DERIVATIONS OF MLE AND MAP

1. MLE - Maximum likelihood estimator

Assume that we have samples, \( x_1, \ldots, x_m \) independent from some distribution, which depends on a parameter \( \theta \), but we don't know the value of that parameter (for instance, we have results of biased coin flips, with the probability of heads being \( \theta \)). We want to get some estimation for the value of that parameter from the samples we get. One of the ways to do this is using MLE (maximum likelihood estimator). For any value for the \( \theta \), we shall define

\[
L'(\theta) = \mathbb{P}(x_1, \ldots, x_m | \theta) = \prod_{i=1}^{m} \mathbb{P}(x_i | \theta)
\]

The maximum likelihood estimator is the \( \theta \) that maximizes \( L' \). \( L' \) is called the likelihood function. We would like to derive \( L' \) and compare its derivative to 0. It will usually be easier to derive \( L \triangleq \log(L') \), and we will find the maximizer of \( L \) which is the maximizer of \( L' \). \( L \) is called the log-likelihood function.

2. MLE for Biased Coin Flip

Let \( X_1, \ldots, X_m \in \{0, 1\} \) be the outcome of iid random coin flips with unknown probability \( \theta \) of being 1. Let \( n_i = |\{j : X_j = i\}| \text{ for } i = 0, 1 \). For any \( \hat{\theta} \in [0, 1] \), the log-likelihood of the observation is

\[
L(X_1, \ldots, X_m | \hat{\theta}) = \log(\prod_{i=1}^{m} \mathbb{P}(x_i | \hat{\theta})) = \sum_{i=1}^{m} \log(\mathbb{P}(x_i | \hat{\theta})) = \sum_{i=1}^{m} \log(1 - \hat{\theta}) + \sum_{i=1}^{m} \log(\hat{\theta}) = n_1 \log \hat{\theta} + n_0 \log(1 - \hat{\theta}) ,
\]

Deriving with respect to \( \hat{\theta} \):

\[
\frac{dL(X_1, \ldots, X_m; \hat{\theta})}{d\theta} = \frac{n_1}{\hat{\theta}} - \frac{n_0}{1 - \hat{\theta}} .
\]

The derivative vanishes when \( n_1 - n_1 \hat{\theta} = n_0 \hat{\theta} \), equivalently when \( \hat{\theta} = \frac{n_1}{n_0 + n_1} \), in other words when \( \hat{\theta} \) is the empirical probability of \( n_1 \). It remains as an exercise to show that this value of \( \hat{\theta} \) is indeed the maximizer of the log-likelihood, or the MLE.
3. MAP - Maximum a Posteriori

Another way to estimate the parameter is using the MAP estimator. It goes like this: Assume that before drawing the samples, we assign probabilities to the different thetas. Each \( \theta \) has \( P(\theta) \). Now after drawing the samples, we calculate

\[
L'(\theta) \triangleq \frac{P(x_1, ..., x_m|\theta)P(\theta)}{P(x_1, ..., x_m)}
\]

We want to find the \( \theta \) that maximizes this value. Note that maximizing \( L' \) is equivalent to maximizing \( P(x_1, ..., x_m|\theta)P(\theta) \), and this in turn is equivalent to maximizing

\[
L(\theta) \triangleq \log(P(x_1, ..., x_m|\theta)P(\theta)) = \log(P(x_1, ..., x_m)) + \log(P(\theta))
\]

When \( P(\theta) \) is equal for all \( \theta \) (\( \theta \) is distributed uniformly) then our MAP estimator is the MLE estimator. We can use the density function if our thetas are from a continuous space, and we will use \( P \) for the density function.

4. MAP for Coin Flips

Assume that we have the same problem as in the section of the coin-flips when we calculated the MLE. But here we will use the MAP estimator, with \( P(\theta) = 2\theta \), for \( 0 \leq \theta \leq 1 \).

\[
L(\theta) = \log(P(x_1, ..., x_m|\theta)) + \log(P(\theta))
\]

Deriving, the \( \log(2) \) vanishes, and we are left with an equation similar to the MLE section. Our MAP will now be

\[
\frac{n_1 + 1}{(n_1 + 1) + n_0}
\]

5. MLE for 1D Gaussian

The density function of a one dimensional Gaussian with expectation \( \mu \) and variance \( \sigma^2 \) is

\[
f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.
\]

Given \( m \) iid observations \( X_1, ..., X_m \), the corresponding log-likelihood function is

\[
L(X_1, ..., X_m; \hat{\mu}, \hat{\sigma}^2) = -\sum_{i=1}^{m} \frac{(X_i - \hat{\mu})^2}{2\hat{\sigma}^2} - m \log \hat{\sigma} - \frac{1}{2} m \log 2\pi.
\]

Derivation with respect to \( \hat{\mu} \):

\[
\frac{\partial L(X_1, ..., X_m; \hat{\mu}, \hat{\sigma}^2)}{\partial \hat{\mu}} = -2 \sum_{i=1}^{m} (X_i - \hat{\mu})/2\hat{\sigma}^2.
\]

\( \partial L/\partial \hat{\mu} = 0 \) implies \( \hat{\mu} = \sum_{i=1}^{m} X_i/m \). Derivation with respect to \( \hat{\sigma}^2 \):

\[
\frac{\partial L(X_1, ..., X_m; \hat{\mu}, \hat{\sigma}^2)}{\partial \hat{\sigma}^2} = \frac{\sum_{i=1}^{m} (X_i - \hat{\mu})^2}{\hat{\sigma}^3} - \frac{m}{\hat{\sigma}}.
\]

\( \partial L/\partial \hat{\sigma} = 0 \) implies \( \hat{\sigma}^2 = \sum_{i=1}^{m} (X_i - \hat{\mu})^2/m \).