Machine Learning 1 - Homework 3

Selene Baez Santamaria

1 Naive Bayes Spam Classification

Answer the following:

1. Write down the likelihood for the general two class naive Bayes classifier.
   Solution:

$$p(\mathbf{T}, \mathbf{X}|\boldsymbol{\theta}) = \prod_{n=1}^{N}(\pi_1 \prod_{d=1}^{D} p(x_n d|\theta_1))^{1-t_n}(\pi_2 \prod_{d=1}^{D} p(x_n d|\theta_2))^{t_n}$$

2. Write down the likelihood for the Poisson model.
   Solution:
   We substitute the general probability expression for the Poisson distribution:

$$p(\mathbf{T}, \mathbf{X}|\boldsymbol{\theta}) = \prod_{n=1}^{N}(\pi_1 \prod_{d=1}^{D} \frac{\lambda_{d1}^{x_n d}}{x_n d!} \exp(-\lambda_{d1}))^{1-t_n}(\pi_2 \prod_{d=1}^{D} \frac{\lambda_{d2}^{x_n d}}{x_n d!} \exp(-\lambda_{d2}))^{t_n}$$

3. Write down the log-likelihood for the Poisson model.
   Solution:
   Taking the logarithm of the likelihood, we get:

$$\begin{aligned}\ln p(\mathbf{T},\mathbf{X}|\boldsymbol{\theta}) &= \sum_{n\in C_1}^{N}(\ln \pi_1 + \sum_{d=1}^{D} x_{nd}\ln\lambda_{d1} - \ln(x_{nd}!) - \lambda_{d1}) \\ &+ \sum_{n\in C_2}^{N}(\ln \pi_2 + \sum_{d=1}^{D} x_{nd}\ln\lambda_{d2} - \ln(x_{nd}!) - \lambda_{d2})\end{aligned}$$

4. Solve for the MLE estimators for λ_{dk}
Solution:
The MLE estimator is found by finding the partial derivative of the log-likelihood over λ_{dk} and setting it to 0. In this case it is simple because most of the terms are independent from λ_{dk}

$$\frac{\partial \ln p(\mathbf{T},\mathbf{X}|\boldsymbol{\theta})}{\partial\lambda_{dk}} = \sum_{n\in C_k}^{N}(\frac{x_{nd}}{\lambda_{dk}} - 1) = 0$$

Solving for λ_{dk}, we get:

$$\begin{aligned}\sum_{n\in C_k}^{N}\frac{x_{nd}}{\lambda_{dk}} &= \sum_{n\in C_k}^{N} 1 \\ \frac{1}{\lambda_{dk}}\sum_{n\in C_k}^{N} x_{nd} &= N_k \\ \lambda_{dk} &= \frac{1}{N_k}\sum_{n\in C_k}^{N} x_{nd}\end{aligned}$$

λ_{dk} represents the average number of word d per class k.

5. Write $p(\mathcal{C}_1|\mathbf{x})$ for the general two class naive Bayes classifier.
Solution:

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{\pi_1\prod_{d=1}^{D}p(x_{nd}\theta_1)}{\pi_1\prod_{d=1}^{D}p(x_{nd}\theta_1) + \pi_2\prod_{d=1}^{D}p(x_{nd}\theta_2)}$$

6. Write $p(\mathcal{C}_1|\mathbf{x})$ for the Poisson model.
Solution:
Again, substituting the Poisson distribution model:

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{\pi_1 \prod_{d=1}^{D} \frac{\lambda_{d1}^{x_{nd}}}{x_nd!} \exp(-\lambda_{d1})}{\pi_1 \prod_{d=1}^{D} \frac{\lambda_{d1}^{x_{nd}}}{x_nd!} \exp(-\lambda_{d1}) + \pi_2 \prod_{d=1}^{D} \frac{\lambda_{d2}^{x_{nd}}}{x_nd!} \exp(-\lambda_{d2})}$$

$$= \frac{\left(\prod_{d=1}^{D} \frac{1}{x_nd!}\right) \pi_1 \prod_{d=1}^{D} \lambda_{d1}^{x_{nd}} \exp(-\lambda_{d1})}{\left(\prod_{d=1}^{D} \frac{1}{x_nd!}\right) \pi_1 \prod_{d=1}^{D} \lambda_{d1}^{x_{nd}} \exp(-\lambda_{d1}) + \pi_2 \prod_{d=1}^{D} \lambda_{d2}^{x_{nd}} \exp(-\lambda_{d2})}$$

$$= \frac{\pi_1 \prod_{d=1}^{D} \lambda_{d1}^{x_{nd}} \exp(-\lambda_{d1})}{\pi_1 \prod_{d=1}^{D} \lambda_{d1}^{x_{nd}} \exp(-\lambda_{d1}) + \pi_2 \prod_{d=1}^{D} \lambda_{d2}^{x_{nd}} \exp(-\lambda_{d2})}$$

7. Rewrite $p(\mathcal{C}_1|\mathbf{x})$ as a sigmoid $\sigma(a) = \frac{1}{1+\exp(-a)}$; solve for a for the Poisson model.
Solution:

$$\begin{aligned} p(\mathcal{C}_1|\mathbf{x}) &= \frac{1}{1 + \exp(-\alpha)} \\ &= \frac{1}{1 + \frac{\pi_2 \prod_{d=1}^{D} \lambda_{d2}^{x_{nd}} \exp(-\lambda_{d2})}{\pi_1 \prod_{d=1}^{D} \lambda_{d1}^{x_{nd}} \exp(-\lambda_{d1})}} \end{aligned}$$

Solving for α

$$\exp(-\alpha) = \frac{\pi_2 \prod_{d=1}^{D} \lambda_{d2}^{x_{nd}} \exp(-\lambda_{d2})}{\pi_1 \prod_{d=1}^{D} \lambda_{d1}^{x_{nd}} \exp(-\lambda_{d1})}$$

$$\ln\left(\exp(-\alpha)\right) = \ln\left(\frac{\pi_2 \prod_{d=1}^{D} \lambda_{d2}^{x_{nd}} \exp(-\lambda_{d2})}{\pi_1 \prod_{d=1}^{D} \lambda_{d1}^{x_{nd}} \exp(-\lambda_{d1})}\right)$$

$$-\alpha = \ln\left(\pi_2 \prod_{d=1}^{D} \lambda_{d2}^{x_{nd}} \exp(-\lambda_{d2})\right) - \ln\left(\pi_1 \prod_{d=1}^{D} \lambda_{d1}^{x_{nd}} \exp(-\lambda_{d1})\right)$$

$$\alpha = \ln\left(\pi_1\right) + \sum_{d=1}^{D} \ln\left(\lambda_{d1}^{x_{nd}} \exp(-\lambda_{d1})\right) - \ln\left(\pi_2\right) - \sum_{d=1}^{D} \ln\left(\lambda_{d2}^{x_{nd}} \exp(-\lambda_{d2})\right)$$

$$\alpha = \ln\left(\pi_1\right) + \sum_{d=1}^{D} \ln\left(\lambda_{d1}^{x_{nd}}\right) + \sum_{d=1}^{D} -\lambda_{d1} - \ln\left(\pi_2\right) - \sum_{d=1}^{D} \ln\left(\lambda_{d2}^{x_{nd}}\right) - \sum_{d=1}^{D} -\lambda_{d2}$$

$$\alpha = \ln\frac{\pi_1}{\pi_2} + \sum_{d=1}^{D} x_{nd} \ln\frac{\lambda_{d1}}{\lambda_{d2}} - \sum_{d=1}^{D} (\lambda_{d1} - \lambda_{d2})$$

8. Assume $a = \mathbf{w}^T x + w_0$; solve for $\mathbf{w}$ and $w0$.
   Solution:
   Gathering terms dependent on x, and rearranging:

$$\alpha = \underbrace{\sum_{d=1}^{D} x_{nd} \ln\frac{\lambda_{d1}}{\lambda_{d2}}}_{w^T x_n} + \underbrace{\ln\frac{\pi_1}{\pi_2} - \sum_{d=1}^{D} (\lambda_{d1} - \lambda_{d2})}_{w_0}$$

Then:

$$w_d = \ln\frac{\lambda_{d1}}{\lambda_{d2}}$$

9. Is the decision boundary a linear function of $\mathbf{x}$? Why?
Solution:
Since w_0 is a constant for all x, then the decision boundary is linear

2 Multi-class Logistic Regression

For $K > 2$ the posterior probabilities take a generalized form of the sigmoid called the softmax:

$$y_k(\boldsymbol{\phi}) = p(\mathcal{C}_k|\boldsymbol{\phi}) = \frac{\exp(a_k)}{\sum_i \exp(a_i)}$$

where $a_k = \mathbf{w}_k^T \boldsymbol{\phi}$

Answer the following:

1. Derive $\frac{\partial y_k}{\partial \mathbf{w}_j}$. Bishop uses an indicator function $\mathbf{I}_{kj}$, entries of the identity matrix; previously we used $[k = j]$ —they are the same thing.
Solution:

$$\frac{\partial y_k(\boldsymbol{\phi})}{\partial \mathbf{w}_j} = \frac{\partial}{\partial \mathbf{w}_j}\left(\frac{\exp(a_k)}{\sum_i \exp(a_i)}\right)$$

$$= \frac{\exp(a_k)}{\sum_i \exp(a_i)}\frac{\partial a_k}{\partial \mathbf{w}_j} - \frac{\exp(a_k)\exp(a_j)}{\left(\sum_i \exp(a_i)\right)^2}\frac{\partial a_k}{\partial \mathbf{w}_j}$$

$$= [k = j]\frac{\exp(a_k)}{\sum_i \exp(a_i)}\boldsymbol{\phi} - \frac{\exp(a_k)}{\sum_i \exp(a_i)}\frac{\exp(a_j)}{\sum_i \exp(a_i)}\boldsymbol{\phi}$$

$$= \frac{\exp(a_k)}{\sum_i \exp(a_i)}\left([k = j] - \frac{\exp(a_j)}{\sum_i \exp(a_i)}\right)\boldsymbol{\phi}$$

$$= y_k(\boldsymbol{\phi})\left(\mathbf{I}_{kj} - y_j(\boldsymbol{\phi})\right)\boldsymbol{\phi}$$

2. Write down the likelihood as a product over N and K then write down the log-likelihood. Use the entries of $\mathbf{T}$ as selectors of the correct class.

Solution:
The likelihood is written as:

$$p(\mathbf{T}|\boldsymbol{\phi}, \mathbf{W}) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathcal{C}_k|\boldsymbol{\phi}_n)^{t_{nk}}$$

$$= \prod_{n=1}^{N}\prod_{k=1}^{K} (y_k(\boldsymbol{\phi}))^{t_{nk}}$$

$$= \prod_{n=1}^{N}\prod_{k=1}^{K} \left(\frac{\exp(a_k)}{\sum_i^K \exp(a_i)}\right)^{t_{nk}}$$

The log-likelihood is written as:

$$\ln p(\mathbf{T}|\boldsymbol{\phi}, \mathbf{W}) = \ln\left(\prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathcal{C}_k|\boldsymbol{\phi}_n)^{t_{nk}}\right)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} (t_{nk} \ln y_k(\boldsymbol{\phi}))$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \left(t_{nk} \ln \frac{\exp(a_k)}{\sum_i^K \exp(a_i)}\right)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \left(t_{nk} \ln \exp(a_k) - \ln \sum_{i}^{K} \exp(a_i)\right)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \left(t_{nk} \left(a_k - \ln \sum_{i}^{K} \exp(a_i)\right)\right)$$

3. Derive the gradient of the log-likelihood with respect to $\mathbf{w}_j$.
   Solution:

$$\frac{\partial \ln p(\mathbf{T}|\boldsymbol{\phi}, \mathbf{W})}{\partial \mathbf{w}_j} = \frac{\partial \left(\sum_{n=1}^{N} \sum_{k=1}^{K} \left(t_{nk} \ln y_k(\boldsymbol{\phi}_n) \right) \right)}{\partial \mathbf{w}_j}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \left(t_{nk} \frac{1}{y_k(\boldsymbol{\phi}_n)} \frac{\partial y_k(\boldsymbol{\phi}_n)}{\partial \mathbf{w}_j} \right)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\frac{t_{nk}}{y_k(\boldsymbol{\phi}_n)} y_k(\boldsymbol{\phi}_n) \left(\mathbf{I}_{kj} - y_j(\boldsymbol{\phi}_n) \right) \boldsymbol{\phi}_n \right)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \left(t_{nk} \left(\mathbf{I}_{kj} - y_j(\boldsymbol{\phi}_n) \right) \boldsymbol{\phi}_n \right)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \left(t_{nk} \mathbf{I}_{kj} \boldsymbol{\phi}_n \right) - \sum_{n=1}^{N} \sum_{k=1}^{K} \left(t_{nk} y_j(\boldsymbol{\phi}_n) \boldsymbol{\phi}_n \right)$$

$$= \sum_{n=1}^{N} \boldsymbol{\phi}_n \sum_{k=1}^{K} \left(t_{nk} \mathbf{I}_{kj} \right) - \sum_{n=1}^{N} y_j(\boldsymbol{\phi}_n) \boldsymbol{\phi}_n \sum_{k=1}^{K} t_{nk}$$

$$= \sum_{n=1}^{N} \boldsymbol{\phi}_n t_{nj} - \sum_{n=1}^{N} y_j(\boldsymbol{\phi}_n) \boldsymbol{\phi}_n$$

$$= \sum_{n=1}^{N} \left(\left(t_{nj} - y_j(\boldsymbol{\phi}_n) \right) \boldsymbol{\phi}_n \right)$$

4. What is the objective function we minimize that is equivalent to maximizing the log-likelihood?
   Solution:
   The objective function is the *cross-entropy error* ($E(\mathbf{W})$) and is equal

to the negative log-likelihood. I.e.:

$$E(\mathbf{W}) = -\ln p(\mathbf{T}|\mathbf{\Phi}, \mathbf{W}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_k(\phi_n)$$

Sometimes we may write y_{nk} for $y_k(\phi_n)$; it is useful sometimes to give clutter-free solutions. Minimizing the cross-entropy requires the same gradients as maximizing the log-likelihood, except there is a change in sign:

$$\begin{aligned}\frac{\partial E(\mathbf{W})}{\partial \mathbf{w}_j} &= \sum_{n=1}^{N}\sum_{k=1}^{K}(y_{nj} - t_{nj})\phi_n \\ &= \sum_{n=1}^{N} e_n\end{aligned}$$

5. Write a stochastic gradient algorithm for logistic regression using this objective function. Make sure to include indices for time and to define the learning rate. The gradients may differ in sign switching from maximizing to minimizing; don't overlook this.
Solution:
The single step update for SGD is:

$$w_j^{t+1} = w_j^t - \eta^t \nabla e_n$$

```
(a) Initialize $\mathbf{W}$
(b) Initialize $\eta$
(c) For $t = 1$ to $T$ do:
      (i) Randomly choose $n$ from [1,$N$]
     (ii) $\mathbf{w}_j = \mathbf{w}_j - \eta \nabla e_n (forallj)$
    (iii) Decrease $\eta$
(d) Return $\mathbf{W}$
```