Psuedocounts

When estimating the parameters of a categorical or multinomial distribution, a common practice is to add **pseudocounts** to the observed counts in the data. This process, known as **additive smoothing**, regulates the maximum-likelihood estimate in order to avoid overfitting. That is, it prevents extreme assignments to the parameters that are due to a lack of data that would be used to estimate that parameter.

Standard psuedocounts

A psuedocount c is a pre-determined value that is added to the counts of occurrence of each category/class in the data when estimating the parameters of a multinomial.

First, let's look an example of estimating parameters without psuedocount. If we are estimating the parameter θ that a coin lands heads and we have observed counts H and T for heads and tails respectively, then our maximum likelihood estimate of θ is

$$\hat{\theta} = \frac{H}{H+T}$$

. If we use a psuedocount value of c, then our estimate would be

$$\begin{aligned}\hat{\theta} &= \frac{H+c}{(H+c)+(T+c)} \\ &= \frac{H+c}{H+T+2c}\end{aligned}$$

. By adding equal psuedocounts to each outcome, we push our estimate of the parameter closer to the uniform distribution than the maximum likelihood estimate of the parameter. In essence, a common psuedocount across categories/classes implies a hypothetical scenario in which we have already observed equal numbers of categories/classes prior to observing the data. Thus, the larger the psuedocount, the more data will be needed to push the estimate of the parameters away from the uniform distribution.

We note that a psuedocount of $c = 1$ is often called **Laplace smoothing**.

M-estimates

M-estimates generalize psueudocounts that are constant across categories/classes to psuedocounts that differ for each category/class. We imagine that we have c total counts and we distribute them amongst the observed class counts in the data according according to some distribution specified by $\mathbf{q} := q_1, q_2, \ldots, q_k$ where k is the number of categories/classes and $\sum_{i=1}^{k} q_i = 1$. That is, we pretend that we have already observed

$q_1c, q_2c, \ldots, q_kc$ items of each category before observing the data. Thus, we would estimate our parameters by

$$\theta_i = \frac{x_i + q_i c}{\sum_{j=1}^{k}(x_j + q_j c)}$$

The distribution specified by $\mathbf{q}$ encodes our prior knowledge about the parameters and the count c encodes the emphasis that will place on this prior knowledge.

Interpretation of psuedocounts as Bayesian prior knowledge

Psuedocounts can be theoretically justified within a Bayesian framework for estimating parameters. More specifically, using psuedocounts to augment the maximum likelihood estimate of the parameters is equivalent to finding a maximum a posteriori estimate or a posterior-mean estimate of the parameters.

Psuedocounts as a maximum a posteriori (MAP) estimate

Recall that given a counts vector $\mathbf{x}$ generated by a multinomial distribution, the posterior distribution over the parameters $\mathbf{\Theta}$ with a Dirichlet prior is a Dirichlet. More specifically, given the prior over the parameters

$$\mathbf{\Theta} \sim \text{Dir}(\alpha_1, \ldots, \alpha_d)$$

the posterior distribution after observing $\mathbf{x}$ is

$$\mathbf{\Theta} \mid \mathbf{x} \sim \text{Dir}(x_1 + \alpha_1, \ldots, x_d + \alpha_d)$$

. The MAP estimate is then

$$\theta_i^{\text{MAP}} = \frac{x_i + \alpha_i - 1}{\sum_{j=1}^{d}(x_j + \alpha_j - 1)}$$

Thus, using psuedocounts of c_i added to each x_i when estimating our parameters as follows:

$$\hat{\theta}_i = \frac{x_i + c_i}{\sum_{j=1}^{d}(x_j + c_j)}$$

is equivalent to the MAP estimate assuming a Dirichlet prior parameterized by $(c_1 + 1, c_2 + 1, \ldots, c_k + 1)$.

Psuedocounts as a mean of the posterior (MOP) estimate

Again assuming that $\mathbf{x}$ was generated by a multinomial distribution assuming a Dirichlet prior, the mean of the posterior is

$$\theta_i^{\text{MOP}} = \frac{x_i + \alpha_i}{\sum_{j=1}^{d}(x_j + \alpha_j)}$$

Thus, psuedocounts of c_i added to each x_i when estimating our parameters, as follows is equivalent to a mean of the posterior estimate assuming the prior distribution was a Dirichlet parameterized by $c_1, c_2, \ldots, c_k$.

$$\hat{\theta}_i = \frac{x_i + c_i}{\sum_{j=1}^{d}(x_j + c_j)}$$

we see that

$$\theta_i^{\text{MOP}} = \hat{\theta}_i \implies c_i = \alpha_i$$

Adding psuedocounts is akin to finding the mean of a Dirichlet posterior given a Dirichlet prior with parameters $c_1, c_2, \ldots, c_k$.