Student Name: Pengcheng Xu

Collaboration Statement:

Total hours spent: 15 hours

I discussed ideas with these individuals:

- I did it on my own

I consulted the following resources:

- The course's day4 PDF

By submitting this assignment, I affirm this is my own original work that abides by the course collaboration policy.

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

First, according to the definition of the expectation, we could write ρ's expectation as:

$$E[\rho] = \int_0^1 \rho \cdot C(a,b)\rho^{a-1}(1-\rho)^{b-1}\, d\rho = C(a,b) \cdot \int_0^1 \rho^{(a+1)-1}(1-\rho)^{b-1}\, d\rho$$

The first part: $C(a, b)$ we know is the constant w.r.t. ρ, it equals to $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$.

The second part: $\int_0^1 \rho^{(a+1)-1}(1-\rho)^{b-1}\, d\rho$, it equals to $\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)}$ based on Hint(3).

So, we could simplify the Expectation:

$$E[\rho] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{a \cdot \Gamma(a)\Gamma(b)}{\Gamma(a+b)(a+b)} = \frac{a}{a+b}$$

Here, we use the Hint(2) in the simplification process.

1b: Problem Statement

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

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

First, based on the definitioin of the expectation, we could write μ_w's expectation as:

$$E[\mu_w] = \int_0^1 \mu_w \cdot C(a_1, a_2, ...a_V) \cdot \mu_1^{a_1-1} \mu_2^{a_2-1} ... \mu_w^{a_w-1} ... \mu_V^{a_V-1} \, d\mu_w$$
$$= C(a_1, a_2, ...a_V) \cdot \int_0^1 \mu_1^{a_1-1} \mu_2^{a_2-1} ... \mu_w^{a_w+1-1} ... \mu_V^{a_V-1} \, d\mu_w$$

The first part: $C(a_1, a_2, ...a_V)$, equals to $\frac{\Gamma(a_1+a_2+...+a_V)}{\prod_{v=1}^{V} \Gamma(a_v)}$, according to Dirichlet's PDF.

The second part: $\int_0^1 \mu_1^{a_1-1} \mu_2^{a_2-1} ... \mu_w^{a_w+1-1} ... \mu_V^{a_V-1} \, d\mu_w$, using Hint(4) above, is transformed to $\frac{\Gamma(a_w+1) \cdot \prod_{v=1, v \neq w}^{V} \Gamma(a_v)}{\Gamma(a_1+a_2+...+a_w+1+...+a_V)} = \frac{a_w \cdot \prod_{v=1}^{V} \Gamma(a_v)}{\Gamma(a_1+a_2+...+a_V+1)}$. Here we use $\Gamma(x+1) = x \cdot \Gamma(x)$

So, we could simplify the Expectation (Using Hint(2) above):

$$E[\mu_w] = \frac{\Gamma(a_1 + a_2 + ... + a_V)}{\prod_{v=1}^{V} \Gamma(a_v)} \cdot \frac{a_w \cdot \prod_{v=1}^{V} \Gamma(a_v)}{\Gamma(a_1 + a_2 + ... + a_V + 1)}$$
$$= \frac{a_w}{a_1 + a_2 + ... + a_V} = \frac{a_w}{\sum_1^V a_v}$$

2a: Problem Statement

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

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

2a: Solution

Next, we'll conclude our result step-by-step, and also comment the basis of our derivation on each line:

/* each observation is conditionally independent of others on μ */

$$p(X_1 = x_1, ..., X_N = x_N|\mu) = \prod_{n=1}^{N} p(X_n = x_n|\mu)$$

/* each observation is identically distributed from a categorical distribution μ */

$$= \prod_{n=1}^{N} Cat(X_n = x_n|\mu)$$

/* we define x_{nv} as a one-hot function. i.e. x_{nv} = 1 if x_n is in type v; x_{nv} = 0 otherwise */

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

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

$$= \prod_{v=1}^{V} \mu_v^{n_v}$$

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:

$$p(X_* = v|X_1 = x_1 \dots X_N = x_n) = \int p(X_* = v, \mu|X_1 = x_1 \dots X_N = x_n)d\mu$$
$$= \frac{n_v + \alpha}{N + V\alpha} \tag{6}$$

2b: Solution

Next, we'll conclude our result step-by-step, and also comment the basis of our derivation on each line:

$$p(X_* = v|X_1 = x_1...X_N = x_N) = \int p(X_* = v, \mu|X_1 = x_1...X_N = x_n)\, du$$

/* Joint prob = Margin prob x conditional prob */

$$= \int p(X_* = v|\mu, X_1 = x_1...X_N = x_n) \cdot p(\mu|X_1 = x_1...X_N = x_n)\, du$$

/* X_* is i.i.d and conditionally independent on μ*/

$$= \int p(X_* = v|\mu) \cdot p(\mu|X_1 = x_1...X_N = x_n)\, du$$

$$= \int \mu_v \cdot p(\mu|X_1 = x_1...X_N = x_n)\, du$$

/* posterior of μ is under Dirichlet distribution*/

$$= \int \mu_v \cdot C(\hat{a_1}...\hat{a_V}) \cdot \mu_1^{\hat{a_1}-1}\mu_2^{\hat{a_2}-1}...\mu_V^{\hat{a_V}-1}\, du$$

$$= C(\hat{a_1}...\hat{a_V}) \cdot \int \mu_1^{\hat{a_1}-1}\mu_2^{\hat{a_2}-1}...\mu_v^{\hat{a_v}}...\mu_V^{\hat{a_V}-1}\, du$$

/* Using the definition of Dirichlet's PDF, Hint (4), and $\Gamma(x+1) = x \cdot \Gamma(x)$ */

$$= \frac{\Gamma(\hat{a_1} + \hat{a_2} + ... + \hat{a_V})}{\prod_{i=1}^{V} \Gamma(\hat{a_i})} \cdot \frac{\hat{a_v} \cdot \prod_{i=1}^{V} \Gamma(\hat{a^i})}{\Gamma(\hat{a_1} + \hat{a_2} + ... + \hat{a_V} + 1)}$$

$$= \frac{\hat{a_v}}{\sum_{i=1}^{V} \hat{a_i}}$$

/* Using μ is under symmetric Dirichlet Distribution (i.e.$p(\mu) = Dir(a, a, ...a)$), and the denotation of $\hat{a_v} = a + n_v$ */

$$= \frac{a + n_v}{\sum_{i=1}^{V}(a + n_i)}$$

$$= \frac{a + n_v}{Va + N}$$

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:

$$p(X_1 = x_1 \dots X_N = x_N) = \int p(X_1 = x_1, \dots X_N = x_N, \mu) d\mu \tag{7}$$

$$= \frac{\Gamma(V\alpha) \prod_{v=1}^{V} \Gamma(n_v + \alpha)}{\Gamma(N + V\alpha) \prod_{v=1}^{V} \Gamma(\alpha)} \tag{8}$$

2c: Solution

We'll derive our result step-by-step, and also comment the basis of our derivation on each line:

/* According to the relationship among joint prob, marginal prob, and conditional prob */

$$p(X_1 ... X_N) = \int p(X_1 ... X_N, \mu)\, du$$

$$= \int p(\mu) \cdot p(X_1 ... X_N | \mu)\, du$$

/* Plug in $\mu's$ PDF and Categorical distribution's PDF */

$$= \int C(a_1 ... a_V) \cdot \mu_1^{a_1 - 1} ... \mu_V^{a_V - 1} \cdot \prod_{v=1}^{V} \mu_v^{n_v}\, du$$

$$= C(a_1 ... a_V) \cdot \int \mu_1^{a_1 + n_1 - 1} ... \mu_V^{a_V + n_V - 1}\, du$$

/* In following equations, define $\hat{a}_i = a_i + n_i$ */

$$= C(a_1 ... a_V) \cdot \frac{1}{C(\hat{a}_1 ... \hat{a}_V)}$$

/* Expanding $C(a_1 ... a_V)$ using Gamma function */

$$= \frac{\Gamma(a_1 + ... + a_V)}{\prod_{v=1}^{V} \Gamma(a_v)} \cdot \frac{\prod_{v=1}^{V} \Gamma(\hat{a}_v)}{\Gamma(\hat{a}_1 + ... + \hat{a}_V)}$$

$$= \frac{\Gamma(a_1 + ... + a_V)}{\prod_{v=1}^{V} \Gamma(a_v)} \cdot \frac{\prod_{v=1}^{V} \Gamma(n_v + a_v)}{\Gamma(a_1 + n_1 + ... + a_V + n_V)}$$

/* Using the given assumption that μ is in symmetric Dirichlet distribution, and $N = n_1 + ... + n_V$ */

$$= \frac{\Gamma(Va)}{\prod_{v=1}^{V} \Gamma(a)} \cdot \frac{\prod_{v=1}^{V} \Gamma(n_v + a_v)}{\Gamma(Va + N)}$$