SPR Day 4

The Dirichlet-Discrete Model for Unigrams

Builds on: Day 3, the Beta-Bernoulli joint model

Readings: Bishop PRML Sec 2.2
Covers Dirichlet distribution
Discrete/Multinomial distributions

Bishop PRML Appendix E Lagrange Multipliers
Needed for ML estimator derivation

Goals

Introduce Discrete distrib.
ML estimation for unigrams
Introduce Dirichlet distrib.
Emphasize two parts of PDF
- norm. const.
- function of r.v.
Emphasize generalization of Beta to K-dimensions

Dirichlet-Multinomial joint model for unigrams
- posterior
- MAP estimation
- predictive posterior
- evidence

From binary to discrete random variables

Previously, we've talked about modeling many binary outcomes
e.g. coin tosses,
win-loss outcomes for games, etc.

Now, consider modeling many outcomes with more than two possibilities (but still a finite number).
e.g. which word will appear next in an email
(could use to build auto-suggestion system)

Binary (aka Bernoulli) random variable

Sample space: 2 possible values
0 or 1

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

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

V-ary (or "Categorical" or "Discrete") random variable

Sample space: V possible values
$\{1, 2, \dots V\}$
V: positive integer ≥ 2
finite, known in advance

Parameter: $\mu \in \Delta^V \subseteq \mathbb{R}^V$
the V-dimensional probability simplex
$\mu = [\mu_1, \dots \mu_{V-1}, \mu_V]$
each entry ≥ 0, $\sum_{v=1}^{V} \mu_v = 1$

$PMF(X=w) = \prod_{v=1}^{V} \mu_v^{[v=w]} = \mu_w$

Writing a likelihood for Discrete rand. vars.

See Bishop 2.25 - 2.33

Consider observing N distinct words from vocab of size V

Denote by random variables $X_1, \cdots X_N$

where each one is a <u>one hot</u> vector indicating which one of V possible words

$$X_n = [0, 0, \cdots \underset{\uparrow}{1}, 0, \cdots 0]$$

Only one entry "on" or "hot"

If we model the $\{X_n\}_{n=1}^N$ as <u>independent</u>, and <u>identically distrib.</u> from same Discrete distribution with parameter μ, we have likelihood

$$p(X_1, \cdots X_N | \mu) = \prod_{n=1}^{N} \text{Disc}(x_n | \mu) = \prod_{n=1}^{N} \prod_{v=1}^{V} \mu_v^{x_{nv}}$$

$$= \prod_{v=1}^{V} \mu_v^{\sum_n x_{nv}}$$

define $m_v = \sum_{n=1}^{N} x_{nv}$

counts how many times word type v appears in our dataset

Then our likelihood is:

$$p(x_1 \cdots x_N | \mu) = \prod_{v=1}^{V} \mu_v^{m_v}$$

← could write as $m_v(x_1 \cdots x_N)$ this is a function of our observed data

Finding an ML estimate for μ

See Bishop 2.25 - 2.33

Again, use log likelihood to avoid numerical problems.

Goal: Find value of parameter μ that maximizes log likelihood of N observations, but is still a valid parameter (belongs to V-dimensional probability simplex)

$$\mu^{ML} = \underset{\mu \in \Delta^V}{\operatorname{argmax}} \sum_{v=1}^{V} m_v \log \mu_v$$

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

$\mu_v \in (0,1)$ ← address this with "ignore, then check"

$\sum_v \mu_v = 1$ ← address this with Lagrange multiplier

another way to write: $1 - \sum_v \mu_v = 0$

Expanding to obtain our Lagrangian, we have

$$\mathcal{L}(\mu_1 \cdots \mu_V, \lambda) = \sum_{v=1}^{V} m_v \log \mu_v + \lambda \left(1 - \sum_{v=1}^{V} \mu_v\right)$$

$\lambda \neq 0$ is the Lagrange multiplier

To find optimal values, set up $V+1$ equations and solve

$$\frac{\partial \mathcal{L}}{\partial \mu_1} = 0 \qquad \frac{m_1}{\mu_1} - \lambda = 0 \qquad \frac{m_1}{\lambda} = \mu_1$$

$$\frac{\partial \mathcal{L}}{\partial \mu_2} = 0 \quad \Rightarrow \quad \frac{m_2}{\mu_2} - \lambda = 0 \Rightarrow \qquad \vdots$$

$$\vdots \qquad \vdots \qquad \frac{m_V}{\lambda} = \mu_V$$

$$\frac{\partial \mathcal{L}}{\partial \mu_V} = 0 \qquad \frac{m_V}{\mu_V} - \lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \qquad 1 - \sum_v \mu_v = 0 \qquad 1 = \sum_v \mu_v$$

Sum these up

by summing the first V eqns up, we get

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

next plug in the $(V+1)^{th}$ eqn, which says $\sum_v \mu_v = 1$

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

Solve for λ gives $\lambda = \sum_v m_v = N$

by defn of m_v the total # of observed words

now, returning the first V eqns, we can find

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

$$\vdots$$

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

Done! Thus, the ML estimate is

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

assumes $N \geq 1$

Quick check: does this obey the $\mu_v \in (0,1)$ constraint we ignored?

yes! $0 \leq \frac{m_v}{N} \leq 1$ for all possible m_v values

Dirichlet distribution

Random variable $\mu = [\mu_1 \cdots \mu_V]$

Sample space: Set of V-length vectors that are
non-negative entries and sum to one
$\mu_v \geq 0$ and $\sum_v \mu_v = 1$

Probability density function

$$\text{DirPDF}(\mu \mid a_1 \cdots a_V) = \frac{\Gamma(a_1 + a_2 + \cdots a_V)}{\prod_{v=1}^{V} \Gamma(a_v)} \prod_{v=1}^{V} \mu_v^{a_v - 1}$$

Parameters $a_1, a_2, \cdots a_V$, $a_v > 0$

$\frac{\Gamma(a_1 + a_2 + \cdots a_V)}{\prod_{v=1}^{V} \Gamma(a_v)}$: const. wrt. our rv. μ, $c(a)$

$\prod_{v=1}^{V} \mu_v^{a_v - 1}$: function that depends on μ, $f(\mu, a)$

Remember Gamma function
$\Gamma : \mathbb{R} \to \mathbb{R}$
if given positive, returns positive

Like any PDF, must integrate to 1 over sample space

$$\int_{\mu \in \Delta^V} \text{DirPDF}(\mu \mid a) d\mu = 1$$

$$\int_{\mu \in \Delta^V} c(a) f(\mu, a) d\mu = 1$$

$$\int_{\mu \in \Delta^V} f(\mu, a) d\mu = \frac{1}{c(a)} = \frac{\prod_v \Gamma(a_v)}{\Gamma(\sum_v a_v)}$$

Use this as an identity!

Connection:

Dirichlet with $V=2$ is equivalent to Beta

Beta r.v. $\mu \in (0,1)$ has PDF

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

Dir r.v. $[\mu, 1-\mu]$ with parameter vector $[a, b]$ has PDF

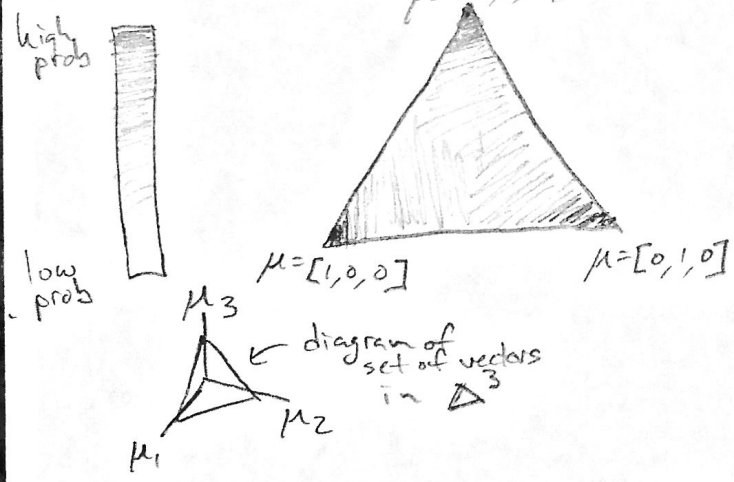
$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

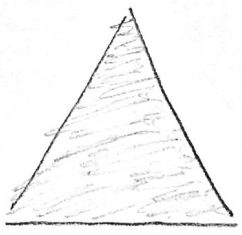
Can say Dirichlet is generalization of Beta to V-dimensions

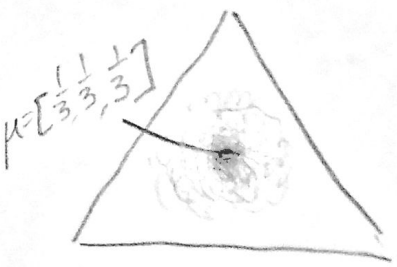
Understanding the Dirichlet

See Bishop Figure 2.5

$a = [0.1, 0.1, 0.1]$ $\quad$ $a = [1, 1, 1]$ $\quad$ $a = [10, 10, 10]$







Exercise:
match parameters to most likely observed samples

Parameter		Observed value
(i) $a = [20,20,20,20]$	→ (a)	$[.25, .25, .25, .25]$ $[.23, .27, .25, .25]$
(ii) $a = [1,1,1,1]$	→ (b)	$[.25, .25, .25, .25]$ $[.8, .1, .05, .05]$

The Dirichet-Discrete Joint Model

Consider these random variables

$\mu = [\mu_1 \cdots \mu_V] \in \Delta^V$ unknown probability vector

$X_1, X_2, \cdots X_N$ N observed unigrams

each one $x_n \in \{1, 2, \cdots V\}$

Can define a complete model (a joint distribution)

$$P(X_1, \cdots X_N, \mu) = \left[\prod_{n=1}^{N} \text{DiscretePMF}(x_n|\mu)\right] \cdot \text{DirPDF}(\mu|a_1 \cdots a_V)$$

discrete likelihood — Dirichlet "prior"

Just like our Beta-Binary model can ask about relevant probabilities that can be derived from this joint distribution

"Posterior"

$$P(\mu|X_1 \cdots X_N) = \frac{P(X_1 \cdots X_N|\mu)P(\mu)}{P(X_1 \cdots X_N)}$$

likelihood, prior, evidence — derived by Bayes rule

"Evidence"

$$P(X_1 \cdots X_N) = \int_{\mu \in \Delta^V} P(X_1 \cdots X_N, \mu) d\mu$$

derived by sum rule

"Predictive Posterior"

$$P(X_N|X_1 \cdots X_{N-1}) = \int_{\mu \in \Delta^V} P(X_N|\mu) P(\mu|X_1 \cdots X_{N-1}) d\mu$$

likelihood, posterior — derived by sum rule + Bayes rule

Posterior of μ for Dirichlet-Discrete Model

assume one hot representation

$$p(\mu | x_1 \cdots x_N) = \frac{\prod_{n=1}^{N} \text{DiscretePMF}(x_n|\mu) \cdot \text{DirPDF}(\mu|a)}{p(x_1 \cdots x_N)}$$

$p(x_1 \cdots x_N)$ ← const wrt μ

$$= \text{Const}_1 \cdot \prod_{n=1}^{N} \prod_{v=1}^{V} \mu_v^{x_{nv}} \cdot \frac{\Gamma(\sum_v a_v)}{\prod_v \Gamma(a_v)} \cdot \prod_{v=1}^{V} \mu_v^{a_v - 1}$$

$\frac{\Gamma(\sum_v a_v)}{\prod_v \Gamma(a_v)}$ ← const wrt μ

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

remember $m_v \geq 0$

m_v = count of term v in the observed N words

We have written the posterior distribution PDF as a constant wrt μ times a function of μ

We know it must be a valid distribution that integrates to 1

so

$$\int_{\mu \in \Delta^V} \prod_{v=1}^{V} \mu_v^{m_v + a_v - 1} \, d\mu = \frac{1}{\text{const}} = \frac{\prod_{v=1}^{V} \Gamma(m_v + a_v)}{\Gamma(N + \sum_v a_v)}$$

where we plugged in what we remember about Gamma function identities from defn of Dirichlet distrib.

Thus, we recognize our posterior is Dirichlet

$$p(\mu | x_1 \cdots x_N) = \text{DirichletPDF}(\mu | m_1 + a_1, \cdots m_V + a_V)$$

$a_1 \cdots a_V$

MAP estimates for Dirichlet-Discrete

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

(Equivalent because posterior = joint / const wrt μ)

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

$$= \underset{\mu \in \Delta^V}{\operatorname{argmax}} \; \sum_v (m_v + a_v - 1) \log \mu_v$$

will be ≥ 0 if $a_v \geq 1$

If so, can apply same derivation as ML estimator (Lagrangian)

Thus,

$$\mu^{MAP} = \left[\frac{m_1 + a_1 - 1}{N + \sum_v (a_v - 1)}, \cdots \frac{m_V + a_V - 1}{N + \sum_v (a_v - 1)} \right]$$

NB: only exists when $a_1 \geq 1$, $a_2 \geq 1$, $\vdots$, $a_V \geq 1$