APM120, quiz #2, 2016, Solution
Applied Linear Algebra and Big Data
Last updated: April 10, 2019

your name: ____________________________

Read these instructions carefully: Please solve all problems, **deriving, calculating and showing explicitly all stages of your solution and explaining each step of each question**. The number of points for each question is noted below, the total number of points is 115 and the final score is min(100, your points). Use the space below and attach additional pages if needed. Use a non-programmable calculator to convert your answers to decimal number format, carrying out calculations to four significant digits. Whenever relevant, use the numerical checks provided to verify your solution.

Start time: 7:00, end time: 9:00.

Good luck!

1. (35 pts) Consider $A\mathbf{x} = \mathbf{b}$ for A and $\mathbf{b}$ given below.

   (a) Do you expect to find a solution that satisfies all of these equations? Explain. If no solution exists, what would you need to assume to find a value of $\mathbf{x}$ that can be helpful in applications? Write down an expression for the general solution for $\mathbf{x}$ in such a case.

   (b) Solve for $\mathbf{x}$ and for the value of the residual $\mathbf{r} = A\mathbf{x} - \mathbf{b}$. *Numerical check:* $r_3 = 3$.

   (c) How is the QR decomposition of A calculated? How you would use the QR decomposition of A to solve for $\mathbf{x}$? When is this decomposition helpful?

   (d) Find the QR decomposition of A explicitly. *Numerical checks:* $Q_{3,2} = 0.123$, $R_{2,2} = 1.354$.

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 6 \\ 5 \\ 0 \end{pmatrix}.$$

Reminder: the inverse of a matrix A is given by $A^{-1} = \frac{1}{|A|}C^T$, where the elements of the matrix of cofactors, C, are given by $c_{ij} = (-1)^{i+j}|M_{ij}|$ where the matrix M_{ij} is obtained from A by eliminating the ith row and jth column.

Solution: (A) there are more equations than unknowns and the equations may not be consistent, so no solution to this problem exists without additional assumptions or requirements. We can therefore look for a solution that minimizes the norm squared, $\mathbf{r}^T\mathbf{r}$, of the residual $\mathbf{r} = A\mathbf{x} - \mathbf{b}$. We showed in class that the $\mathbf{x}$ that minimizes this norm squared solves $A^TA\mathbf{x} = A^T\mathbf{b}$ so that $\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b}$. We have

```
A=[1 2; 2 1; 1 1]; b=[ 6;5;0];
```

(B) Therefore

```
A^TA=A'*A=[6      5;     5      6];
(A^TA)^{-1}=inv(A'*A)=[  0.5455   -0.4545; -0.4545    0.5455];
x=inv(A'*A)*A'*b=[1;2];
r=A*x-b=[-1;-1;3];
```

(C) The QR decomposition corresponds to the Gram-Schmidt orthogonalization process of the columns of A. The result of the process is the orthonormal matrix Q for which $Q^TQ = I$, and R may then be calculated as $R = Q^TA$. Using $A = QR$ we can rewrite $A^TAx = A^Tb$ as $(QR)^T(QR)x = A^Tb$ and therefore $R^TRx = A^Tb$. We can break this into two diagonal problems,

$$R^Ty = A^Tb$$
$$Rx = y$$

where each of the two steps involves solving efficiently via forward or backward substitution. It therefore makes sense to use this approach and calculate the QR decomposition when the problem needs to be solved for many values of the rhs b.

(D) Calculating $A = QR$,

```
Q(:,1)=A(:,1)/norm(A(:,1));
Q(:,2)=A(:,2)-(A(:,2)'*Q(:,1))*Q(:,1);
Q(:,2)=Q(:,2)/norm(Q(:,2));
R=Q'*A; Q,R
```

The results are,

```
Q =[0.4082    0.8616;
    0.8165   -0.4924;
    0.4082    0.1231];
R =[2.4495    2.0412;
   -0.0000    1.3540];
```

The two step solution is given by (this was not required in the exam),

```
A'*b=[16;17];
y=R'\A'*b=[6.5320; 2.7080];
x=R\y=[1;2];
```

which is consistent with the solution obtained without QR decomposition.

2. (35 pts) Consider the following two data sets representing the size of two mountain glaciers in Switzerland (X) and of two glaciers in the Canadian Rockies (Y) during $N = 5$ years (the time average has been subtracted off, and you do not need to normalize by the standard deviation in this case). Analyze the relation between the two using maximum covariance analysis and address the following points.

   (a) What does the covariance matrix $C = XY^T/N$ represent? Calculate it and interpret each of its elements.

   (b) Calculate and interpret its singular values. *Numerical check:* $\sigma_1 = 10.47$.

   (c) Let $\mathbf{v}_i$ be the ith V singular vector of the covariance matrix C. What is $C\mathbf{v}_i$ equal to? Show this starting from $C = U\Sigma V^T$. Let the V vector corresponding to the largest singular value be $\mathbf{v}_1 = [0.2298; -0.9732]$. Calculate $\mathbf{u}_1$.

   (d) Interpret these U and V singular vectors corresponding to the largest singular value σ_1.

   (e) Write down a formula for the total variance of X, what does it represent?

   (f) Write down a formula for the total covariance of X and Y, what does it represent? Calculate the percent covariance explained by the first SVD mode of C.

   (g) Explain the difference between total variance and total covariance. Can the total covariance vanish while the total variance of X or Y is not zero? Can the total variance of X or Y vanish while the total covariance is not zero?

```
X =[-1     2     0    -2     1;
     3     4     0    -4    -3];
Y =[-1     2     0    -2     1;
    -3    -4     0     4     3];
```

Solution: (A) The covariance matrix is given by $C = XY^T/5$, or

$$C_{ij} = \frac{1}{5}\sum_{n=1}^{5} x_{in}y_{jn}.$$

It corresponds to the correlation of the different glacier sizes between the two sites.

```
C=X*Y'/5=[2   -2;
          2   -10];
```

The interpretation of C_{11} is that variable X_1 is positively correlated with Y_1 which means that both increase and decrease together. Denoting time mean with an overbar, this may be written as $C_{11} = \overline{X_1Y_1} > 0$. Similarly, $\overline{X_1Y_2} < 0$, $\overline{X_2Y_1} > 0$ and $\overline{X_2Y_2} < 0$. A negative covariance means that one variable increases the other decreases. The singular values are given by the square root of the eigenvalues of C^TC,

```
C*C'=[ 8   24;
      24   104];
sqrt(eig(C'*C))=[1.5279, 10.4721]
```

(B) The singular values are the square root of the eigenvalues of C^TC which are easily calculated for this two by two matrix, while this was not required, the SVD of C is given by,

```
[U,S,V]=svd(C)
U =[0.2298    0.9732;
    0.9732   -0.2298];
S =[10.4721        0;
          0    1.5279];
V =[0.2298    0.9732;
   -0.9732    0.2298];
```

The singular values suggest that there is one dominant mode represented by the first eigenvector of U and first of V.

(C) The singular values and vectors satisfy $A\mathbf{v}_i = \sigma_i \mathbf{u}_i$. We can therefore calculate,

```
u1=C*V(:,1)/S(1,1)
so that
u1=[0.2298; 0.9732]
```

consistent with the first column of U.

(D) The values of the corresponding first columns of U and V indicate that the second element of X (see second element of the first column of U) varies in the opposite direction from the second elements of Y (second element of the first column of V). The second mode (second column of U, V) suggests that the first element of X and the first element of Y vary together and are positively correlated, although this mode is less dominant than the first one as the second singular value is significantly smaller than the first. The singular values represent the covariance explained:

(E) The total variance of X is $\frac{1}{5}\sum_{n=1}^{5} X_{1n}^2 + X_{2n}^2)$. It is therefore equal to the sum over the diagonal elements of XX^T/N, and reflects how much variability there is in this variable.

(F) The total covariance is the sum of squares of all of the elements of $C = XY^T/N$ and reflects the degree to which X and Y are correlated with each other rather than how much they vary. The total covariance is also equal to the sum of squares of the singular values of C. Based on this the first mode explains $100 * \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$ percent of the total covariance,

```
percent1=100*S(1,1)^2/(S(1,1)^2+S(2,2)^2)
percent2=100*S(2,2)^2/(S(1,1)^2+S(2,2)^2)
 97.9157, 2.0843
```

(G) It is possible for the total covariance to vanish while the total variance is not zero if the two data sets vary independently of each other. The opposite situation is not possible, because if the variance vanishes there is also no variability and therefore no correlation.

3. (35 pts) Consider the following sets of items bought by $N = 6$ different customers on eBay.

   (a) Find the frequent patterns (that is, frequent items, pairs of items and triplets of items) and their support, based on a threshold support of $s \geq 0.5$.

   (b) Explain what are association rules, how confidence and interest of such rules are calculated and what they represent.

   (c) Find the association rules of the form $\{i, j\} \rightarrow k$ that can be deduced from frequent triplets and calculate their confidence and interest.

   (d) Write the data matrix for the N sets given below.

   (e) Suppose you were comparing purchase histories of different customers. Given that the number of products is very large, explain in a few sentences (without actual calculations) how you could efficiently quantify the similarity between buying habits of different customers using a signature matrix; explain in 2-3 sentences how one would create a signature matrix for this data matrix using random hash functions.

```
1  {shirt,pants,socks,dress}
2  {pants,dress,lipstick,perfume,shirt}
3  {shirt,pants,socks,phone}
4  {skirt,dress}
5  {shirt,soap,socks}
6  {pants,dress,shirt,lipstick,perfume}
```

Solution:

   (a) Start with finding the support (percent occurrence) of individual items, (multiply each value by $100/N$ to get percentage)

```
dress: 4
lipstick: 2
pants: 4
perfume: 2
shirt: 5
skirt: 1
socks: 3
phone: 1
soap: 1
```

   Frequent items are:

```
dress
pants
shirt
socks
```

Possible frequent pairs and their counts are, therefore,

```
shirt,socks: 3
shirt,pants: 4
shirt,dress: 3
socks,pants: 2
socks,dress: 1
pants,dress: 3
```

The items that appear in the frequent sets are therefore,

```
shirt,socks,pants,dress
```

the possible triplets are

```
shirt,socks,pants: 2
shirt,socks,dress: 1
shirt,pants,dress: 3
socks,pants,dress: 1
```

the only one for which all possible sub-pairs are frequent and which therefore needs to be tested is

```
shirt,pants,dress:
```

and it is indeed a frequent triplet:

```
shirt,pants,dress: 3
```

(b) The possible association rules and their support (fraction of occurrence), note that `{I,J}` support means the fraction of sets that include both `I` and `j`; the confidence (fraction of baskets with I that also contain j); and interest (confidence in $I \rightarrow j$ minus fraction of baskets that contain j),

(c) Finding the association rules:

```
I             ->  j    : I support, {I,J} support, j fraction, confidence, interes
shirt,pants -> dress: 4            3              2/3          3/4          1/12
pants,dress -> shirt: 3            3              5/6          3/3          1/6
shirt,dress -> pants: 3            3              2/3          3/3          1/3
```

The third rule is the most interesting one...

(d) Data matrix:

	1	2	3	4	5	6
Shirt	1	1	1	0	1	1
pants	1	1	1	0	0	1
socks	1	0	1	0	1	0
skirt	0	0	0	1	0	0
dress	1	1	0	1	0	1
lipstick	0	1	0	0	0	1
perfume	0	1	0	0	0	1
phone	0	0	1	0	0	0
soap	0	0	0	0	1	0

(e) Signature matrices: $\ldots$

4. (10 pts) Comparing sets via a signature matrix.

   (a) Explain briefly the steps used to compare many text documents efficiently using a signature matrix.

   (b) Write the data matrix for the sets given below, explain how such sets are derived from text documents.

   (c) Calculate the signature matrix for the sets given by the columns of the data matrix using the MinHash algorithm and the two hash functions of the form $h_i(r) = mod(a_i r + b_i, N)$ also specified below.

   (d) Compare the actual Jaccard similarity of all pairs of sets to the Jaccard similarities calculated from the columns of the signature matrix.

   (e) When would you expect the similarities obtained from the original data and from the signature matrix to be nearly equal.

The sets are,

```
S1={0, 2}
S2={0, 1, 3}
S3={1, 2}
```

The hash functions are given by,

$$h_1(r) = 3r + 1 \ mod \ 5$$
$$h_2(r) = r + 3 \ mod \ 5$$

Numerical check: the third column of the signature matrix is

h	S1	S2	S3
h1			2
h2			0

Solution: (A) steps are: form k-shingles from words, starting with a stop word; convert shingles to numbers using crc32 or similar; form a data matrix; create a signature of each set using MinHash algorithm; compare signatures to find similar documents. (B) The data matrix is,

r	S1	S2	S3
0	1	1	0
1	0	1	1
2	1	0	1
3	0	1	0

(C) add hash values to data matrix, (in Matlab, e.g., `r=0:4;mod(2*r+1,5)`

r	S1	S2	S3	h1	h2
0	1	1	0	1	3
1	0	1	1	4	4
2	1	0	1	2	0
3	0	1	0	0	1

initialize signature matrix (* indicates infinity)

hi	S1	S2	S3
h1	*	*	*
h2	*	*	*

scan the row 0 of data matrix to update signature,

h	S1	S2	S3
h1	1	1	*
h2	3	3	*

row 1 of data matrix,

h	S1	S2	S3
h1	1	1	4
h2	3	3	4

row 2 of data matrix,

h	S1	S2	S3
h1	1	1	2
h2	0	3	0

row 3 of data matrix,

h	S1	S2	S3
h1	1	0	2
h2	0	1	0

row 4 of data matrix (no update necessary):

h	S1	S2	S3
h1	1	0	2
h2	0	1	0

(D) Jaccard similarity of columns of signature matrix:

```
sim12=0
sim13=0
sim23=0
```

Jaccard similarity of columns of data matrix:

```
sim12=1/4
sim13=1/4
sim23=1/4
```

(E) Approximation of similarity obtained from the signature matrix should get better for very large sparse data sets and with many more hash functions (although still many less hash functions than number of rows in data matrix).