CS246
Mining Massive Data Sets
Winter 2018

Handouts

Assignments

Gradiance (no late periods allowed, class token 79D9D7F3):

- GHW 1: Due on 1/25 at 11:59pm (Pacific Standard time).
- GHW 2: Due on 1/25 at 11:59pm.
- GHW 3: Due on 2/01 at 11:59pm.
- GHW 4: Due on 2/08 at 11:59pm.
- GHW 5: Due on 2/15 at 11:59pm.
- GHW 6: Due on 2/22 at 11:59pm.
- GHW 7: Due on 3/01 at 11:59pm.
- GHW 8: Due on 3/08 at 11:59pm.
- GHW 9: Due on 3/15 at 11:59pm.

Homeworks (2 late periods allowed):

- HW0 (Spark tutorial) to help you set up Spark: Due on 1/25 at 11:59pm. Solutions: [code]
- HW1: Due on 1/25 at 11:59pm. Submission Templates: [pdf | tex]. Solutions: [pdf | code]
- HW2: Due on 2/08 at 11:59pm. Submission Templates: [pdf | tex]. Solutions: [pdf | code]
- HW3: Due on 2/22 at 11:59pm. Submission Templates: [pdf | tex]. Solutions: [pdf | code]
- HW4: Due on 3/08 at 11:59pm. Submission Templates: [pdf | tex]. Solutions: [pdf | code]
- Final exam with solutions. For problem 1, see the code in [Python] [Java] [Scala].

Lecture notes (Future Schedule is tentative)

- **01/09: Introduction; MapReduce**
  Slides: [pdf][pptx]
  Reading: Ch1: Data Mining and Ch2: Large-Scale File Systems and Map-Reduce (Sect. 2.1-2.4)
- **01/11: Frequent Itemsets Mining**
  Slides: [pdf] [pptx]
  Reading: Ch6: Frequent itemsets
- **01/16: Locality-Sensitive Hashing I**
  Slides: [pdf] [pptx]
  Reading: Ch3: Finding Similar Items (Sect. 3.1-3.4)
- **01/18: Locality-Sensitive Hashing II**
  Slides: [pdf] [pptx]
  Reading: Ch3: Finding Similar Items (Sect. 3.5-3.8)
- **01/23: Clustering**
  Slides: [pdf] [pptx]
  Reading: Ch7: Clustering (Sect. 7.1-7.4)
- **01/25: Dimensionality Reduction**
  Slides: [pdf] [pptx]
  Reading: Ch11: Dimensionality Reduction (Sect. 11.4)
- **01/30: Recommender Systems I**
  Slides: [pdf] [pptx]
  Reading: Ch9: Recommendation systems
- **02/01: Recommender Systems II**
  Slides: [pdf] [pptx]
  Reading: Ch9: Recommendation systems
- **02/06: PageRank**
  Slides: [pdf] [pptx]
  Reading: Ch5: Link Analysis (Sect. 5.1-5.3, 5.5)
- **02/08: Link Spam and Introduction to Social Networks**
  Slides: [PageRank.pdf] [PageRank.pptx], [graph.pdf] [graph.pptx]
  Reading: Ch5: Link Analysis (Sect. 5.4)
  Reading: Ch10: Analysis of Social Networks (Sect. 10.1-10.2, 10.6)
- **02/13: Social Networks: Community Detection and Trawling**
  Slides: [pdf] [pptx]
  Reading: Ch10: Analysis of Social Networks (Sect. 10.3-10.5)
- **02/15: Algorithms on Large Graphs**
  Slides: [pdf] [pptx]
  Reading: Ch10: Analysis of Social Networks (Sect. 10.7-10.8)
- **02/20: Large-Scale Machine Learning I**
  Slides: [pdf], [pptx]
  Reading: Ch12: Large-Scale Machine Learning

- **02/22: Large-Scale Machine Learning II**
  Slides: [pdf], [pptx]
  Reading: Ch12: Large-Scale Machine Learning
- **02/27: Mining Data Streams I**
  Slides: [pdf] [pptx]
  Reading: Ch4: Mining data streams (Sect. 4.1-4.3)
- **03/01: Mining Data Streams II**
  Slides: [pdf] [pptx]
  Reading: Ch4: Mining data streams (Sect. 4.4-4.7)
- **03/06: Computational Advertising**
  Slides: [pdf] [pptx]
  Reading: Ch8: Advertising on the Web
- **03/08: Learning through Experimentation**
  Slides: [pdf] [pptx]
  Reading: A Contextual-Bandit Approach to Personalized News Article Recommendation by Li, Chu, Langford, Schapier. WWW 2010.
- **03/13: Optimizing Submodular Functions**
  Slides: [pdf] [pptx]
  Reading: Turning Down the Noise in the Blogosphere by El-Arini, Veda, Shahaf, Guestrin. KDD 2009.
- **03/15: Review**
  Slides: [pdf] [pptx]

All readings have been derived from the Mining Massive Datasets by J. Leskovec, A. Rajaraman and J. Ullman.

Recitation sessions documents

- **Probability review notes (courtesy CS 229)**
- **Probability review slides**
- **Proof techniques review (TBA)**
- **Linear algebra review (courtesy CS 229)**
- **Linear algebra review slides (TBA)**
- **Linear algebra, probability, and proof techniques (from CS224W)**
- **Spark tutorial slides**