1 Definitions Ahoy!

1.1 MLE/MAP

1. **MLE:** Finds the best parameters for a specific dataset, $\mathcal{D}$. Specifically, we want to find the parameters $\hat{\theta}_{MLE}$ that maximize the likelihood for $\mathcal{D}$.

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta)$$

2. **MAP:** Finds the best parameters given $\mathcal{D}$ and a prior belief about the parameters. Specifically, we want to find the parameters $\hat{\theta}_{MAP}$ that maximize the posterior distribution $p(\theta \mid \mathcal{D})$ of parameters θ.

$$\begin{aligned}
\hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmax}}\, p(\theta \mid \mathcal{D}) \\
&= \underset{\theta}{\operatorname{argmax}}\, \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\text{Normalizing Constant}} \\
&= \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta)p(\theta) \\
&= \underset{\theta}{\operatorname{argmin}} - \log\left(p(\mathcal{D} \mid \theta)p(\theta)\right) \\
&= \underset{\theta}{\operatorname{argmin}} - \log p(\mathcal{D} \mid \theta) - \log p(\theta)
\end{aligned}$$

3. **MLE and MAP for conditional likelihood:** When we want to predict the output y given the input x using our supervised dataset, we have to reformulate the MLE and MAP optimizations to use the conditional likelihood (and conditional posterior) instead:

$$\begin{aligned}
\hat{\theta}_{MLE} &= \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta) \\
&= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right) \\
&= \underset{\theta}{\operatorname{argmin}} - \log \prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right) \\
&= \underset{\theta}{\operatorname{argmin}} - \sum_{i=1}^{N} \log p\left(y^{(i)} \mid x^{(i)}, \theta\right)
\end{aligned}$$

$$\begin{aligned}
\hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmax}}\, p(\theta \mid \mathcal{D}) \\
&= \underset{\theta}{\operatorname{argmax}} \left(\prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right)\right) p(\theta) \\
&= \underset{\theta}{\operatorname{argmin}} - \log \prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right) - \log p(\theta) \\
&= \underset{\theta}{\operatorname{argmin}} - \sum_{i=1}^{N} \log p\left(y^{(i)} \mid x^{(i)}, \theta\right) - \log p(\theta)
\end{aligned}$$

2 Anybody have a MAP?

Imagine you are a data scientist working for an advertising company. The advertising company has recently run an ad and they want you to estimate its performance. The ad was shown to N people. $y^{(i)} = 1$ if person i clicked on the ad and 0 otherwise. Thus $\sum_i^N y^{(i)} = N_1$ people decided to click on the ad. Assume that the probability that the i-th person clicks on the ad is ϕ and the probability that the i-th person does not click on the ad is $1 - \phi$.
Note

$$p(\mathcal{D} \mid \theta) = p\left(y^{(1)}, y^{(2)}, ..., y^{(N)} \mid \phi\right) = \prod_{i=1}^{N} p\left(y^{(i)} \mid \phi\right) = \phi^{N_1}(1-\phi)^{N-N_1}$$

1. Calculate $\hat{\phi}_{MLE}$.

$$\begin{aligned}
\hat{\phi}_{MLE} &= \operatorname*{argmin}_{\phi} -\log\left(p(\mathcal{D} \mid \phi)\right) \\
&= \operatorname*{argmin}_{\phi} -\log\left(\phi^{N_1}(1-\phi)^{N-N_1}\right) \\
&= \operatorname*{argmin}_{\phi} -N_1 \log(\phi) - (N - N_1)\log(1-\phi)
\end{aligned}$$

Setting the derivative equal to zero yields

$$\begin{aligned}
0 &= \frac{-N_1}{\phi} + \frac{(N - N_1)}{1-\phi} \\
\implies \phi_{MLE} &= \frac{N_1}{N}
\end{aligned}$$

2. Your coworker tells you that $\phi \sim \text{Beta}(\alpha, \beta)$. That is:

$$p(\phi) = \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{B(\alpha, \beta)}$$

Note that $B(\alpha, \beta)$ is not a function of ϕ and can be treated as a constant. Formulate the optimization of the log posterior, $\operatorname{argmin}_\phi -\log p(\phi \mid \mathcal{D})$, in terms of N, N_1, ϕ, α, and β.

$$\begin{aligned}
&\operatorname*{argmin}_{\phi} -\log p(\phi \mid \mathcal{D}) \\
&= \operatorname*{argmin}_{\phi} -\log\left(p(\mathcal{D} \mid \phi)p(\phi)\right) && \text{By Bayes rule and dropping } 1/p(\mathcal{D}) \\
&= \operatorname*{argmin}_{\phi} -\log p(\mathcal{D} \mid \phi) - \log p(\phi) \\
&= \operatorname*{argmin}_{\phi} -\log\left(\phi^{N_1}(1-\phi)^{N-N_1}\right) - \log\frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{B(\alpha,\beta)} \\
&= \operatorname*{argmin}_{\phi} -\log\phi^{N_1} - \log(1-\phi)^{N-N_1} - \log\phi^{\alpha-1} - \log(1-\phi)^{\beta-1} && \text{Drop } B(\alpha,\beta) \\
&= \operatorname*{argmin}_{\phi} -N_1\log\phi - (N-N_1)\log(1-\phi) - (\alpha-1)\log\phi - (\beta-1)\log(1-\phi) \\
&= \operatorname*{argmin}_{\phi} -(N_1+\alpha-1)\log\phi - (N-N_1+\beta-1)\log(1-\phi)
\end{aligned}$$

3. Now, calculate $\hat{\phi}_{MAP}$.

$$\begin{aligned}\hat{\phi}_{MAP} &= \text{argmin}_\phi - \log\left(p(\mathcal{D} \mid \phi)p(\phi)\right) \\ &= \text{argmin}_\phi - (N_1 + \alpha - 1)\log(\phi) - (N - N_1 + \beta - 1)\log(1-\phi)\end{aligned}$$

Setting the derivative equal to zero yields

$$\begin{aligned}0 &= \frac{-N_1 - \alpha + 1}{\phi} + \frac{(N - N_1 + \beta - 1)}{1-\phi} \\ \implies \phi_{MAP} &= \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2}\end{aligned}$$

4. Suppose $N = 100$ and $N_1 = 10$. Furthermore, you believe that in general people click on ads about 6 percent of the time, so you, somewhat naively, decide to set $\alpha = 6 + 1 = 7$, and $\beta = 100 - 6 + 1 = 95$. Calculate $\hat{\phi}_{MAP}$.

$\hat{\phi}_{MAP} = \frac{N_1+\alpha-1}{N+\alpha+\beta-2} = \frac{10+7-1}{100+102-2} = \frac{16}{200} = 0.08$

5. How do $\hat{\phi}_{MLE}$ and $\hat{\phi}_{MAP}$ differ? Argue which estimate you think is better.

Both estimates are reasonable given the available information. If you believe that this advertisement is similar to those advertisements that averaged a 6 percent click rate, then $\hat{\phi}_{MAP}$ may be a superior estimate, but if the circumstances under which the advertisement was shown were different from the usual, then $\hat{\phi}_{MLE}$ might be a better choice.

3 Conceptual MLE/MAP Questions

1. When calculating the MAP estimates, we rely on the Bayes formula and then argue we can ignore $p(\mathcal{D})$. Why do we usually ignore calculating $p(\mathcal{D})$?

   We are taking the argmax over θ and since the value of $P(\mathcal{D})$ is not related to θ we can ignore this value.

2. As the amount of data increases, how are MLE and MAP affected?

   MLE finds the θ that maximizes the likelihood term $p(D|\theta)$. $p(D|\theta)$ becomes more informative about the true parameter values because there is more data.

   MAP finds the θ that maximizes the posterior, $p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$. The influence of the prior term $p(\theta)$ decreases relative to the likelihood term because the prior becomes less significant in comparison to the increasing amount of data.

3. Can MLE and MAP estimates be the same? If so, when?

   As the amount of data increases, the MLE and MAP estimates converge to the same value. Therefore, with enough data, MLE and MAP will be approximately equal.

   Additionally, if MAP uses a uniform distribution as the prior, the MLE and MAP estimates will be equal. In other words, you can think of MLE as a special case of MAP, where the prior is uniform!