TTIC 31020: Introduction to Machine Learning

Problem Set 5

Hung Le Tran

02 Feb 2024

Gaussian Mixtures

Problem 5.1

(a) Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Denote the map $g(y) = \frac{y+1}{2}$ which maps -1 to 0 and 1 to 1, and $h(y) = 1 - g(y)$ which maps -1 to 1 and 1 to 0. Then

$$\begin{aligned}
&\mathbb{P}[(x_i, y_i); p_+, \mu_+, \mu-, \Sigma_-, \Sigma_+] \\
&\qquad = \left(p_+ \frac{1}{\sqrt{(2\pi)^d |\Sigma_+|}} \exp\left(-\frac{1}{2}(x - \mu_+)^T \Sigma_+^{-1} (x - \mu_+)\right)\right)^{g(y_i)} \\
&\qquad\qquad \left((1 - p_+) \frac{1}{\sqrt{(2\pi)^d |\Sigma_-|}} \exp\left(-\frac{1}{2}(x - \mu_-)^T \Sigma_-^{-1} (x - \mu_-)\right)\right)^{h(y_i)}
\end{aligned}$$

so

$$\begin{aligned}
&\log \mathbb{P}[(x_i, y_i); p_+, \mu_+, \mu-, \Sigma_-, \Sigma_+] \\
&\qquad = g(y_i)\left(\log p_+ - \log\sqrt{(2\pi)^d} - \frac{1}{2}\log|\Sigma_+| - \frac{1}{2}(x_i - \mu_+)^T \Sigma_+^{-1} (x_i - \mu_+)\right) \\
&\qquad\quad + h(y_i)\left(\log(1 - p_+) - \log\sqrt{(2\pi)^d} - \frac{1}{2}\log|\Sigma_-| - \frac{1}{2}(x_i - \mu_-)^T \Sigma_-^{-1} (x_i - \mu_-)\right)
\end{aligned}$$

Abuse of notation: $\mathbb{P}((x_i, y_i)) = \mathbb{P}[(x_i, y_i); p_+, \mu_+, \mu-, \Sigma_-, \Sigma_+]$, $P(S) = \mathbb{P}[S; p_+, \mu_+, \mu-, \Sigma_-, \Sigma_+]$.

Since training samples are drawn i.i.d,

$$\mathbb{P}(S) = \prod_{i=1}^{m} \mathbb{P}((x_i, y_i)) \Rightarrow \log \mathbb{P}(S) = \sum_{i=1}^{m} \log \mathbb{P}((x_i, y_i))$$

Let $m_1 = \sum_{i=1}^{m} \mathbb{1}\{y_i = 1\}, m_0 = m - m_1$. Then we can perform MLE:

For p_+:

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}p_+}\sum_{i=1}^{m}\log\mathbb{P}((x_i, y_i)) \\
&= \sum_{y_i=1}\frac{1}{p_+} + \sum_{y_i=-1}\frac{-1}{1-p_+} \\
&= \frac{m_1}{p_+} - \frac{m-m_1}{1-p_+} \\
\Rightarrow p_+ &= \frac{m_1}{m}
\end{aligned}$$

For μ_+:

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}u}\sum_{i=1}^{m}\log\mathbb{P}((x_i, y_i)) \\
&= -\frac{1}{2}\sum_{y_i=1}2\Sigma_+^{-1}(x_i-\mu) \\
\Rightarrow 0 &= \sum_{y_i=1}\Sigma_+^{-1}(x_i-\mu_+) \\
\Rightarrow \hat{\mu}_+ &= \frac{1}{m_1}\sum_{y_i=1}x_i
\end{aligned}$$

Similarly

$$\hat{\mu}_- = \frac{1}{m_0}\sum_{y_i=-1}x_i$$

For Σ_+, note that Σ_+ is diagonal so its inverse is diagonal too. We have:

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\Sigma_+^{-1}}\sum_{i=1}^{m}\log\mathbb{P}((x_i, y_i)) \\
&= \frac{\mathrm{d}}{\mathrm{d}\Sigma_+^{-1}}\sum_{y_i=1}\left[\frac{-1}{2}\log\left(\frac{1}{|\Sigma_+^{-1}|}\right) - \frac{1}{2}(x_i-\mu_+)^T\Sigma_+^{-1}(x_i-\mu_+)\right] \\
&= \frac{1}{2}\sum_{y_i=1}\frac{\mathrm{d}}{\mathrm{d}\Sigma_+^{-1}}\left[\log(|\Sigma_+^{-1}|) - (x_i-\mu_+)^T\Sigma_+^{-1}(x_i-\mu_+)\right] \\
&= \frac{1}{2}\sum_{y_i=1}\left[\Sigma - (x_i-\mu_+)(x_i-\mu_+)^T\right] \\
\Rightarrow \hat{\Sigma}_+ &= \frac{1}{m_1}\sum_{y_i=1}(x_i-\hat{\mu}_+)(x_i-\hat{\mu}_+)^T
\end{aligned}$$

Similarly,

$$\hat{\Sigma}_- = \frac{1}{m_0}\sum_{y_i=-1}(x_i-\hat{\mu}_-)(x_i-\hat{\mu}_-)^T$$

(b) Computing posterior:

$$
\begin{aligned}
\eta(x) &= \mathbb{P}(Y=1 \mid x) \\
&= \frac{\mathbb{P}(X=x \mid Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x \mid Y=1)\mathbb{P}(Y=1)+\mathbb{P}(X=x \mid Y=-1)\mathbb{P}(Y=-1)} \\
&= \frac{p_+ \frac{1}{\sqrt{(2\pi)^d|\Sigma_+|}} \exp\left(-\frac{1}{2}(x-\mu_+)^T\Sigma_+^{-1}(x-\mu_+)\right)}{p_+ \frac{1}{\sqrt{(2\pi)^d|\Sigma_+|}} \exp\left(-\frac{1}{2}(x-\mu_+)^T\Sigma_+^{-1}(x-\mu_+)\right) + (1-p_+)\frac{1}{\sqrt{(2\pi)^d|\Sigma_-|}} \exp\left(-\frac{1}{2}(x-\mu_-)^T\Sigma_-^{-1}(x-\mu_-)\right)}
\end{aligned}
$$

We know that

$$
\begin{aligned}
\mathbb{P}(Y=1 \mid x) &= \frac{\mathbb{P}(X=x \mid Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x \mid Y=1)\mathbb{P}(Y=1)+\mathbb{P}(X=x \mid Y=-1)\mathbb{P}(Y=-1)} \\
&= \frac{1}{1+\frac{\mathbb{P}(X=x|Y=-1)\mathbb{P}(Y=-1)}{\mathbb{P}(X=x|Y=+1)\mathbb{P}(Y=+1)}}
\end{aligned}
$$

so

$$
\begin{aligned}
r(x) &= -\log\frac{\mathbb{P}(X=x \mid Y=-1)\mathbb{P}(Y=-1)}{\mathbb{P}(X=x \mid Y=+1)\mathbb{P}(Y=+1)} \\
&= \log(X=x \mid Y=+1)+\log\mathbb{P}(Y=+1)-\log(X=x \mid Y=-1)-\log(Y=-1) \\
&= \left(\log p_+ - \log\sqrt{(2\pi)^d} - \frac{1}{2}\log|\Sigma_+| - \frac{1}{2}(x-\mu_+)^T\Sigma_+^{-1}(x-\mu_+)\right) \\
&- \left(\log(1-p_+) - \log\sqrt{(2\pi)^d} - \frac{1}{2}\log|\Sigma_-| - \frac{1}{2}(x-\mu_-)^T\Sigma_-^{-1}(x-\mu_-)\right) \\
&= \frac{1}{2}x^T(\Sigma_-^{-1}-\Sigma_+^{-1})x + x^T(\Sigma_+^{-1}\mu_+ - \Sigma_-^{-1}\mu_-) \\
&+ \left(\log\frac{p_+}{1-p_+} - \frac{1}{2}\mu_+^T\Sigma_+^{-1}\mu_+ + \frac{1}{2}\mu_-^T\Sigma_-^T\mu_- - \frac{1}{2}\log|2\pi\Sigma_+| + \frac{1}{2}\log|2\pi\Sigma_-|\right)
\end{aligned}
$$

The Bayes predictor is then $h_{Bayes}(x) = \operatorname{sign}(r(x))$

(c) $r(x)$ has a leading quadratic term in x, but since both Σ_+ and Σ_- are diagonal, we only have to be concerned with terms $x[i]^2, x[i], 1$, which total to $D = d+d+1 = 2d+1$. The mapping is therefore (1-index):

$$
\phi(x) = [x[1]^2 \cdots x[d]^2 \quad x[1] \cdots x[d] \quad 1]
$$

then the corresponding w would be

$$
w[i] = \begin{cases}
(\Sigma_+^{-1}-\Sigma_-^{-1})[i,i] & \text{for } 1 \le i \le d \\
(\Sigma_+^{-1}\mu_+ - \Sigma_-^{-1}\mu_-)[i-n] & \text{for } d+1 \le i \le 2d \\
\left(\log\frac{p_+}{1-p_+} - \frac{1}{2}\mu_+^T\Sigma_+^{-1}\mu_+ + \frac{1}{2}\mu_-^T\Sigma_-^T\mu_- - \frac{1}{2}\log|2\pi\Sigma_+| + \frac{1}{2}\log|2\pi\Sigma_-|\right) & \text{for } i = 2d+1
\end{cases}
$$

(d) Let $L = \max\{|w[i]|+1\}$ then we can choose $\Sigma_- = \frac{1}{L}I$ so that $\Sigma_-^{-1} = LI$, hence from the first d equations, we have

$$
\Sigma_+^{-1} = diag(w+L) \Rightarrow \Sigma_+ = diag\left(\frac{1}{w+L}\right)
$$

where arithmetic operations $w+L$ and $\frac{1}{w+L}$ are element-wise. This ensures that the covariance matrices are semi positive definite, since they are diagonal and all diagonal entries are positive (at least 1).

Then, for $i \in [n]$, we have

$$(w[i]+L)\mu_+[i] - L\mu_-[i] = w[i+n]$$

Choose $\mu_-[i] = 0 \ \forall\, i \in [n]$ then

$$\mu_+[i] = \frac{w[i+n]}{w[i]+L}$$

well-defined, again, because $L > |w[i]| \ \forall\, i \in [n]$.

For the last equation,

$$w[2d+1] = \log \frac{p_+}{1-p_+} - \frac{1}{2}\sum_{i=1}^{d}\left(\frac{w[i+n]}{w[i]+L}\right)^2 (w[i]+L) \\ - \frac{1}{2}2\pi \sum_{i=1}^{d} \log\left(\frac{1}{w[i]+L}\right) - \frac{1}{2} + \frac{1}{2}2\pi \sum_{i=1}^{d}\frac{1}{L}$$

and one can solve for p_+, as image of $\log\left(\frac{x}{1-x}\right)$ for $x \in [0,1]$ is $\mathbb{R}$.

(e) Geometrically it is a hyperplane in D-dimensional space/quadratic curve in d-dimensional space.

Problem 5.2

Note: In this problem onward, I've made the unfortunate mistake of letting D be the dictionary and only realized that it might be confusing too late until the pset... Hope it's not too confusing, as I do not make explicit reference to the D as in dimension of the feature space.

(a) For now we have 2 topics, with, say $\mathcal{Y} = \{0,1\}$ and $\mathbb{P}(Y=+1) = p_{topic}, \mathbb{P}(Y=0) = 1-p_{topic}$. Let D be the dictionary and $t \in D$ be a typical word. Then

$$\mathbb{P}((x_i,y_i); p_{topic}, \{p_y\}) = \left(p_{topic}\prod_{t\in D} p_1[t]^{m(t,x_i)}\right)^{y_i}\left((1-p_{topic})\prod_{t\in D} p_0[t]^{m(t,x_i)}\right)^{1-y_i}$$

where $m(t,x_i)$ counts the number of appearances of word t in string x_i.

Therefore

$$\log \mathbb{P}((x_i,y_i); p_{topic}, \{p_y\}) = y_i\left(\log p_{topic} + \sum_{t\in D} m(t,x_i)\log(p_1[t])\right) \\ + (1-y_i)\left(\log(1-p_{topic}) + \sum_{t\in D} m(t,x_i)\log(p_0[t])\right)$$

Since $S = \{(x_i,y_i) : i \in [m]\}$ are drawn i.i.d, we have that

$$\log \mathbb{P}(S; p_{topic}, \{p_y\}) = \sum_{i=1}^{m} \log \mathbb{P}((x_i,y_i); p_{topic}, \{p_y\})$$

Let $m_1 = \sum_{i=1}^m \mathbb{1}\{y_i = 1\}, m_0 = m - m_1 = \sum_{i=1}^m \mathbb{1}\{y_i = 0\}$. Then for p_{topic}:

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}p_{topic}} \log \mathbb{P}(S; p_{topic}, \{p_y\}) \\
&= \frac{m_1}{p_{topic}} - \frac{m_0}{1 - p_{topic}} \\
\Rightarrow \hat{p}_{topic} &= \frac{m_1}{m}
\end{aligned}$$

For p_1, we have constraint: $\sum_{t \in D} p_1[t] = 1$, while having maximize $\log(S; p_{topic}, \{p_y\})$. Use Lagrange multiplier:

$$\begin{aligned}
0 &= \sum_{y_i=1} \sum_{t \in D} \frac{m(t, x_i)}{p_1[t]} - \lambda \\
&= \frac{n_1[t]}{p_1[t]} - \lambda \\
\Rightarrow \hat{\lambda} &= \frac{n_1[t]}{\hat{p}_1[t]}
\end{aligned}$$

where $n_1[t]$ is the number appearances of word t in all positive-labeled training samples.

It follows that for $t \in D$,

$$\hat{p}_1[t] = \frac{n_1[t]}{\sum_{t \in D} n_1[t]} = \frac{n_1[t]}{100 m_1}$$

Similarly,

$$\hat{p}_0[t] = \frac{n_0[t]}{100 m_0}$$

(b)

$$\begin{aligned}
r(x) &= \log\left(\frac{\mathbb{P}(Y = +1)}{\mathbb{P}(Y = 0)}\right) + \log(\mathbb{P}(X = x \mid Y = 1)) - \log(\mathbb{P}(X = x \mid Y = 0)) \\
&= \log\left(\frac{p_{topic}}{1 - p_{topic}}\right) + \sum_{t \in D} m(t, x) \log(p_1[t]) - \sum_{t \in D} m(t, x) \log(p_0[t]) \\
&= \log\left(\frac{p_{topic}}{1 - p_{topic}}\right) + \sum_{t \in D} m(t, x) \log\left(\frac{p_1[t]}{p_0[t]}\right)
\end{aligned}$$

(c) The feature map is

$$\phi(x) = [m(t_1, x) \cdots m(t_{|D|}, x) \quad 1]^T$$

where $D = \{t_1, \ldots, t_{|D|}\}$, with the weight corresponding to $r(x)$ being:

$$w = \left[\log\left(\frac{p_1[t_1]}{p_0[t_1]}\right) \quad \cdots \quad \log\left(\frac{p_1[t_{|D|}]}{p_0[t_{|D|}]}\right) \quad \log\left(\frac{p_{topic}}{1 - p_{topic}}\right)\right]^T$$

Dimension is $|D| + 1$.

(d) Computing the weight that corresponds to MLE parameters:

For $i \in [|D|]$,

$$\begin{aligned} w[i] &= \log\left(\frac{\hat{p}_1[t_i]}{\hat{p}_0[t_i]}\right) \\ &= \log\left(\frac{n_1[t_i]m_0}{n_0[t_i]m_1}\right) \end{aligned}$$

and

$$\begin{aligned} w[|D|+1] &= \log\left(\frac{\hat{p}_{topic}}{1-\hat{p}_{topic}}\right) \\ &= \log\frac{m_1}{m_0} \end{aligned}$$

Problem 5.3

(a) We have to maximize: $\mathbb{P}(p_{topic}, \{p_y\} \mid S)$. We also have

$$\begin{aligned} \arg\max \mathbb{P}(p_{topic}, \{p_y\} \mid S) &= \arg\max \frac{1}{C}\mathbb{P}(S \mid p_{topic}, \{p_y\})\mathbb{P}(p_{topic}, \{p_y\}) \\ &= \arg\max \mathbb{P}(S \mid p_{topic}, \{p_y\})\mathbb{P}(p_{topic}, \{p_y\}) \\ &= \arg\max \log \mathbb{P}(S \mid p_{topic}, \{p_y\})\mathbb{P}(p_{topic}, \{p_y\}) \end{aligned}$$

We know that $p_{topic} \sim Dir(1)$ so

$$\mathbb{P}(p_{topic} = p) = \frac{1}{Z(1)}$$

Then

$$\begin{aligned} &\log \mathbb{P}(S \mid p_{topic}, \{p_y\})\mathbb{P}(p_{topic}, \{p_y\}) \\ =&\log \mathbb{P}(p_{topic}, \{p_y\}) + \sum_{i=1}^{m} \log \mathbb{P}((x_i, y_i) \mid p_{topic}, \{p_y\}) \\ =&\log\left(\frac{1}{Z(1)}\frac{1}{Z(\alpha)}\prod_{t\in D} p_1[t]^{\alpha-1}\frac{1}{Z(\alpha)}\prod_{t\in D} p_0[t]^{\alpha-1}\right) + \sum_{y_i=1}\left(\log p_{topic} + \sum_{t\in D} m(t, x_i)\log(p_1[t])\right) \\ &+ \sum_{y_i=0}\left(\log(1-p_{topic}) + \sum_{t\in D} m(t, x_i)\log(p_0[t])\right) \\ =&C + (\alpha-1)\sum_{t\in D}(\log(p_1[t]) + \log(p_0[t])) + m_1 \log p_{topic} \\ &+ n_1[t]\log(p_1[t]) + m_0\log(1-p_{topic}) + n_0[t]\log(p_0[t]) \end{aligned}$$

Now we can do MAP estimation: For p_{topic}:

$$\begin{aligned} 0 &= \frac{m_1}{p_{topic}} - \frac{m_0}{1-p_{topic}} \\ \Rightarrow \hat{p}_{topic} &= \frac{m_1}{m} \end{aligned}$$

For $p_1[t]$, we can use Lagrange multipliers again:

$$0 = \frac{\alpha - 1}{p_1[t]} + \frac{n_1[t]}{p_1[t]} - \lambda_1$$
$$\Rightarrow \hat{\lambda}_1 = \frac{\alpha - 1 + n_1[t]}{p_1[t]}$$

It follows that

$$\hat{p}_1[t] = \frac{\alpha - 1 + n_1[t]}{|D|(\alpha - 1) + 100m_1}$$

Similarly

$$\hat{p}_0[t] = \frac{\alpha - 1 + n_0[t]}{|D|(\alpha - 1) + 100m_0}$$

(b) We found from Problem 2 that

$$w = \left[\log\left(\frac{p_1[t_1]}{p_0[t_1]}\right) \quad \cdots \quad \log\left(\frac{p_1[t_{|D|}]}{p_0[t_{|D|}]}\right) \quad \log\left(\frac{p_{topic}}{1 - p_{topic}}\right)\right]^T$$

So for $i \in [|D|]$:

$$w[i] = \log\left(\frac{(\alpha - 1 + n_1[t])(|D|(\alpha - 1) + 100m_1)}{(\alpha - 1 + n_0[t])(|D|(\alpha - 1) + 100m_0)}\right)$$

and

$$w[|D| + 1] = \log\left(\frac{m_1}{m_0}\right)$$

Problem 5.4

(a) Bayes' rule gives us

$$\mathbb{P}(Y = y \mid x) = \frac{\mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y)}{\sum_{y \in \mathcal{Y}} \mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y)}$$

(b) We now have $p_{topic} \in \mathbb{R}^k$. Use $y \in \mathcal{Y}$ to index p_{topic}.

Then

$$\begin{aligned}
\mathbb{P}(Y = y \mid x) &= \frac{\mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y)}{\sum_{y \in \mathcal{Y}} \mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y)} \\
&= \frac{\left(\prod_{t \in D} p_y[t]^{m(t,x)}\right) p_{topic}[y]}{\sum_{y' \in \mathcal{Y}} \left(\prod_{t \in D} p_{y'}[t]^{m(t,x)}\right) p_{topic}[y']} \\
&= \frac{\exp\left(\log p_{topic}[y] + \sum_{t \in D} m(t,x) \log p_y[t]\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\log p_{topic}[y'] + \sum_{t \in D} m(t,x) \log p_{y'}[t]\right)}
\end{aligned}$$

so we can define $r_y(x) = \log p_{topic}[y] + \sum_{t \in D} m(t,x) \log p_y[t]$ to get the desired form.

Then we can use feature map:

$$\phi(x) = [m(t_1, x) \quad \cdots \quad m(t_{|D|}, x) \quad 1]^T$$

then the corresponding weight would be

$$w_y = [\log p_y[t_1] \quad \cdots \quad \log p_y[t_{|D|}] \quad \log p_{topic}[y]]^T$$

MAP estimation:

$$\begin{aligned}\arg\max \mathbb{P}(p_{topic}, \{p_y\} \mid S) &= \arg\max \mathbb{P}(S \mid p_{topic}, \{p_y\})\mathbb{P}(p_{topic}, \{p_y\}) \\ &= \arg\max \log \mathbb{P}(S \mid p_{topic}, \{p_y\})\mathbb{P}(p_{topic}, \{p_y\})\end{aligned}$$

Let $m_y = \sum_{i=1}^{m} \mathbb{1}\{y_i = y\}, n_y[t] = \sum_{y_i=y} m(t, x_i)$. Then

$$\begin{aligned}&\log \mathbb{P}(S \mid p_{topic}, \{p_y\})\mathbb{P}(p_{topic}, \{p_y\}) \\ =&\log \mathbb{P}(p_{topic}, \{p_y\}) + \sum_{i=1}^{m} \log\{\mathbb{P}((x_i, y_i) \mid p_{topic}, \{p_y\})\} \\ =&\log \left(\frac{1}{Z(1)} \prod_{y' \in \mathcal{Y}} \left(\frac{1}{Z(\alpha)} \prod_{t \in D} p_{y'}[t]^{\alpha-1} \right) \right) + \sum_{y' \in \mathcal{Y}} \sum_{y_i = y'} \left(\log p_{topic}[y'] + \sum_{t \in D} m(t, x_i) \log(p_{y'}[t]) \right) \\ =&C + (\alpha - 1) \sum_{y' \in \mathcal{Y}} \sum_{t \in D} \log(p_{y'}[t]) + \sum_{y' \in \mathcal{Y}} m_{y'} \left(\log p_{topic}[y'] + n_{y'}[t] \log p_{y'}[t] \right)\end{aligned}$$

Use Lagrange multipliers to solve for $p_{topic}[y]$:

$$\begin{aligned}0 &= \frac{m_y}{p_{topic}[y]} - \lambda_{topic} \\ \Rightarrow \hat{p}_{topic}[y] &= \frac{m_y}{m}\end{aligned}$$

And for each $p_y[t]$:

$$0 = \frac{\alpha - 1}{p_y[t]} + \frac{n_y[t]}{p_y[t]} - \lambda_y = \frac{\alpha - 1 + n_y[t]}{p_y[t]} - \lambda_y$$

hence

$$\hat{p}_y[t] = \frac{\alpha - 1 + n_y[t]}{|D|(\alpha - 1) + 100 m_y}$$

then

$$\begin{aligned}\hat{w}_y &= [\log (p_{topic}[y]) \quad \log p_y[t_1] \quad \cdots \quad \log p_y[t_{|D|}]]^T \\ &= [\log \frac{m_y}{m} \quad \log \left(\frac{\alpha - 1 + n_y[t]}{|D|(\alpha - 1) + 100 m_y} \right) \quad \cdots]^T\end{aligned}$$

(d)

$$
\begin{aligned}
& -\log \mathbb{P}(y_i \mid x_i, \{w_y\}) \\
&= -\log\left(\frac{\exp(r_{y_i}(x))}{\sum_{y'\in\mathcal{Y}}\exp(r_{y'}(x))}\right) \\
&= -r_{y_i}(x) + \log\left(\sum_{y'\in\mathcal{Y}}\exp(r_{y'}(x))\right) \\
\Rightarrow & -\log \mathbb{P}(\{y_i\} \mid \{x_i\}, \{w_y\}) \\
&= \sum_{i=1}^{m} -\log \mathbb{P}(y_i \mid x_i, \{w_y\}) \\
&= \sum_{i=1}^{m}\left[-r_{y_i}(x) + \log\left(\sum_{y'\in\mathcal{Y}}\exp(r_{y'}(x))\right)\right]
\end{aligned}
$$

(e) The loss form is:

$$
l(y_i; r_1(x), \ldots, r_k(x)) = -r_i(x) + \log\left(\sum_{j=1}^{k}\exp(r_j(x))\right)
$$

Problem 5.5

State explicitly that $p_{y,tran}[i,j] = \mathbb{P}(w[t+1] = i \mid w[t] = j)$.

Let $N = 100$.

Denote $RL(t, t', x_i)$ as the number of times the word t appears to the immediate right of the word t' in sentence x_i.

Denote $S(t, x_i) = \mathbb{1}\{x_i[1] = t\}$, i.e., if sentence x_i starts with word t.

Then define the total counts $RLT(t, t', y) = \sum_{y_i=y} RL(t, t', x_i)$; $ST(t, y) = \sum_{y_i=y} S(t, x_i)$.

We also make $p_{topic}[k]$ synonymous with $p_{topic}[y]$ where y is the kth label.

$m_y = m_k = \sum_{i=1}^{m} \mathbb{1}\{y_i = y\}$.

(a)

$$
\begin{aligned}
\mathbb{P}((x_i, y_i); p_{topic}, \{p_{y,init}, p_{y,tran}\}) &= p_{topic}[y_i] p_{y_i,init}(x_i[1]) \prod_{l=2}^{N} p_{y_i,tran}(x_i[l], x_i[l-1]) \\
&= p_{topic}[y_i] \prod_{t\in D} (p_{y_i,init}(t))^{S(t,x_i)} \prod_{t,t'\in D} (p_{y_i,tran}(t,t'))^{RL(t,t',x_i)} \\
\Rightarrow \log \mathbb{P}((x_i, y_i); p_{topic}, \{p_{y,init}, p_{y,tran}\}) &= \log(p_{topic}[y_i]) \\
&\quad + \sum_{t\in D} S(t, x_i)\log(p_{y_i,init}(t)) + \sum_{t,t'\in D} RL(t,t',x_i)\log(p_{y_i,tran}(t,t'))
\end{aligned}
$$

therefore

$$\log \mathbb{P}(S; p_{topic}, \{p_{y,init}, p_{y,tran}\}) = \sum_{i=1}^{m} [\log(p_{topic}[y_i]) + \sum_{t \in D} S(t, x_i) \log(p_{y_i,init}(t)) + \sum_{t,t' \in D} RL(t, t', x_i) \log(p_{y_i,tran}(t, t'))]$$

which evaluates to

$$= \sum_{j=1}^{k} m_j \log(p_{topic}[j]) + \sum_{j=1}^{k} \sum_{t \in D} ST(t, y_j) \log(p_{y_j,init}(t)) + \sum_{j=1}^{k} \sum_{t,t' \in D} RLT(t, t', y_j) \log(p_{y_j,tran}(t, t'))$$

$$= \sum_{j=1}^{k} \left[m_j \log(p_{topic}[j]) + \sum_{t \in D} ST(t, y_j) \log(p_{y_j,init}(t)) + \sum_{t,t' \in D} RLT(t, t', y_j) \log(p_{y_j,tran}(t, t')) \right]$$

We can then do MLE:

To find $\hat{p}_{topic}[j]$, subject to constraint: $\sum_{j=1}^{k} p_{topic}[j] = 1$, use Lagrange:

$$\begin{aligned} 0 &= \frac{m_j}{p_{topic}[j]} - \lambda_{topic} \\ \Rightarrow \hat{\lambda}_{topic} &= \frac{m_j}{\hat{p}_{topic}[j]} \\ \Rightarrow \hat{p}_{topic}[j] &= \frac{m_j}{m} \end{aligned}$$

Find $p_{y_j,init}(t)$, subject to constraint $\sum_{t \in D} p_{y_j,init}(t) = 1$, use Lagrange:

$$\begin{aligned} 0 &= \frac{ST(t, y_j)}{p_{y_j,init}(t)} - \lambda_{y_j,init} \\ \Rightarrow \hat{\lambda}_{y_j,init} &= \frac{ST(t, y_j)}{p_{y_j,init}(t)} \\ \Rightarrow \hat{p}_{y_j,init}(t) &= \frac{ST(t, y_j)}{\sum_{t'' \in D} ST(t'', y_j)} \end{aligned}$$

Find $p_{y_j,tran}(t, t')$, subject to constraint $\sum_{t'' \in D} p_{y_j,tran}(t'', t') = 1$, use Lagrange:

$$\begin{aligned} 0 &= \frac{RLT(t, t', y_j)}{p_{y_j,tran}(t, t')} - \lambda_{y_j,tran,t'} \\ \Rightarrow \hat{\lambda}_{y_j,tran,t'} &= \frac{RLT(t, t', y_j)}{p_{y_j,tran}(t, t')} \\ \Rightarrow \hat{p}_{y_j,tran}(t, t') &= \frac{RLT(t, t', y_j)}{\sum_{t'' \in D} RLT(t'', t', y_j)} \end{aligned}$$

(b) With prior, since $p_{topic} \sim Dir(1)$, the prior only contributes into a constant in the log probability. Meanwhile, the prior on $p_{y,init}$ contributes a constant and $ST(t, y_j)(\alpha-1) \log(p_{y_j,init}(t))$ terms for $j \in [k]$. The prior on $p_{y,tran}$ contributes a constant and $RLT(t, t', y_j)(\alpha-1) \log(p_{y_j,tran}(t, t', y_j))$.

Hence, if we perform the Lagrange analysis again:

$$\hat{p}_{topic}[j] = \frac{m_j}{m}$$
$$\hat{p}_{y_j,init}(t) = \frac{ST(t, y_j) + \alpha - 1}{\sum_{t'' \in D}(ST(t'', y_j) + \alpha - 1)}$$
$$\hat{p}_{y_j,tran}(t, t') = \frac{RLT(t, t', y_j) + \alpha - 1}{\sum_{t'' \in D}(RLT(t'', t', y_j) + \alpha - 1)}$$

(c) For $k = 2$, then $p_{topic}[1] = 1 - p_{topic}[0]$

$$\begin{aligned}
r(x) &= \log\left(\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}\right) + \log \mathbb{P}(X = x \mid Y = 1) - \log \mathbb{P}(X = x \mid Y = 0) \\
&= \log\left(\frac{p_{topic}[1]}{1 - p_{topic}[1]}\right) + \sum_{t \in D} S(t, x) \log(p_{1,init}(t)) + \sum_{t,t' \in D} RL(t, t', x) \log(p_{1,tran}(t, t')) \\
&- \sum_{t \in D} S(t, x) \log(p_{0,init}(t)) + \sum_{t,t' \in D} RL(t, t', x) \log(p_{0,tran}(t, t')) \\
&= \log\left(\frac{p_{topic}[1]}{1 - p_{topic}[1]}\right) + \sum_{t \in D} S(t, x)[\log(p_{1,init}(t) - p_{0,init}(t))] \\
&+ \sum_{t,t' \in D} RL(t, t', x)(\log(p_{1,tran}(t, t')) - \log(p_{0,tran}(t, t')))
\end{aligned}$$

Hence define the feature map

$$\phi(x)[i] = \begin{cases} S(t_i, x) & \text{for } 1 \leq i \leq |D| \\ R(t, t', x) \text{ (all combinations of } (t, t')) & \text{for } |D| + 1 \leq i \leq |D| + |D|^2 \\ 1 & \text{for } i = |D|^2 + |D| + 1 \end{cases}$$

with corresponding weight:

$$w[i] = \begin{cases} \log(p_{1,init}(t_i) - p_{0,init}(t_i)) & \text{for } 1 \leq i \leq |D| \\ \log(p_{1,tran}(t, t')) - \log(p_{0,tran}(t, t')) & \text{for } |D| + 1 \leq i \leq |D| + |D|^2 \\ \log\left(\frac{p_{topic}[1]}{1 - p_{topic}[1]}\right) & \text{for } i = |D|^2 + |D| + 1 \end{cases}$$

(e)

$$\hat{w}[i] = \begin{cases} \log\left(\frac{ST(t_i,1)+\alpha-1}{\sum_{t'' \in D}(ST(t'',1)+\alpha-1)}\right) - \log\left(\frac{ST(t_i,0)+\alpha-1}{\sum_{t'' \in D}(ST(t'',0)+\alpha-1)}\right) & \text{for } i \in [1, |D|] \\ \log\left(\frac{RLT(t,t',1)+\alpha-1}{\sum_{t'' \in D}(RLT(t'',t',1)+\alpha-1)}\right) - \log\left(\frac{RLT(t,t',0)+\alpha-1}{\sum_{t'' \in D}(RLT(t'',t',0)+\alpha-1)}\right) & \text{for } i \in [|D| + 1, |D| + |D|^2] \\ \log\left(\frac{m_1}{m_0}\right) & \text{for } i = |D| + |D|^2 + 1 \end{cases}$$