CS 760: Machine Learning Spring 2024

Homework 6: Frequentists vs Bayesians

AUTHORS: Jed Pulley

DO NOT POLLUTE! AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

Problem 6.1. Frequentist (MLE)

To find the MLE of p^*, we first start with the likelihood function:

$$\mathbb{P}(p^*) = \prod_{i=1}^{n} (p^{\mathsf{x_i}}(1-p)^{1-\mathsf{x_i}})$$

Then we take the log of the likelihood function:

$$log(\mathbb{P}(p^*)) = logp\sum_{i=1}^{n} \mathsf{x_i} + log(1-p)\sum_{i=1}^{n}(1-\mathsf{x_i})$$

Using our optimization 101 technique, we get the derivative and set it equal to zero:

$$p_{MLE} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{x_i}$$

Which we recognize to just be the mean.

Problem 6.2. Bayesian (MAP)

We use our MAP formula $p_{MAP} = \arg\max_p \mathbb{P}(X|p)$ as our starting point. Using Bayes Rule, we can rearrange it as such:

$$p_{MAP} = \arg\max_p \frac{\mathbb{P}(X|p)\mathbb{P}(p)}{\mathbb{P}(X)}$$

Since we're maximizing for p, we can ignore the $\mathbb{P}(\mathbb{X})$ term as it doesn't depend on p:

$$p_{MAP} = \arg\max_p \mathbb{P}(X|p)\mathbb{P}(p)$$

This is very similar to our MLE statement above, with the exception of our prior term $\mathbb{P}(p)$ which we're assuming is information gathered about a previous event. Notably, if there is no prior information, our MAP estimate is equal to our MLE.

Given our prior $\mathbb{P}(p)$ being modeled as $Beta(\alpha, \beta)$, we can show our prior below:

$$\mathbb{P}(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

Using the prior information, the likelihood of our data (we'll say $X = [\mathsf{x_1}, \mathsf{x_2}, ..., \mathsf{x_N}]^\mathsf{T}$) can be written as:

$$\mathbb{P}(X|p) = p^{\sum_{i=1}^{N} \mathsf{x}_i}(1-p)^{N-\sum_{i=1}^{N} \mathsf{x}_i} = p^{\mathbf{1}^\mathsf{T}\mathsf{X}}(1-p)^{N-\mathbf{1}^\mathsf{T}\mathsf{X}}$$

Simplifying and matching with the form of a Beta distribution, we find our MAP estimator to be:

$$\hat{p}_{MAP} = \frac{\sum_{i=1}^{N} \mathsf{x_i} + \alpha - 1}{N + \alpha + \beta - 2}$$

Problem 6.3. (Correct prior)

Here I copied down the code from the lecture notes, modified them into Python, and then ran through a bunch of iterations to find the below numbers:

(a) Under the correct prior, MLE converges around $N = 115$.

(b) MAP converges within 0.01 faster than the MLE, around $N = 85$

In this instance, MAP performs better since we have a really accurate prior.

Problem 6.4. (Incorrect Prior)

(a) For the incorrect prior, MLE converges at $N = 115$.

(b) MAP takes much longer, and finally converges around $N = 550$.

These results make sense for the MLE since it is not affected by the incorrect prior. So in both cases (incorrect and correct), it remains the same. However, our MAP estimate varies wildly. It performs better in 6.3 since our prior is really accurate. However, in the incorrect prior case, it is way off.

Problem 6.5

Based on my results, which approach you use is highly dependant on how accurate your prior is. In the case where it was dead nuts accurate, MAP converaged within our confidence interval much faster than the MLE. However, when it was WAY off, MLE got there much faster. So in my opinion, it's entirely dependant on the situation.

Personally, I prefer the MLE approach since it doesn't rely on a prior, which in many instances is a naive assumption that you would have one to begin with. That being said, when you actually do have a prior (that is, a correct one), it's incredibly helpful.

The advantage of MLE is its simplicity and ease of use. Especially since in a lot of cases the MLE just turns out to be the mean. The disadvantage is the complement of the above: it's too simple and often just winds up being the mean. Using MLE can constrain you depending on the situation.

The advantage of MAP lies entirely in the assumption of a prior, especially given that MLE and MAP are exactly the same thing if there is no prior at all. Having the prior allows us to affectively leverage previous information to influence our decision. The disadvantage is that often times we don't have a prior or we need to make naive assumptions to use one.