Assignment 3
Machine Learning 1, Fall 2016

Dana Kianfar
University of Amsterdam

January 6, 2018

1 Naive Bayes Classification

1. We have the following class-conditional probability for a single observation x_n.

$$p\left(x_n|C_k\right) = \prod_{d=1}^{D} p\left(x_{nd}|C_k\right)$$

Therefore the likelihood for all observations, given number of classes $k = 2$ is as follows. We define $x_n d \in \mathbb{R}$ to be the dth feature of the nth row of matrix $\mathbf{X}$.

$$p(\mathbf{t}|\mathbf{X}, C_k) = \Big[\prod_{n=1}^{N}\prod_{d=1}^{D}\left[p\left(C_1\right)\cdot p\left(x_{nd}|C_1\right)\right]^{t_n}\left[p(C_2)\cdot p(x_{nd}|C_2)\right]^{(1-t_n)}$$

2.

$$\begin{aligned} p(\mathbf{t}|\mathbf{X}, C_k) &= \prod_{n=1}^{N}\prod_{d=1}^{D}\left[\pi_1\cdot\frac{\lambda_{d1}^{x_{nd}}e^{-\lambda_{d1}}}{x_{nd}!}\right]^{t_n}\left[\pi_2\cdot\frac{\lambda_{d2}^{x_{nd}}e^{-\lambda_{d2}}}{x_{nd}!}\right]^{(1-t_n)} \\ &= \prod_{d=1}^{D}\prod_{k=1}^{2}\prod_{n\in N_k}\pi_k\cdot\frac{\lambda_{dk}^{x_{nd}}e^{-\lambda_{dk}}}{x_{nd}!} \end{aligned}$$

3.

$$\log p(\mathbf{t}|\mathbf{X}, C_k) = \sum_{d=1}^{D}\sum_{k=1}^{2}\sum_{n\in N_k}\log\pi_k + x_{nd}\log\lambda_{dk} - \log(x_{nd}!) - \lambda_{dk}$$

4.

$$
\begin{aligned}
\frac{\partial \log p(\mathbf{t}|\mathbf{X}, C_k)}{\partial \lambda_{dk}} &= \sum_{d=1}^{D}\sum_{k=1}^{2}\sum_{n\in N_k} 0 + \frac{x_{nd}}{\lambda_{dk}} - 0 - 1 \\
&= \sum_{d=1}^{D}\sum_{k=1}^{2}\sum_{n\in N_k} \frac{x_{nd}}{\lambda_{dk}} - 1 \\
\frac{\partial \log p(\mathbf{t}|\mathbf{X}, C_k)}{\partial \lambda_{dk}} &= 0 \\
\sum_{d=1}^{D}\sum_{k=1}^{2}\sum_{n\in N_k} \frac{x_{nd}}{\lambda_{dk}} - 1 &= 0 \\
\sum_{d=1}^{D}\sum_{k=1}^{2}\left(\sum_{n\in N_k} \frac{x_{nd}}{\lambda_{dk}}\right) - |N_k| &= 0 \\
\text{It follows that} \\
\frac{\sum_{n\in N_k} x_{nd}}{\lambda_{dk}} &= |N_k| \\
\lambda_{dk} &= \frac{\sum_{n\in N_k} x_{nd}}{|N_k|}
\end{aligned}
$$

5.

$$
\begin{aligned}
p(C_1|\mathbf{x}) &= \frac{\prod_{d=1}^{D} p(x_d|C_1)p(C_1)}{\sum_{k=1}^{2}\prod_{d=1}^{D} p(x_d|C_k)p(C_k)} \\
&= \frac{\prod_{d=1}^{D} p(x_d|C_1)p(C_1)}{\prod_{d=1}^{D} p(x_d|C_1)p(C_1) + \prod_{d=1}^{D} p(x_d|C_2)p(C_2)} \\
&= \frac{1}{1 + \frac{\prod_{d=1}^{D} p(x_d|C_2)p(C_2)}{\prod_{d=1}^{D} p(x_d|C_1)p(C_1)}} \\
&= \frac{1}{1 + \exp(a-)}
\end{aligned}
$$

Where we define $a = -\log\left(\frac{\prod_{d=1}^{D} p(x_d|C_2)p(C_2)}{\prod_{d=1}^{D} p(x_d|C_1)p(C_1)}\right)$.

6.

$$
\begin{aligned}
p(C_1|\mathbf{x}) &= \frac{\prod_{d=1}^{D} \pi_1 \cdot \dfrac{\lambda_{d1}^{x_d} e^{-\lambda_{d1}}}{x_d!}}{\sum_{k=1}^{2}\prod_{d=1}^{D} \pi_k \cdot \dfrac{\lambda_{dk}^{x_d} e^{-\lambda_{dk}}}{x_d!}} \\
&= \frac{1}{1 + \dfrac{\prod_{d=1}^{D} \pi_2 \cdot \dfrac{\lambda_{d2}^{x_d} e^{-\lambda_{d2}}}{x_d!}}{\prod_{d=1}^{D} \pi_1 \cdot \dfrac{\lambda_{d1}^{x_d} e^{-\lambda_{d1}}}{x_d!}}} \\
&= \frac{1}{1 + \exp(-a)}
\end{aligned}
$$

Where we define $a = -\log\left(\frac{\prod_{d=1}^{D} \pi_2 \cdot \lambda_{d2}^{x_d} e^{-\lambda_{d2}}}{\prod_{d=1}^{D} \pi_1 \cdot \lambda_{d1}^{x_d} e^{-\lambda_{d1}}}\right)$.

7. See solution for question 6.

8.

$$
\begin{aligned}
a &= \mathbf{w}^T x + w_0 \\
&= -\log\left(\frac{\prod_{d=1}^{D} \pi_2 \cdot \lambda_{d2}^{x_d} e^{-\lambda_{d2}}}{\prod_{d=1}^{D} \pi_1 \cdot \lambda_{d1}^{x_d} e^{-\lambda_{d1}}}\right) \\
&= -\sum_{d=1}^{D} x_d \log(\lambda_{d2}) - \lambda_{d2} + \log \pi_2 - x_d \log(\lambda_{d1}) + \lambda_{d1} - \log \pi_1 \\
&= -\sum_{d=1}^{D} x_d \left(\log(\lambda_{d2}) - \log(\lambda_{d1})\right) - \lambda_{d2} + \log \pi_2 + \lambda_{d1} - \log \pi_1
\end{aligned}
$$

Therefore we have:

$$
\begin{aligned}
w_0 &= -\lambda_{d2} + \log \pi_2 + \lambda_{d1} - \log \pi_1 \\
\mathbf{w} &= (\log \lambda_{01} - \log \lambda_{02}, \cdots, \log \lambda_{D1} - \log \lambda_{D2})^T \in \mathbb{R}^D
\end{aligned}
$$

9. The decision boundary is linear in $\mathbf{x}$, as shown below. The decision boundary occurs where $p(C_1|x_d) = p(C_2|x_d) = 0.5$. We define $\phi = (1, x_1, x_2, \cdots, x_d)^T$ and $\widetilde{\mathbf{w}} = (w_0, w_1, \cdots, w_d)^T$.Therefore we have the following.

$$
\begin{aligned}
\frac{1}{1 + \exp(-a)} &= \frac{1}{2} \\
\exp(-a) &= 1 \\
a &= 0 \\
\widetilde{\mathbf{w}}^T \phi &= 0
\end{aligned}
$$

Which shows that the decision boundary is linear in $\mathbf{x}$.

2 Multi-class Logistic Regression

1. We define $a_k = w_k^T \phi$. Using the quotient rule we have the following.

$$
\begin{aligned}
y_k(\phi) = p(C_k|\phi) &= \frac{\exp(a_k)}{\sum_{i=1}^{K} \exp(a_i)} \\
\frac{\partial y_k}{\partial w_j} &= \frac{\phi \cdot \exp(a_k)\, \mathbb{I}_{jk} \cdot \left(\sum_{i=1}^{K} \exp(a_i)\right) - \phi \cdot \exp(a_j) \exp(a_k)}{\left(\sum_{i=1}^{K} \exp(a_i)\right)^2} \\
&= \frac{\phi \cdot \exp(a_k)\, \mathbb{I}_{jk} \cdot \left(\sum_{i=1}^{K} \exp(a_i)\right)}{\left(\sum_{i=1}^{K} \exp(a_i)\right)^2} - \frac{\phi \cdot \exp(a_j) \exp(a_k)}{\left(\sum_{i=1}^{K} \exp(a_i)\right)^2} \\
&= \frac{\phi \cdot \exp(a_k)\, \mathbb{I}_{jk}}{\sum_{i=1}^{K} \exp(a_i)} - \frac{\phi \cdot \exp(a_j)}{\sum_{i=1}^{K} \exp(a_i)} \cdot \frac{\exp(a_k)}{\sum_{i=1}^{K} \exp(a_i)} \\
&= \phi \cdot y_k \left(\mathbb{I}_{jk} - y_j\right)
\end{aligned}
$$

2.

$$
\begin{aligned}
p(\mathbf{T}|\mathbf{W}, \Phi) &= \prod_{n=1}^{N} p(\mathbf{T}_n|\phi_n, \mathbf{W}) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} p(C_k|\phi_n, \mathbf{W})^{t_n k} \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_n k} \\
\log p(\mathbf{T}|\mathbf{W}, \Phi) &= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log \left(\exp\left(a_k\right)\right) - log \left(\sum_{i=1}^{K} \exp\left(a_k\right) \right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(w_k^T \phi_n - log \left(\sum_{i=1}^{K} \exp\left(\mathbf{w}_i^T \phi_n\right) \right) \right)
\end{aligned}
$$

3.

$$
\begin{aligned}
\frac{\partial \log p(\mathbf{T}|\mathbf{W}, \Phi)}{\partial \mathbf{w}_j} &= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\mathbb{I}_{kj} \cdot \phi - \frac{\phi_n \cdot \exp(\mathbf{w}_j^T \phi_n)}{\sum_{i=1}^{K} \exp\left(\mathbf{w}_i^T \phi_n\right)} \right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \cdot \phi_n \left(\mathbb{I}_{kj} - y_{nj}\right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_n \left(t_{nj} - y_{nj}\right)
\end{aligned}
$$

4. Maximizing the log-likelihood is equivalent to minimizing the negative log-likelihood, which in the case of logistic regression is the Cross Entropy error function. We minimize this error function with respect to $\mathbf{w}_j$.

$$
\mathbb{E}\left(\mathbf{W}\right) = -\log p(\mathbf{T}|\mathbf{W}, \Phi) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk}
$$

5. (a) Define and initialize algorithm variables. $\tau = 0$ is the current step of the algorithm. τ_{max} is the maximum number of steps that the algorithm can take, after which it must terminate. η is the learning rate, which can be initialized within $(0, 1)$. Optionally define ϵ as the acceptable error rate and initialize it to an appropriately small value.

   (b) Initialize $\mathbf{W}^0$ with random values. One possible choice is to use an isotropic Gaussian prior such as $\mathcal{N}(\vec{0}, \beta\mathbb{I})$, where $\beta \in \mathbb{R}$.

   (c) While $\tau < \tau_{max}$ or $\mathbb{E} > \epsilon$, execute the following for all $\phi_n \in \Phi$ and $\mathbf{w}_j \in \mathbf{W}$:

$$
\begin{aligned}
\mathbf{w}_j^{\tau+1} &= \mathbf{w}_j^{\tau} - \eta \nabla \mathbb{E}(\mathbf{w}_j) \\
&= \mathbf{w}_j^{\tau} - \eta \nabla \left(\phi_n \left(y_{nj} - t_{nj}\right)\right) \\
&= \mathbf{w}_j^{\tau} - \eta \phi_n \left(y_{nj} - t_{nj}\right)
\end{aligned}
$$

   And then set

$$
\tau = \tau + 1
$$

(d) Terminate and return $\mathbf{W}$.