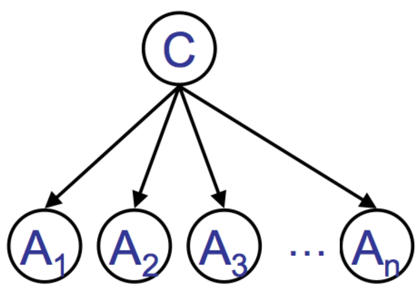
Maximum Likelihood and EM for Naïve Bayes

Jesse Hoey
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, CANADA, N2L3G1
jhoey@cs.uwaterloo.ca

We have the following model:



Where each of C, A_i are binary variables taking on values 0 or 1. The parameters of this model are

$$\begin{aligned}\theta_c &= P(C = 1)\\ \theta_{i1} &= P(A_i = 1|C = 1)\\ \theta_{i0} &= P(A_i = 1|C = 0)\end{aligned}$$

collectively, we will call $\theta_i = \{\theta_{i1}, \theta_{i0}\}$ and $\theta = \{\theta_c, \theta_1, \ldots \theta_N\}$.

To make predictions with this model, we compute

$$\begin{aligned}P(C = 1|a_1, a_2, \ldots, a_N) &\propto P(a_1, a_2, \ldots, a_N|C = 1)P(C = 1)\\ &= \left[\prod_{i=1}^{N} \theta_{i1}^{a_i}(1 - \theta_{i1})^{1-a_i}\right]\theta_c\end{aligned}$$

Similarly, we compute $P(C = 0|a_1, a_2, \ldots, a_N) \propto \left[\prod_{i=1}^{N} \theta_{i0}^{a_i}(1 - \theta_{i0})^{1-a_i}\right](1 - \theta_c)$ and normalise to find the actual probabilities.

To find the maximum likelihood values for the parameters given a set of M data $\mathbf{d} =$

$\{d_1, d_2, \ldots, d_M\}$, where each $d_j = \{a_{j1}, a_{j2}, \ldots, a_{jN}, c_j\}$, we compute:

$$\begin{aligned}
\theta_{ML} &= \arg\max_\theta P(\mathbf{d}|\theta) \\
&= \arg\max_\theta \left[\prod_{j=1}^{M} P(d_j|\theta)\right] \\
&= \arg\max_\theta \left[\log \prod_{j=1}^{M} P(d_j|\theta)\right] \\
&= \arg\max_\theta \sum_{j=1}^{M} \log P(d_j|\theta) \\
&= \arg\max_\theta \sum_{j=1}^{M} \log P(a_{j1}, a_{j2}, \ldots, a_{jN}, c_j|\theta) \\
&= \arg\max_\theta \sum_{j=1}^{M} \log \left[\left(\prod_{i=1}^{N} \theta_{ic_j}^{a_{ji}} (1-\theta_{ic_j})^{1-a_{ji}}\right) \theta_c^{c_j} (1-\theta_c)^{1-c_j}\right] \\
&= \arg\max_\theta \sum_{j=1}^{M} \left[\sum_{i=1}^{N} \big(a_{ji} \log \theta_{ic_j} + (1-a_{ji}) \log(1-\theta_{ic_j})\big) + c_j \log \theta_c + (1-c_j)\log(1-\theta_c)\right]
\end{aligned} \tag{1}$$

Taking derivatives we get

$$\begin{aligned}
\frac{\partial}{\partial \theta_c} \sum_j \log P(d_j|\theta) &= \sum_j \frac{c_j}{\theta_c} - \frac{(1-c_j)}{(1-\theta_c)} \\
&= \frac{1}{\theta_c(1-\theta_c)} \sum_j (-\theta_c + c_j)
\end{aligned}$$

setting to zero allows us to find $\theta_c = \frac{\sum_j c_j}{M}$. Simillary, taking the derivative with respect to θ_{i1}, we find

$$\begin{aligned}
\frac{\partial}{\partial \theta_{i1}} \sum_j \log P(d_j|\theta) &= \sum_{j|c_j=1} \frac{a_{ji}}{\theta_{i1}} - \frac{(1-a_{ji})}{(1-\theta_{i1})} \\
&= \frac{1}{\theta_{i1}(1-\theta_{i1})} \sum_{j|c_j=1} (a_{ji} - \theta_{i1})
\end{aligned}$$

again setting to zero allows us to find $\theta_{i1} = \frac{\sum_{j|c_j=1} a_{ji}}{M_1}$ where M_1 is the number of datapoints with $c_j = 1$. Similarly, we can find $\theta_{i0} = \frac{\sum_{j|c_j=0} a_{ji}}{M_0}$ where M_0 is the number of datapoints with $c_j = 0$.

Laplace correction:
There may be cases where the sums in the numerators are zero (so $\sum_{j|c_j=1} a_{ji} = 0$ or $\sum_{j|c_j=0} a_{ji} = 0$). In such cases, we will find that θ_{i1} or θ_{i0} are identically zero. This is a case

where a certain feature is always zero in the dataset. When this happens, the predictions of C will be identically zero for any test data where that feature does occur. However, the absence of the feature from the dataset may not be all that significant (e.g. all the other features may be predicting a certain class C, but the addition of this one feature that never happened in the training set will make the probability go to zero). To correct for this, we can use the *Laplace correction*, which essentially ensures that no probability is identically zero. We do this by adding 1 to numerator and d to the denominator, where d is the number of values the variable can take on ($d = 2$ in this case).

$$\theta_{i1} = \frac{\sum_{j|c_j=1} a_{ji} + 1}{M_1 + 2}$$

and

$$\theta_{i0} = \frac{\sum_{j|c_j=0} a_{ji} + 1}{M_0 + 2}$$

essentially we are "imagining" one more datum for each class that has every feature present. Of course, this is actually just placing a very weak prior (Beta distribution with $a = b = 2$) on each feature parameter.

Hidden class variable

Now suppose the class variable C is hidden (*i.e.* not present in any of the data), so we now have $d_j = \{a_{j1}, a_{j2}, \ldots, a_{jN}\}$. We could "guess" at a set of values for θ, and then do the following (super easy):

1. compute for each data point $c_j^* = \arg\max_{c_j} P(c_j|a_{j1}, a_{j2}, a_{j3}, \ldots, a_{jN})$
2. fill in the data with these computed values,
3. compute θ_{ML} using the equations derived above
4. set $\theta \leftarrow \theta_{ML}$ and goto step 1

Or, we could use Expectation Maximization in full. To do this, we note that, if we had a guess for θ, say θ_g, we could maximize the following expression over θ and get a new, better (always closer to the local maximum) value for θ_g:

$$\sum_j \sum_{c_j} P(c_j|d_j, \theta_g) \log P(d_j, c_j|\theta)$$

That is, we can do

$$\theta_g' = \arg\max_\theta \sum_j \sum_{c_j} P(c_j|d_j, \theta_g) \log P(d_j, c_j|\theta)$$

and we are guaranteed that $P(\mathbf{d}|\theta_g') \geq P(\mathbf{d}|\theta_g)$.

Expanding the right side out, we have the equivalent of Equation (1):

$$\theta_{ML} = \arg\max_\theta \sum_{j=1}^{M} \sum_{c_j} P(c_j|d_j\theta_g) \left[\sum_{i=1}^{N} \left(a_{ji} \log \theta_{ic_j} + (1 - a_{ji}) \log(1 - \theta_{ic_j}) \right) + c_j \log \theta_c + (1 - c_j) \log(1 - \theta_c) \right]$$

So, if we take derivatives with respect to θ_c, we get

$$\begin{aligned} \frac{\partial}{\partial \theta_c} \sum_j \sum_{c_j} P(c_j|d_j, \theta_g) \log P(d_j, c_j|\theta) &= \sum_j \sum_{c_j} P(c_j|d_j, \theta_g) \left(\frac{c_j}{\theta_c} - \frac{(1 - c_j)}{(1 - \theta_c)} \right) \\ &= \sum_j \sum_{c_j} P(c_j|d_j, \theta_g)(-\theta_c + c_j) \end{aligned}$$

setting to zero and solving for θ_c gives

$$\theta_c = \frac{\sum_j \sum_{c_j} c_j P(c_j|d_j, \theta_g)}{\sum_j \sum_{c_j} P(c_j|d_j, \theta_g)}$$

which, since c_j is Boolean, is:

$$\begin{aligned}\theta_c &= \frac{\sum_j P(c_j = 1|d_j, \theta_g)}{\sum_j \sum_{c_j} P(c_j|d_j, \theta_g)} \\ &= \frac{\sum_j P(c_j = 1|d_j, \theta_g)}{M}\end{aligned} \tag{2}$$

The updated estimate is simply the average probability of C given the data! Note that if you don't calculate $P(c_j|d_j, \theta_g)$ exactly, but something proportional to it (so you dont' normalise first), then you can't use the simplified Equation (2), and must sum the values you compute to normalise.

The other parameters are similarly computed as

$$\theta_{i1} = \frac{\sum_j a_{ji} P(c_j = 1|d_j, \theta_g)}{\sum_j P(c_j = 1|d_j, \theta_g)}$$

and

$$\theta_{i0} = \frac{\sum_j a_{ji} P(c_j = 0|d_j, \theta_g)}{\sum_j P(c_j = 0|d_j, \theta_g)}$$