Computer Science Department - Rutgers University **Spring 2021**

Statistics and Bandit States of Knowledge

16:198:605

Suppose you have k bandits. For bandit i, you have collected samples $X^i_1, \ldots, X^i_{k_i}$. What do you really know about the bandits?

If you have the additional assumption that each bandit has a normal distribution with some (unknown) mean μ_i and some (known) variance σ_i^2, then for any choice of $(\mu_1, \ldots, \mu_k)$, we can assess the *likelihood* of those means based on the data:

$$\begin{aligned}
\text{lik}(\mu_1, \ldots, \mu_k) &= \prod_{i=1}^{k} \prod_{t=1}^{k_i} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}\left(\mu_i - X^i_t\right)^2\right) \\
&= \prod_{i=1}^{k} \left[\frac{1}{\sigma_i\sqrt{2\pi}}\right]^{k_i} \exp\left(-\frac{1}{2\sigma_i^2}\left(\sum_{t=1}^{k_i}\left(\mu_i - X^i_t\right)^2\right)\right) \\
&= \prod_{i=1}^{k} \left[\frac{1}{\sigma_i\sqrt{2\pi}}\right]^{k_i} \exp\left(-\frac{1}{2\sigma_i^2}\left(k_i\mu_i^2 - 2\mu_i\sum_{t=1}^{k_i} X^i_t + \sum_{t=1}^{k_i}\left(X^i_t\right)^2\right)\right) \\
&= \prod_{i=1}^{k} \left[\frac{1}{\sigma_i\sqrt{2\pi}}\right]^{k_i} \exp\left(-\frac{1}{2\sigma_i^2}\left(k_i\mu_i^2 - 2\mu_i\hat{\mu}_i k_i + \sum_{t=1}^{k_i}\left(X^i_t\right)^2\right)\right),
\end{aligned} \tag{1}$$

where in the above, we are replacing the sum of the X^i_t with $\hat{\mu}_i * k_i$.

At this point, it is convenient to complete the square:

$$\begin{aligned}
\text{lik}(\mu_1, \ldots, \mu_k) &= \prod_{i=1}^{k} \left[\frac{1}{\sigma_i\sqrt{2\pi}}\right]^{k_i} \exp\left(-\frac{1}{2\sigma_i^2}\left(k_i\mu_i^2 - 2\mu_i\hat{\mu}_i k_i + \sum_{t=1}^{k_i}\left(X^i_t\right)^2\right)\right) \\
&= \prod_{i=1}^{k} \left[\frac{1}{\sigma_i\sqrt{2\pi}}\right]^{k_i} \exp\left(-\frac{1}{2\sigma_i^2}\left(k_i\mu_i^2 - 2\mu_i\hat{\mu}_i k_i + k_i\hat{\mu}_i^2 - k_i\hat{\mu}_i^2 + \sum_{t=1}^{k_i}\left(X^i_t\right)^2\right)\right) \\
&= \prod_{i=1}^{k} \left[\frac{1}{\sigma_i\sqrt{2\pi}}\right]^{k_i} \exp\left(-\frac{1}{2\sigma_i^2}\left(k_i(\mu_i - \hat{\mu}_i)^2 - k_i\hat{\mu}_i^2 + \sum_{t=1}^{k_i}\left(X^i_t\right)^2\right)\right)
\end{aligned} \tag{2}$$

At this point, we can factor out all terms that do not depend on the unknown $\mu_1, \ldots, \mu_k$, leaving

$$\text{lik}(\mu_1, \ldots, \mu_k) \propto \prod_{i=1}^{k} \exp\left(-\frac{k_i}{2\sigma_i^2}(\mu_i - \hat{\mu}_i)^2\right) \tag{3}$$

The interpretation of this is effectively two-fold, maybe three-fold depending on who is counting:

- The 'posterior' distribution for each μ_i is independent (factors nicely), normally distributed, centered at $\hat{\mu}_i$ with variance σ_i^2/k_i.
- The full 'state of knowledge' of $(\mu_1, \ldots, \mu_k)$ at this time is described completely in terms of the vector of sample means $(\hat{\mu}_1, \ldots, \hat{\mu}_k)$ and sample counts $(k_1, \ldots, k_k)$.

This suggests that these two vectors taken together are sufficient to describe the state of knowledge at any point in time for this bandit problem. Note that the variances $\{\sigma_i^2\}$ are effectively taken as 'known' and thus constant throughout, and do not need to be included. If the variances were *unknown*, then a similar factoring would show that the full state of knowledge is given by the sample means, the sample variances, and the counts - but again, these statistics are only sufficient based on the assumption that the bandits are normal.

One thing to consider here (and I have some thoughts on the matter) is rather than declaring in advance that a collection of statistics are useful / representative - can we train a network that collects/computes relevant statistics on the fly? Imagine an RNN taking a new sample each time and updating some internal state to reflect aspects of all samples collected so far, and using this as the 'state of knowledge' for the bandit problem.

Additional Information

What if we had additional information, for instance that all the means had to sum to 0?

In this case, the likelihood of a given vector needs to be modified accordingly:

$$\text{lik}(\mu_1, \ldots, \mu_k) \propto \begin{cases} \prod_{i=1}^{k} \exp\left(-\frac{k_i}{2\sigma_i^2}(\mu_i - \hat{\mu}_i)^2 \right) & \text{if } \mu_1 + \ldots + \mu_k = 0 \\ 0 & \text{else.} \end{cases} \tag{4}$$

This is a much harder thing to summarize in a single vector / knowledge state - we 'know' the means must sum to 0, but how can we represent this as input to a network?

One possibility would be that instead of thinking about sufficient statistics (that give full information), we think about *maximum likelihood estimators*: What are the μ_i that are most likely given this data?

Without the additional information, the maximum likelihood estimators for $(\mu_1, \ldots, \mu_k)$ are simply $(\hat{\mu}_1, \ldots, \hat{\mu}_k)$ (*why?*). But given that there are no requirements that the sample means sum to 0, this cannot be the maximum likelihood estimator in this new situation.

So consider the problem of maximizing the likelihood subject to the constraint that $\sum_i \mu_i = 0$. To simplify, consider maximizing the log of the likelihood, or equivalently:

$$\text{minimize } \sum_{i=1}^{k} \frac{k_i}{2\sigma_i^2}(\mu_i - \hat{\mu}_i)^2 \text{ subject to } \sum_i \mu_i = 0. \tag{5}$$

Using Lagrange multipliers, we want to find $\mu_1, \ldots, \mu_k, \lambda$ such that for each i, we have

$$\frac{k_i}{\sigma_i^2}(\mu_i - \hat{\mu}_i) + \lambda 1 = 0, \tag{6}$$

and $\mu_1 + \ldots + \mu_k = 0$.

From the first condition, we get that $\mu_i = \hat{\mu_i} - \lambda\sigma_i^2/k_i$. Plugging this into the second condition, we get that we want

$$\sum_{i=1}^{k} \left(\hat{\mu} - \lambda\sigma_i^2/k_i \right) = 0, \tag{7}$$

or

$$\sum_i \hat{\mu}_i - \lambda \sum_i \frac{\sigma_i^2}{k_i} = 0, \tag{8}$$

or

$$\lambda = \frac{\sum_i \hat{\mu}_i}{\sum_i \sigma_i^2/k_i} \tag{9}$$

This suggests that for any i, the maximum likelihood estimator for μ_i is given by

$$\mu_i^* = \hat{\mu_i} - \frac{\sigma_i^2}{k_i} \frac{\sum_j \hat{\mu}_j}{\sum_j \sigma_j^2/k_j} \tag{10}$$

You can see from the above that these estimators modify the base naive estimators $\hat{\mu}_i$ to something that necessarily satisfies the constraint (all estimators summing to 0). Additionally of interest, this reflects the fact that the bandits are no longer independent - the best estimator for any one bandit depends on data from all the other bandits. Hence data from any bandit is informative about all bandits.

This leads naturally to the idea then that the 'state of knowledge' in this case might be best summarized in the collection $(\mu_1^*, \ldots, \mu_k^*)$ and the sample counts $(k_1, \ldots, k_k)$.