CS 5350/6350: Machine Learining Spring 2015

Homework 6 solution

Handed out: Apr 13, 2015
Due date: Apr 27, 2015

1 Naive Bayes Classification

In the class, we saw how we can build a naive Bayes classifier for discrete variables. In this question, you will explore the case when features are not discrete. Suppose instances, represented by $\mathbf{x}$, are d dimensional real vectors and the labels, represented by y, are either 0 or 1.

Recall from class that the naive Bayes classifier assumes that all features are conditionally independent of each other, given the class label. That is

$$p(\mathbf{x}|y) = \prod_j p(x_j|y)$$

Now, each x_j is a real valued feature. Suppose we assume that these drawn from a class specific normal distribution. That is,

1. When $y = 0$, each x_j is drawn from a normal distribution with mean $\mu_{0,j}$ and standard deviation $\sigma_{0,j}$, and
2. When $y = 1$, each x_j is drawn from a normal distribution with mean $\mu_{1,j}$ and standard deviation $\sigma_{1,j}$

Now, suppose we have a training set $S = \{(\mathbf{x}_i, y_i)\}$ with m examples and we wish to learn the parameters of the classifier, namely the prior $p = P(y = 1)$ and the μ's and the σ's. For brevity, let the symbol θ denote all these parameters together.

(a) Write down $P(S|\theta)$, the likelihood of the data in terms of the parameters. Write down the log-likelihood.

Solution:

$$\begin{aligned} P(S|\theta) &= \prod_{i=1}^{m} P(\{(\mathbf{x}_i, y_i)\}|\theta) \\ &= \prod_{i=1}^{m} \left[P(y_i|\theta) \prod_{j=1}^{d} P(x_{i,j}|y_i, \theta) \right] \end{aligned}$$

Since all features are conditionally independent and

$$x_j \sim \begin{cases} \mathcal{N}(\mu_{0,j}, \sigma_{0,j}^2) & \text{if} \quad y = 0 \\ \mathcal{N}(\mu_{1,j}, \sigma_{1,j}^2) & \text{if} \quad y = 1 \end{cases}$$

We have

$$P(x_{i,j}|y_i, \theta) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_{0,j}} \exp\left(-\frac{(x_{i,j}-\mu_{0,j})^2}{2\sigma_{0,j}^2}\right) & \text{if} \quad y = 0 \\ \\ \frac{1}{\sqrt{2\pi}\sigma_{1,j}} \exp\left(-\frac{(x_{i,j}-\mu_{1,j})^2}{2\sigma_{1,j}^2}\right) & \text{if} \quad y = 1 \end{cases}$$

and

$$P(y_i|\theta) = \begin{cases} 1-p & \text{if} \quad y = 0 \\ p & \text{if} \quad y = 1 \end{cases}$$

Taking log and substituting $P(x_{i,j}|y_i, \theta)$, $P(y_i|\theta)$, we have

$$\begin{aligned} \log P(S|\theta) =& \sum_{i=1}^{m} \log P(y_i|\theta) + \sum_{i=1}^{m}\sum_{j=1}^{d} \log P(x_{i,j}|y_i, \theta) \\ =& Count(y_i = 0)\log(1-p) + Count(y_i = 1)\log p + \\ & Count(y_i = 0)\sum_{j=1}^{d}\left(\log\frac{1}{\sqrt{2\pi}\sigma_{0,j}} - \frac{(x_{i,j}-\mu_{0,j})^2}{2\sigma_{0,j}^2}\right) + \\ & Count(y_i = 1)\sum_{j=1}^{d}\left(\log\frac{1}{\sqrt{2\pi}\sigma_{1,j}} - \frac{(x_{i,j}-\mu_{1,j})^2}{2\sigma_{1,j}^2}\right) \end{aligned}$$

(b) What is the prior probability p? You can derive this by taking the derivative of the log-likelihood with respect to the prior and setting it to zero.

Solution:

First, we take partial derivative of log-likelihood with respect to p

$$\frac{\partial \log P(S|\theta)}{\partial p} = \frac{Count(y_i = 1)}{p} - \frac{Count(y_i = 0)}{1-p}$$

Setting the derivative to zero, we can get

$$p = \frac{Count(y_i = 1)}{Count(y_i = 0) + Count(y_i = 1)}$$

(c) By taking the derivative of the log-likelihood with respect to the μ_j's and the σ_j's,

derive expressions for the μ_j's and the σ_j's.

Solution:

The idea of solving μ and σ is to take the partial derivative and set it to zero.

$$\begin{aligned}\frac{\partial \log P(S|\theta)}{\partial \mu_{1,j}} &= -Count(y=1)\frac{\partial}{\partial \mu_{1,j}}\sum_{j=1}^{d}\frac{(x_{i,j}-\mu_{1,j})^2}{2\sigma_{1,j}^2} \\ &= Count(y=1)\frac{x_{i,j}-\mu_{1,j}}{\sigma_{1,j}^2}\end{aligned}$$

Therefore

$$\mu_{1,j} = \frac{\sum_i x_{i,j}[y_i=1]}{Count(y_i=1)}$$

Similarly, we can get

$$\mu_{0,j} = \frac{\sum_i x_{i,j}[y_i=0]}{Count(y_i=0)}$$

$$\sigma_{1,j} = \sqrt{\frac{\sum_i (x_{i,j}-\mu_{i,j})^2[y_i=1]}{Count(y_i=1)}}$$

$$\sigma_{0,j} = \sqrt{\frac{\sum_i (x_{i,j}-\mu_{i,j})^2[y_i=0]}{Count(y_i=0)}}$$

2 Logistic Regression

We looked maximum a posteriori learning of the logistic regression classifier in class. In particular, we showed that learning the classifier is equivalent to the following optimization problem:

$$\min_{\mathbf{w}} \sum_{i=1}^{m} \log\left(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i)\right) + \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{w}$$

In this question, you will derive the stochastic gradient descent algorithm for the logistic regression classifier.

(a) What is the derivative of the function $\log\left(1+\exp(-y\mathbf{w}_i^T\mathbf{x}_i)\right)$ with respect to the weight vector?

Solution:

For each training example $(\mathbf{x}_i, y_i)$ we have

$$\begin{aligned}\frac{\partial}{\partial w_j}\log\left(1+\exp(-y_i\mathbf{w}_i^T\mathbf{x}_i)\right) =& \frac{-y_i x_{i,j}\exp(-y_i\mathbf{w}_i^T\mathbf{x}_i)}{1+\exp(-y_i\mathbf{w}_i^T\mathbf{x}_i)} \\ =& -\frac{y_i x_{i,j}}{1+\exp(y_i\mathbf{w}_i^T\mathbf{x}_i)}\end{aligned}$$

Therefore,

$$\frac{d}{d\mathbf{w}}\log\left(1+\exp(-y\mathbf{w}_i^T\mathbf{x}_i)\right) = -\frac{y_i\mathbf{x}_i}{1+\exp(y_i\mathbf{w}_i^T\mathbf{x}_i)}$$

(b) The inner most step in the SGD algorithm is the gradient update where we use a single example instead of the entire dataset to compute the gradient. Write down the objective where the entire dataset is composed of a single example, say $(\mathbf{x}_i, y_i)$. Derive the gradient of this objective with respect to the weight vector.

Solution:

If the entire dataset only has a single example $(\mathbf{x}_i, y_i)$, then the objective function will be

$$f(\mathbf{w}) = \min_{\mathbf{w}}\log\left(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i)\right) + \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{w}$$

And the gradient of the objective function is

$$\bigtriangledown f = -\frac{y_i\mathbf{x}_i}{1+\exp(y_i\mathbf{w}^T\mathbf{x}_i)} + \frac{2\mathbf{w}}{\sigma^2}$$

(c) Write down the pseudo code for the stochastic gradient algorithm using the gradient from part (b) above.

Solution:

The pseudo code is showing as follows.

Initialize $\mathbf{w} = \mathbf{0}$
for epoch = 1 to T **do**
    Shuffle the dataset
    for each traning example $(\mathbf{x}_i, y_i)$ **do**
        update $\mathbf{w}$: $\mathbf{w} = \mathbf{w} - r\left(\frac{2\mathbf{w}}{\sigma^2} - \frac{y_i\mathbf{x}_i}{1+\exp(y_i\mathbf{w}^T bx_i)}\right)$
    end for
end for
return w

---

3 The EM algorithm

The two local newspapers The Times and The Gazette publish n articles everyday. The article length in the newspapers is distributed based on the Exponential Distribution with parameter λ. That is, for an non-negative integer x:

$$P(wordcount = x|\lambda) = \lambda e^{-\lambda x}.$$

with parameters λ_T, λ_G for the Times and the Gazette, respectively.

(**Note**: Technically, using the exponential distribution is not correct here because the exponential distribution applies to real valued random variables, whereas here, the word counts can only be integers. However, for simplicity, we will use the exponential distribution instead of, say a Poisson.)

(a) Given an issue of one of the newspapers $(x_1, \ldots x_n)$, where x_i denotes the length of the ith article, what is the most likely value of λ?

Solution:

The goal is to maximize $P(\lambda|x_1, \ldots, x_n)$. If we use maximum-likelihood estimation and assume prior distribution of λ is uniform, then we can define the most likely value of λ as

$$\begin{aligned}
\hat{\lambda} &= \operatorname*{argmax}_{\lambda} P(\lambda|x_1, \ldots, x_n) \\
&= \operatorname*{argmax}_{\lambda} P(x_1, \ldots, x_n|\lambda) \\
&= \operatorname*{argmax}_{\lambda} \prod_{i=1}^{n} P(x_i|\lambda) \\
&= \operatorname*{argmax}_{\lambda} \prod_{i=1}^{n} \lambda e^{-\lambda x_i}
\end{aligned}$$

Taking log and set derivative with respect λ to zero, we have

$$\frac{d}{d\lambda} \log P(x_1, \ldots, x_n|\lambda) = \frac{d}{d\lambda} \sum_{i=1}^{n} \log(\lambda e^{-\lambda x_i}) = 0$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

(b) Assume now that you are given a collection of m issues $\{(x_1, \ldots x_n)\}_1^m$ but you do not know which issue is a Times and which is a Gazette issue. Assume that the probability

of a document is generated from the Times is η. In other words, it means that the probability that a document is generated from the Gazette is $1 - \eta$.

Explain the generative model that governs the generation of this data collection. In doing so, name the parameters that are required in order to fully specify the model.

Solution:

First, η, λ_T and λ_G are three required parameters for specifying the model. The generative model can be described using the following probabilities:

$$\begin{aligned} P(\mathbf{x}, y = Times) &= P(y = Times)P(\mathbf{x}|y = Times) \\ &= P(y = Times)\prod_{j=1}^{n} P(x_j|y = Times) \\ &= \eta \prod_{j=1}^{n} \lambda_T \exp(-\lambda_T x_j) \\ &= \eta \lambda_T^n \exp(-\lambda_T \sum_{j=1}^{n} x_j) \end{aligned}$$

Similarly, we have

$$P(\mathbf{x}, y = Gazette) = (1 - \eta)\lambda_G^n \exp(-\lambda_G \sum_{j=1}^{n} x_j)$$

(c) Assume that you are given the parameters of the model described above. How would you use it to cluster issues to two groups, the Times issues and the Gazette issues?

Solution:

Clearly, the label should be

$$y = \underset{y}{\operatorname{argmax}}\, P(y|\mathbf{x}) = \underset{y}{\operatorname{argmax}}\, P(\mathbf{x}, y)$$

Since we already know $P(\mathbf{x}, y = Times)$ and $P(\mathbf{x}, y = Gazette)$, we can just assign y to a label that results a greater probability.

(d) Given the collection of m issues without labels of which newspaper they came from, derive the update rule of the EM algorithm. Show all of your work.

Solution:

(i) E-step, we need to estimate $P(y|x_i, \theta^t)$.

$$
\begin{aligned}
P(y = Times|\mathbf{x}_i, \theta^t) &= \frac{P(\mathbf{x}_i, y = Times|\theta^t)}{P(\mathbf{x}_i, y = Times|\theta^t) + P(\mathbf{x}_i, y = Gazette|\theta^t)} \\
&= \frac{\eta^t(\lambda_T^t)^n \exp(-\lambda_T^t \sum_{j=1}^n x_j)}{\eta^t(\lambda_T^t)^n \exp(-\lambda_T^t \sum_{j=1}^n x_j) + (1-\eta^t)(\lambda_G^t)^n \exp(-\lambda_G^t \sum_{j=1}^n x_j)} \\
P(y = Gazette|\mathbf{x}_i, \theta^t) &= \frac{P(\mathbf{x}_i, y = Gazette|\theta^t)}{P(\mathbf{x}_i, y = Times|\theta^t) + P(\mathbf{x}_i, y = Gazette|\theta^t)} \\
&= \frac{(1-\eta^t)(\lambda_G^t)^n \exp(-\lambda_G^t \sum_{j=1}^n x_j)}{\eta^t(\lambda_T^t)^n \exp(-\lambda_T^t \sum_{j=1}^n x_j) + (1-\eta^t)(\lambda_G^t)^n \exp(-\lambda_G^t \sum_{j=1}^n x_j)}
\end{aligned}
$$

(ii) M-step, we first find the expectation of the log-likelihood for the ith example.

$$
\begin{aligned}
&\sum_i E_{y\ Q_i^t}[\log P(\mathbf{x}, y|\theta)] \\
=&\sum_i [P(y = Times|\mathbf{x}_i, \theta^t) \log P(\mathbf{x}_i, y = Times|\theta) + \\
&P(y = Gazette|\mathbf{x}_i, \theta^t) \log P(\mathbf{x}_i, y = Gazette|\theta)]
\end{aligned} \tag{1}
$$

Take the derivative of above expression with respect to η, λ_T and λ_G and set them to zero, we get

$$
\begin{aligned}
\eta^{t+1} &= \frac{Softcount(y = Times)}{m} \\
\lambda_T^{t+1} &= \frac{n \sum_i P(y = Times|\mathbf{x}_t, \theta^t)}{\sum_i P(y = Times|\mathbf{x}_t, \theta^t \sum_j x_j} \\
\lambda_G^{t+1} &= \frac{n \sum_i P(y = Gazette|\mathbf{x}_t, \theta^t)}{\sum_i P(y = Gazette|\mathbf{x}_t, \theta^t \sum_j x_j}
\end{aligned}
$$

When running the EM algorithm, in each iteration, update η, λ_T, λ_G until all parameters converge.