Final Exam - 10-605

Dec 7, 2017

10-605 **Name:**_____
Fall 2017
Final Exam Answer Key **Andrew ID:**_____
Time Limit: 80 Minutes

Grade Table (for teacher use only)

Question	Points	Score
1	10	
2	12	
3	10	
4	12	
5	10	
6	10	
7	6	
8	8	
9	6	
10	4	
11	2	
Total:	90	

1. (10 points) TA: Anant Circle the correct answers.

   (a) **(4 points)** Consider the scenario of employing the hash-trick with binary features indicating if a word w is present in a document. Let $|w|$ denote the number of documents in the corpus for which feature w is true. If two words w_1 and w_2 hash into the same bucket, then the bucket represents

   A. $|w_1$ and $w_2|$

   B. $|w_1$ or $w_2|$

   which should be a useful set of features for classification if

   A. $|w_1| << |w_2|$ or $|w_2| << |w_1|$

   B. $|w_1| \approx |w_2|$

   which we hope is true, since a large number of words occur very few times in the corpus.

   (b) **(2 points)** In minibatch stochastic gradient descent, increasing the minibatch size results in a

   A. larger

   B. smaller

   number of gradient updates per epoch.

   (c) **(2 points)** The gradient descent update step in Logistic Regression cannot be made sparse when there is L_2 regularization.

   A. True

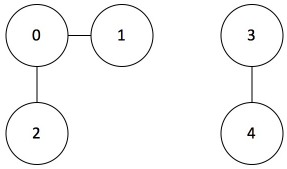
   B. False

   (d) **(2 points)** Hadoop is more suitable than Spark to implement PageRank, as there is iteration involved.

   A. True

   B. False

2. (12 points) TA: Ning Dong **Signal/Collect** Two graph nodes are in the same connected component if there is some path that connects them. Here we will write a signal/collect program to find connected components in a undirected graph (where undirected edges are encoded using 2 directed edges). Below is an example graph with two components, containing the nodes {0,1,2} and {3,4}.



After convergence, the state for each vertex v should contain an identifier for the connected component it is part of: i.e., if after convergence, it should be that v_1.state=v_2.state if and only iff v_1 and v_2 are in the same connected component.

Assume getId() returns a numeric node identifier (e.g., "3"), and notice that you can use these identifiers to construct the cluster identifier (e.g., the final state for the two nodes $\{3, 4\}$ in the second component above might be 3).

Provide a function to fill in each blank below.

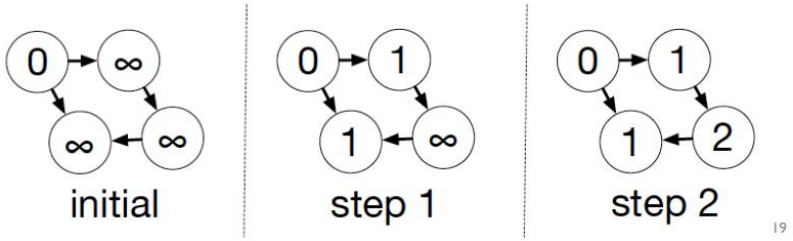
initialState getId()

collect min(oldState.min(signals))

signal source.state

Below, as a hint, is the code for single-source shortest path:

```
initialState | if (isSource) 0 else infinity
collect()    | return min(oldState, min(signals))
signal()     | return source.state + edge.weight
```



3. (10 points) TA: Rose **Perceptrons and Linear Classifiers**:

   (a) **(2 points)** Which logical operations can *not* be implemented by a perceptron? Circle all that apply.

   A. OR
   B. AND
   C. NOT
   D. XOR

   (b) **(3 points)** In class, we derived the mistake bound for Perceptrons as $(\frac{R}{\gamma})^2$. What does this imply? Circle all that apply.

   A. As the number of mistakes increases, the margin γ will shrink
   B. For a dataset with radius R and margin γ, a perceptron will make at most $(\frac{R}{\gamma})^2$ mistakes while processing that dataset
   C. For a dataset with radius R and margin γ, a perceptron will make at least $(\frac{R}{\gamma})^2$ mistakes while processing that dataset

   (c) **(5 points)** Assume binary training data containing m documents with a vocabulary size of V and a total size of n words. Assume the vocabulary fits in memory. Which of the algorithms below can be implemented so that it trains in time $O(n)$ and outputs a classifier of size $O(|V|)$? Circle all that apply.

   A. Naive Bayes
   B. unregularized logistic regression, assuming a constant number of epochs
   C. a perceptron, assuming a constant number of epochs
   D. the averaged perceptron, assuming a constant number of epochs
   E. a kernel perceptron with a polynomial kernel of degree 2, assuming a constant number of epochs

4. (12 points) TA: Chen Hu **Autodiff**
In a binary classification task, we have a Wengert list as follows:

$$z_1 = dot(\mathbf{x}, W)$$
$$z_2 = sigmoid(z_1)$$
$$z_3 = crossEntropy(z_2, y)$$
$$z_4 = frobeniusNorm(W)$$
$$loss = add(z_3, z_4)$$

Recall the cross-entropy loss criterion for binary classification tasks is:

$$crossEntropy(z, y) = -y \log z - (1 - y) \log(1 - z)$$

where $y \in \{0, 1\}$ and $z \in [0, 1]$, and the Frobenius norm is:

$$frobeniusNorm(W) = \sqrt{\sum_i \sum_j w_{i,j}^2}$$

(a) (4 points)
Derive the partial derivative $\frac{\partial loss}{\partial z_2}$.
$\frac{\partial loss}{\partial z_2} = \frac{\partial z_3}{\partial z_2} = \frac{\partial(-y \log z_2 - (1-y)\log(1-z_2))}{\partial z_2}$
$= \frac{-y}{z_2} + \frac{1-y}{1-z_2} = \frac{z_2 - y}{z_2(1-z_2)}$

(b) (4 points) Recall that $sigmoid(a) = \frac{1}{1+e^{-a}}$. Derive the partial derivative of $loss$ w.r.t. z_1, i.e. $\frac{\partial loss}{\partial z1}$, and express this in terms of z_2 and y.
$\frac{\partial loss}{\partial z_1} = \frac{\partial loss}{\partial z_2}\frac{\partial z_2}{\partial z_1} = (\frac{-y}{z_2} + \frac{1-y}{1-z_2})\frac{\partial z_2}{\partial z_1} = (\frac{-y}{z_2} + \frac{1-y}{1-z_2}) \cdot z_2(1 - z_2)$
$= -y(1 - z_2) + (1 - y)z_2 = z_2 - y$

(c) (2 points) For a binary classification task, one could use either the sigmoid function, or the softmax function:

$$softmax(a) = \langle \frac{e^{a_1}}{e^{a_1} + e^{a_2}}, \frac{e^{a_2}}{e^{a_1} + e^{a_2}} \rangle \quad (1)$$

Which one will lead to a more accurate classifier?
sigmoid / **softmax** / **same**
Justify your answer with one sentence.

In binary classification, sigmoid and softmax are identical. Sigmoid is a special case of softmax in binary classification.

(d) (2 points) Which algorithm is most similar to the Wengert list above?
Unregularized logistic regression/Logistic regression with lazy L2 regularization / **Logistic regression with L2 regularization** / **Single Layer Perceptron**

5. (10 points) TA: Yifan Yang **GPUs and Deep learning**

   (a) (2 points) Compared with CPUs, GPUs have **more** / **less** total memory and **more** / **fewer** cores.

   (b) (4 points) Which of the following are advantages of GPUs for machine learning? (Circle all that apply.)

   (A) They can automatically differentiate numerical operations.
   (B) They can efficiently perform numeric operations on arbitrary-precision integers.
   (C) They can efficiently perform matrix/vector operations.
   (D) They have native support for hash tables, which can be used to represent feature vectors.

   (c) (4 points) Which units are part of an LSTM? Circle all that apply.

   (A) Forget gate
   (B) Hidden state
   (C) ReLu gate
   (D) Cell state

6. (10 points) TA: Tao Lin **Bloom Filters & Count-Min Sketches**

   (a) (2 points) Like a set in Python, we can add and remove the elements in a Bloom Filter
   True / **False**

   (b) (4 points) Two strings "*Cucumber*" and "*Melon*" are added into an empty Bloom Filter:

   ```
   bf = BloomFilter()
   bf.add("Cucumber")
   bf.add("Melon")
   ```

   Which of the following assertions are **possible**? Select all that apply.
   A, C, D

   A. `bf.contains("Cucumber") = True`
   B. `bf.contains("Cucumber") = False`
   C. `bf.contains("University") = True`
   D. `bf.contains("University") = False`

   (c) (4 points) Two counters are added into an empty Count-Min Sketch.

   ```
   cm = CountMinSketch()
   cm["Cucumber"] += 5
   cm["Melon"] += 3
   cm["University"] += 7
   ```

   Which of the following are **possible** answers retrieved as the value of the counter of "*Cucumber*"? Select all that apply.
   3 / **5** / **7** / **8**

7. (6 points) TA: Bo Chen **Locality Sensitive Hashing**

   (a) (2 points) One application of LSH is very fast clustering. More concretely, all instances with same **bx** vector will be mapped to the same cluster. If the original feature vector is of dimension **d**, mapped to **b** bits, then there will be at most 2^b clusters.

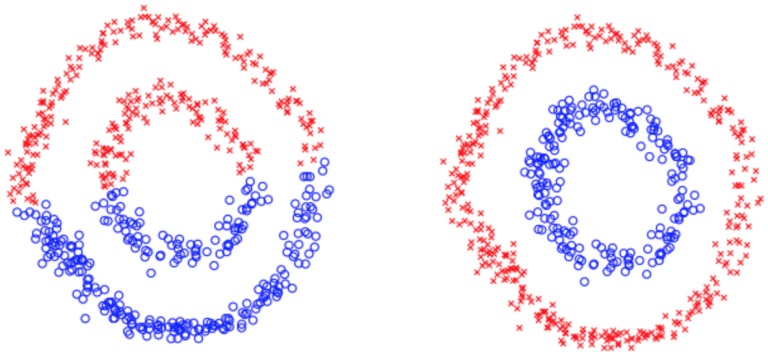
   (b) (2 points) In LSH, we map feature vectors of dimension **d** to **b** bits. In *Big-O notation*, storing the random hyper-planes that define the LSH encoder requires $O(bd)$ space.

   (c) (2 points) We have two feature vectors originally of dimension **24**, and the cosine of the angle between the two vectors is $\pi/\mathbf{6}$. Using LSH, they are both mapped to **12** bit codes. What do you expect the Hamming distance between the two codes to be?

   A. about 0 bits
   B. about 1 bit
   C. about 2 bits
   D. about 11 bits
   E. about 12 bits

8. (8 points) TA: Janani Spectral clustering.

(a) (4 points) A dataset is clustered using Spectral Clustering and K-Means clustering, the following are the clusters formed in each case.



A B

Which of the graphs above might be the output Spectral Clustering? **A** / **B**
Which corresponds to K-means clustering? **A** / **B**

(b) (4 points) Consider adapting PIC to document clustering. Let N and D be the diagonal matrices used for normalization and the matrix F is defined so that $F(i,k)$ is the TFIDF weight of word w_k in document v_i. Which of the following order of updates is more efficient?

$$v^t = Wv^{t-1} = D^{-1}Av^{t-1} = D^{-1}(N^{-1}(F^T(F(N^{-1}v^{t-1})))) \quad (2)$$

$$v^t = Wv^{t-1} = D^{-1}Av^{t-1} = ((((D^{-1}N^{-1})F^T)F)N^{-1})v^{t-1} \quad (3)$$

In one or two sentences, discuss why.

The order of updates for the first equation (2) is more efficient as it involves matrix x vector operations

9. (6 points) Prof. Cohen **SSL on Graphs**

   (a) (2 points) True or false: in using graph-based SSL algorithms on graph data, it is common to construct a graph in which edges connect nodes v_1 and v_2 only if v_1 and v_2 are highly similar.
   True **False**
   e.g. for a k-nn or ϵ-ball graph

   (b) (2 points) Let $|S|$ be the number of labeled seeds and $|V|$ indicate the number of vertices in the graph. The performance of an SSL algorithm always increases with an increase in the $\frac{|S|}{|V|}$ ratio.
   True **False**
   more unlabeled data often helps

   (c) (2 points) In implementing the Harmonic Field algorithm for large datasets in Spark, it is useful to keep the edges in memory. Why? Circle the best answer.

   (A) They are modified in every iteration, and writing to disk is slow.
   (B) They are accessed in every iteration, and reading from disk is slow.
   (C) Spark requires all parallel data structures to implemented with RDDs, which must be stored in memory.

   A: edges are not modified
   C:"which must be stored in memory" is not true

10. (4 points) TA: Minxing Liu Parameter Server Concepts

(a) (2 points) Circle the correct answer. In the SSP (Stale Synchronous Parallel) model, if we set staleness = 0, it will be similar to a **fully synchronous** / **asynchronous** Parameter Server; if we set staleness to a very large value, it will be similar to a **fully synchronous** / **asynchronous** Parameter Server.

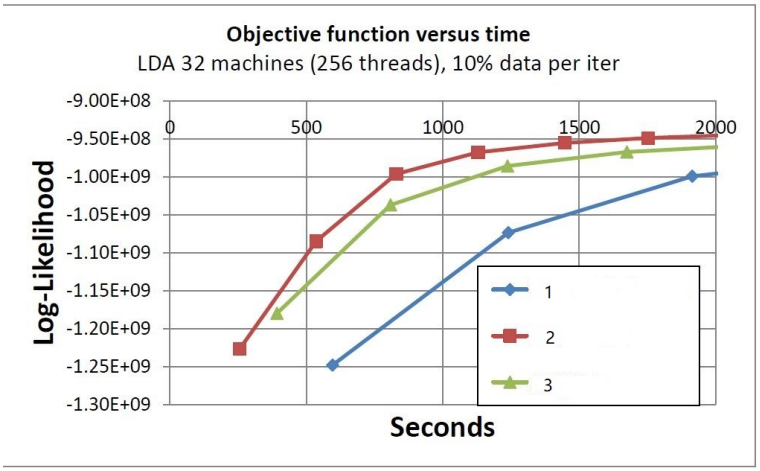


Figure 1: Experimental results using SSP, BSP and asynchronous parameter servers running LDA.

(b) (2 points) The figure shows a plot between the likelihood and running time using SSP (Stale Synchronous Parallel), BSP (Bulk Synchronous Parallel) and asynchronous parameter servers, running LDA. As you can see, the three models have different convergence trends.

Which line corresponds to the SSP model, with a well-tuned staleness parameter? **1** / **2** / **3**

11. (2 points) Multiple choice - choose the single best answer. I plan to have

- a Happy Holiday Season
- a Merry Christmas
- a Happy Hannukah
- a Excellent Exam Week
- a Fantastic Festivus
- a Carefree Krampus
- a Wonderful Winter Break
- a Very Nice Visa-Renewal Trip Home

That's all Folks!!! Have a good one!!!