CS5487 Machine Learning

Semester A 2024/25
Midterm Solutions

Problem 1 MLE for Binomial Distribution [35 marks]

(a) [5 marks] The data log-likelihood is

$$\begin{aligned}
\log p(\mathcal{D}|\pi) &= \sum_{i=1}^{N} \log p(x_i|\pi) && (1)\\
&= \sum_{i=1}^{N} \log[\binom{n}{x_i}\pi^{x_i}(1-\pi)^{n-x_i}] && (2)\\
&= \sum_{i=1}^{N} \left[x_i \log\pi + (n-x_i)\log(1-\pi) + \log\binom{n}{x_i}\right] && (3)\\
&= \sum_{i=1}^{N} x_i \log\pi + (Nn - \sum_{i=1}^{N} x_i)\log(1-\pi) + \sum_{i=1}^{N}\log\binom{n}{x_i} && (4)\\
&= S\log\pi + (Nn-S)\log(1-\pi) + C && (5)
\end{aligned}$$

where $S = \sum_{i=1}^{N} x_i$ and $C = \sum_{i=1}^{N} \log\binom{n}{x_i}$.

(b) [5 marks] The MLE optimization problem is:

$$\begin{aligned}
\hat{\pi} &= \operatorname*{argmax}_{\pi}\ \log p(\mathcal{D}|\pi) && (6)\\
&= \operatorname*{argmax}_{\pi}\ S\log\pi + (Nn-S)\log(1-\pi) + C && (7)\\
&= \operatorname*{argmax}_{\pi}\ S\log\pi + (Nn-S)\log(1-\pi) && (8)
\end{aligned}$$

(c) [20 marks] Taking the derivative of the objective and setting to zero,

$$\frac{\partial}{\partial\pi}\{S\log\pi + (Nn-S)\log(1-\pi)\} = S\frac{1}{\pi} + (Nn-S)\frac{-1}{1-\pi} = 0 \quad (9)$$

Multiplying both sides by $\pi(1-\pi)$,

$$\begin{aligned}
S(1-\pi) - (Nn-S)\pi &= 0 && (10)\\
S - S\pi - Nn\pi + S\pi &= 0 && (11)\\
S - Nn\pi &= 0 && (12)\\
\Rightarrow \hat{\pi} = \frac{S}{Nn} &= \frac{1}{Nn}\sum_{i=1}^{N} x_i && (13)
\end{aligned}$$

(d) [5 marks] In the MLE solution, $S = \sum_i x_i$ is the total number of successes observed in N sequences, and Nn is the total number of trials in N sequences. Thus, $\pi = \frac{S}{Nn}$ is the fraction of successes observed in all trials from all sequences.

. . . .

Problem 2 MAP for Binomial distribution [35 marks]

(a) [5 marks] The MAP optimization problem is

$$\hat{\pi} = \underset{\pi}{\operatorname{argmax}} \ \log p(\pi|\mathcal{D}) \tag{14}$$

$$= \underset{\pi}{\operatorname{argmax}} \ \log p(\mathcal{D}|\pi) + \log p(\pi) \tag{15}$$

$$= \underset{\pi}{\operatorname{argmax}} \ S \log \pi + (Nn - S) \log(1 - \pi) + C + (\alpha - 1) \log \pi + (\beta - 1) \log(1 - \pi) + D \tag{16}$$

$$= \underset{\pi}{\operatorname{argmax}} \ (S + \alpha - 1) \log \pi + (Nn - S + \beta - 1) \log(1 - \pi) \tag{17}$$

where S and C are defined as in Problem 1, and $D = \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$.

(b) [15 marks] Taking the derivative and setting to zero,

$$\frac{\partial}{\partial \pi} \{(S + \alpha - 1) \log \pi + (Nn - S + \beta - 1) \log(1 - \pi)\} \tag{18}$$

$$= (S + \alpha - 1)\frac{1}{\pi} + (Nn - S + \beta - 1)\frac{-1}{1 - \pi} = 0. \tag{19}$$

Multiplying both sides by $\pi(1 - \pi)$,

$$(S + \alpha - 1)(1 - \pi) - (Nn - S + \beta - 1)\pi = 0 \tag{20}$$

$$(S + \alpha - 1) - (Nn + \beta - 1 + \alpha - 1)\pi = 0 \tag{21}$$

$$\Rightarrow \hat{\pi} = \frac{S + \alpha - 1}{Nn + \alpha + \beta - 2} \tag{22}$$

where $S = \sum_i x_i$.

(c) [10 marks] The MAP estimator has an additional term $(\alpha - 1)$ in the numerator and $(\beta + \alpha - 2)$ in the denominator. We can obtain the following interpretation of adding a set of *virtual sequences* to the data before computing the MLE solution:

- $(\alpha - 1)$ is the number of "success" trials in the virtual sequences.
- $(\beta - 1)$ is the number of "failure" trials in the virtual sequences.
- $(\alpha + \beta - 2)$ is the total number of trials in all the virtual sequences.

(d) [5 marks] As N increases, then S also increases. The N and S in the numerator and denominator will override the virtual samples. In particular, if $N \gg \alpha$ and $S \gg \beta + \alpha$ then we obtain the MLE solution.

....

Problem 3 Bayesian estimation for Negative Binomial distribution [30 marks]

(a) [5 marks] Using Bayes' rule,

$$p(\pi|\mathcal{D}) = \frac{p(\mathcal{D}|\pi)p(\pi)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\pi)p(\pi)}{\int p(\mathcal{D}|\pi)p(\pi)d\pi} \tag{23}$$

$$\propto \left[\prod_{i=1}^{N} \pi^{x_i}(1 - \pi)^{n - x_i}\right] \pi^{\alpha - 1}(1 - \pi)^{\beta - 1} \tag{24}$$

where we have ignored constants that are not a function of π.

(b) [15 marks] It suffices to look at the $p(\pi|\mathcal{D})$ as a function of π to obtain the form of the posterior distribution,

$$\begin{aligned} p(\pi|\mathcal{D}) &\propto \left[\prod_{i=1}^{N} \pi^{x_i}(1-\pi)^{n-x_i}\right] \pi^{\alpha-1}(1-\pi)^{\beta-1} &(25)\\ &= \pi^{\sum_i x_i}(1-\pi)^{Nn-\sum_i x_i}\pi^{\alpha-1}(1-\pi)^{\beta-1} &(26)\\ &= \pi^{S+\alpha-1}(1-\pi)^{Nn-S+\beta-1} &(27)\end{aligned}$$

This has the form of an unnormalized Beta distribution with parameters $\hat{\alpha} = S + \alpha$ and $\hat{\beta} = Nn - S + \beta$. Thus the posterior is

$$p(\pi|\mathcal{D}) = \text{Beta}(\pi|S + \alpha, Nn - S + \beta), \tag{28}$$

where $S = \sum_i x_i$.

(c) [5 marks] The parameters of the posterior are $\hat{\alpha} = S + \alpha$ and $\hat{\beta} = Nn - S + \beta$. We have the following interpretation based on *virtual samples*:

- $\hat{\alpha} = S + \alpha$ is the number of successful trials, where S is the number of observed successful trials in the data, and α is the number of successful virtual trials.
- $\hat{\beta} = Nn - S + \beta$ is the number of failure trials: $Nn - S$ is the number of observed unsuccessful trials in the data, and β is the number of unsuccessful virtual trials.

(d) [10 marks] What happens to the posterior as the number of samples N increases?

We have the following statistics of the posterior of $\hat{\pi}$,

- The mean is: $E[\hat{\pi}] = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}} = \frac{S+\alpha}{Nn+\alpha+\beta}$
- The mode is: $\text{mode}(\hat{\pi}) = \frac{\hat{\alpha}-1}{\hat{\alpha}+\hat{\beta}-2} = \frac{S+\alpha-1}{Nn+\alpha+\beta-2}$, which is the same as the MAP solution (the maximum of the posterior).
- The variance is

$$\text{var}(\hat{\pi}) = \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha}+\hat{\beta})^2(\hat{\alpha}+\hat{\beta}+1)} = \frac{(S+\alpha)(Nn-S+\beta)}{(Nn+\alpha+\beta)^2(Nn+\alpha+\beta+1)}. \tag{29}$$

If N increases and $Nn \gg \alpha, \beta$, then $E[\hat{\pi}] \approx \frac{S}{Nn}$ and $\text{mode}(\hat{\pi}) \approx \frac{S-1}{Nn-2} \approx \frac{S}{Nn}$. Thus as N increases, the posterior mean and mode converge to the MLE solution in Problem 1.

The variance is a function like $\frac{O(NnS))}{O((Nn)^3)}$. Thus, the variance decays as $\text{var}(\hat{\pi}) = 1/O(N^2)$. Thus, as N increases, the variance of the posterior decreases and eventually converges to 0.

Putting the two together, as N increases, the posterior converges to a delta function on the MLE solution (the mean converges to the MLE value, and the variance converges to 0).

....

— End of Midterm —