Machine Learning I

Homework III

Rafael Kiesel

Answer 1.1

Likelihood for single datapoint

$$\begin{aligned}
x &= (x_1, \dots, x_D) \\
p(x, t \mid \theta) &= [p(x \mid \mathcal{C}_1, \theta_1)p(\mathcal{C}_1)]^{1-t}[p(x \mid \mathcal{C}_2, \theta_2)p(\mathcal{C}_2)]^t \\
&= [p(\mathcal{C}_1)\prod_{d=1}^{D} p(x_d \mid \mathcal{C}_1, \theta_{d1})]^{1-t}[p(\mathcal{C}_2)\prod_{d=1}^{D} p(x_d \mid \mathcal{C}_2, \theta_{d2})]^t
\end{aligned}$$

Likelihood for multiple datapoints

$$\begin{aligned}
X &= (x^{(1)}, \dots, x^{(N)}) \\
x^{(n)} &= (x_1^{(n)}, \dots, x_D^{(n)}) \\
t &= (t_1, \dots, t_N) \\
p(X, t \mid \theta) &= \prod_{n=1}^{N} p(x^{(n)}, t_n \mid \theta) \\
&= \prod_{n=1}^{N}[p(\mathcal{C}_1)\prod_{d=1}^{D} p(x_d^{(n)} \mid \mathcal{C}_1, \theta_{d1})]^{1-t_n}[p(\mathcal{C}_2)\prod_{d=1}^{D} p(x_d^{(n)} \mid \mathcal{C}_2, \theta_{d2})]^{t_n}
\end{aligned}$$

Answer 1.2

Likelihood for single datapoint

$$\begin{aligned}
p(x, t \mid \theta) &= [p(\mathcal{C}_1)\prod_{d=1}^{D} p(x_d \mid \mathcal{C}_1, \theta_{d1})]^{1-t}[p(\mathcal{C}_2)\prod_{d=1}^{D} p(x_d \mid \mathcal{C}_2, \theta_{d2})]^t \\
&= [p(\mathcal{C}_1)\prod_{d=1}^{D} \frac{\lambda_{d1}^{x_d}}{x_d!}\exp(-\lambda_{d1})]^{1-t}[p(\mathcal{C}_2)\prod_{d=1}^{D} \frac{\lambda_{d2}^{x_d}}{x_d!}\exp(-\lambda_{d2})]^t
\end{aligned}$$

Likelihood for multiple datapoints

$$p(X, t \mid \theta) = \prod_{n=1}^{N}[p(\mathcal{C}_1)\prod_{d=1}^{D} \frac{\lambda_{d1}^{x_d^{(n)}}}{x_d^{(n)}!}\exp(-\lambda_{d1})]^{1-t_n}[p(\mathcal{C}_2)\prod_{d=1}^{D} \frac{\lambda_{d2}^{x_d^{(n)}}}{x_d^{(n)}!}\exp(-\lambda_{d2})]^{t_n}$$

Answer 1.3

Log-likelihood for single datapoint

$$
\begin{aligned}
\log(p(x, t \mid \theta)) =& \log([p(\mathcal{C}_1) \prod_{d=1}^{D} \frac{\lambda_{d1}^{x_d}}{x_d!} \exp(-\lambda_{d1})]^{1-t} [p(\mathcal{C}_2) \prod_{d=1}^{D} \frac{\lambda_{d2}^{x_d}}{x_d!} \exp(-\lambda_{d2})]^{t}) \\
=& (1-t) \log(p(\mathcal{C}_1)) + (1-t) \sum_{d=1}^{D} \log\left(\frac{\lambda_{d1}^{x_d}}{x_d!} \exp(-\lambda_{d1})\right) \\
& + t \log(p(\mathcal{C}_2)) + t \sum_{d=1}^{D} \log\left(\frac{\lambda_{d2}^{x_d}}{x_d!} \exp(-\lambda_{d2})\right) \\
=& (1-t) \left(\log(p(\mathcal{C}_1)) + \sum_{d=1}^{D} (x_d \log(\lambda_{d1}) - \log(x_d!) - \lambda_{d1})\right) \\
& + t \left(\log(p(\mathcal{C}_2)) + \sum_{d=1}^{D} (x_d \log(\lambda_{d2}) - \log(x_d!) - \lambda_{d2})\right)
\end{aligned}
$$

Log-likelihood for multiple datapoints

$$
\begin{aligned}
\log(p(X, t \mid \theta)) =& \sum_{n=1}^{N} \log(p(x^{(n)}, t_n \mid \theta)) \\
=& \sum_{n=1}^{N} (1-t_n) \left(\log(p(\mathcal{C}_1)) + \sum_{d=1}^{D} (x_d^{(n)} \log(\lambda_{d1}) - \log(x_d^{(n)}!) - \lambda_{d1})\right) \\
& + \sum_{n=1}^{N} t_n \left(\log(p(\mathcal{C}_2)) + \sum_{d=1}^{D} (x_d^{(n)} \log(\lambda_{d2}) - \log(x_d^{(n)}!) - \lambda_{d2})\right)
\end{aligned}
$$

Answer 1.4

δ_{ij} is the Kronecker delta.

$$
\begin{aligned}
\frac{\partial}{\partial \lambda_{dk}} \log(p(X, t \mid \theta)) &= \sum_{n=1}^{N} \frac{\partial}{\partial \lambda_{dk}} (1 - t_n) \left(\log(p(\mathcal{C}_1)) + \sum_{\hat{d}=1}^{D} (x_{\hat{d}}^{(n)} \log(\lambda_{\hat{d}1}) - \log(x_{\hat{d}}^{(n)}!) - \lambda_{\hat{d}1}) \right) \\
&\quad + \sum_{n=1}^{N} \frac{\partial}{\partial \lambda_{dk}} t_n \left(\log(p(\mathcal{C}_2)) + \sum_{\hat{d}=1}^{D} (x_{\hat{d}}^{(n)} \log(\lambda_{\hat{d}2}) - \log(x_{\hat{d}}^{(n)}!) - \lambda_{\hat{d}2}) \right) \\
&= \sum_{n=1}^{N} (1 - t_n) \left(\sum_{\hat{d}=1}^{D} (\frac{x_{\hat{d}}^{(n)}}{\lambda_{\hat{d}1}} - 1) \delta_{d\hat{d}} \right) \delta_{k1} \\
&\quad + \sum_{n=1}^{N} t_n \left(\sum_{\hat{d}=1}^{D} (\frac{x_{\hat{d}}^{(n)}}{\lambda_{\hat{d}2}} - 1) \delta_{d\hat{d}} \right) \delta_{k2} \\
&= \sum_{n=1}^{N} (1 - t_n) (\frac{x_d^{(n)}}{\lambda_{d1}} - 1) \delta_{k1} + \sum_{n=1}^{N} t_n (\frac{x_d^{(n)}}{\lambda_{d2}} - 1) \delta_{k2} \\
&= \left\{ \begin{array}{ll} \sum_{n=1}^{N} (1 - t_n)(\frac{x_d^{(n)}}{\lambda_{d1}} - 1), & \text{for } k = 1 \\ \sum_{n=1}^{N} t_n (\frac{x_d^{(n)}}{\lambda_{d2}} - 1), & \text{for } k = 2 \end{array} \right\} = 0 \\
\iff & \left\{ \begin{array}{ll} \sum_{n=1}^{N} (1 - t_n) x_d^{(n)}, & \text{for } k = 1 \\ \sum_{n=1}^{N} t_n x_d^{(n)}, & \text{for } k = 2 \end{array} \right\} = \left\{ \begin{array}{ll} \sum_{n=1}^{N} (1 - t_n) \lambda_{d1}, & \text{for } k = 1 \\ \sum_{n=1}^{N} t_n \lambda_{d2}, & \text{for } k = 2 \end{array} \right\} \\
\iff & \forall n : (t_n = 0 \land k = 2) \lor (t_n = 1 \land k = 1) \\
&\quad \lor \left\{ \begin{array}{ll} \frac{\sum_{n=1}^{N} (1-t_n) x_d^{(n)}}{\sum_{n=1}^{N} (1-t_n)}, & \text{for } k = 1 \\ \frac{\sum_{n=1}^{N} t_n x_d^{(n)}}{\sum_{n=1}^{N} t_n}, & \text{for } k = 2 \end{array} \right\} = \left\{ \begin{array}{ll} \lambda_{d1}, & \text{for } k = 1 \\ \lambda_{d2}, & \text{for } k = 2 \end{array} \right\} \\
\iff & \forall n : (t_n = 0 \land k = 2) \lor (t_n = 1 \land k = 1) \\
&\quad \lor \lambda_{dk} = \left\{ \begin{array}{ll} \frac{\sum_{n=1}^{N} (1-t_n) x_d^{(n)}}{\sum_{n=1}^{N} (1-t_n)}, & \text{for } k = 1 \\ \frac{\sum_{n=1}^{N} t_n x_d^{(n)}}{\sum_{n=1}^{N} t_n}, & \text{for } k = 2 \end{array} \right\}
\end{aligned}
$$

Therefore if the probability of both classes is nonzero (or rather we have samples for both classes) our MLE's are

$$
\lambda_{dk} = \left\{ \begin{array}{ll} \frac{\sum_{n=1}^{N} (1-t_n) x_d^{(n)}}{\sum_{n=1}^{N} (1-t_n)}, & \text{for } k = 1 \\ \frac{\sum_{n=1}^{N} t_n x_d^{(n)}}{\sum_{n=1}^{N} t_n}, & \text{for } k = 2 \end{array} \right\}.
$$

Answer 1.5

$$
\begin{aligned}
p(\mathcal{C}_1 \mid x) &= \frac{p(x \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(x)} \\
&= \frac{p(x \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(x \mid \mathcal{C}_1)p(\mathcal{C}_1) + p(x \mid \mathcal{C}_2)p(\mathcal{C}_2)}
\end{aligned}
$$

Answer 1.6

$$
\begin{aligned}
p(\mathcal{C}_1 \mid x) &= \frac{p(x \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(x \mid \mathcal{C}_1)p(\mathcal{C}_1) + p(x \mid \mathcal{C}_2)p(\mathcal{C}_2)} \\
&= \frac{p(\mathcal{C}_1)\prod_{d=1}^{D} \frac{\lambda_{d1}^{x_d}}{x_d!}\exp(-\lambda_{d1})}{p(\mathcal{C}_1)\prod_{d=1}^{D} \frac{\lambda_{d1}^{x_d}}{x_d!}\exp(-\lambda_{d1}) + p(\mathcal{C}_2)\prod_{d=1}^{D} \frac{\lambda_{d2}^{x_d}}{x_d!}\exp(-\lambda_{d2})}
\end{aligned}
$$

Answer 1.7

$$
\begin{aligned}
p(\mathcal{C}_1 \mid x) =& \frac{p(x \mid \mathcal{C}_1)p(\mathcal{C}_1)}{p(x \mid \mathcal{C}_1)p(\mathcal{C}_1) + p(x \mid \mathcal{C}_2)p(\mathcal{C}_2)} \\
=& \frac{1}{1 + \frac{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}} \\
=& \frac{1}{1 + \exp(- - \log(\frac{p(x|\mathcal{C}_2)p(\mathcal{C}_2)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}))} \\
\log\left(\frac{p(x \mid \mathcal{C}_2)p(\mathcal{C}_2)}{p(x \mid \mathcal{C}_1)p(\mathcal{C}_1)}\right) =& \log\left(\frac{p(\mathcal{C}_2)\prod_{d=1}^{D} \frac{\lambda_{d2}^{x_d}}{x_d!}\exp(-\lambda_{d2})}{p(\mathcal{C}_1)\prod_{d=1}^{D} \frac{\lambda_{d1}^{x_d}}{x_d!}\exp(-\lambda_{d1})}\right) \\
=& \log(p(\mathcal{C}_2)) - \log(p(\mathcal{C}_1)) \\
& + \sum_{d=1}^{D} x_d \log(\lambda_{d2}) - x_d \log(\lambda_{d1}) - \lambda_{d2} + \lambda_{d1} \\
=& - a
\end{aligned}
$$

Answer 1.8

$$
\begin{aligned}
a &= -\log(p(\mathcal{C}_2)) + \log(p(\mathcal{C}_1)) + \sum_{d=1}^{D} -x_d \log(\lambda_{d2}) + x_d \log(\lambda_{d1}) + \lambda_{d2} - \lambda_{d1} \\
&= w^T x + w_0 \\
\iff & \sum_{d=1}^{D} x_d(\log(\lambda_{d1}) - \log(\lambda_{d2})) + \log(p(\mathcal{C}_1)) - \log(p(\mathcal{C}_2)) + \sum_{d=1}^{D} \lambda_{d2} - \lambda_{d1} \\
&= w^T x + w_0 \\
\iff & w = (w_d)_{d=1,\dots,D} = (\log(\lambda_{d1}) - \log(\lambda_{d2}))_{d=1,\dots,D} \\
& \wedge w_0 = \log(p(\mathcal{C}_1)) - \log(p(\mathcal{C}_2)) + \sum_{d=1}^{D} \lambda_{d2} - \lambda_{d1}
\end{aligned}
$$

Answer 1.9

The decision boundary is given by

$$
\{x \in \mathbb{R}^D \mid \sigma(w^T x + w_0) = 1/2\} = \{x \in \mathbb{R}^D \mid w^T x + w_0 = 0\}.
$$

Therefore our decision boundary is an (affine) linear subspace of $\mathbb{R}^D$ of dimension $D-1$. This is the case because our feature functions give us the components of our vector x (or more generally because they are linear in x).

Answer 2.1

$$
\begin{aligned}
\frac{\partial y_k}{\partial w_i} =& \frac{\partial}{\partial w_i} \frac{\exp(w_k^T \phi)}{\sum_{j=1}^{K} \exp(w_j^T \phi)} \\
=& -\frac{\exp(w_i^T \phi)\exp(w_k^T \phi)}{(\sum_{j=1}^{K} \exp(w_j^T \phi))^2} \phi + \delta_{ik} \frac{\exp(w_i^T \phi)}{\sum_{j=1}^{K} \exp(w_j^T \phi)} \phi
\end{aligned}
$$

Answer 2.2

T is seen as a $K \times N$ matrix, with $T_{kn} = 1 \iff \phi_n$ has class $\mathcal{C}_k$.

$$\begin{aligned}
T &= (t_1, \dots, t_N) \\
p(T \mid \Phi, w_1, \dots, w_K) &= \prod_{n=1}^{N} p(t_n \mid \phi_n, w_1, \dots, w_K) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} [p(\mathcal{C}_k \mid \phi_n, w_1, \dots, w_K)]^{T_{kn}} \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} y_k(\phi_n)^{T_{kn}} \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\frac{\exp(w_k^T \phi_n)}{\sum_{j=1}^{K} \exp(w_j^T \phi_n)} \right)^{T_{kn}}
\end{aligned}$$

$$\begin{aligned}
\log(p(T \mid \Phi, w_1, \dots, w_K)) &= \log \left(\prod_{n=1}^{N} \prod_{k=1}^{K} \left(\frac{\exp(w_k^T \phi_n)}{\sum_{j=1}^{K} \exp(w_j^T \phi_n)} \right)^{T_{kn}} \right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \left[T_{kn} \log(\exp(w_k^T \phi_n)) - T_{kn} \log(\sum_{j=1}^{K} \exp(w_j^T \phi_n)) \right] \\
&= \sum_{n=1}^{N} \left[-\log(\sum_{j=1}^{K} \exp(w_j^T \phi_n)) + \sum_{k=1}^{K} T_{kn} w_k^T \phi_n \right]
\end{aligned}$$

Answer 2.3

$$\begin{aligned}
\frac{\partial}{\partial w_j} \log(p(T \mid \Phi w_1, \dots, w_K)) &= \frac{\partial}{\partial w_j} \sum_{n=1}^{N} \left[-\log(\sum_{i=1}^{K} \exp(w_i^T \phi_n)) + \sum_{k=1}^{K} T_{kn} w_k^T \phi_n \right] \\
&= \sum_{n=1}^{N} \left[-\frac{\exp(w_j^T \phi_n)}{\sum_{i=1}^{K} \exp(w_i^T \phi_n)} \phi_n + \sum_{k=1}^{K} T_{kn} \phi_n \delta_{kj} \right] \\
&= \sum_{n=1}^{N} -y_j(\phi_n)\phi_n + T_{jn}\phi_n
\end{aligned}$$

Answer 2.4

By switching signs we get

$$\sum_{n=1}^{N}(y_j(\phi_n) - T_{jn})\phi_n$$

which is the derivative of the cross-entropy

$$E(w_1, \dots, w_K) = -log(p(T \mid w_1, \dots, w_K)).$$

Therefore the cross-entropy has to be minimized in order to maximize the log-likelihood.

Answer 2.5

Given a learning rate $\eta > 0$ and start vectors (one for each class) $w_1^{(0)}, \dots, w_K^{(0)}$, we can compute better vectors iteratively:

$$w_k^{(i)} := w_k^{(i-1)} - \eta \frac{\partial}{\partial w_k} E(w_1^{(i-1)}, \dots, w_K^{(i-1)}) \qquad (i > 0, k = 1, \dots K)$$

This is the standard stochastic gradient descent.