CS 6250 Homework 2

Tian Tan

Feb.1, 2017

1 Logistic Regression

1.1 Batch Gradient Descent

a. the gradient of the negative log-likelihood in terms of w for this setting.

$$NLL(D,w) = -\sum_{i=1}^{N}[(1-y_i)log(1-\sigma(w^Tx_i)) + y_i log\sigma(w^Tx_i)] \tag{1}$$

$$= -\sum_{i=1}^{N}[(1-y_i)log(\frac{e^{-w^Tx_i}}{e^{-w^Tx_i}+1}) + y_i log(\frac{1}{e^{-w^Tx_i}+1}) \tag{2}$$

$$= -\sum_{i=1}^{N}[(1-y_i)(-log(e^{w^Tx_i}+1)) + y_i(-log(e^{w^tx_i}+1) + w^Tx_i)] \tag{3}$$

$$= -\sum_{i=1}^{N}[-log(e^{w^Tx_i}+1) + y_i w^Tx_i] \tag{4}$$

$$\nabla = \sum_{i=1}^{N}(\frac{x_i e^{w^Tx_i}}{e^{w^Tx_i}+1} - y_i x_i) \tag{5}$$

$$= \sum_{i=1}^{N} x_i(\frac{1}{e^{-w^Tx_i}+1} - y_i) \tag{6}$$

1.2 Stochastic Gradient Descent

a. the log likelihood, l, of a single $(x_t, y_t) pair$

$$l = (1-y_t)log(1-\sigma(w^Tx_t)) + y_t log\sigma(w^Tx_t) \tag{7}$$

b. update the coefficient vector at time t using w_{t-1}

$$w_t = w_{t-1} - \eta[\frac{x_t e^{w_{t-1}^Tx_t}}{e^{w_{t-1}^Tx_t}+1} - y_t x_t] \tag{8}$$

$$= w_{t-1} - \eta[\sigma(w_{t-1}^Tx_t) - y_t]x_t \tag{9}$$

c. time complexity if x_t is very sparse:
O(n_t) where n_t is the number of nonzeros in x_t
d. consequence of using a very large η and very small η :
It is easy to miss the local optimal using large η since the gradient decent speed is too fast; It takes too long time to converge when using very small η

e. update w_t under the penality of L2 norm regularization

$$w_t = w_{t-1} - \eta[(\sigma(w_{t-1}^T x_t) - y_t)x_t + 2\mu w_{t-1}] \tag{10}$$

time complexity is O(Nnf) where N is the number of iteration, n is the number of training examples, and f is the average nonzero x_t per example

f.

k: number of output categories;
d: number of input dimensions;
σ_i^2: prior varianves;
x_j: training data;
c_j: index from 0 to k-1;
m: maximum number of training epochs;
ϵ: minimum relative error improvement;
η_0: initial learning rate;
σ: annealing rate;
initialization;

```
for e = 0 to m-1 do
    η_e = η_0 / (1+e/δ);
    for j=0 to n-1 do
        Z=1+Σ_{c<k-1} exp(β_c · x_j);
        for i such that x_{j,i} ≠ 0 do
            for c=0 to k-2 do
                β_{c,i} = βc, i + η_e (u_i−q)/n ∇_{c,i} Err_R(β_c, i, σ^2) ;
            end
            ;
            μ_i =q;
        end
        ;
        for c=0 to k-2 do
            p(c|x_j, β) = exp(β_c · x_j)/Z;
            β_c = β_c + η_e ∇_c Err_R(β_{c,i}, σ^2);
        end
        ;
        q=q+1
    end
    ;
    l_e = − Σ_{j<n−1} logp(c_j|x_j, β) + Err_R(β, σ^2) ;
    if relDiff(l_e, l_{e−1})¡ϵ then
        return β;
    end
end
```

Algorithm 1: lazy update

Metric	Deceased patients	Alive patients	Function to complete
Event Count			event_count_matrics
1. Average Event Count	1029.059	682.65	
2. Max Event Count	16829	12627	
3. Min Event Count	2	1	
Encounter Count			encounter_count_matrics
1. Average Encounter Count	24.861	18.6694	
2. Max Encounter Count	375	391	
3. Min Encounter Count	1	1	
Record Length			record_length_matrics
1. Average Encounter Count	151.397	194.65	
2. Max Encounter Count	2601	3103	
3. Min Encounter Count	0	0	
Common Diagnosis	1. DIAG320128 2. DIAG319835 3. DIAG313217 4. DIAG197320 5. DIAG132797	1. DIAG320128 2. DIAG319835 3. DIAG317576 4. DIAG42872402 5. DIAG313217	
Common Laboratory Test	1. LAB3009542 2. LAB3023103 3. LAB3000963 4. LAB3018572 5. LAB3016723	1. LAB3009542 2. LAB3000963 3. LAB3023103 4. LAB3018572 5. LAB3007461	
Common Medication	1. DRUG19095164 2. DRUG43012825 3. DRUG19049105 4. DRUG956874 5. DRUG19122121	1. DRUG19095164 2. DRUG43012825 3. DRUG19049105 4. DRUG19122121 5. DRUG956874	

Table 1: Descriptive statistics for alive and dead patients

2 Programming

2.1 Descriptive Statistics

2.2 Transform data

2.3 SGD Logistic Regression

Metric	eta	c	roc
Default	0.01	0	0.6
test1	0.1	0	0.66
test2	0.5	0	0.65
test3	0.1	0.05	0.60
test4	0.1	0.0001	0.66

Table 2: Descriptive statistics for alive and dead patients

Discussion: when learning rate η value increases from 0.01 to 0.1 and μ remains 0, roc value increases from 0.6 to 0.66; Moreover, when η value increases from 0.1 to 0.35, μ remains 0, roc value remains the same as 0.66; As η increases to 0.5, roc value decreases. Here an appropriate η value is important: if η is too small, learning speed will be slow, or even too slow that it can barely reach the optimal; if η is too large, it is easy to miss the local optimal; here an appropriate η value can be chosen as 0.1; As for the regulation parameter μ, we can tell from the table that if μ is too large, the accuracy of model will decrease, which is obvious since μ regulates the trade-off between maximizing likelihood and parameter values to be close to zero, so μ should be close to zero.

2.4 Hadoop

When using η =0.1 and μ = 0.0001, the roc value is 0.64, which is lower than the unensemble one. When using η = 0.3 and 0.5 with the same μ value roc auc value becomes larger(0.66), which means that there might exist overfitting in the single logistic regression where η is 0.1, and η=0.3 might be a better learning rate compared to previous one.