

## Gaussian Mixtures

1. .

(a) **Parameter Estimation** Our unknown parameters are  $\theta = \{p_+, \mu_-, \mu_+, \text{diag } \Sigma_-, \text{diag } \Sigma_+\}$ .

First we determine the log likelihood of a given sample  $S$ . We denote the indicator function to be

$$[[y_i = 1]] = (1 + y_i)/2$$

and

$$[[y_i = -1]] = (1 - y_i)/2$$

Additionally, we denote the density of a multivariate Gaussian with mean  $\mu$  and covariance  $\Sigma$  to be

$$f(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

We derive the log-likelihood as follows:

$$\begin{aligned} \ell(\theta|S) &= \log P(S|\theta) = \log \prod_{i=1}^m P(x_i, y_i|\theta) = \log \prod_{i=1}^m P(y_i|\theta)P(x_i|y_i, \theta) \\ &= \sum_{i=1}^m \log(P(y_i|\theta)) + \sum_{i=1}^m \log(P(x_i|y_i, \theta)) \\ &= \sum_{i=1}^m [[y_i = 1]] \log(p_+) + [[y_i = -1]] \log(1 - p_+) + \sum_{i=1}^m [[y_i = 1]] \log f(x_i|\mu_+, \Sigma_+) + [[y_i = -1]] \log f(x_i|\mu_-, \Sigma_-) \\ &= \sum_{i=1}^m [[y_i = 1]](\log(p_+) + \log f(x_i|\mu_+, \Sigma_+)) + [[y_i = -1]](\log(1 - p_+) + \log f(x_i|\mu_-, \Sigma_-)) \\ &= \sum_{i=1}^m [[y_i = 1]](\log(p_+) - \frac{1}{2}(x_i - \mu_+)^\top \Sigma_+^{-1}(x_i - \mu_+) - \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_+|) \\ &\quad + [[y_i = -1]](\log(1 - p_+) - \frac{1}{2}(x_i - \mu_-)^\top \Sigma_-^{-1}(x_i - \mu_-) - \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_-|) \end{aligned}$$

From here, we can take the derivatives with respect to each parameter.

(a)  $p_+$  is the probability of a positive sample. We then take the derivative of the log likelihood w.r.t.  $p_+$  and set it to 0, which yields

$$\begin{aligned} \frac{\partial \ell}{\partial p_+} &= \sum_{i=1}^m [[y_i = +1]] \frac{1}{p_+} - \sum_{i=1}^m [[y_i = -1]] \frac{1}{1 - p_+} = 0 \\ \implies \frac{p_+}{1 - p_+} &= \frac{\sum_{i=1}^m [[y_i = +1]]}{\sum_{i=1}^m [[y_i = -1]]} \\ \implies p_+ &= \frac{\sum_{i=1}^m [[y_i = +1]]}{\sum_{i=1}^m [[y_i = +1]] + \sum_{i=1}^m [[y_i = -1]]} \\ \hat{p}_+ &= \frac{\sum_{i=1}^m [[y_i = +1]]}{m} \end{aligned}$$

(b) To find  $\mu_+$ , we take the gradient with respect to  $\mu_+$  and set it to 0.

$$\nabla_{\mu_+} \ell = \sum_{i=1}^m [[y_i = 1]](-1)(\Sigma_+^{-1} + \Sigma_+^{-1\top})(x_i - \mu_+) = 0$$

Since  $\Sigma_+$  is a diagonal matrix, the inverse is symmetric.

$$\begin{aligned}
 0 &= \sum_{i=1}^m [[y_i = 1]] \Sigma_+^{-1} (x_i - \mu_+) \\
 \implies \sum_{i=1}^m [[y_i = 1]] x_i &= \mu_+ \sum_{i=1}^m [[y_i = 1]] \\
 \hat{\mu}_+ &= \frac{\sum_{i=1}^m [[y_i = 1]] x_i}{\sum_{i=1}^m [[y_i = 1]]}
 \end{aligned}$$

(c) The process to find  $\mu_-$  is the same as above, so we have

$$\hat{\mu}_- = \frac{\sum_{i=1}^m [[y_i = -1]] x_i}{\sum_{i=1}^m [[y_i = -1]]}$$

(d) In the cases of  $\Sigma_+$  and  $\Sigma_-$  we thankfully rely on the fact that  $\Sigma$  is diagonal,

$$\begin{aligned}
 \frac{\partial}{\partial \Sigma_+} \ell(\theta|S) &= -\frac{1}{2} \sum_{i=1}^m [[y_i = 1]] \frac{\partial}{\partial \Sigma_+} ((x_i - \mu_+)^\top \Sigma_+^{-1} (x_i - \mu_+) + \log |\Sigma_+|) \\
 &= -\frac{1}{2} \sum_{i=1}^m [[y_i = 1]] (-\Sigma_+^{-\top} (x_i - \mu_+) (x_i - \mu_+)^\top \Sigma_+^{-\top} + \Sigma_+^{-1}) \\
 \implies \sum_{i=1}^m [[y_i = 1]] \Sigma_+^{-1} &= \sum_{i=1}^m [[y_i = 1]] \Sigma_+^{-1} (x_i - \mu_+) (x_i - \mu_+)^\top \Sigma_+^{-1}
 \end{aligned}$$

These derivatives are elementary<sup>1</sup> matrix calculus operations<sup>2</sup>. From here, we simplify further.

$$\begin{aligned}
 \sum_{i=1}^m [[y_i = 1]] I_d &= \sum_{i=1}^m [[y_i = 1]] \Sigma_+^{-1} (x_i - \mu_+) (x_i - \mu_+)^\top \\
 \Sigma_+ \sum_{i=1}^m [[y_i = 1]] &= \sum_{i=1}^m [[y_i = 1]] (x_i - \mu_+) (x_i - \mu_+)^\top \\
 \hat{\Sigma}_+ &= \frac{\sum_{i=1}^m [[y_i = 1]] (x_i - \mu_+) (x_i - \mu_+)^\top}{\sum_{i=1}^m [[y_i = 1]]}
 \end{aligned}$$

(e) The process to find  $\Sigma_-$  is the same as above, so we have

$$\hat{\Sigma}_- = \frac{\sum_{i=1}^m [[y_i = -1]] (x_i - \mu_-) (x_i - \mu_-)^\top}{\sum_{i=1}^m [[y_i = -1]]}$$

To summarize, our MLE estimators are:

$$\begin{aligned}
 \hat{p}_+ &= \frac{\sum_{i=1}^m [[y_i = +1]]}{m} \\
 \hat{\mu}_+ &= \frac{\sum_{i=1}^m [[y_i = +1]] x_i}{\sum_{i=1}^m [[y_i = +1]]} \\
 \hat{\mu}_- &= \frac{\sum_{i=1}^m [[y_i = -1]] x_i}{\sum_{i=1}^m [[y_i = -1]]} \\
 \hat{\Sigma}_+ &= \frac{\sum_{i=1}^m [[y_i = 1]] (x_i - \mu_+) (x_i - \mu_+)^\top}{\sum_{i=1}^m [[y_i = 1]]} \\
 \hat{\Sigma}_- &= \frac{\sum_{i=1}^m [[y_i = -1]] (x_i - \mu_-) (x_i - \mu_-)^\top}{\sum_{i=1}^m [[y_i = -1]]}
 \end{aligned}$$

<sup>1</sup>Wikipedia matrix calculus

<sup>2</sup>MSE post differentiating quadratic form

(b) **Prediction**

$$\begin{aligned}
P(Y = 1|x) &= \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x)} = \frac{P(X = x|Y = 1)p_+}{P(X = x|Y = 1)P(Y = 1) + P(X = x|Y = 0)P(Y = 0)} \\
&= \frac{1}{1 + \frac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=1)P(Y=1)}} = \frac{1}{1 + \frac{1-p_+}{p_+} \frac{f(x|\mu_-, \Sigma_-)}{f(x|\mu_+, \Sigma_+)}}
\end{aligned}$$

We obtain the following discriminant:

$$\begin{aligned}
r(x) &= \log \left( \frac{p_+}{1-p_+} \right) + \log \left( \frac{f(x|\mu_+, \Sigma_+)}{f(x|\mu_-, \Sigma_-)} \right) \\
&= \log \left( \frac{p_+}{1-p_+} \right) + \log \left( \frac{\sqrt{|\Sigma_-|}}{\sqrt{|\Sigma_+|}} \right) - \frac{1}{2}(x - \mu_+)^T \Sigma_+^{-1} (x - \mu_+) + \frac{1}{2}(x - \mu_-)^T \Sigma_-^{-1} (x - \mu_-) \\
&= \log \left( \frac{p_+}{1-p_+} \right) + \frac{1}{2} \log \left( \frac{|\Sigma_-|}{|\Sigma_+|} \right) + \frac{1}{2}(\mu_+^T \Sigma_+^{-1} \mu_+ - \mu_-^T \Sigma_-^{-1} \mu_-) + \frac{1}{2}x^T (\Sigma_-^{-1} - \Sigma_+^{-1})x + x^T (\Sigma_+^{-1} \mu_+ - \Sigma_-^{-1} \mu_-)
\end{aligned}$$

The Bayes predictor is simply

$$h(x) = \text{sign}(r(x))$$

Since, when  $r(x) > 0$ , we have  $P(Y = 1|x) > \frac{1}{2}$ , and when  $r(x) < 0$ , we have  $P(Y = 1|x) < \frac{1}{2}$ .

(c) **As a Linear Predictor** Letting

$$\begin{aligned}
b &= \log \left( \frac{p_+}{1-p_+} \right) + \frac{1}{2} \log \left( \frac{|\Sigma_-|}{|\Sigma_+|} \right) + \frac{1}{2}(\mu_+^T \Sigma_+^{-1} \mu_+ - \mu_-^T \Sigma_-^{-1} \mu_-) \\
\text{diag}(a_1, \dots, a_d) &= \frac{1}{2}(\Sigma_-^{-1} - \Sigma_+^{-1}) \\
v &= \Sigma_+^{-1} \mu_+ - \Sigma_-^{-1} \mu_-
\end{aligned}$$

We can write our discriminant as

$$r(x) = b + x^T A x + x^T v$$

Let  $v = (v_1, \dots, v_d)^T$ . Then we can write

$$r(x) = b + \sum_{i=1}^d a_i x_i^2 + \sum_{i=1}^d v_i x_i$$

Thus, it is clear that with the feature map:

$$\phi : x \mapsto (1, x_1, \dots, x_d, x_1^2, \dots, x_d^2)^T$$

$r$  is a linear predictor. Namely:

$$\begin{aligned}
r(x) &= \langle w, \phi(x) \rangle \\
w &= (b, v_1, \dots, v_d, a_1, \dots, a_d)^T
\end{aligned}$$

This shows that  $D = 2d + 1$  is good enough.

## Modeling Text Documents

## 2. A Simple Model.

- (a) We shall denote  $p_{\text{topic}}$  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  $i$ th coordinate of the sample  $x$ .

Given a sample

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

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

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

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

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

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

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

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

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

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 minimum 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:

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

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:

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

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

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

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:

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

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

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

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

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

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,

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

So we have the following simplification for  $w$ :

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

### 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.

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

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

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

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.

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

Now we focus on the first term.

$$\begin{aligned} p(X|Y, \{p_y\}) &= P(X_1 = x_1, \dots, X_n = x_n|Y_1 = y_1, \dots, Y_n = y_n, \{p_y\}) \\ &= \prod_{i=1}^n P(X_i = x_i|Y_1 = y_1, \dots, Y_n = y_n, \{p_y\}) \\ &= \prod_{i=1}^n P(X_i = x_i|Y_i = y_i, \{p_y\}) \\ &= \prod_{i=1}^n P(X_i[1] = x_i[1], \dots, X_i[N] = x_i[N]|Y_i = y_i, \{p_y\}) \end{aligned}$$



$$= \prod_{i=1}^n \prod_{j=1}^N P(X_i[j] = x_i[j] | Y_i = y_i, \{p_y\})$$

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.

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

Where  $n_j(x, y)$  is the number of  $(x_i, y_i)$ 's such that  $x_i[j] = x$  and  $y_i = y$ . We differentiate with respect to  $p_y[j]$  and set equal to zero.

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

(b) Recall that the discriminant is given by

$$r(x) = \log \frac{p}{1-p} + \sum_{i=1}^N \log \frac{p_1[i]^{x[i]} (1-p_1[i])^{1-x[i]}}{p_0[i]^{x[i]} (1-p_0[i])^{1-x[i]}}$$

We can simplify this

$$\begin{aligned}
r(x) &= \log \frac{p}{1-p} + \sum_{i=1}^N \sum_{x \in \{0,1\}} [[x[i] = x]] \left( x \log \frac{p_1[i]}{p_0[i]} + (1-x) \log \frac{1-p_1[i]}{1-p_0[i]} \right) \\
&= \log \frac{p}{1-p} + \sum_{i=1}^N [[x[i] = 1]] \log \frac{p_1[i]}{p_0[i]} + [[x[i] = 0]] \log \frac{1-p_1[i]}{1-p_0[i]}
\end{aligned}$$

Taking a slightly modified feature map from before, i.e.

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

We have the following  $w \in \mathbb{R}^N$  and  $b \in \mathbb{R}$  such that  $r(x) = \langle w, \phi(x) \rangle + b$ :

$$\begin{aligned}
w[i] &= \log \frac{p_1[i]}{p_0[i]} - \log \frac{1-p_1[i]}{1-p_0[i]} \\
b &= \log \frac{p}{1-p} + \sum_{i=1}^N \log \frac{1-p_1[i]}{1-p_0[i]}
\end{aligned}$$

Now we can easily plug in our MAP estimators.

$$\begin{aligned}
w[i] &= \log \frac{n_i(1, 1) + (\alpha - 1)}{n(1) + (\alpha - 1)} \frac{n(0) + (\alpha - 1)}{n_i(1, 0) + (\alpha - 1)} - \log \frac{n_i(0, 1)}{n(1) + (\alpha - 1)} \frac{n(0) + (\alpha - 1)}{n_i(0, 0)} \\
&= \log \frac{n_i(1, 1) + (\alpha - 1)}{n_i(1, 0) + (\alpha - 1)} - \log \frac{n_i(0, 1)}{n_i(0, 0)} \\
b &= \log \frac{n(1)/n}{n(0)/n} + \sum_{i=1}^N \log \frac{n_i(0, 1)}{n(1) + (\alpha - 1)} \frac{n(0) + (\alpha - 1)}{n_i(0, 0)} \\
&= \log \frac{n(1)}{n(0)} + \sum_{i=1}^N \log \frac{n_i(0, 1)}{n(1) + (\alpha - 1)} \frac{n(0) + (\alpha - 1)}{n_i(0, 0)}
\end{aligned}$$

#### 4. Multiple Classes.

(a) Let  $p(k) = P(Y = k)$  and  $p_i(x|y) = P(X[i] = x|Y = y)$ .

$$\begin{aligned} P(Y = y|x) &= \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)} = \frac{P(X[1] = x[1], \dots, X[N] = x[N]|Y = y)p(y)}{P(X = x|Y = 1)P(Y = 1) + \dots + P(X = x|Y = k)P(Y = k)} \\ &= \frac{p(y) \prod_{i=1}^N p_i(x[i]|y)}{\sum_{j=1}^k p(j) \prod_{i=1}^N p_i(x[i]|j)} \end{aligned}$$

(b) We solve the equation:

$$\exp(\langle w_y, \phi(x) \rangle) = p(y) \prod_{i=1}^N p_i(x[i]|y)$$

We take the log of both sides and simplify.

$$\begin{aligned} \langle w_y, \phi(x) \rangle &= \log(p(y)) + \sum_{i=1}^N \log(p_i(x[i]|y)) \\ &= \log(p(y)) + \sum_{i=1}^N [[x[i] = 1]] \log(p_i(1|y)) + [[x[i] = 0]] \log(p_i(0|y)) \end{aligned}$$

We let the feature map be

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

And we let  $w_y \in \mathbb{R}^{N+1}$  be

$$w_y = \left( (\log(p_i(1|y)) - \log(p_i(0|y)))_{i=1}^N, \log(p(y)) + \sum_{i=1}^N \log(p_i(0|y)) \right)$$

(c) For convenience, we write the bias term separate from  $w$ . The process for computing the MAP estimators for the parameters is almost entirely the same as before. Our objective function, for a sample  $S = X \times Y$  is as follows:

$$\begin{aligned} p(X|Y, \{p_y\})p(Y|p)p(p, \{p_y\}) &= \prod_{i=1}^n p(X = x_i|Y = y_i, \{p_y\})p(Y = y_i|p)p(p)p(\{p_y\}) \\ &= \prod_{i=1}^n \prod_{j=1}^N \prod_{l=1}^k p(X[j] = x_i[j]|Y_i = y_i, \{p_y\})p(Y = y_i|p) \frac{1}{Z(1)} p_l^{1-1} \frac{1}{Z(\alpha)^k} p_l[j]^{\alpha-1} \\ &= \frac{1}{Z(\alpha)^k} \prod_{i=1}^n \prod_{j=1}^N \prod_{l=1}^k p_j(x_i[j]|y_i)p(y_i)p_l[j]^{\alpha-1} \end{aligned}$$

Again, we can disregard the constant, and take log, to get the following:

$$\begin{aligned} &\sum_{i=1}^n \sum_{j=1}^N \log p_j(x_i[j]|y_i) + \sum_{i=1}^n \log(p(y_i)) + \sum_{l=1}^k \sum_{j=1}^N (\alpha - 1) \log(p_l[j]) \\ &= \sum_{j=1}^N \sum_{l=1}^k \sum_{x \in \{0,1\}} \sum_{i=1}^n [[x_i[j] = x \wedge y_i = l]] \log(p_j(x|l)) \\ &+ \sum_{l=1}^k \sum_{i=1}^n [[y_i = l]] \log(p(l)) + \sum_{l=1}^k \sum_{j=1}^N (\alpha - 1) \log(p_l[j]) \\ &= \sum_{j=1}^N \sum_{l=1}^k \sum_{x \in \{0,1\}} n_j(x, l) \log(p_j(x|l)) + \sum_{l=1}^k n(l) \log(p(l)) + \sum_{l=1}^k \sum_{j=1}^N (\alpha - 1) \log(p_l[j]) \end{aligned}$$

$$= \sum_{j=1}^N \sum_{l=1}^k n_j(1, l) \log(p_l[j]) + n_j(0, l) \log(1 - p_l[j]) + \sum_{l=1}^k n(l) \log(p(l)) + \sum_{l=1}^k \sum_{j=1}^N (\alpha - 1) \log(p_l[j])$$

Since we no longer have two  $p(l)$ 's, we need to apply Lagrange multipliers, with the constraint:

$$\sum_{l=1}^k p(l) = 1$$

We can write the Lagrangian, together with the first order conditions, as follows:

$$\begin{aligned} \mathcal{L} &= \sum_{l=1}^k n(l) \log(p(l)) - \lambda \left( \sum_{l=1}^k p(l) - 1 \right) \\ [p(l)] : \frac{n(l)}{p(l)} &= \lambda \\ [\lambda] : \sum_{l=1}^k p(l) &= 1 \\ \implies 1 &= p(l) + \sum_{m \neq l} \frac{n(m)}{n(l)} p(l) \\ \implies n(l) &= p(l) \sum_{l=1}^k n(l) = p(l)n \\ \implies \hat{p}(l) &= \frac{n(l)}{n} \end{aligned}$$

We can now differentiate with respect to  $p_l[j]$  and set equal to zero.

$$\begin{aligned} 0 &= \frac{n_j(1, l)}{p_l[j]} - \frac{n_j(0, l)}{1 - p_l[j]} + \frac{\alpha - 1}{p_l[j]} \\ \frac{p_l[j]}{1 - p_l[j]} &= \frac{n_j(1, l) + (\alpha - 1)}{n_j(0, l)} \\ \hat{p}_l[j] &= \frac{n_j(1, l) + (\alpha - 1)}{n_j(0, l) + n_j(1, l) + (\alpha - 1)} = \frac{n_j(1, l) + (\alpha - 1)}{n(l) + (\alpha - 1)} \end{aligned}$$

Plugging  $\hat{p}(l)$  and  $\hat{p}_l[j]$  into  $w_y$ , we have:

$$\begin{aligned} w_y[j] &= \log p_y[j] - \log(1 - p_y[j]) = \log \frac{n_j(1, y) + (\alpha - 1)}{n(y) + (\alpha - 1)} - \log \frac{n_j(0, y) + (\alpha - 1)}{n(y) + (\alpha - 1)} \\ &= \log \frac{n_j(1, y) + (\alpha - 1)}{n_j(0, y) + (\alpha - 1)} \\ b &= \log(p(y)) + \sum_{j=1}^N \log(1 - p_y[j]) = \log \frac{n(y)}{n} + \sum_{j=1}^N \log \frac{n_j(0, y) + (\alpha - 1)}{n(y) + (\alpha - 1)} \end{aligned}$$

(d) We write  $-\log P(\{y_i\}|\{x_i\}, \{w_y\}) = -\log P(Y|X, W)$ . Recall the posterior:

$$P(Y = y|x) = \frac{\exp(r_y(x))}{\sum_{l=1}^k \exp(r_l(x))} = \frac{\exp(\langle w_y, \phi(x) \rangle)}{\sum_{l=1}^k \exp(\langle w_l, \phi(x) \rangle)}$$

Noting that  $(x_i, y_i)$  are i.i.d., we have:

$$P(Y|X, w) = \prod_{i=1}^n P(Y_i = y_i|X_i = x_i, w) = \prod_{i=1}^n \frac{\exp(\langle w_{y_i}, \phi(x_i) \rangle)}{\sum_{l=1}^k \exp(\langle w_l, \phi(x_i) \rangle)}$$

$$-\log P(Y|X, w) = \sum_{i=1}^n \log \sum_{l=1}^k \exp(\langle w_l, \phi(x_i) \rangle) - \sum_{i=1}^n \langle w_{y_i}, \phi(x_i) \rangle$$

(e)

$$\begin{aligned} -\log P(Y|X, w) &= \sum_{i=1}^n \left( \log \left( \sum_{l=1}^k \exp(\langle w_l, \phi(x_i) \rangle) \right) - \langle w_{y_i}, \phi(x_i) \rangle \right) \\ \ell(y_i; r_1(x), \dots, r_k(x)) &= \log \left( \sum_{l=1}^k \exp(r_l(x)) \right) - \log(\exp(r_{y_i}(x))) \\ &= \log \left( \frac{\exp(r_{y_i}(x))}{\sum_{l=1}^k \exp(r_l(x))} \right) = \log \left( \sum_{l=1}^k \exp(r_l(x) - r_{y_i}(x)) \right) \\ &= \log \left( \sum_{l=1}^k \exp(\langle w_l - w_{y_i}, \phi(x) \rangle) \right) \end{aligned}$$

### 5. Adding Dependencies: A Markov Model.