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

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, $\text{argmin}_\phi - \log p(\phi \mid \mathcal{D})$, in terms of N, N_1, ϕ, α, and β.

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

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 $\hat{\phi}_{MAP}$.

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

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 $p(\mathcal{D})$?

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

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