Homework 3

JD Gallego Posada, *University of Amsterdam* 28/09/2016

The notations $\mathbf{A}_{\cdot j}$ and $\mathbf{A}_{i\cdot}$ represent the j-th column and i-th row of the matrix $\mathbf{A}$, resp.

1 Naive Bayes Spam Classification

1. Write down the likelihood for the general two class naive Bayes classification.

Define S_k as the set of indexes n for which t_n belongs to class k.

$$\begin{aligned} p(\{(\mathbf{x}_n, t_n)\}_{n=1}^N|\mathbf{\Theta}) &= \prod_{n=1}^N p(\mathbf{x}_n|t_n, \mathbf{\Theta})\, p(t_n) = \prod_{n=1}^N \prod_{d=1}^D p(x_{nd}|t_n, \mathbf{\Theta})\, p(t_n) \\ &= \left(\prod_{n\in S_1} \pi_1 \prod_{d=1}^D p(x_{nd}|\mathscr{C}_1, \theta_{d1})\right)\left(\prod_{n\in S_2} \pi_2 \prod_{d=1}^D p(x_{nd}|\mathscr{C}_2, \theta_{d2})\right) \end{aligned}$$

2. Write down the likelihood for the Poisson model.

$$p(\{(\mathbf{x}_n, t_n)\}_{n=1}^N|\mathbf{\Lambda}) = \left(\prod_{n\in S_1} \pi_1 \prod_{d=1}^D \frac{{\lambda_{d1}}^{x_{nd}}}{x_{nd}!} e^{-\lambda_{d1}}\right)\left(\prod_{n\in S_2} \pi_2 \prod_{d=1}^D \frac{{\lambda_{d2}}^{x_{nd}}}{x_{nd}!} e^{-\lambda_{d2}}\right)$$

3. Write down the log-likelihood for Poisson model.

$$\begin{aligned} \log p(\{(\mathbf{x}_n, t_n)\}_{n=1}^N|\mathbf{\Lambda}) &= |S_1|\log\pi_1 + \sum_{n\in S_1}\sum_{d=1}^D (x_{nd}\log\lambda_{d1} - \lambda_{d1} - \log x_{nd}!) \\ &\quad + |S_2|\log\pi_2 + \sum_{n\in S_2}\sum_{d=1}^D (x_{nd}\log\lambda_{d2} - \lambda_{d2} - \log x_{nd}!) \end{aligned}$$

4. Solve for the MLE estimators for λ_{dk}.

$$\frac{\partial \log p(\{(\mathbf{x}_n, t_n)\}_{n=1}^N|\mathbf{\Lambda})}{\partial \lambda_{dk}} = \sum_{n\in S_k}\left(\frac{x_{nd}}{\lambda_{dk}} - 1\right) = \frac{1}{\lambda_{dk}}\left(\sum_{n\in S_k} x_{nd}\right) - |S_k| = 0$$

$$\lambda_{dk\,MLE} = \frac{\sum_{n\in S_k} x_{nd}}{|S_k|}$$

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

$$\begin{aligned} p(\mathscr{C}_1|\mathbf{x}, \mathbf{\Theta}) &= \frac{p(\mathbf{x}|\mathscr{C}_1, \mathbf{\Theta})p(\mathscr{C}_1)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathscr{C}_1, \mathbf{\Theta}_{\cdot 1})p(\mathscr{C}_1)}{p(\mathbf{x}|\mathscr{C}_1, \mathbf{\Theta}_{\cdot 1})p(\mathscr{C}_1) + p(\mathbf{x}|\mathscr{C}_2, \mathbf{\Theta}_{\cdot 2})p(\mathscr{C}_2)} \\ &= \frac{p(\mathscr{C}_1)\prod_{d=1}^D p(x_d|\mathscr{C}_1, \theta_{d1})}{p(\mathscr{C}_1)\prod_{d=1}^D p(x_d|\mathscr{C}_1, \theta_{d1}) + p(\mathscr{C}_2)\prod_{d=1}^D p(x_d|\mathscr{C}_2, \theta_{d2})} \end{aligned}$$

6. Write $p(\mathscr{C}_1|\mathbf{x})$ for the Poisson model.

$$p(\mathscr{C}_1|\mathbf{x}, \mathbf{\Lambda}) = \frac{\pi_1\prod_{d=1}^D \frac{{\lambda_{d1}}^{x_d}}{x_d!} e^{-\lambda_{d1}}}{\pi_1\prod_{d=1}^D \frac{{\lambda_{d1}}^{x_d}}{x_d!} e^{-\lambda_{d1}} + \pi_2\prod_{d=1}^D \frac{{\lambda_{d2}}^{x_d}}{x_d!} e^{-\lambda_{d2}}}$$

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

$$p(\mathscr{C}_1|\mathbf{x}, \mathbf{\Lambda}) = \frac{1}{1 + \frac{\pi_2 \prod_{d=1}^{D} \frac{\lambda_{d2}{}^{x_d}}{x_d!} e^{-\lambda_{d2}}}{\pi_1 \prod_{d=1}^{D} \frac{\lambda_{d1}{}^{x_d}}{x_d!} e^{-\lambda_{d1}}}} = \frac{1}{1+\exp(-a)}$$

where $a = \log \frac{\pi_1 \prod_{d=1}^{D} \frac{\lambda_{d1}{}^{x_d}}{x_d!} e^{-\lambda_{d1}}}{\pi_2 \prod_{d=1}^{D} \frac{\lambda_{d2}{}^{x_d}}{x_d!} e^{-\lambda_{d2}}}$.

8. Assume $a = \mathbf{w}^T\mathbf{x} + w_0$; solve for $\mathbf{w}$ and w_0.

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

where $w_0 = \log \frac{\pi_1}{\pi_2} + \left(\sum_{d=1}^{D} -\lambda_{d1} + \lambda_{d2} \right)$ and $\mathbf{w} = \begin{bmatrix} \log \frac{\lambda_{11}}{\lambda_{12}} & \cdots & \log \frac{\lambda_{D1}}{\lambda_{D2}} \end{bmatrix}^T$.

9. Is the decision boundary a linear function of $\mathbf{x}$? Why?

Yes. Note that the sigmoid function is increasing monotonous. The region boundary is the set $\{\mathbf{x} : \sigma(\mathbf{w}^T\mathbf{x} + w_0) = 0.5\}$, which is equivalent to the set $\{\mathbf{x} : \mathbf{w}^T\mathbf{x} + w_0 = 0\}$, which is a decision boundary linear on $\mathbf{x}$.

2 Multi-class Logistic Regression

Denote by $\mathbf{W}$ the MxK matrix which contains each $\mathbf{w}_j$ vector in its columns and by $\mathbf{\Phi}$ the NxM matrix which contains the features for each example $\boldsymbol{\phi}^T$ in its rows.

1. Derive $\frac{\partial y_k}{\partial \mathbf{w}_j}$.

$$\frac{\partial y_k}{\partial a_k} = \frac{e^{a_k} \sum_i e^{a_i} - e^{a_k} e^{a_k}}{(\sum_i e^{a_i})^2} = \frac{e^{a_k}}{\sum_i e^{a_i}} - \left(\frac{e^{a_k}}{\sum_i e^{a_i}} \right)^2 = y_k - y_k^2 = y_k(1 - y_k)$$

$$\frac{\partial y_k}{\partial a_j} = \frac{-e^{a_k} e^{a_j}}{(\sum_i e^{a_i})^2} = -\frac{e^{a_k}}{\sum_i e^{a_i}} \frac{e^{a_j}}{\sum_i e^{a_i}} = -y_k y_j = y_k(0 - y_j) \quad \text{for } k \neq j$$

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$$

$$\frac{\partial y_k}{\partial \mathbf{w}_j} = \frac{\partial y_k}{\partial a_j} \frac{\partial a_j}{\partial \mathbf{w}_j} = y_k(I_{kj} - y_j)\boldsymbol{\phi} \quad \rightarrow \quad \frac{\partial y_{nk}}{\partial \mathbf{w}_j} = y_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi}_n$$

where $y_{nk} = y_k(\boldsymbol{\phi}_n)$.

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.

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

$$\log p(\mathbf{T}|\mathbf{W}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk}$$

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

$$
\begin{aligned}
\frac{\partial \log p(\mathbf{T}|\mathbf{W})}{\partial \mathbf{w}_j} &= \frac{\partial \log p(\mathbf{T}|\mathbf{W})}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial \mathbf{w}_j} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk}(I_{kj} - y_{nj}) \boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N} (t_{nj} - y_{nj}) \boldsymbol{\phi}_n \\
&= \boldsymbol{\Phi}^T (\mathbf{T}_{\cdot j} - \mathbf{Y}_{\cdot j})
\end{aligned}
$$

4. Which is the objective function we minimize that is equivalent to maximizing the log-likelihood?

Since $\max f \equiv \min -f$, it is equivalent to minimize the negative of the log-likelihood, which in this case coincides with the cross entropy error, denoted by E.

5. Write down a stochastic gradient algorithm for logistic regression using this objective function.

Note that using the comment from point 4, the expression in point 3 can be generalized to:

$$
\frac{\partial E}{\partial \mathbf{W}} = \frac{\partial - \log p(\mathbf{T}|\mathbf{W})}{\partial \mathbf{W}} = \boldsymbol{\Phi}^T (\mathbf{Y} - \mathbf{T})
$$

After observing a particular example, say n,

$$
\frac{\partial E}{\partial \mathbf{W}} = \boldsymbol{\Phi}_{n\cdot}^T (\mathbf{Y}_{n\cdot} - \mathbf{T}_{n\cdot}) = \boldsymbol{\phi}_n (\mathbf{Y}_{n\cdot} - \mathbf{T}_{n\cdot})
$$

Algorithm 1: Stochastic Gradient Descent for Multi-class Logistic Regression

Data: Φ matrix of features and $\mathbf{T}$ matrix of targets

Result: Approximate minimizer $\mathbf{W}^*$ of E

Carefully choose initialization $\mathbf{W}^{(0)}$;

Carefully choose learning rate $\eta > 0$;

while *Not convergence* **do**

  Randomly choose observation n;

  $\mathbf{W}^{(\tau+1)} := \mathbf{W}^{(\tau)} - \eta \boldsymbol{\Phi}_{n\cdot}^T (\mathbf{Y}_{n\cdot}{}^{(\tau)} - \mathbf{T}_{n\cdot})$

end

return $\mathbf{W}^*$

The loop will make the algorithm converge (if η is not too big) since E is a convex function (combination of convex functions).