INSTRUCTOR SOLUTION for HW1

Collaboration Statement:

Total hours spent: 3 hours

I consulted the following resources:

- Bishop's PRML textbook

Links: [HW1 instructions] [collab. policy]

Contents

1a: Problem Statement

Let $\rho \in (0.0, 1.0)$ be a Beta-distributed random variable: $p \sim \text{Beta}(a, b)$.

Show that $\mathbb{E}[\rho] = \frac{a}{a+b}$.

Hint: You can use these identities, which hold for all $a > 0$ and $b > 0$:

$$\Gamma(a) = \int_{t=0}^{\infty} e^{-t} t^{a-1} dt \tag{1}$$

$$\Gamma(a+1) = a\Gamma(a) \tag{2}$$

$$\int_0^1 \rho^{a-1}(1-\rho)^{b-1} d\rho = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{3}$$

1a: Solution

$\mathbb{E}_{\text{Beta}(\rho \mid a,b)}[\rho]$	
$= \int_{\rho=0}^{1} \rho \text{BetaPDF}(\rho \mid a, b) d\rho$	by the definition of expectation
$= \int_{\rho=0}^{1} \rho \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \rho^{a-1}(1-\rho)^{b-1} d\rho$	substituting in definition of Beta PDF
$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{\rho=0}^{1} \rho^{a}(1-\rho)^{b-1} d\rho$	group the ρ terms, move const wrt ρ outside integral
$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}$	use identity for integral of unnormalized Beta PDFs with $a' = a+1, b' = b$
$= \frac{\Gamma(a+b)}{\Gamma(a+b+1)} \frac{\Gamma(a+1)}{\Gamma(a)}$	simplify and rearrange like terms
$= \frac{\Gamma(a+b)}{(a+b)\Gamma(a+b)} \cdot \frac{a\Gamma(a)}{\Gamma(a)}$	use identity $\Gamma(x+1) = x\Gamma(x)$
$= \frac{a}{a+b}$	

1b: Problem Statement

Let μ be a Dirichlet-distributed random variable: $\mu \sim \text{Dir}(a_1, \dots a_V)$.

Show that $\mathbb{E}[\mu_w] = \frac{a_w}{\sum_{v=1}^{V} a_v}$, for any integer w that indexes a vocabulary word.

** Hint:** You can use the identity:

$$\int \mu_1^{a_1-1} \mu_2^{a_2-1} \dots \mu_V^{a_V-1} d\mu = \frac{\prod_{v=1}^{V} \Gamma(a_v)}{\Gamma(a_1 + a_2 \dots + a_V)} \tag{4}$$

1b: Solution

$\mathbb{E}_{\text{Dir}(\mu|a_1,\dots a_V)}[\mu_w]$

$= \int_{\mu \in \Delta^V} \; \mu_w \cdot \text{DirPDF}(\mu_1, \dots \mu_V | a) d\mu$

by the definition of expectation for a vector r.v.

$= \int_{\mu \in \Delta^V} \mu_w \cdot \frac{\Gamma(\sum_v a_v)}{\prod_v \Gamma(a_v)} (\mu_1^{a_1-1} \dots \mu_w^{a_w-1-1} \dots \mu_V^{a_V-1}) d\mu$

substituting in definition of Dirichlet PDF, remembering that by our sum-to-one constraint that $\mu_V = 1 - \sum_{v=1}^{V-1} \mu_v$

$= \frac{\Gamma(\sum_v a_v)}{\prod_v \Gamma(a_v)} \int (\mu_1^{a_1-1} \dots \mu_w^{a_w} \dots \mu_V^{a_V-1}) d\mu$

group the μ_v terms, move const wrt μ outside integral

$= \frac{\Gamma(\sum_v a_v)}{\prod_v \Gamma(a_v)} \frac{\Gamma(a_w+1) \prod_{v \neq w} \Gamma(a_v)}{\Gamma(1+\sum_v a_v)}$

use identity for integral of unnormalized Dirichlet PDF with vector $a' = [a_1, a_2, \dots, a_w + 1, \dots a_V]$

$= \frac{\Gamma(\sum_v a_v)}{\Gamma(1+\sum_v a_v)} \frac{\Gamma(a_w+1)}{\Gamma(a_w)}$

simplify and rearrange like terms

$= \frac{\Gamma(\sum_v a_v)}{(\sum_v a_v)\Gamma(\sum_v a_v)} \cdot \frac{a_w \Gamma(a_w)}{\Gamma(a_w)}$

use identity $\Gamma(x+1) = x\Gamma(x)$

$= \frac{a_w}{\sum_v a_v}$

2a: Problem Statement

Show that the likelihood of all N observed words can be written as:

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N|\mu) = \prod_{v=1}^{V} \mu_v^{n_v} \tag{5}$$

2a: Solution

We begin with the statement that the joint probability is the product of conditionally-independent and identically distributed Categorical random variables.

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N|\mu) = \prod_{n=1}^{N} p(X_n = x_n|\mu) \tag{6}$$

We recall the Categorical PMF can be written using indicator notation as:

$$p(X_n = x_n|\mu) = \prod_{v=1}^{V} \mu_v^{[x_n=v]} \tag{7}$$

Substituting the above into our first equation, we get:

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N|\mu) = \prod_{n=1}^{N} \prod_{v=1}^{V} \mu_v^{[x_n=v]} \tag{8}$$

Next, we reverse the order of the two products (which we can always do by the commutativity of multiplication), so we multiply over v first and then over n.

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N|\mu) = \prod_{v=1}^{V} \prod_{n=1}^{N} \mu_v^{[x_n=v]} \tag{9}$$

Finally, we move the product over n inside the exponent, where it becomes a sum, and we simplify to arrive at the desired expression (plug in definition of n_v):

$$\begin{aligned} p(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N|\mu) &= \prod_{v=1}^{V} \mu_v^{\sum_{n=1}^{N}[x_n=v]} && (10) \\ &= \prod_{v=1}^{V} \mu_v^{n_v} && (11) \end{aligned}$$

2b: Problem Statement

Derive the next-word posterior predictive, after integrating away parameter μ.

That is, show that after seeing the N training words, the probability of the next word X_* being vocabulary word v is:

$$
\begin{aligned}
p(X_* = v | X_1 = x_1 \ldots X_N = x_n) &= \int p(X_* = v, \mu | X_1 = x_1 \ldots X_N = x_n) d\mu \\
&= \frac{n_v + \alpha}{N + V\alpha}
\end{aligned}
\tag{12}
$$

2b: Solution

$p(X_* = w \| X_1 = x_1 \ldots X_N = x_n)$	
$= \int_{\mu \in \Delta^V} p(X_* = w, \mu \| x_1, \ldots x_N) d\mu$	by the sum rule, applied to the joint probability $p(X_*, \mu \| X)$.
$= \int_{\mu \in \Delta^V} p(X_* = w \| \mu, x_1, \ldots x_N) p(\mu \| x_1, \ldots x_N) d\mu$	by the product rule
$= \int_{\mu \in \Delta^V} p(X_* = w \| \mu) p(\mu \| x_1, \ldots x_N, \alpha) d\mu$	because X_* is conditionally independent of $X_1 \ldots X_N$ given μ
$= \int_{\mu \in \Delta^V} \mu_w p(\mu \| x_1, \ldots x_N, \alpha) d\mu$	because $p(X_* \| \mu)$ is a Categorical PMF
$= \int_{\mu \in \Delta^V} \mu_w \text{Dir}(\mu \| n_1 + \alpha, \ldots n_V + \alpha) d\mu$	because $p(\mu \| X)$ is a Dirichlet by Bishop PRML Eq. 2.41
$= \mathbb{E}_{\text{Dir}(\mu \| n_1 + \alpha, \ldots n_V + \alpha)}[\mu_w]$	by the definition of expectations of vector-valued random variables
$= \frac{n_w + \alpha}{N + V\alpha}$	via the identity proved earlier about expectations of Dirichlet random variables in Problem 1b

2c: Problem Statement

Derive the marginal likelihood of observed training data, after integrating away the parameter μ.

That is, show that the marginal probability of the observed N training words has the following closed-form expression:

$$p(X_1 = x_1 \dots X_N = x_N) = \int p(X_1 = x_1, \dots X_N = x_N, \mu) d\mu \tag{13}$$

$$= \frac{\Gamma(V\alpha) \prod_{v=1}^{V} \Gamma(n_v + \alpha)}{\Gamma(N + V\alpha) \prod_{v=1}^{V} \Gamma(\alpha)} \tag{14}$$

2c: Solution

$p(X_1 = x_1 \dots X_N = x_N) =$

$= \int_{\mu \in \Delta^V} p(X_1 = x_1, X_2 = x_2, \dots X_N = x_N, \mu) d\mu$

by the sum rule, applied to the joint probability $p(X, \mu)$.

$= \int_{\mu \in \Delta^V} p(X_1 = x_1, \dots X_N = x_N | \mu) p(\mu) d\mu$

by the product rule

$= \int_{\mu \in \Delta^V} \prod_{v=1}^{V} \mu_v^{n_v} \cdot \mathrm{Dir}(\mu | \alpha, \dots \alpha) d\mu$

substitute in the prior and the likelihood, using the simplifying expression for a likelihood of N iid categoricals from Problem 2a

$= \int_{\mu \in \Delta^V} \prod_{v=1}^{V} \mu_v^{n_v} \cdot \frac{\Gamma(V\alpha)}{\prod_{v=1}^{V} \Gamma(\alpha)} \prod_{v=1}^{V} \mu_v^{\alpha - 1} d\mu$

using the definition of the Dirichlet PDF

$= \frac{\Gamma(V\alpha)}{\prod_{v=1}^{V} \Gamma(\alpha)} \int_{\mu \in \Delta^V} \prod_{v=1}^{V} \mu_v^{n_v + \alpha - 1} d\mu$

grouping like exponents, moving terms const. wrt μ out of the integral

$= \frac{\Gamma(V\alpha)}{\prod_{v=1}^{V} \Gamma(\alpha)} \frac{\prod_{v=1}^{V} \Gamma(n_v + \alpha)}{\Gamma(N + V\alpha)}$

Using the identity for the unnormalized Dirichlet integral from Problem 1b