Assignment 6 - Machine Learning [*]

Anirudh Narasimhamurthy(u0941400)

April 27, 2015

1 Naive Bayes Classification

In this question we are asked to build a Naive Bayes Classifier where the given input features are real valued and not discrete. The assumptions and the given information is summarized below and the notations which I use here will be used in the rest of the problem.

Given:

- Feature vector **x** where $\mathbf{x} = \{x_1, x_2, x_3, ...x_d\}$. They are d-dimensional real vectors.
- The labels are given by **y** and y $\in \{0, 1\}$
- S=$\{x^i, y^i\}^m$ is the training set of m examples.
- p , the prior proabability = P(y=1).
- $\sigma_{1,j}, \sigma_{0,j}, \mu_{1,j}, \mu_{0,j}$ are the parameters of the distribution from which x_j is drawn based on the value of the label y
- $\theta \in \{\ \sigma_{1,j}, \sigma_{0,j}, \mu_{1,j}, \mu_{0,j}, p\}$

1.a

In this part we are asked to derive an expression for likelihood of data in terms of the parameters and also find the log -likelihood. The first part involves deriving the value for $P(S|\theta)$. From the notation detailed above, we will have

$P(S|\theta) = P(\{x^i, y^i\}^m|\theta)$ (Since S=$\{x^i, y^i\}^m$)
$P(S|\theta) = P(\{x^i_1, x^i_2, x^i_3, ...x^i_d, y^i\}^m|\theta)$
(Since each feature vector is of d-dimensions)

Now using product rule we can split the above probabilities as product of the following two probabilities. Also I am first giving the expression for one such example (x_i, y_i) below:

$P(S|\theta) = P(y^i|\theta) \cdot P(x^i_1, x^i_2, x^i_3, ...x^i_d|y^i, \theta)$

[*] CS 6350 ; Spring 2015

We know we have m-such examples. And assuming each of them is drawn from independent and identically distributed distribution and with the Naive Bayes assumption of the conditional independenece of the individual features, the expression for the likelihood becomes:

$$\boxed{P(S|\theta) = \prod_{i=1}^{m} P(y^i|\theta) \cdot P(x^i|y^i, \theta)}$$

Each of our feature vector is of d-dimensions and so our likelihood parameter would now become:

$$P(S|\theta) = \prod_{j=1}^{d} \cdot \prod_{i=1}^{m} P(x_j^i|y^i, \theta) P(y^i|\theta)$$

- We also know that the $P(y^i = 1) = p$. Assuming there are k such examples which have label 1, then there will be (m-k) examples which would have label 0 [since our labels are only 0 and 1] and its probability would be given by (1-p) ()$P(y^i = 0)$].
- We also know that when y=1 , each of the x_j's are drawn from a particular distribution with particular parameters and when y=0, each of the x_j's are drawn from the same distribution with different parameters. Plugging those into the equation, and taking the log of the likelihood we would get:

  <u>Log Likelihood</u>

$$= \sum_{i=1}^{k}(\sum_{j=1}^{d}(log(\frac{1}{\sqrt{2\pi}\sigma_{1,j}} \exp \frac{-(x_j^i - \mu_{1,j})^2}{2\sigma_{1,j}^2}))+logp)+\sum_{i=1}^{m-k}(\sum_{j=1}^{d}(log(\frac{1}{\sqrt{2\pi}\sigma_{0,j}} \exp \frac{-(x_j^i - \mu_{0,j})^2}{2\sigma_{0,j}^2}))+log(1-p))$$

$$= k.logp+\sum_{i=1}^{k}\sum_{j=1}^{d} -log(\sqrt{2}\pi\sigma_{1,j})+\frac{-(x_j^i - \mu_{1,j})^2}{2\sigma_{1,j}^2}+(m-k).log(1-p)+\sum_{i=1}^{m-k}\sum_{j=1}^{d} -log(\sqrt{2}\pi\sigma_{0,j})+\frac{-(x_j^i - \mu_{0,j})^2}{2\sigma_{0,j}^2}$$

This would be the required expression for log-likelihood.

<u>1.b</u>

In this part of the question we are asked to derive an expression for the prior probability by taking the derivative of the log-likelihood with respect to prior and equating it to zero.

We had derived the log-likelihood in the previous question and taking the derivative with respect to p would be:

$$\frac{d}{dp}\left(k.logp + \sum_{i=1}^{k}\sum_{j=1}^{d} -log(\sigma_{1,j}) + \frac{-(x_j^i - \mu_{1,j})^2}{2\sigma_{1,j}^2} + (m-k).log(1-p) + \sum_{i=1}^{m-k}\sum_{j=1}^{d} -log(\sigma_{0,j}) + \frac{-(x_j^i - \mu_{0,j})^2}{2\sigma_{0,j}^2}\right)$$

Implies

$$\frac{k}{p} + 0 + \frac{m-k}{1-p} * -1 + 0 = 0$$

$\frac{k}{p} + \frac{-m+k}{1-p} = 0$

$k(1-p) + (-m+k)p = 0$

$k \cancel{-kp} - mp + \cancel{kp} = 0$

$mp = k$

$$\boxed{p = \frac{k}{m}}$$

Intuitively this expression for prior makes sense too. This is because the prior p in our case represents $P(y = 1)$ and k represents the number of times or the count of the label y=1 occuring in the examples and m represents the total number of examples.

If our result was to be expressed in terms of Iverson bracket instead of the notation of m and k, the prior proabability would be given by :

$$\boxed{p = \frac{[y^i = 1]}{[y^i = 1] + [y^i = 0]}}$$

1.c

Deriving expression for $\sigma_{1,j}$

To derive the expression we find the derivative of the log-likelihood with respect to one particular $\sigma_{1,j}$. Each of the j is different and when we differentiate with respect to the individual j, the summation over j would go out and the derivative would look like:

$$\frac{d}{d\sigma_{1,j}}[k.logp + \sum_{i=1}^{k}\sum_{j=1}^{d} -log(\sigma_{1,j}) + \frac{-(x_j^i - \mu_{1,j})^2}{2\sigma_{1,j}^2} + (m-k).log(1-p) + \sum_{i=1}^{m-k}\sum_{j=1}^{d} -log(\sigma_{0,j}) + \frac{-(x_j^i - \mu_{0,j})^2}{2\sigma_{0,j}^2}]$$

$\sum_{i=1}^{k}\left(\frac{-1}{\sigma_{1,j}} - \frac{(x_j^i - \mu_{1,j})^2}{2}\frac{-2}{\sigma_{1,j}^3}\right) + 0 + 0 + 0 = 0$

$\sum_{i=1}^{k} -\sigma_{1,j}^2 + \sum_{i=1}^{k}(x_j^i - \mu_{1,j})^2 = 0$

$-k \cdot \sigma_{1,j}^2 + \sum_{i=1}^{k}(x_j^i - \mu_{1,j})^2 = 0 => \sigma_{1,j}^2 = \frac{\sum_{i=1}^{k}(x_j^i - \mu_{1,j})^2}{k}$

$\boxed{\sigma_{1,j} = \sqrt{\frac{\sum_{i=1}^{k}(x_j^i - \mu_{1,j})^2}{k}}}$

This is nothing but the standard deviation of the distribution which again makes sense intuitively. And given any j, we can plug into in this expression and obtain the value of $\sigma_{1,j}$

Deriving expression for $\sigma_{0,j}$

$$\frac{d}{d\sigma_{0,j}}[k.logp + \sum_{i=1}^{k}\sum_{j=1}^{d} -log(\sigma_{1,j}) + \frac{-(x_j^i - \mu_{1,j})^2}{2\sigma_{1,j}^2} + (m-k).log(1-p) + \sum_{i=1}^{m-k}\sum_{j=1}^{d} -log(\sigma_{0,j}) + \frac{-(x_j^i - \mu_{0,j})^2}{2\sigma_{0,j}^2}]$$

$\sum_{i=1}^{m-k}\left(\frac{-1}{\sigma_{0,j}} - \frac{(x_j^i-\mu_{0,j})^2}{2}\frac{-2}{\sigma_{0,j}^3}\right)+0+0+0=0$

$\sum_{i=1}^{m-k} -\sigma_{0,j}^2 + \sum_{i=1}^{m-k}(x_j^i-\mu_{0,j})^2 = 0$

$-(m-k)\cdot\sigma_{0,j}^2 + \sum_{i=1}^{m-k}(x_j^i-\mu_{0,j})^2 = 0 => \sigma_{0,j}^2 = \frac{\sum_{i=1}^{m-k}(x_j^i-\mu_{0,j})^2}{m-k}$

$$\boxed{\sigma_{0,j} = \sqrt{\frac{\sum_{i=1}^{m-k}(x_j^i-\mu_{0,j})^2}{(m-k)}}}$$

Deriving expression for $\mu_{1,j}$

$$\frac{d}{d\mu_{1,j}}[k.logp+\sum_{i=1}^{k}\sum_{j=1}^{d}-log(\sigma_{1,j})+\frac{-(x_j^i-\mu_{1,j})^2}{2\sigma_{1,j}^2}+(m-k).log(1-p)+\sum_{i=1}^{m-k}\sum_{j=1}^{d}-log(\sigma_{0,j})+\frac{-(x_j^i-\mu_{0,j})^2}{2\sigma_{0,j}^2}]$$

$\sum_{i=1}^{k}\frac{-1}{2\sigma_{1,j}^2}\cdot(2\mu_{1,j}-2x_j^i)+0+0+0=0$

$\sum_{i=1}^{k}\frac{x_j^i}{\sigma_{1,j}^2} - k\cdot\frac{\mu_{1,j}}{\sigma_{1,j}^2} = 0 => k\mu_{1,j} = \sum_{i=1}^{k}x_j^i$

$$\boxed{\mu_{1,j} = \frac{\sum_{i=1}^{k}x_j^i}{k}}$$

Deriving expression for $\mu_{0,j}$

$$\frac{d}{d\mu_{0,j}}[k.logp+\sum_{i=1}^{k}\sum_{j=1}^{d}-log(\sigma_{1,j})+\frac{-(x_j^i-\mu_{1,j})^2}{2\sigma_{1,j}^2}+(m-k).log(1-p)+\sum_{i=1}^{m-k}\sum_{j=1}^{d}-log(\sigma_{0,j})+\frac{-(x_j^i-\mu_{0,j})^2}{2\sigma_{0,j}^2}]$$

$\sum_{i=1}^{m-k}\frac{-1}{2\sigma_{0,j}^2}\cdot(2\mu_{0,j}-2x_j^i)+0+0+0=0$

$\sum_{i=1}^{m-k}\frac{x_j^i}{\sigma_{0,j}^2} - k\cdot\frac{\mu_{0,j}}{\sigma_{0,j}^2} = 0 => (m-k)\mu_{0,j} = \sum_{i=1}^{m-k}x_j^i$

$$\boxed{\mu_{0,j} = \frac{\sum_{i=1}^{m-k}x_j^i}{(m-k)}}$$

Thus we find that the expression for μ_j's and σ_j's correspond to the actual mean and variance/ standard deviation depending on whether we represent it as σ or σ^2 of the distribution from which they were drawn.

2 Logistic Regression

In this question we are asked to derive the stochastic gradient descent algorithm for the logistic regression classifier.

2.a

Finding the derivative of the function $log(1+exp(-y_iw^Tx_i))$ with respect to the weight vector

$\frac{d}{dw}\left(log(1 + exp(-y_i w^T x_i))\right)$

Let $1 + exp(-y_i w^T x_i) = u$. Then the derivative becomes

$\frac{d}{du}log(u) \cdot \frac{du}{dw} = \frac{1}{u} \cdot -y_i x_i.exp(-y_i w^T x_i)$

Plugging back the value of u into the above expression we get the required derivative to be:

$$\boxed{\frac{-y_i x_i.exp(-y_i w^T x_i)}{1 + exp(-y_i w^T x_i)}}$$

A more generic expression would have just the term w instead of w^T and it would look like:

$$\boxed{\frac{-y_i x_i.exp(-y_i w x_i)}{1 + exp(-y_i w^T x_i)}}$$

2.b

In Stochastic Gradient Descent algorithm, the gradient is computed by using a single example (x_i, y_i) instead of the entire dataset.

The objective function where the entire dataset is composed of a single example (x_i, y_i) is given by :

$$log(1 + exp(-y_i w^T x_i)) + \frac{1}{\sigma^2} w^T w$$

The σ^2 term corresponds to the regularization term and the log term corresponds to the empirical loss term. We have to minimize the entire expression fover the w's.

Gradient of the objective with respect to this weight vector is the derivative of the objective function when the entire dataset is composed of single example :

$= \frac{d}{dw}\left(log(1 + exp(-y_i w^T x_i)) + \frac{1}{\sigma^2} w^T w\right)$

$$\boxed{\text{Gradient or}\nabla J(w) = \frac{-y_i x_i.exp(-y_i w x_i)}{1 + exp(-y_i w x_i)} + \frac{2w}{\sigma^2}}$$

2. c

Pseudocode for Stochastic Gradient Descent Algorithm

```
1. Initialize $w^0 = 0, t = 0 \in R^d$
2. For epoch 1...T :
    2.1 Pick a random example $(x_i, y_i)$ from the training set S.
    2.2 Treat $(x_i, y_i)$ as the full dataset and take the derivative of the SVM objective at the
    current $w^{t-1}$ to be $\nabla J^t(w^{t-1})$
    $J^t(w) = \frac{1}{\sigma^2} w^t w + log(1 + exp(-y_i w^t x_i))$
```

2.3 Update w as follows:
$w^t = w^{t-1} - r\nabla J^t(w^{t-1})$
i.e $\boxed{w^t = w^{t-1} - r.\frac{-y_i x_i.exp(-y_i w x_i)}{1+exp(-y_i w x_i)} + \frac{2w}{\sigma^2}}$
where r is the learning rate.

3. Return the final **w**

3 The EM Algorithm

In this problem we are given the information that there are two local newspapers Times and Gazette and each of them publish 'n' articles daily. We are given the detail about the article length drawn from an exponential distribution. In our case it is the Poisson distribution and is given by:

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

Here λ is the parameter of the distribution and for Times and Gazette they are λ_T and λ_G respectively. Also x in our case is a integer and not real valued as it represents the article length.

3.a Finding the most likely value of λ

From the explanation provided by Vivek during TA hours and in the canvas posts:

" Each example is a vector of n numbers $(x_1, x_2, ...x_n)$, where x_i denotes the length of the i^{th} article. Let us call this vector capital X. Each element of X is generated using the same distribution that is defined by the single parameter lambda. The lambda itself depends on which newspaper the example is from."

The most likely value of λ corresponds to finding the λ which maximizes $P(X|\lambda)$

This is equivalent to finding the log-likelihood and taking the derivative with respect to parameter λ and setting it to 0.

Since each of the issues are independent of each other and each of the article word count or length is independent of one another, in this assumption we have :

$$\prod_{i=1}^{n} P(x_i|\lambda)$$

From our distribution we know that this is equivalent to

$$\prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \prod_{i=1}^{n} e^{-\lambda x_i}$$

Taking the log of the likelihood we get,

$$log(\lambda^n \prod_{i=1}^{n} e^{-\lambda x_i}) = nlog\lambda - \lambda \sum_{i=1}^{n} x_i$$

To find the most likely value of λ, we take the derivative with respect to λ and set it to 0.

$$\frac{d}{d\lambda}\left(nlog\lambda - \lambda\sum_{i=1}^{n} x_i\right) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

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

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

The λ which we have obtained represents the inverse of mean. This is an interesting result.

3.b Generative model that governs the generation of this data collection

We are given the collection of m issues $(x_1, x_2, x_3.....)_1^m$ and the generative model that governs the data generation is the E-Step or the expectation step in the Expectation Maximization algorithm. A short description of that is given below:

1. Draw a label from a multinomial distribution at random. In our case since the labels are given by probabilities η and $1 - \eta$ and we have two possible labels, our distribution will be binomial distribution.
$P(y = T) = \eta$
$P(y = G) = 1 - \eta$

2. The example 'x' is drawn from the exponential distribution with the parameter λ. In our case corresponding to the example being picked it would either correspond to λ_T and λ_G

The parameters that are required to fully specify the model are $\{\lambda_T, \lambda_G, \eta\}$

3.c Clustering the issues to two groups, the Times issues and the Gazette issues

Given the parameters of the model namely $\lambda_T, \lambda_G, \eta$, the prediction is given by :

$$h = \text{arg max}_{y \in T,G} P(y)P(x|\lambda_T, \lambda_G, y)$$

which in other words can be given by :

$$h = \begin{cases} T, & \text{if } P(y = T).P(X|y = T, \lambda_T, \lambda_G) \geq P(y = G).P(X|y = G, \lambda_T, \lambda_G) \\ G, & \text{otherwise} \end{cases}$$

or in other words h= T, if $\frac{P(y=T).P(X|y=T,\lambda_T,\lambda_G)}{P(y=G).P(X|y=G,\lambda_T,\lambda_G)} \geq 1$

This is equivalent to classifying issues as Times and Gazette issues and this can be considered as the equivalence of clustering.

3.d Deriving update rule for EM algorithm and showing the work

The update rule or M-Step in Expectation Maximization corresponds to finding θ^{t+1} by maximizing with respect to θ.

It corresponds to

$\theta^{t+1} = max_\theta \sum_i E_{y\sim Q_t^i}[logP(x_i, y|\theta)]$

We would have to first derive an expression for $P(S|\theta)$ and this is given as follows:

$P(S|\theta) = P(x_i, y|\theta)$

Applying the product rule, we would have :

$P(S|\theta) = P(x|y, \theta).P(y|\theta)$

We have 'm' examples so we would have :

$P(S|\theta) = \prod_{i=1}^{m} P(x^i|y^i, \theta) \cdot P(y^i|\theta)$

And in each of our example, our feature vector x is of n-dimensions. So we would have:

$P(S|\theta) = \prod_{i=1}^{m} \prod_{j=1}^{n} P(x_j^i|y^i, \theta) \cdot P(y^i|\theta)$

We know our x_j's come from the exponential distribution according to the value of y. Also the probability of a document generated from Times is η and from the Gazette is $1-\eta$. Assuming there are k examples which have Times as the label and (m-k) examples which have Gazette as label, when we take log of the likelihood or $P(S|\theta)$ would have:

$log(P(S|\theta)) = \sum_{i=1}^{k}(\sum_{j=1}^{n}(ln\lambda_T - \lambda_T x_j^i) + ln(\eta)) + \sum_{i=1}^{m-k}(\sum_{j=1}^{n}(ln\lambda_G - \lambda_G x_j^i) + ln(1-\eta))$

Maximizing with respect to θ in our case would correspond to maximizing in terms of $\lambda_T, \lambda_G \text{and} \eta$. This is equivalent to taking the derivative and setting it to zero.

Expression for λ_T:

$\frac{d}{d\lambda_T} log(P(S|\theta)) = \sum_{i=1}^{k}\sum_{j=1}^{n}(\frac{1}{\lambda_T} - x_j^i) = 0$

$\frac{k*n}{\lambda_T} = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_j^i)$

$$\boxed{\lambda_T = \frac{kn}{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_j^i)}}$$

Expression for λ_G:

$\frac{d}{d\lambda_G} log(P(S|\theta)) = \sum_{i=1}^{m-k}\sum_{j=1}^{n}(\frac{1}{\lambda_G} - x_j^i) = 0$

$\frac{(m-k)*n}{\lambda_G} = \sum_{i=1}^{m-k}\sum_{j=1}^{n}(x_j^i)$

$$\boxed{\lambda_G = \frac{(m-k)n}{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_j^i)}}$$

Expression for η:

$\frac{d}{d\eta} log(P(S|\theta)) = kln(\eta) + (m-k)ln(1-\eta) = 0$

$\frac{k}{\eta} - \frac{m-k}{1-\eta} = 0$ $k - k\eta - m\eta + k\eta = 0$

$$\boxed{\eta = \frac{k}{m}}$$

The expectation maximization algorithm for our problem can thus be summarized as follows:

1. Initialize the parameters $\eta, \lambda_G, \lambda_T$ with random values from the binomial and exponential distribution respectively.

2. Repeat until convergence (t=1, 2 ,3 ...)

   - **E-Step**: For every example x_i, we estimate the value of y and the probability of y being equal to a particular label is shown in the derivation of expression for η
   - **M-step** : Find θ^{t+1} which would correspond to the maximum values of $\lambda_T, \lambda_G and \eta$ at that point.

3. When the θ^t values do not change much from two iterations, we can understand that convergence has happened.

4. Return the final θ