HW1: Beta, Dirichlet, and Estimators of Unigram Probability

Last modified: 2023-01-31 13:48

Due date: Thu Feb 09 at 11:59pm ET

Status: **RELEASED.**

How to turn in: Submit PDF to https://www.gradescope.com/courses/496674/assignments/2582224/

Jump to: Problem 1 Problem 2

Questions?: Post to the hw1 topic on the Piazza discussion forums.

Instructions for Preparing your PDF Report

What to turn in: PDF of typeset answers via LaTeX. No handwritten solutions will be accepted, so that grading can be speedy and you get prompt feedback.

Please use provided LaTeX Template: https://github.com/tufts-ml-courses/cs136-23s-assignments/blob/main/unit1_HW/hw1_template.tex

Your PDF should include (in order):

- Cover page with your full name and Collaboration statement
- Problem 1 answer
- Problem 2 answer

When you turn in the PDF to gradescope, mark each part via the in-browser Gradescope annotation tool)

How to write your solutions

Throughout this homework, we are practicing the skills necessary to derive, analyze, and apply formal mathematical statements involving probabiility.

Each step of a mathematical derivation that you turn in should be:

- legible
- justified by at least an accompanying short phrase (e.g. "using Bayes rule" or "by the identity 2.15 in the textbook")

Solutions that lack justifications or skip key steps without showing work will receive poor marks.

Problem 1: Mean for the Beta and Dirichlet

1a

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

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

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

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

$$\Gamma(a+1) = a\Gamma(a)$$

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

If you are curious, the derivation of the last identity can be found in the textbook in Exercise 2.5. But you can use these identities without understanding the derivation.

1b

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

Show that $\mathbb{E}[\mu_v] = \frac{a_v}{\sum_{w=1}^{V} a_w}$.

Hint: You can use the identity:

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

Where the integral is over the V-dimensional probability simplex (the set of vectors of size V that are non-negative and sum to one). We denote this as $\Delta^V \subset \mathbb{R}^V$.

Problem 2: Bayesian estimation for unigram probabilities

Consider a model with the following random variables:

- $X_1, \dots X_N$: N observable words, each one a known term in a finite vocabulary of size V.

Each observation can be represented as an integer index into the vocabulary: $x_n \in \{1, 2, \dots V\}$.

- $\mu = [\mu_1 \dots \mu_V]$: a parameter vector indicating the probability of each of the V possible terms

The vector $\mu \in \Delta^V$ has V non-negative entries, and the sum of this vector must equal one: $\sum_v \mu_v = 1$

Prior distribution on our parameter

We assume a symmetric Dirichlet prior distribution on the vector μ:

$$p(\mu) = \text{Dir}(\alpha, \alpha, \dots \alpha)$$

where we assume a known scalar hyperparameter $\alpha > 0$.

Likelihood of word data given parameter

We model any list of words by making the assumption that each word is **conditionally independent** of the other words given a parameter vector μ:

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

We further assume that each individual word is **identically distributed** from a Categorical distribution that defines its probabilities via parameter vector μ

$$p(X_n = x_n | \mu) = \text{Cat}(X_n = x_n | \mu)$$

which is equivalent to enumerating the probability of each possible term in our vocabulary as follows:

$$\begin{aligned} p(X_n = 1 | \mu) &= \mu_1 \\ &\vdots \\ p(X_n = v | \mu) &= \mu_v \\ &\vdots \\ p(X_n = V | \mu) &= \mu_V \end{aligned}$$

Posterior distribution over parameters given data

As shown in Section 2.2 of the textbook, when we have a Dirichlet prior $p(\mu)$ and a Categorical likelihood $p(X_1, \dots X_N | \mu)$, the posterior over μ is ALSO Dirichlet distribution (see Eq. 2.41).

Next-word prediction

In addition to the N observed training words, we also model a "new" word X_* that appears immediately after the N original words.

You should assume:

- word X_* is also *conditionally independent* of other words given μ
- word X_* is *identically distributed* like the other words given μ: $p(X_* = v | \mu) = \mu_v$

2a

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

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

where n_v is the non-negative integer that counts how often vocabulary term v appears in the training data: $n_v = \sum_{n=1}^{N} [x_n = v]$. Where the bracket expression is 1 if the expression inside is true, and 0 otherwise. This is commonly called Iverson bracket notation: https://en.wikipedia.org/wiki/Iverson_bracket

2b

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

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

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

Hint: Use the known definition of the posterior $p(\mu|x_1, \ldots x_N)$ given in the textbook Eq. 2.41.

2c

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:

$$\begin{aligned} p(X_1 = x_1 \ldots X_N = x_N) &= \int p(X_1 = x_1, \ldots X_N = x_N, \mu) d\mu \\ &= \frac{\Gamma(V\alpha) \prod_{v=1}^{V} \Gamma(n_v + \alpha)}{\Gamma(N + V\alpha) \prod_{v=1}^{V} \Gamma(\alpha)} \end{aligned}$$
