1 Topics in Big Data Analytics - Lecture Notes 21 (April 10, 2014) by Sudipta Pathak

PROBLEM : From out of a repository of research articles identify those that pertain to a specific topic.
IDEA 1: Correlations Based. This problem can be thought of as a classification problem. Every paper in the repository has to be classified as either pertinent to the topic or not pertinent to the topic.

The classification can be treated as a learning process. There are two phases in any learning process:
1. Training
2. Testing

Training: We have a collection P of positive examples and a collection of negative examples. Let W be a vector of keywords.

```
    For each $a \in P$ do
    For each word $w \in a$ do
        Compute the enrichment of $w$ as the number of occurrences of $w$ in $a$;
    EndFor
EndFor
For each work $w$ do
    Compute $w'$s average enrichment across all the documents in $P$
EndFor
```

Let W_p be the positive enrichment vector. W_p is nothing but a vector of average enrichments for the keywords computed using the positive examples. Along the same lines compute an enrichment vector W_n for the articles in Negative set.

Let $W = w_1, w_2, \ldots, w_k$. We define an enrichment vector S as follows. $S[i] = W_p[i] - W_n[i]$ where $W_p[i]$ and $W_n[i]$ refer to enrichments of w_i, $1 \leq i \leq k$. Each value in $S \in [-1, +1]$

For any unknown document q, we compute an enrichment vector in the same way. Let this vector be S'. Compute the Pearson's correlation between S and S'.

Pearson's correlation coefficient between X and Y is defined as $\frac{E(X-\mu_X)E(Y-\mu_Y)}{\sigma_X \sigma_Y} \in [-1, +1]$.

If this coefficient is $< \tau$ (a threshold) we will classify it as negative, otherwise classify it as positive. Pick a threshold τ that gives the best accuracy in the training data.

Gene Selection :

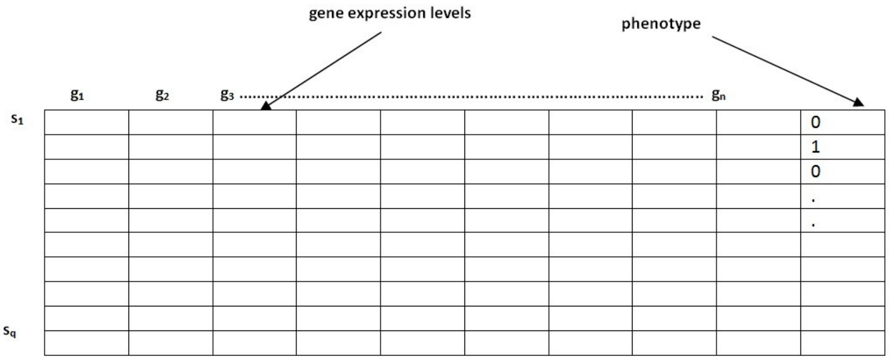

Figure 1: Gene Selection

Problem Definition: Assume that there are n genes: $g_1, g_2, \ldots, g_n$. The input to the gene selection problem has a sequence of vectors $S_1, S_2, \ldots, S_q$ where each vector is the data collected from one microarray experiment. Vector $S_i = x_i^1, x_i^2, \ldots, x_i^n, y_i$. Here x_i^j is the expression level (a real number) of the jth gene (g_j) in experiment i (for $1 \leq i \leq q$ and $1 \leq j \leq n$). Also, y_i is 1 if the event of interest is present in experiment i and y_i is -1 if the event is absent in experiment i (for $1 \leq i \leq q$). The y_i's can be thought of as representing a phenotype. The problem is to identify a minimum set of genes $g_{i_1}, g_{i_2}, \ldots, g_{i_m}$ that are enough to predict y_i in each experiment i, $1 \leq i \leq q$. We are also required to infer a prediction function f.

Clearly, the above problem can also be thought of as a *classification problem* and the function f can be considered as the *classifier*. Vectors for which $y_i = 1$ form one class and the other vectors form another class.

Techniques used for feature selection can also be used for gene selection. For example, principal component analysis (PCA) used in feature selection (actually better called feature reduction in this case) is one of the current methods for gene selection in microarray data.

Support Vector Machine (SVM) can also be used to identify a subset of genes that can explain the phenotype.

Singular Value Decomposition : (SVD)

INPUT : A matrix $A_{m \times n}$ with $m \geq n$

OUTPUT : A decomposition of A such that $A = U \sum V^T$. U is an $n \times m$ orthogonal matrix, i.e., $U^T U = I$. V is a $n \times n$ orthogonal matrix. $\sum$ is an $m \times n$ diagonal matrix $= (\sigma_1, \sigma_2, \ldots, \sigma_n)$. Here $\sigma_1, \sigma_2, \ldots, \sigma_n$ are singular values.

JACOBI ITERATIVE ALGORITHM :

IDEA : We have rotation as a basic operation. A rotation is nothing but pre and post multiplying by orthogonal matrices. We apply a series of rotations on A to get $J_1^T A J_1, J_2^T J_1^T A J_1 J_2, \ldots$. In any rotation we zero out an off diagonal element. We apply a rotation for each off diagonal element once. This series will constitute a SWEEP. We do as many sweeps as needed to obtain a diagonal matrix.

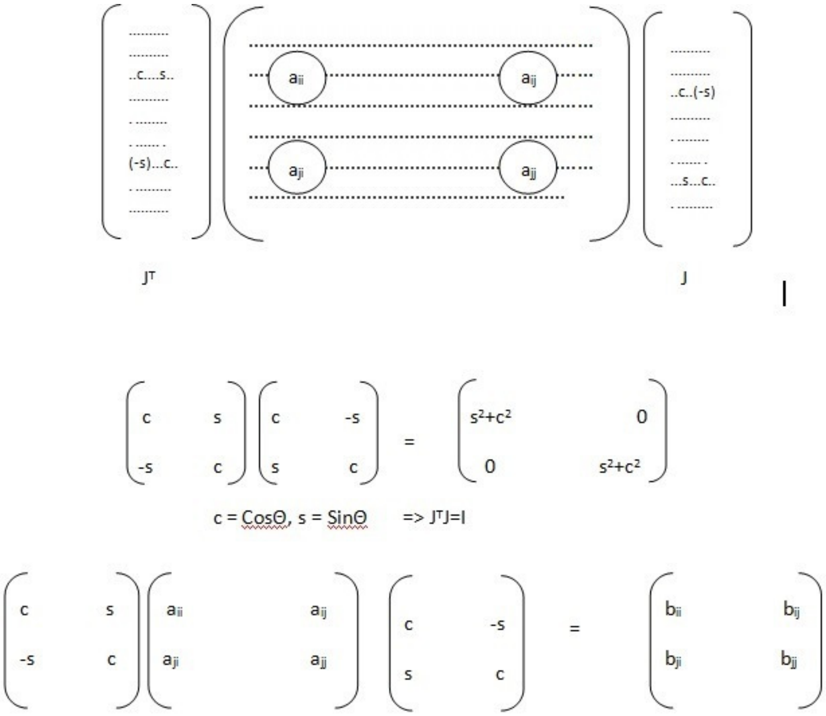

Figure 2: Jacobi Iterative Algorithm

We want $b_{ij} = 0$. $c = \frac{1}{\sqrt{1+t^2}}$; $s = \frac{t}{\sqrt{1+t^2}}$ where $t = \frac{sign(\tau)}{|\tau|+\sqrt{1+\tau^2}}$ and $\tau = \frac{a_{jj}-a_{ii}}{2a_{ij}}$.

ORDER OF ELIMINATION:

1. CLASSICAL: Use the off diagonal element with the largest absolute value.
2. Cyclic: As an example, when $n = 3$ the following order can be used: $12, 13, 23$.

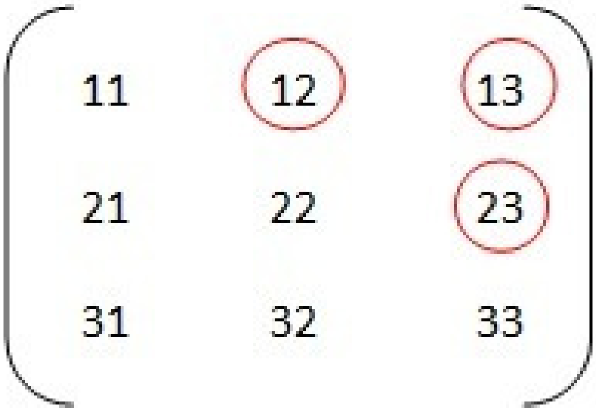

Figure 3: Cyclic order of elimination

CONVERGENCE:

CLAIM: If a_{ij} is zeroed out in any rotation, then the norm of the off diagonal elements decreases by $2a_{ij}^2$.

PROOF:

Fact: The Frobenius norm of a matrix doesn't change with orthogonal transformations. This means that

$$a_{ii}^2 + a_{jj}^2 + 2a_{ij}^2 = b_{ii}^2 + b_{jj}^2 + 2b_{ij}^2.$$

Notation: *off*(A) = norm of the off diagonal element of A. Let J be the rotation operation that targets the element a_{ij}. Let $B = J^T A J$.

off$(B) = norm(B) - \sum b_{ii}^2 = norm(A) - \sum_i a_{ii}^2 + (a_{ii}^2 + a_{ij}^2 - b_{ii}^2 - b_{jj}^2) =$ *off*$(A) - 2a_{ij}^2$.

Sequential and Parallel Rotations

We have to decompose A as: $A = U \sum V^T$.

SEQUENTIAL :

$B_1 = J_1^T A J_1$

$B_2 = J_2^T B_1 J_2$

$B_3 = J_3^T B_2 J_3$

.

.
.
.

PARALLEL :
$B_1 = J_1^T A J_1$
$B_2 = J_2^T A J_2$
$B_3 = J_3^T A J_3$
.
.
.
.

$$B' = (J_1 J_2 J_3.......J_k)^T A(J_1 J_2 J_3........J_k).$$

JACOBI RELAXATION SCHEME(JRS) (Rajasekaran and Song 2008)

The idea is to target an off-diagonal element in each rotation, but to decrease its value only by a fraction (instead of decreasing it to zero). If J is the rotation let $B = J^T AJ$. We let $b_{ij} = \lambda a_{ij}$ for some $0 < \lambda < 1$. The following values for the parameters ensure this: $c = \frac{1}{\sqrt{1+t^2}}$; $s = \frac{t}{\sqrt{1+t^2}}$; $t = \frac{Sin(\tau)(1-\lambda)}{|\tau|+\sqrt{\tau^2+(1-\lambda^2)}}$; and $\tau = \frac{a_{jj}-a_{ii}}{2a_{ij}}$.