COMS W4903: Machine Learning for Data Science - Homework 1

*Homework 1: Due February 5, 2017 by midnight**

Danning Sui†
Data Science Institute

February 5, 2017

Contents

1 Problem 1 (written) - *25 points*

Imagine we have a sequence of N observations $(x_1, ..., x_N)$, where each $x_i \in \{0, 1, 2, ..., \infty\}$. We model this sequence as *i.i.d.* random variables from a negative binomial distribution with unknown parameter $\pi \in [0, 1]$and known parameter $r > 0$, where

$$p(x_i = j|\pi, r) = \binom{x_i + r - 1}{x_i} \pi^{x_i} (1 - \pi)^r$$

(a) What is the joint likelihood of the data $(x_1, ..., x_N)$?

*Columbia University, Spring 2017
†uni:ds3516

Joint Likelihood

$$
\begin{aligned}
p(x_1, x_2, ..., x_N|\pi, r) &= \prod_{i=1}^{N} p(x_i|\pi, r) \\
&= \prod_{i=1}^{N} [\binom{x_i + r - 1}{x_i} \pi^{x_i}(1-\pi)^r] \\
&= \prod_{i=1}^{N} [\frac{(x_i + r - 1)!}{x_i!(r-1)!} \pi^{x_i}(1-\pi)^r] \\
&= [\prod_{i=1}^{N} \frac{(x_i + r - 1)!}{x_i!(r-1)!}] \pi^{\sum x_i}(1-\pi)^{Nr}
\end{aligned}
$$

(b) Derive the maximum likelihood estimate $\hat{\pi}_{ML}$ for π .

We've already known parameter r, so the joint parameter can be seen as $p(x_1, x_2, ..., x_N|\pi)$ with r's value plugged in.

Maximum Likelihood Estimate

$$
\hat{\pi}_{ML} = argmax_{\pi} \prod_{i=1}^{N} p(x_i|\pi) = argmax_{\pi} ln(\prod_{i=1}^{N} p(x_i|\pi)) = argmax_{\pi} \sum_{i=1}^{N} lnp(x_i|\pi)
$$

To then solve for $\hat{\pi}_{ML}$, find

$$
\nabla_{\pi} \sum_{i=1}^{N} lnp(x_i|\pi) = \sum_{i=1}^{N} \nabla_{\pi} lnp(x_i|\pi) = 0
$$

First take the gradient with respect to π

$$\begin{aligned}
0 &= \nabla_\pi \sum_{i=1}^{N} ln[\frac{(x_i+r-1)!}{x_i!(r-1)!}\pi^{x_i}(1-\pi)^r] \\
&= \sum_{i=1}^{N} \nabla_\pi ln[\frac{(x_i+r-1)!}{x_i!(r-1)!}\pi^{x_i}(1-\pi)^r] \\
&= \sum_{i=1}^{N}[x_i\frac{1}{\pi} - r\frac{1}{1-\pi}] \\
&= \sum_{i=1}^{N} x_i\frac{1}{\pi} - Nr\frac{1}{1-\pi} \\
&= \sum_{i=1}^{N} x_i(1-\pi) - Nr\pi \\
&= (-Nr - \sum_{i=1}^{N} x_i)\pi + \sum_{i=1}^{N} x_i \\
&\implies \hat{\pi}_{ML} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i + Nr} = \frac{\bar{x}}{\bar{x}+r}
\end{aligned}$$

To help learn π, you use a prior distribution. You select the distribution $p(\pi) = Beta(a,b)$.

(c) Derive the maximum a posteriori (MAP) estimate $\hat{\pi}_{MAP}$ for π?

$$\begin{aligned}
\hat{\pi}_{MAP} &= argmax_\pi \ p(\pi|x) \\
&= argmax_\pi \ p(x|\pi)p(\pi) \\
&= argmax_\pi \prod_{i=1}^{N} p(x_i|\pi)p(\pi) \\
&= argmax_\pi \sum_{i=1}^{N} lnp(x_i|\pi) + lnp(\pi)
\end{aligned}$$

To then solve for $\hat{\pi}_{MAP}$, find

$$\nabla_\pi[\sum_{i=1}^{N} lnp(x_i|\pi) + lnp(\pi)] = 0$$

First take the gradient with respect to π

$$
\begin{aligned}
0 &= \nabla_\pi \sum_{i=1}^{N} ln[\frac{(x_i+r-1)!}{x_i!(r-1)!}\pi^{x_i}(1-\pi)^r] + \nabla_\pi ln[\frac{1}{B(a,b)}\pi^{a-1}(1-\pi)^{b-1}] \\
&= \sum_{i=1}^{N} x_i \frac{1}{\pi} - Nr\frac{1}{1-\pi} + (a-1)\frac{1}{\pi} - (b-1)\frac{1}{1-\pi} \\
&= (\sum_{i=1}^{N} x_i + a - 1)\frac{1}{\pi} + (-Nr-b+1)\frac{1}{1-\pi} \\
&= (\sum_{i=1}^{N} x_i + a - 1)(1-\pi) + (-Nr-b+1)\pi \\
&= (\sum_{i=1}^{N} x_i + a - 1) + (2 - Nr - \sum_{i=1}^{N} x_i - b - a)\pi
\end{aligned}
$$

$$
\implies \hat{\pi}_{MAP} = \frac{\sum_{i=1}^{N} x_i + a - 1}{\sum_{i=1}^{N} x_i + Nr + a + b - 2} = \frac{\bar{x} + \frac{a-1}{N}}{\bar{x} + r + \frac{a+b-2}{N}}
$$

(d) Use Bayes rule to derive the posterior distribution of π and identify the name of this distribution. The posterior distribution of π is

$$
\begin{aligned}
p(\pi|x) &\propto p(x|\pi)p(\pi) \\
&\propto \prod_{i=1}^{N}[\binom{x_i+r-1}{x_i}\pi^{x_i}(1-\pi)^r]\frac{1}{B(a,b)}\pi^{a-1}(1-\pi)^{b-1} \\
&\propto \pi^{\sum x_i + a - 1}(1-\pi)^{Nr+b-1}
\end{aligned}
$$

$$
\implies p(\pi|x) \sim Beta(\sum_{i=1}^{N} x_i + a, Nr + b)
$$

(e) What is the mean and variance of π under this posterior? Discuss how it relates to $\hat{\pi}_{ML}$ and $\hat{\pi}_{MAP}$. Since π belongs to a $Beta$ distribution,

$$
E[\pi|x] = \frac{\sum_{i=1}^{N} x_i + a}{\sum_{i=1}^{N} x_i + Nr + a + b}
$$

$$
Var[\pi|x] = \frac{(\sum_{i=1}^{N} x_i + a)(Nr+b)}{(\sum_{i=1}^{N} x_i + Nr + a + b)^2(\sum_{i=1}^{N} x_i + Nr + a + b + 1)}
$$

Here we can find that $\hat{\pi}_{ML}$ is the mean of distribution $Beta(0,0)$, while the $\hat{\pi}_{MAP}$ is the mean of distribution $Beta(a-1, b-1)$.