CSE 312: Foundations of Computing II
Instructor: Alex Tsun
Date: 2/25/22
Lecture Topics: 7.5 Maximum a Posteriori Estimation

[**Tags:** MLE**,** MAP, Beta]

1. In PSet2, you implemented a Naive Bayes classifier as a spam filter. We estimated the quantities $P(Y = y)$ and $P(\text{word} \mid Y = y)$ for $y \in \{\text{spam}, \text{ham}\}$ and all words. Explain in a few sentences which estimation techniques we used (MLE or MAP) for each, and make sure to include which distribution's parameter we were estimating, and the prior distribution if we had one.

Solution: We used MLE to estimate $P(Y = y)$ and MAP to estimate $P(\text{word} \mid Y = y)$ with a $Beta(2,2)$ prior. They were both Bernoulli parameters.

[**Tags:** MAP, Beta]

2. Suppose $\boldsymbol{x} = (x_1, \ldots, x_n)$ are iid samples from $Geo(\Theta)$ where Θ is a random variable (not fixed).
   a. Using the prior $\Theta \sim Beta(\alpha, \beta)$ (for some arbitrary but known parameters $\alpha, \beta \geq 1$), show that the posterior distribution $\Theta|\boldsymbol{x}$ also follows a Beta distribution and identify its parameters (by computing $\pi_\Theta(\theta|\boldsymbol{x})$). Then, explain this sentence: "The Beta distribution is the conjugate prior for the rate parameter of the Geometric distribution". Hint: This can be done in just a few lines!
   b. Now derive the MAP estimate for Θ. Recall the mode of $W \sim Beta(\gamma, \eta)$ is $\frac{\gamma-1}{\gamma-1+\eta-1}$ (pretend you saw $\gamma - 1$ heads and $\eta - 1$ tails). Hint: This should be just one line using your answer to part (a).
   c. Explain how this MAP estimate differs from the MLE/MoM estimate (recall for the Poisson distribution it was just the inverse of the sample mean $\frac{n}{\sum_{i=1}^n x_i}$) and provide an interpretation of α and β as to how they affect the estimate.

Solution:

a. We know the posterior is proportional (in LaTeX, \propto) to likelihood times prior:

$$\pi_\Theta(\theta|x) \propto L(x|\theta)\pi_\Theta(\theta)$$

Then, the likelihood is just the product of geometric PMFs, and the prior is just the beta PDF (note again the $\propto$ because we drop the normalizing constant for Beta):

$$\propto \left(\prod_{i=1}^{n}(1-\theta)^{x_i-1}\theta\right) \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$= \left((1-\theta)^{\sum_{i=1}^n x_i - n}\theta^n\right) \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$= \theta^{(n+\alpha)-1}(1-\theta)^{(\sum_{i=1}^n x_i - n + \beta)-1}$$

Hence, our posterior $\Theta|\boldsymbol{x} \sim Beta(n + \alpha, \sum_{i=1}^n x_i - n + \beta)$.

b. The MAP estimate just comes from taking the mode which is

$$\frac{(n+\alpha)-1}{((n+\alpha)-1)+(\sum_{i=1}^{n} x_i - n + \beta - 1)} = \frac{n+\alpha-1}{\sum_{i=1}^{n} x_i + \alpha + \beta - 2}$$

c. The interpretation is: pretend you had $\alpha - 1$ extra samples (waiting for $\alpha - 1$ heads), where it took a total of $\alpha + \beta - 2$ trials to wait for (pretend we saw $\beta - 1$ tails in the process). This comes from staring at the MLE/MoM estimate and looking at the difference:

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