Solutions by **Andrew Lys** Collaborated with **Sam Fine** `andrewlys (at) u.e.`

Gaussian Mixtures

1.

Modeling Text Documents

2. A Simple Model.

(a) We shall denote p_{topic} as p, since it is given that this is a single probability. For simplicity, we assume that $y \in \{0, 1\}$, and that $x \in \{0, 1\}^N$. We denote $x[i]$ to be the ith coordinate of the sample x.

Given a sample

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

We define the following sample statistics. For $x \in \{0, 1\}$, $y \in \{0, 1\}$:

$$n_j(y, x) = |\{i : (x_i, y_i) \in S, x_i[j] = x, y_i = y\}|$$

$$n(y) = |\{i : (x_i, y_i) \in S, y_i = y\}|$$

We want to find estimators for p and for

$$P(x[1] = x_1, \ldots, x[N] = x_N | y = y)$$

By the independence of $x[i]|y$, we can simplify this expression:

$$P(x[1] = x_1, \ldots, x[N] = x_N | y = y) = \prod_{i=1}^{N} P(x[i] = x_i | y = y)$$

Thus, we can focus on estimators of p and $P(x[i] = x|y = y) := p_i(x|y)$ (I know that this swaps the arguments of $n_i(y, x)$, it's too much to change now). We should expect our MLEs for p and $p_i(x|y)$ to be the sample means, i.e.

$$\hat{p} = \frac{n(1)}{n}$$
$$\hat{p}_i(x|y) = \frac{n_i(y, x)}{n(y)}$$

We define our log-likelihood function as

$$\ell(\theta|S) = \sum_{i=1}^{n} \log(P(y = y_i, x[1] = x_i[1], \ldots, x[N] = x_i[N]))$$

Given that S was drawn i.i.d., we can simplify.

$$\begin{aligned}
\ell(\theta|S) &= \sum_{i=1}^{n} \log(P(x[1] = x_i[1], \ldots, x[N] = x_i[N] | y = y_i) P(y = y_i)) \\
&= \sum_{i=1}^{n} \log(P(y = y_i) \prod_{j=1}^{N} P(x[j] = x_i[j] | y = y_i)) \\
&= \sum_{i=1}^{n} \log(P(y = y_i)) + \sum_{j=1}^{N} \log(P(x[j] = x_i[j] | y = y_i)) \\
&= \sum_{i=1}^{n} \log(P(y = y_i)) + \sum_{i=1}^{n} \sum_{j=1}^{N} \log(p_j(x_i[j] | y_i))
\end{aligned}$$

Writing the parameters explicitly, we have:

$$\ell(\theta|S) = \sum_{i=1}^{n} \log(P(y = y_i|p)) + \sum_{i=1}^{n}\sum_{j=1}^{N} \log(P(x[i] = x_j[i]|y_i, p_i(x|y)))$$

To solve for the minmimum of $\ell(\theta|S)$, we use the method of Lagrange multipliers. First, we can split the problem into two steps. It's clear that that right sum does not depend on p, so we can begin by finding the optimal p.

We note:

$$\begin{aligned}
P(y = y_i|p) &= P(y = y_i|y_i = 1, p)P(y_i = 1|p) + P(y = y_i|y_i = 0, p)P(y_i = 0|p)\\
&= P(y = 1|p)[[y_i = 1]] + P(y = 0|p)[[y_i = 0]]\\
&= p^{y_i}(1-p)^{1-y_i}
\end{aligned}$$

Plugging this into our log-likelihood, we have:

$$\begin{aligned}
\ell(\theta|S) &= \sum_{i=1}^{n} \log(p^{y_i}(1-p)^{1-y_i}) + \sum_{i=1}^{n}\sum_{j=1}^{N} \log(P(x[i] = x_j[i]|y_i, p_i(x|y)))\\
&= \sum_{i=1}^{n} y_i \log(p) + (1-y_i)\log(1-p) + \sum_{i=1}^{n}\sum_{j=1}^{N} \log(P(x[i] = x_j[i]|y_i, p_i(x|y)))
\end{aligned}$$

Taking the derivative with respect to p and setting it to zero, we have:

$$\begin{aligned}
\frac{d}{dp}\ell(\theta|S) = \sum_{i=1}^{n} \frac{y_i}{p} - \frac{1-y_i}{1-p} &= 0\\
\sum_{i=1}^{n} \frac{y_i}{p} &= \sum_{i=1}^{n} \frac{1-y_i}{1-p}\\
\frac{1-p}{p} &= \frac{\sum_{i=1}^{n} 1-y_i}{\sum_{i=1}^{n} y_i}\\
p &= \frac{\sum_{i=1}^{n} y_i}{n}
\end{aligned}$$

Thus, we have that $\hat{p} = \frac{n(1)}{n}$.

Now, we solve for $\hat{p}_i(x|y)$, by using the method of Lagrange multipliers. Our objective function is as follows:

$$\sum_{i=1}^{n}\sum_{j=1}^{N} \log(P(x[j] = x_i[j]|y_i))$$

We can write this in a nicer form.

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{N} \log(P(x[j] = x_i[j]|y_i)) &= \sum_{i=1}^{n}\sum_{j=1}^{N} \log(p_j(x_i[j]|y_i))\\
&= \sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{x\in\{0,1\}} [[x_i[j] = x]] \log(p_j(x|y_i))\\
&= \sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}} [[x_i[j] = x \wedge y_i = y]] \log(p_j(x|y))\\
&= \sum_{j=1}^{N}\sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}} \log(p_j(x|y)) \sum_{i=1}^{n}[[x_i[j] = x \wedge y_i = y]]\\
&= \sum_{j=1}^{N}\sum_{y\in\{0,1\}}\sum_{x\in\{0,1\}} \log(p_j(x|y))n_j(y,x)
\end{aligned}$$

We now have the following constraints:

$$\sum_{x\in\{0,1\}} p_j(x|y) = 1 \qquad \forall y \in \{0,1\}, j \in [N]$$

This gives us the following Lagrangian:

$$\mathcal{L} = \sum_{j=1}^{N} \sum_{y\in\{0,1\}} \sum_{x\in\{0,1\}} \log(p_j(x|y))n_j(y,x) + \sum_{j=1}^{N} \sum_{y\in\{0,1\}} \lambda_j(y) \left(\sum_{x\in\{0,1\}} p_j(x|y) - 1 \right)$$

Taking the derivatives with respect to $p_j(x|y)$, we have:

$$\begin{aligned}
[p_j(x|y)] &: \frac{n_j(y,x)}{p_j(x|y)} = \lambda_j(y) \\
[\lambda_j(y)] &: \sum_{x\in\{0,1\}} p_j(x|y) = 1
\end{aligned}$$

Since we have equality, in $\lambda_j(y)$, for $x \in \{0,1\}$ we can solve for $p_j(x|y)$:

$$\begin{aligned}
\frac{n_j(y,x)}{p_j(x|y)} &= \frac{n_j(y,1-x)}{p_j(1-x|y)} \\
p_j(1-x|y) &= \frac{n_j(y,1-x)}{n_j(y,x)} p_j(x|y) \\
\implies 1 &= p_j(x|y) + \frac{n_j(y,1-x)}{n_j(y,x)} p_j(x|y) \\
n_j(y,x) &= p_j(x|y)n_j(y,x) + n_j(y,1-x)p_j(x|y) \\
&= p_j(x|y)(n_j(y,x) + n_j(y,1-x)) \\
p_j(x|y) &= \frac{n_j(y,x)}{n_j(y,x) + n_j(y,1-x)} \\
\hat{p}_j(x|y) &= \frac{n_j(y,x)}{n(y)}
\end{aligned}$$

(b) Using Baye's Law, and conditional independence we have:

$$\begin{aligned}
P(Y=1|X=x) &= \frac{P(X=x|Y=1)P(Y=1)}{P(X=x)} \\
&= \frac{P(X[1]=x[1],\ldots,X[N]=x[N]|Y=1)P(Y=1)}{P(X[1]=x[1],\ldots,X[N]=x[n])} \\
&= \frac{P(Y=1)\prod_{i=1}^{N} P(X[i]=x[i]|Y=1)}{P(X[1]=x[1],\ldots,X[N]=x[n]|Y=1)P(Y=1) + P(X[1]=x[1],\ldots,X[n]=x[n]|Y=0)P(Y=0)} \\
&= \frac{p\prod_{i=1}^{N} p_i(x[i]|1)}{p\prod_{i=1}^{N} p_i(x[i]|1) + (1-p)\prod_{i=1}^{N} p_i(x[i]|0)}
\end{aligned}$$

Now we can reduce this into the form of a logistic function.

$$\begin{aligned}
P(Y=1|X=x) &= \frac{p\prod_{i=1}^{N} p_i(x[i]|1)}{p\prod_{i=1}^{N} p_i(x[i]|1) + (1-p)\prod_{i=1}^{N} p_i(x[i]|0)} \\
&= \frac{1}{1 + \frac{1-p}{p}\frac{\prod_{i=1}^{N} p_i(x[i]|0)}{\prod_{i=1}^{N} p_i(x[i]|1)}} \\
&= \frac{1}{1 + e^{-(\log(\frac{p}{1-p}) + \sum_{i=1}^{N} \log(\frac{p_i(x[i]|1)}{p_i(x[i]|0)}))}}
\end{aligned}$$

Therefore, we can get our discriminant as follows:

$$r(x) = \log\left(\frac{p}{1-p}\right) + \sum_{i=1}^{N} \log\left(\frac{p_i(x[i]|1)}{p_i(x[i]|0)}\right)$$

(c) We can simplify the discriminant by noting

$$p_i(x|y) = p_i(1|y)^x p_i(0|y)^{1-x}$$

Giving us

$$\begin{aligned}
r(x) &= \log\left(\frac{p}{1-p}\right) + \sum_{i=1}^{N} \log\left(\frac{p_i(1|1)^{x[i]} p_i(0|1)^{1-x[i]}}{p_i(1|0)^{x[i]} p_i(0|0)^{1-x[i]}}\right) \\
&= \log\left(\frac{p}{1-p}\right) + \sum_{i=1}^{N} \left(x[i] \log\left(\frac{p_i(1|1)}{p_i(1|0)}\right) + (1-x[i]) \log\left(\frac{p_i(0|1)}{p_i(0|0)}\right)\right) \\
&= \log\left(\frac{p}{1-p}\right) + \sum_{i=1}^{N} x[i] \log\left(\frac{p_i(1|1)}{p_i(1|0)}\right) + -x[i] \log\left(\frac{p_i(0|1)}{p_i(0|0)}\right) + \log\left(\frac{p_i(0|1)}{p_i(0|0)}\right) \\
&= \log\left(\frac{p}{1-p}\right) + \sum_{i=1}^{N} x[i] \left(\log\left(\frac{p_i(1|1)}{p_i(1|0)}\right) - \log\left(\frac{p_i(0|1)}{p_i(0|0)}\right)\right) + \log\left(\frac{p_i(0|1)}{p_i(0|0)}\right) \\
&= \log\left(\frac{p}{1-p}\right) + \sum_{i=1}^{N} \log\left(\frac{p_i(0|1)}{p_i(0|0)}\right) + \sum_{i=1}^{N} x[i] \left(\log\left(\frac{p_i(1|1)}{p_i(0|1)} \frac{p_i(0|0)}{p_i(1|0)}\right)\right)
\end{aligned}$$

The feature map must include a constant 1 to account for the term on the left, and must have N more features for each of $x[i]$. Thus, our feature map is simply:

$$\phi : x \mapsto (1, x[1], \ldots, x[N])$$

Therefore, our vector w, such that $r(x) = \langle w, \phi(x) \rangle$, is:

$$w = \left(\log\left(\frac{p}{1-p}\right) + \sum_{i=1}^{N} \log\left(\frac{p_i(0|1)}{p_i(0|0)}\right), \log\left(\frac{p_1(1|1)}{p_1(0|1)} \frac{p_1(0|0)}{p_1(1|0)}\right), \ldots, \log\left(\frac{p_N(1|1)}{p_N(0|1)} \frac{p_N(0|0)}{p_N(1|0)}\right)\right)$$

(d) The log odds term in the bias has a simple interpretation.

$$\frac{\hat{p}}{1-\hat{p}} = \frac{n(1)/n}{n(0)/n} = \frac{n(1)}{n(0)}$$
$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \log\left(\frac{n(1)}{n(0)}\right)$$

Similarly,

$$\frac{\hat{p}_i(x|y)}{\hat{p}_i(x'|y')} = \frac{n_i(y,x)/n(y)}{n_i(y',x')/n(y')}$$

So,

$$\frac{\hat{p}_i(0|1)}{\hat{p}_i(0|0)} = \frac{n_i(1,0)}{n_i(0,0)} \frac{n(0)}{n(1)}$$
$$\frac{\hat{p}_i(1|1)\hat{p}_i(0|0)}{\hat{p}_i(0|1)\hat{p}_i(1|0)} = \frac{n_i(1,1)/n(1) n_i(0,0)/n(0)}{n_i(1,0)/n(1) n_i(0,1)/n(0)}$$

$$= \frac{n_i(1,1)n_i(0,0)}{n_i(1,0)n_i(0,1)}$$

So we have the following simplification for w:

$$w = \left((N-1)\log\frac{n(0)}{n(1)} + \sum_{i=1}^{N}\log\frac{n_i(1,0)}{n_i(0,0)}, \log\frac{n_1(1,1)n_1(0,0)}{n_1(1,0)n_1(0,1)}, \ldots, \log\frac{n_N(1,1)n_N(0,0)}{n_N(1,0)n_N(0,1)}\right)$$

3. Adding a Prior.

(a) The MAP estimate is defined as follows:

$$\hat{\theta} = \arg\max_{\theta} p(\theta|S)$$

In our case,

$$\theta = (p, \{p_y\})$$

Where we define:

$$p = P(y = 1)$$
$$p_y[i] = P(x[i] = 1|y)$$

and p_y is a vector of N elements. Let S be a sample of n i.i.d. points.

$$S = ((x_1, y_1), \ldots, (x_n, y_n))$$

our posterior distribution, $p(\theta|S)$ is given by:

$$\begin{aligned} p(p, \{p_y\}|S) &= \frac{p(S|p, \{p_y\})p(p, \{p_y\})}{p(S)} \\ &= \frac{p(X|Y, p, \{p_y\})p(Y|p, \{p_y\})p(p, \{p_y\})}{p(S)} \\ &= \frac{p(X|Y, \{p_y\})p(Y|p)p(p, \{p, p_y\})}{p(X|Y)p(Y)} \end{aligned}$$

where X is the vector of x_i's and Y is the vector of y_i's.

Note that, we are not conditioning the denominator with respect to the parameters we are optimizing over. The denominator is the distribution over the distributions of p and $\{p_y\}$. Therefore, we can ignore it in the optimization problem.

$$\hat{\theta} = \arg\max_{p, \{p_y\}} p(X|Y, \{p_y\})p(Y|p)p(p, \{p_y\})$$

We break this expression down, term by term, first focusing on the last term.

$$\begin{aligned} p(p, \{p_y\}) &= p(p)p(\{p_y\}) = f_{Dir(1)}(p)p(p_1)p(p_0) \\ &= f_{Dir(\alpha)}(p_1)f_{Dir(\alpha)}(p_0) \\ &= \frac{1}{Z(\alpha)^2}\prod_{i=1}^{N} p_1[i]^{\alpha-1}p_0[i]^{\alpha-1} \end{aligned}$$

Since $Z(\alpha)^2$ is fixed, we can ignore it in the expression for $\hat{\theta}$. Now we focus on the second term.

$$\begin{aligned} p(Y|p) = P(Y_1 = y_1, \ldots, Y_n = y_n|p) &= \prod_{i=1}^{n} P(Y_i = y_i|p) \\ = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \end{aligned}$$

Now we focus on the first term.

$$p(X|Y, \{p_y\}) = P(X_1 = x_1, \ldots, X_n = x_n|Y_1 = y_1, \ldots, Y_n = y_n, \{p_y\})$$

$$\begin{aligned}
&= \prod_{i=1}^{n} P(X_i = x_i | Y_1 = y_1, \dots, Y_n = y_n, \{p_y\}) \\
&= \prod_{i=1}^{n} P(X_i = x_i | Y_i = y_i, \{p_y\}) \\
&= \prod_{i=1}^{n} P(X_i[1] = x_i[1], \dots, X_i[N] = x_i[N] | Y_i = y_i, \{p_y\}) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{N} P(X_i[j] = x_i[j] | Y_i = y_i, \{p_y\})
\end{aligned}$$

Since log is monotone, we can take the log of our expression to get the arg max.

$$\begin{aligned}
\hat{\theta} = \arg\max_{p, \{p_y\}} &\sum_{i=1}^{n} \sum_{j=1}^{N} \log P(X_i[j] = x_i[j] | Y_i = y_i, \{p_y\}) \\
&+ \sum_{i=1}^{n} y_i \log(p) + (1 - y_i) \log(1 - p) \\
&+ \sum_{i=1}^{N} \sum_{y \in \{0,1\}} (\alpha - 1) \log(p_y[i])
\end{aligned}$$

First, we get $\hat{p}$ by differentiating with respect to p and setting it to zero.

$$\begin{aligned}
\frac{d}{dp}\hat{\theta} &= \sum_{i=1}^{n} \frac{y_i}{p} - \frac{1 - y_i}{1 - p} = 0 \\
\sum_{i=1}^{n} \frac{y_i}{p} &= \sum_{i=1}^{n} \frac{1 - y_i}{1 - p} \\
\frac{1 - p}{p} &= \frac{\sum_{i=1}^{n} 1 - y_i}{\sum_{i=1}^{n} y_i} \\
p &= \frac{\sum_{i=1}^{n} y_i}{n} = \frac{n(1)}{n}
\end{aligned}$$

Where $n(y)$ is the number of y_i's that are equal to y. Before we try and solve for $p_y[i]$, we can do a better job at simplifying the first term.

$$\begin{aligned}
\log P(X_i[j] = x_i[j] | Y_i = y_i, \{p_y\}) &= [[x_i[j] = 1]] \log(p_{y_i}[j]) + [[x_i[j] = 0]] \log(1 - p_{y_i}[j]) \\
&= \log(p_{y_i}[j]^{x_i[j]} (1 - p_{y_i}[j])^{1 - x_i[j]}) \\
&= \sum_{y \in \{0,1\}} [[y_i = y]] \log(p_y[j]^{x_i[j]} (1 - p_y[j])^{1 - x_i[j]}) \\
\implies &\sum_{i=1}^{n} \sum_{j=1}^{N} \log P(X_i[j] = x_i[j] | Y_i = y_i, \{p_y\}) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{N} \sum_{y \in \{0,1\}} [[y_i = y]] \log(p_y[j]^{x_i[j]} (1 - p_y[j])^{1 - x_i[j]}) \\
&= \sum_{j=1}^{N} \sum_{i=1}^{n} \sum_{y \in \{0,1\}} \sum_{x \in \{0,1\}} [[y_i = y \wedge x_i[j] = x]] \log(p_y[j]^{x} (1 - p_y[j])^{1 - x})
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{N} \sum_{y \in \{0,1\}} \sum_{x \in \{0,1\}} \log(p_y[j]^x (1-p_y[j])^{1-x}) \sum_{i=1}^{n} [[y_i = y \wedge x_i[j] = x]] \\
&= \sum_{j=1}^{N} \sum_{y \in \{0,1\}} \sum_{x \in \{0,1\}} \log(p_y[j]^x (1-p_y[j])^{1-x}) n_j(x,y) \\
&= \sum_{j=1}^{N} \sum_{y \in \{0,1\}} \sum_{x \in \{0,1\}} n_j(x,y)(x \log(p_y[j]) + (1-x)\log(1-p_y[j]))
\end{aligned}
$$

Now we differentiate with respect to $p_y[j]$ and set equal to zero.

$$
\begin{aligned}
\frac{d}{dp_y[j]}\hat{\theta} &= \sum_{x \in \{0,1\}} n_j(x,y) \left(\frac{x}{p_y[j]} - \frac{1-x}{1-p_y[j]} \right) + (\alpha - 1)\frac{1}{p_y[j]} \\
&= n_j(1,y)\frac{1}{p_y[j]} - n_j(0,y)\frac{1}{1-p_y[j]} + (\alpha-1)\frac{1}{p_y[j]} = 0
\end{aligned}
$$

$$
\begin{aligned}
\implies \frac{1}{1-p_y[j]} n_j(0,y) &= \frac{1}{p_y[j]}(n_j(1,y) + (\alpha - 1)) \\
\implies \frac{1-p_y[j]}{p_y[j]} &= \frac{n_j(0,y)}{n_j(1,y) + (\alpha-1)} \\
\implies p_y[j] &= \frac{n_j(1,y) + (\alpha-1)}{n_j(0,y) + n_j(1,y) + (\alpha - 1)} \\
\implies \hat{p}_y[j] &= \frac{n_j(1,y) + (\alpha-1)}{n(y) + (\alpha-1)}
\end{aligned}
$$

4. Multiple Classes.

5. Adding Dependencies: A Markov Model.