Machine Learning 1 Homework 3 Solutions

1 Naive Bayes Spam Classification

Naive Bayes is a particular form of classification that makes strong independence assumptions regarding the features of the data, conditional on the classes (see Bishop section 4.2.3). Specifically, NB assumes each feature is independent given the class label. In contrast, when we looked at probabilistic generative models for classification in the lecture, we used a full-covariance Gaussian to model data from each class, which incorporates correlation between all the input features (i.e. they are not conditionally independent).

If correlated features are treated independently, the evidence for a class gets overcounted. However, Naive Bayes is very simple to construct because, by ignoring correlations, the class-conditional likelihood is a product of D univariate distributions, each of which is simple to learn:

$$p(\mathbf{x}|\mathcal{C}_k) = \prod_{d=1}^{D} p(x_d|\mathcal{C}_k)$$

Consider a spam filter that classifies your emails into two classes $\mathcal{C}_1$ (spam) and $\mathcal{C}_2$ (non-spam). To do this you first make a *bag-of-words* (BoW) representation of your entire training set (a bunch of spam emails and non-spam emails). A BoW is a vector of dimension D of word counts, one for each document (i.e. the words go into a bag and are shaken, losing their order so only their count matters). You can think of D as the vocabulary size of the training set, but it may also be tokens or special features you think are important for spam detection. Your training set is therefore an N by D matrix of word counts $\mathbf{X}$, and $\mathbf{t}$, where $t_n = 0$ if $n \in \mathcal{C}_1$ and $t_n = 1$ if $n \in \mathcal{C}_2$. Assume we know $p(\mathcal{C}_1) = \pi_1$ (and $\pi_2 = 1 - \pi_1$). We can model the word counts using different distributions, for this question we will use a Poisson distribution model:

$$\begin{aligned} p(x_d|\mathcal{C}_k, \theta_{dk}) &= \mathcal{P}\left(x_d|\lambda_{dk}\right) \\ &= \frac{\lambda_{dk}^{x_d}}{x_d!}\exp(-\lambda_{dk}) \end{aligned}$$

i.e. the parameters $\theta_{dk} = \lambda_{dk}$.

With this information answer the following questions:

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

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

2. Write down the likelihood for the Poisson model.

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

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

$$\begin{aligned} \ln p(\mathbf{T}, \mathbf{X}|\boldsymbol{\theta}) = &\sum_{n \in \mathcal{C}_1}^{N} \left(\ln \pi_1 + \sum_{d=1}^{D} x_{nd} \ln \lambda_{d1} - \ln(x_{nd}!) - \lambda_{d1} \right) \\ &+ \sum_{n \in \mathcal{C}_2}^{N} \left(\ln \pi_2 + \sum_{d=1}^{D} x_{nd} \ln \lambda_{d2} - \ln(x_{nd}!) - \lambda_{d2} \right) \end{aligned}$$

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

$$\begin{aligned} \frac{\ln p(\mathbf{T}, \mathbf{X}|\boldsymbol{\theta})}{\partial \lambda_{dk}} &= \sum_{n \in \mathcal{C}_k}^{N} \left(\frac{x_{nd}}{\lambda_{dk}} - 1 \right) = 0 \\ \sum_{n \in \mathcal{C}_k}^{N} \frac{x_{nd}}{\lambda_{dk}} &= \sum_{n \in \mathcal{C}_k}^{N} 1 \\ \lambda_{dk} &= \frac{1}{N_k} \sum_{n \in \mathcal{C}_k}^{N} x_{nd} \end{aligned}$$

I.e. λ_{dk} is the average number of word/token d per email for class k.

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

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

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

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

Notice the product over count factorials cancels because it is a constant for each class.

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.

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

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

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

See above. $w_d = \ln \frac{\lambda_{d1}}{\lambda_{d2}}$

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

Yes: a is a linear function of $\mathbf{x}$ (w_0 is a constant for all $\mathbf{x}$).

2 Multi-class Logistic Regression

In class we saw the binary classification version of logistic regression. Here you will derive the gradients for the general case $K > 2$. Much of the preliminaries are in Bishop 4.3.4. This will be useful for Lab 2.

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 $a_k = \mathbf{w}_k^T \boldsymbol{\phi}$. NB: The posterior for class k depends on all the other classes i; keep this in mind when working out the derivatives for $\mathbf{w}_k$. The training set is a pair of matrices $\boldsymbol{\Phi}$ and $\mathbf{T}$. Each row of $\mathbf{T}$ uses a one-hot encoding of the class labeling for that training example.

Answer the following questions:

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

$$\begin{aligned}\frac{\partial y_k(\boldsymbol{\phi})}{\partial \mathbf{w}_j} &= \frac{\exp(a_k)}{\sum_i \exp(a_i)} \frac{\partial a_k}{\partial \mathbf{w}_j} - \frac{\exp(a_k)\exp(a_j)}{\left(\sum_i \exp(a_i)\right)^2} \frac{\partial a_j}{\partial \mathbf{w}_j} \\ &= y_k(\boldsymbol{\phi})\mathbf{I}_{kj}\boldsymbol{\phi} - y_k(\boldsymbol{\phi})y_j(\boldsymbol{\phi})\boldsymbol{\phi} \\ &= y_k(\boldsymbol{\phi})(\mathbf{I}_{kj} - y_j(\boldsymbol{\phi}))\boldsymbol{\phi}\end{aligned}$$

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.

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

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

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

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{T}|\mathbf{\Phi}, \mathbf{W})}{\partial \mathbf{w}_j} &= \sum_{n=1}^{N}\sum_{k=1}^{K} \frac{t_{nk}}{y_k(\boldsymbol{\phi}_n)} \frac{\partial y_k}{\partial \mathbf{w}_j} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \frac{t_{nk}}{y_k(\boldsymbol{\phi}_n)} y_k(\boldsymbol{\phi}_n)(\mathbf{I}_{kj} - y_j(\boldsymbol{\phi}_n))\boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(\mathbf{I}_{kj} - y_j(\boldsymbol{\phi}_n))\boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\mathbf{I}_{kj}\boldsymbol{\phi}_n - \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} y_j(\boldsymbol{\phi}_n)\boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N} \boldsymbol{\phi}_n \underbrace{\sum_{k=1}^{K} t_{nk}\mathbf{I}_{kj}}_{=t_{nj}} - \sum_{n=1}^{N} y_j(\boldsymbol{\phi}_n)\boldsymbol{\phi}_n \underbrace{\sum_{k=1}^{K} t_{nk}}_{=1} \\
&= \sum_{n=1}^{N} (t_{nj} - y_j(\boldsymbol{\phi}_n))\boldsymbol{\phi}_n
\end{aligned}
$$

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

   The objective function is the *cross-entropy error* ($E(\mathbf{W})$) and is equal to the negative log-likelihood. I.e.:

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

   Sometimes we may write y_{nk} for $y_k(\boldsymbol{\phi}_n)$; 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:

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

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.

   The single step update for SGD is:

$$\mathbf{w}_j^{t+1} = \mathbf{w}_j^t - \eta^t \nabla e_n$$

   (a) Initialize $\mathbf{W}$

   (b) Initialize η

   (c) For $t = 1$ to T do:

      i. Randomly choose n from $[1, N]$

      ii. $\mathbf{w}_j = \mathbf{w}_j - \eta \nabla e_n$ (for all j)

      iii. Decrease η

   (d) Return $\mathbf{W}$