Notes for 6.864, ML vs. MAP (Sept. 24, 2009)

Prepared by Harr Chen.

Maximum Likelihood

We are looking at a binomial distribution of n trials:

$$p(X = k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

We observe k heads in n trials, and are interested in estimating the parameter θ, which represents the probability of heads. The likelihood function (which is also $p(\text{data}|\theta)$) is:

$$L(\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

We'd like to maximize this expression wrt θ, or equivalently the log of the expression (working in log-space is easier and does not change the location of highest likelihood):

$$\begin{aligned}\operatorname*{argmax}_\theta L(\theta) &= \operatorname*{argmax}_\theta \log L(\theta) \\ &= \operatorname*{argmax}_\theta \left(k \log\theta + (n-k)\log(1-\theta)\right)\end{aligned}$$

Note that we dropped $\binom{n}{k}$ here because it is independent of θ. To find the maximum we differentiate wrt θ and set the result to zero:

$$\frac{\partial \log L(\theta)}{\partial\theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

From here elementary algebra shows that the maximum likelihood estimate of θ is:

$$\theta = \frac{k}{n},$$

as expected.

Maximum a Posteriori

Our objective is different here; rather than maximizing data likelihood given parameter values, we instead maximize parameter values given data D, which we transform using Bayes' rule as follows:

$$\operatorname*{argmax}_\theta p(\theta|D) = \operatorname*{argmax}_\theta \frac{p(D|\theta)p(\theta)}{p(D)} = \operatorname*{argmax}_\theta p(D|\theta)p(\theta)$$

Note we can drop $p(D)$ because it is constant wrt θ. The term $p(D|\theta)$ is the same likelihood as before; the term $p(\theta)$ is a *prior*, which reflects our belief about the parameter values *before* we observe any data. Typically for a binomial likelihood, we will use the *beta distribution* as its prior:

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

The values α and β are fixed *hyperparameters* (set externally as opposed to being estimated). $B(\alpha,\beta)$ is a hairy normalization factor that only depends on the hyperparameters, and ensures the distribution sums to one. The beta distribution is convenient because it is a *conjugate prior* for

the binomial — that is, a binomial likelihood times a beta prior yields a beta posterior. Note then that:

$$p(D|\theta)p(\theta) = \frac{\binom{n}{k}}{B(\alpha,\beta)}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

To find the MAP estimates, we proceed as before, taking the log of the objective and dropping constants (including the new $B(\alpha,\beta)$):

$$\operatorname*{argmax}_\theta \log(p(D|\theta)p(\theta)) = \operatorname*{argmax}_\theta \left((k+\alpha-1)\log\theta + (n-k+\beta-1)\log(1-\theta)\right)$$

From here, we can then take the partial derivative wrt θ and set to zero, solving for θ:

$$\frac{\partial \log p(D|\theta)p(\theta)}{\partial\theta} = \frac{k+\alpha-1}{\theta} - \frac{n-k+\beta-1}{1-\theta} = 0$$

Algebra yields:

$$\theta = \frac{k+\alpha-1}{n+\alpha+\beta-2}$$

Note that in the final result α and β serve as "pseudocounts" that mimic the estimation behavior that we'd get if we had observed $\alpha-1$ additional heads and $\beta-1$ additional tails. In particular, add-one smooothing can be emulated by setting $\alpha,\beta=2$.

This concept generalizes to the multivariate case with a multinomial likelihood and a Dirichlet prior, though the math is slightly trickier (and requires Lagrange multipliers).