CSCI4390/6390 – Data Mining
Fall 2009, Exam I
Total Points: 100 + (10 bonus)

1. (30 points) Let $\mathbf{\Sigma} = \begin{pmatrix} 101/2 & 99/2 \\ 99/2 & 101/2 \end{pmatrix}$ be the covariance matrix for some dataset, with mean $\boldsymbol{\mu} = (2, 5)$. Answer the following questions.

   (a) (10 points) Compute the dominant eigenvector and eigenvalue of $\mathbf{\Sigma}$ by the power method. Carry out at least 3 iterations, i.e., starting with an initial vector $\boldsymbol{x}_0$, iterate until you get $\boldsymbol{x}_3$. Approximate up to 2 decimal places, rounding up when necessary. Don't forget to normalize the eigenvector.

   Answer: let $x_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

   We get $\Sigma \cdot x_0 = \begin{pmatrix} 50.5 \\ 49.5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.98 \end{pmatrix} = x_1$

   Next $\Sigma \cdot x_1 = \begin{pmatrix} 99.01 \\ 98.99 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} = x_2$

   Finally $\Sigma \cdot x_2 = \begin{pmatrix} 100 \\ 100 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} = x_3$

   This implies $u_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\lambda_1 = 100$.

   (b) (5 points) Using the spectral decomposition and/or using the fact that eigenvectors are orthogonal, what is the second eigenvector and eigenvalue of $\mathbf{\Sigma}$? Don't forget to normalize the eigenvector.

   Answer: The spectral decomposition gives us:

$$
\begin{aligned}
\Sigma - \lambda_1 u_1 u_1^T &= \begin{pmatrix} 50.5 & 49.5 \\ 49.5 & 50.5 \end{pmatrix} - 100 \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 50.5 & 49.5 \\ 49.5 & 50.5 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 100 & 100 \\ 100 & 100 \end{pmatrix} \\
&= \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \\
&= 1 \times \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \end{pmatrix}
\end{aligned}
$$

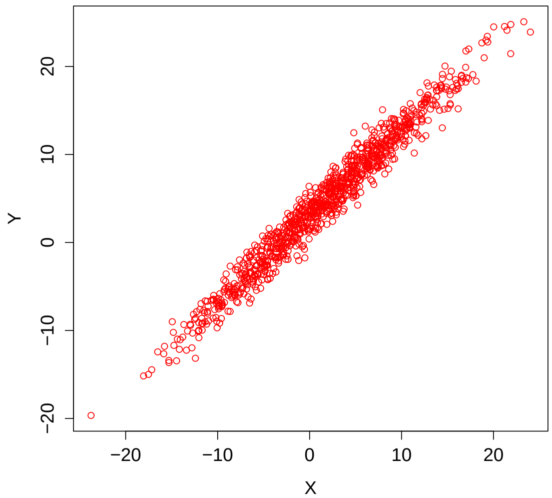
   This implies that $\lambda_2 = 1$ and $u_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

   (c) (5 points) What is the "intrinsic" dimensionality of this dataset (discounting some small amount of variance)? Why?

   Answer: Clearly the intrinsic dimensionality is 1, since most of the variance ($\frac{100}{101} = 99\%$) is captured by the first principal component.

   (d) (10 points) If the $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ from above characterize the normal distribution from which the points were generated, sketch the exact orientation/extent of the 2D normal in the XY plane. Use the contours corresponding to one standard deviation along each principal axis for your sketch.

   Answer: Your sketch should look like this:



2. (15 points) Consider the 3-way contingency table for **x, y, z**:

	z=F		**z**=G	
	y=D	**y**=E	**y**=D	**y**=E
x=A	10	10	10	5
x=B	15	5	5	20
x=C	25	10	25	10

(a) (10 points) Compute the χ^2 measure for the correlation between **y** and **z**.

(b) (5 points) Are they dependent or independent at the 95% confidence level (see the table below for χ^2 values)? Why?

Chi-Square Probabilities: *p-values* for different Chi-Square values are given for various degrees of freedom *df*. For example for *df* = 5, a chi-Square value of $\chi^2 = 11.070$ has a *p-value* of 0.05.

p-value	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
df=1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
df=2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
df=3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
df=4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
df=5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
df=6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548

Answer: Summing out **x**, we have the new 2-way contingency table between **y** and **z**, along with the row/col marginal frequencies:

	z=F	**z**=G	
y=D	50	40	90
y=E	25	35	60
	75	75	150

The expected counts in each cell are then given as follows:

	z=F	**z**=G
y=D	(90*75)/150=45	(90*75)/150= 45
y=E	(60*75)/150=30	(60*75)/150=30

Subtracting the expected and observed values, and squaring them, we get:

	z=F	**z**=G
y=D	5^2=25	-5^2=25
y=E	-5^2=25	5^2=25

Dividing by the expected counts, gives:

	z=F	**z**=G
y=A	0.56	0.56
y=B	0.83	0.83

Finally, summing all these values we obtain $\chi^2 = 0.56 + 0.56 + 0.83 + 0.83 = 2.78$.

Since there is only one degree of freedom, we find that the chi-square value is the left of the critical value, namely 3.841, which has a *p-value* of 0.05. Thus we cannot reject the null hypothesis, and we conclude that the two variables are independent.

3. (20 points) Assume that a unit hypercube is given as $[0,1]^d$, i.e., the domain is $[0,1]$ in each dimension. The main diagonal in the hypercube is defined as the vector from $(\mathbf{0},0) = (\overbrace{0,\cdots,0}^{d-1},0)$ to $(\mathbf{1},1) = (\overbrace{1,\cdots,1}^{d-1},1)$. For example, when $d=2$, the main diagonal goes from $(0,0)$ to $(1,1)$. On the other hand, the main anti-diagonal is defined as the vector from $(\mathbf{1},0) = (\overbrace{1,\cdots,1}^{d-1},0)$ to $(\mathbf{0},1) = (\overbrace{0,\cdots,0}^{d-1},1)$ For example, for $d=2$, the anti-diagonal is from $(1,0)$ to $(0,1)$.

   (a) (10 points) Sketch the diagonal and anti-diagonal in $d=3$ dimensions, and compute the angle between them.
   Answer: The main diagonal is $(1,1,1)$ and the anti-diagonal is $(0,0,1)-(1,1,0) = (-1,-1,1)$. The angle is therefore: $\cos\theta = \frac{1}{\sqrt{3}\times\sqrt{3}} = 1/3$, which implies $\theta = 70.53°$.

   (b) (10 points) What happens to the angle between the main diagonal and anti-diagonal as $d \to \infty$. First compute a general expression for the d dimensions, and then take the limit as $d \to \infty$.
   Answer: The main diagonal is $(1,1,1)$ and the anti-diagonal is $(\overbrace{-1,\cdots,-1}^{d-1},1)$.
   The angle is therefore: $\cos\theta = -(d-2)/d$.
   As $d \to \infty$, $cos(\theta) \to -1 + 2/d = -1$, which implies $\theta = 180°$ or $\theta = 0°$. In other words the diagonal and anti-diagonal are parallel!

4. (15 points) Consider the dataset below, which shows the quantity of each items bought by a customer.

tid	itemset with item quantity
1	2A, 1B, 1C
2	3A, 2B
3	2A, 2B, 1C

Using $minsup = 2$, find all frequent quantitative itemsets, i.e., frequent itemsets where quantity must be explicitly considered. For example, the frequency of A is 3, the frequency of 2A is 3 (since all three customers buy at least 2 A's), but the frequency of 3A is only 1. You may use/adapt any itemset mining method of your choice.

Answer: The level 1 itemsets A(3), B(3), C(2), all are frequent.

Next level 2: AA(3), AB(3), AC(2), BB(2), BC(2), CC(0), only CC or 2C is not frequent.

Next level 3: AAA(1), AAB(3), AAC(2), ABB(2), ABC(2), BBB(0), BBC(1). The only frequent one are AAB, AAC, ABB, ABC

Final level 4: AABB(2), AABC(2).

5. (10 points) Consider the dataset shown below:

	A	B	Class
$\boldsymbol{x}_1$	3.5	4	H
$\boldsymbol{x}_2$	2	4	H
$\boldsymbol{x}_3$	9.1	4.5	L
$\boldsymbol{x}_4$	2	6	H
$\boldsymbol{x}_5$	1.5	7	H
$\boldsymbol{x}_6$	7	6.5	H
$\boldsymbol{x}_7$	2.1	2.5	L
$\boldsymbol{x}_8$	8	4	L

Let us make an "oblique" split, instead of an axis parallel split, given as follows: $A - B \leq 3$. Compute the Information Gain of this oblique split based on Gini Index.

Answer: The gini index for the whole dataset is:
$1 - (5/8)^2 - (3/8)^2 = 1 - (0.625)^2 - (0.375)^2 = 1 - 0.39 - 0.14 = 0.47$.

Based on the oblique split, we have $D_Y = \{x_1, x_2, x_4, x_5, x_6, x_7\}$ with $P_H = 5/6$ and $P_L = 1/6$.
For D_N we have $P_H = 0/2 = 0$ and $P_L = 2/2 = 1$.

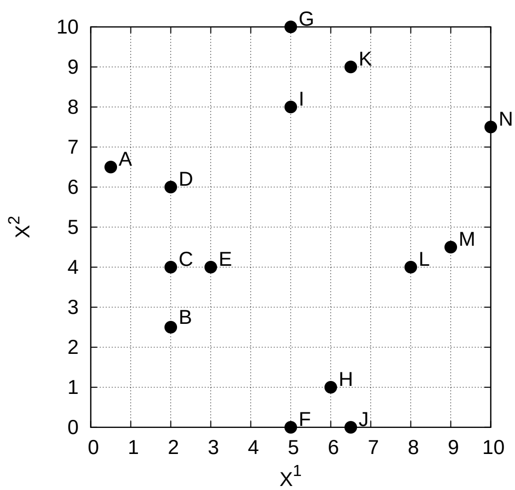
The gini for D_Y is therefore:
$1 - (0.83)^2 - (0.17)^2 = 1 - (0.69 + 0.03) = 1 - 0.72 = 0.28$.

And for D_N it is $1 - 0^2 - 1^2 = 1 - 1 = 0$.

The weighted gini of the split is: $\frac{6}{8}0.28 + \frac{2}{8}0 = 0. = 0.21$.

Thus the Gain is $0.47 - 0.21 = 0.26$.

6. (10 points) Consider the dataset shown below:



Define the L_∞ norm, between two points $\boldsymbol{a} = (a_1, a_2)$ and $\boldsymbol{b} = (b_1, b_2)$ as follows: $L_\infty(\boldsymbol{a}, \boldsymbol{b}) = \max\{|a_1 - b_1|, |a_2 - b_2|\}$. Starting with $\mu_1 = E$ and $\mu_2 = L$, show the clusters after assigning each point to the closest cluster, using the L_∞ distance in the K-means method. In case of ties, assign points to the alphabetically lower center.

Answer: It is clear that A, B, C, D, and E all belong to E. Note that F,G,H,I,J,K all have the same L_∞ distance to both E and L, thus they go to E. The only points that belong to L are: L, M, N.

7. (**Bonus: 10 points**) Draw a sketch of the 4D hypersphere.