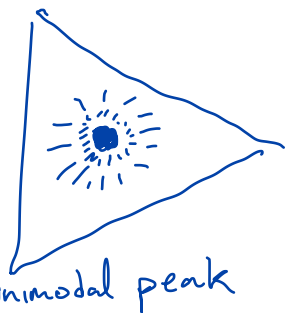
multi-modal, Mass at corners

$\alpha = [1, 1, 1]$



uniform

$\alpha = [50, 50, 50]$



unimodal peak

Dirichlet-Categorical Model

Jointly explain:

$[\mu_1, \mu_2, \cdots \mu_V]$ unknown probability vector

$x_1, x_2, \cdots x_N$ N observed words

$x_n \in \text{onehot}(V)$

Define a complete model as joint distribution:

$$p(x_1, \cdots x_N, \mu) = \left[\prod_{n=1}^{N} \text{CatPMF}(x_n|\mu)\right] \cdot \text{DirPDF}(\mu|\alpha)$$

likelihood prior

Just like Beta-Bernoulli, can derive other relevant distributions from the joint via sum/product rule

posterior	$p(\mu \mid x_1 \cdots x_N)$
evidence	$p(x_1 \cdots x_N)$
predictive posterior	$p(x_{N+1} \mid x_1 \cdots x_N)$

Posterior of μ under Dir-Cat

Bayes rule says:

$$p(\mu | x_1 \cdots x_N) = \frac{1}{p(x_1 \cdots x_N)} \cdot \left[\prod_{n=1}^{N} \left(\text{CatPMF}(x_n | \mu) \right) \cdot \text{DirPDF}(\mu | \alpha) \right]$$

Expanding PDF/PMF definitions ...

$$= \text{Const}_1 \cdot \prod_{n=1}^{N} \prod_{v=1}^{V} \mu_v^{x_{nv}} \cdot c(\alpha) \cdot \prod_{v=1}^{V} \mu_v^{\alpha_v - 1}$$

Defining $m_v = \sum_{n=1}^{N} x_{nv}$ for all v,

$$= \text{Const}_2 \cdot \prod_{v=1}^{V} \mu_v^{m_v + \alpha_v - 1}$$

We have written posterior PDF as

$$\text{constant wrt } \mu \cdot \prod_{v=1}^{V} \mu_v^{\hat{\alpha}_v - 1} \quad \text{with } \hat{\alpha}_v = m_v + \alpha_v$$

recognise form of PDF as function of μ as Dirichlet

Thus, posterior is Dirichlet

$$p(\mu | x_1 \cdots x_N) = \text{DirPDF}(\mu | \hat{\alpha}_1, \cdots \hat{\alpha}_V)$$

where $\hat{\alpha}_v = \alpha_v + m_v$ = prior pseudocount + count of symbol v

MAP Estimation for Dir-Cat

$$\mu^{MAP} = \underset{\mu \in \Delta^V}{\operatorname{argmax}} \log p(\mu | x_1 \cdots x_N)$$

Plug in Dirichlet form of posterior

$$= \underset{\mu \in \Delta^V}{\operatorname{argmax}} \log \left[c(\hat{\alpha}) \prod_{v=1}^{V} \mu_v^{\hat{\alpha}_v - 1} \right]$$

Simplify by applying log of product = sum of logs

$$= \underset{\mu \in \Delta^V}{\operatorname{argmax}} \text{ Const} + \sum_{v=1}^{V} (\hat{\alpha}_v - 1) \log \mu_v$$

Drop const terms

$$= \underset{\mu \in \Delta^V}{\operatorname{argmax}} \sum_{v=1}^{V} (\hat{\alpha}_v - 1) \log \mu_v$$

Can solve just like ML estimator so long as $\hat{\alpha}_v - 1 \geq 0$ for all v

$$\mu^{MAP} = \left[\frac{m_1 + \alpha_1 - 1}{N + \sum_v (\alpha_v - 1)}, \cdots \frac{m_V + \alpha_V - 1}{N + \sum_v (\alpha_v - 1)} \right]$$

only exists when $m_1 + \alpha_1 > 1$
$\vdots$
$m_V + \alpha_V > 1$

sufficient condition: prior's mode exists
$\alpha_1 > 1$
$\vdots$
$\alpha_V > 1$