CSE 255: Final Exam (Spring 2016)

Important Instructions:

1. **There are 10 problems in this exam. Each has a different number of assigned points, for a total of 80 points. You have 3 hours to complete the exam.**
2. **Your final answers the questions have to be entered on the "Final answers" sheet. Only that sheet will be graded. You need to turn in all of the sheets.**
3. **The question sheets have empty spaces and are printed single sided. Please use the available space to work on your answers and only put the final answer into the answer sheet.**
4. **Your final answers should be concise. You do not need to write anything other than what is asked in the question.**

Some spark functions that you may need in your solution:

```
map(f, preservesPartitioning=False)
#Return a new RDD by applying a function to each element
of this RDD.

reduce(f)
#Reduces the elements of RDD using the specified
commutative and associative binary operator.

filter(f)
#Return a new RDD containing only the elements that
satisfy a predicate.

reduceByKey(func, numPartitions=None)
#Merge the values for each key using an associative
reduce function.

sample(withReplacement, fraction, seed=None)
#Return a sampled subset of this RDD.

fullOuterJoin(other, numPartitions=None)
#For each element (k, v) in self, the resulting RDD will
either contain all pairs (k, (v, w)) for w in other, or
the pair (k, (v, None)) if no elements in other have key
k.
```

```
#Similarly, for each element (k, w) in other, the
resulting RDD will either contain all pairs (k, (v, w))
for v in self, or the pair (k, (None, w)) if no elements
in self have key k.

rightOuterJoin(other, numPartitions=None)
#For each element (k, w) in other, the resulting RDD
will either contain all pairs (k, (v, w)) for v in this,
or the pair (k, (None, w)) if no elements in self have
key k.


leftOuterJoin(other, numPartitions=None)
#For each element (k, v) in self, the resulting RDD will
either contain all pairs (k, (v, w)) for w in other, or
the pair (k, (v, None)) if no elements in other have key
k.

distinct(numPartitions=None)
#Return a new RDD containing the distinct elements in
this RDD.

collect()
#Return a list that contains all of the elements in this
RDD.
```

1. (3 pts) Which of the following are properties of Spark transformations?

A. They are not computed right away.
B. They are computed right away.
C. They are vulnerable to machine failures.
D. They define an execution plan, rather than a data structure in memory.

2. (3 pts) Which of the following is not a property of RDDs?

A. They can be changed after they are constructed.
B. They can be created by transformations applied to existing RDDs.
C. They enable parallel operations on collections of distributed data.
D. They track lineage information to enable efficient recomputation of lost data.
E. They always reside in memory.

3. (5 pts) The effect of change of basis on distances.

We are given 2 vectors in R^d : $\vec{v}_1, \vec{v}_2$, and a $d \times d$ orthonormal matrix M .

Which of the following statements is correct:

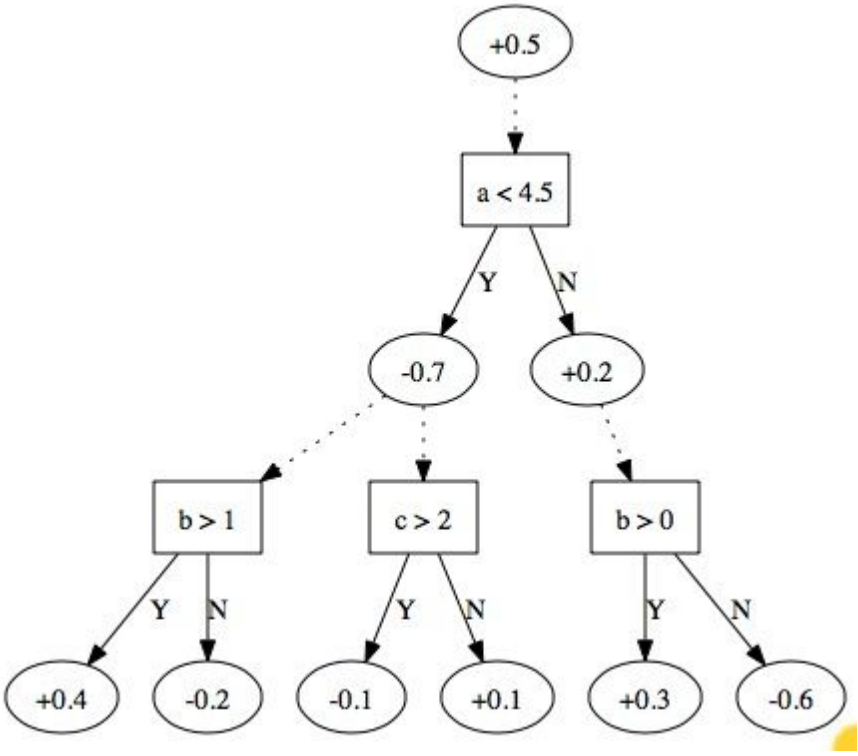
Hint: multiplying a vector by an orthonormal matrix is equivalent to changing an orthonormal basis.

A. $||M\vec{v}_1|| = ||\vec{v}_1||$
B. $M\vec{v}_1$ and $M\vec{v}_2$ are orthogonal to each other.
C. $\vec{v}_1 \cdot \vec{v}_2 = M\vec{v}_1 \cdot M\vec{v}_2$
D. $|| \vec{v}_1 - \vec{v}_2|| = ||M\vec{v}_1 - M\vec{v}_2||$
E. $\vec{v}_1 - \vec{v}_2 = M\vec{v}_1 - M\vec{v}_2$

4. (10 pts) Mark all of the following statements that are true:

A. Bagging reduces the bias of the base classifiers.
B. Boosting reduces the bias of the base classifiers.
C. Increasing the depth of a decision tree increases the variance and reduces the bias.
D. The generalization error of boosted trees is very sensitive to the number of trees that are combined.
E. If more of the training examples have a large margin, the error on an independent test set is reduced.

5. (6 pts) Consider the following Alternating Decision Tree:



What score would the tree associate with an example in which a=4,b=0, c=5
Write the sum as well as the final result.

6. (15 pts) Suppose $\vec{x}$ is a random vector in R^3 and suppose that the PCA analysis yields the mean vector $\vec{\mu}$,the eigenvectors (of length one) $\vec{v}_1, \vec{v}_2, \vec{v}_3$ and the eigenvalues $\lambda_1 > \lambda_2 > \lambda_3$

a. Write expressions for the percentage of the variance explained by the top eigenvector, the top 2 eigenvectors and the top 3 eigenvectors.

b. Write an expression for the approximation of $\vec{x}$ using the top two eigen-vectors.

c. Let **U** be the set of all unit-length vectors. What is $\vec{u} \in U$ which maximizes $Var(\vec{u} \cdot \vec{x})$?

7. (8 pts) Suppose
$\vec{x} \in R^4, \quad F(\vec{x}) = (x_1 + 2)^2 + |x_2 - 1| + (x_3 - 4)^4 + |x_4 + 1|$
Compute the gradient of $F(\vec{x})$ at $\vec{x} = (0, 0, 0, 0)$

8. (10 pts) Consider the following methods for computing the variance of the elements of an RDD X:

Method 1:

```
N, S, S2 = X.map(lambda x: np.array([1, x, x*x])) \
            .reduce(lambda a, b: a+b)
print "variance=", (S2/N)-(S/N)**2
```

Method 2:

```
N, S = X.map(lambda x: np.array([1, x])) \
        .reduce(lambda a, b: a+b)
mean = S / N
S2 = X.map(lambda x: (x-mean) * (x-mean)).reduce(lambda a, b:
a+b)
print 'variance=', S2 / N
```

Which of these methods is faster? Explain why.

9. (14 pts) Given a directed graph in the form of an RDD where each element (i,j) of the graphRDD represents an edge in the graph from node i to node j. Find one of the nodes with maximum out-degree in the graph and print the node and the out-degree.

```
nodes = range(1000)
edges = set()
for e in range(100000):
    i, j = np.random.choice(nodes, 2)
    if i != j:
        edges.add((i, j))
graphRDD=sc.parallelize(list(edges))

#Your answer here


print "Node %d has max degree %d" % (N, D)
```

10. (6 pts) Given an RDD containing details of customers and another RDD containing details of their purchases, generate the following purchase report in the form of an RDD.

[('C3', (('f', 'AZ'), '4')), ('C2', (('f', 'CA'), None)), ('C1', (('m', 'CA'), '2')), ('C5', (('m', 'NJ'), None)), ('C4', (('m', 'NY'), '7'))]

```
customerRDD = sc.parallelize(
    [('C1',('m','CA')),('C2',('f','CA')),('C3',('f','AZ')),
     ('C4',('m','NY')),('C5',('m','NJ'))])
purchaseRDD=sc.parallelize([('C1','2'),('C3','4'),('C4','7')])

#Your answer here
```

Answers sheet - Final Exam CSE255, June 2016

Name: _____ PID: _____

Q1 (3 pts): A B C D	**Q2 (3 pts) : A B C D E**
Q3 (5 pts): A B C D	**Q4 (10 pts): A B C D E**
Q5 (6 pts):	
Q6 (15 pts):	
Q7 (8 pts):	

Name: _____ PID: _____

Q8 (10 pts):
Q9 (14 pts):
Q10 (6 pts):