FAQs

• Term project: Proposal
  • Oct. 23 By 5:00PM via Canvas

Today’s topics

• Term Project
• Predictive analytics: Linear Regression using MapReduce

Cost function (used for the K-Means algorithm)

- This slide has been added to the Week-7-B slide set
- The objective is to find:

\[
\arg\min_{\mu_i} \sum_{x_i \in S_i} |x_i - \mu_i|^2
\]

- Where \(\mu_i\) is the mean of points in \(S_i\)

Term project: Proposal

- Submit via Canvas
  • Team submission
  • 1,000 ~ 1,200 words
  • Document only

- Components of proposal
  • Title
  • Problem formulation
  • Your strategy to solve the problem
  • Your dataset
  • Timeline
  • Bibliography
1. Title

- Title should be concise and self-descriptive
- Examples of the good titles:
  - “Predictive analytics of the agricultural aspect with the wild fire using historical records in California”
  - “Do I pay right for my taxi ride?”
- Examples of the BAD titles:
  - “Term project for the team 20”
  - “Big Data Analytics”

2. Problem formulation

- The proposal should clearly identify the problem
- It should include at least one or two carefully crafted paragraph that states and highlights the problem
- The problem formulation should be able to answer following questions:
  - What is the problem you are solving?
  - What is the goal of this project?
  - Why is your project important?
  - Why is it interesting as a Big Data problem and who would use it if it were solved?

3. Your strategy to solve the problem

- Describe your proposed approach to solve the problem
- The description of the strategy should include,
  - The algorithms/techniques/models you plan to use in this project.
    - Regressions
    - Classifications
    - Graph analytics
    - Machine learning (if you take or took ML course)
    - Correlation analysis
  - The framework you plan to use in this project.
- Please note that you are also required to produce software as the final output of this project.

4. Your dataset

- The proposal should include a dataset to use for your project. Please include the link to the dataset and description.

5. Project timeline (weekly plan)

- You should provide a table with a weekly plan to complete the term project
- If you have teammate, the plan should also include information about the respective roles

6. Bibliography

- Include a bibliography
- All references must be cited (cross-referenced) in the report
7. Datasets: Web data

- Web data commons
  - [http://webdatacommons.org](http://webdatacommons.org)
  - Hyperlink Graph
    - 3.5 billion web pages and 128 billion hyperlinks
  - Product data corpus
    - 5.6 million product records
  - What are the example project?
  - Web analysis
    - Where is your web page?
      - Using in-degree, out-degree, harmonic centrality, PageRank...
    - [https://www.wim.uni-mannheim.de/fileadmin/lehrstuehle/ki/pub/Meusel-etal-GraphStructureOfTheWeb.pdf](https://www.wim.uni-mannheim.de/fileadmin/lehrstuehle/ki/pub/Meusel-etal-GraphStructureOfTheWeb.pdf)

Datasets: Enron Email data

- Enron Email dataset
  - 1,227,255 emails with 493,384 attachments
  - [https://www.cs.cmu.edu/~./enron/](https://www.cs.cmu.edu/~./enron/)
  - Possible project
    - Spam filter

Dataset: Network data

- SNAP
  - Stanford Large Network Dataset Collection
    - [https://snap.stanford.edu/data/index.html](https://snap.stanford.edu/data/index.html)
  - Amazon network
  - Twitter messages
  - Memetracker messages
  - Online community (Reddit, and Flicker)
  - And more

Dataset: Kaggle

- [https://www.kaggle.com/](https://www.kaggle.com/)
- Read Kaggle’s questions
- Pros: Many sets of interesting data
- Cons: Size of the most of datasets is compact.
- Please check the data size and make sure it is a “Big Data” problem

Dataset: etc.

- Reddit comments
  - [https://www.reddit.com/u/balaj7I_have_every_publicly_available_reddit_comments/](https://www.reddit.com/u/balaj7I_have_every_publicly_available_reddit_comments/)
  - ~1.7 billion comments (250GB compressed)
- Wikipedia dump
  - [https://meta.wikimedia.org/wiki/Data_dumps](https://meta.wikimedia.org/wiki/Data_dumps)

Data repositories

- [https://registry.opendata.aws](https://registry.opendata.aws)
- Find your dataset and visit the original web page for the dataset
If your team has \( x \) members,

- Your scope must be large enough to cover activities of \( x \) members

**How?**
- Select the right topic
- Or, consider components such as,
  - Visualization
  - User interface
  - Data collection
  - Data integration/fusion

---

### Example projects

**Spring 2017**
- Product rating and classification using NLP
- Amazon DVD Sales Analysis Based on Reviews and Genre
- Aiming for business success using Yelp and Big Data
- Finding Global Health Outcome Correlations
- Analysis of Bike Sharing Service in the Bay Area
- NBA Team Success and Relation to Hometown
- Using Big Data Analytics to Track the Cost/Benefit Ratio of Higher Education
- Too Many Games, Not Enough Time
- Recommendation System using Amazon Reviews
- Analysis Prediction and Visualization of Chicago Crime Rate by Area in Real-time
- Hashtag/tweet location correlation analysis
- Effectiveness of intrusion detection using old datasets

---

### Example projects

**Spring 2016**
- Combined Fuel Economy of Medium Duty Hybrid Trucks
- Where to Live
- Analysis of Reddit Comments for Word Popularity and Trends Using Hadoop
- Music Recommendation Based on Genre using Million song data
- Trend Analysis to Improve Donations
- Movie recommendation system based on movie attributes and human-movie interaction information
  - Song Genre generation
  - Predicting Wind Speeds Based on Atmospheric Data
  - Derelict Market Survival and Similarity Analysis
  - Trends in Baby Names and Predicting Popular Names
  - Predicting PM2.5
  - Predicting Stock Similarity using K-Means
  - Analyzing Trends on Twitter

---

### Example projects

**Spring 2015**
- Similarity and Clustering on The Million Song Dataset
- Where to Live
- Analysis on Methods Used to Analyze Eclipse and Mozilla Bug Data
- Movie recommendation using Collaborative Filtering
- A Retrospective Study Of Change In The English Language Via Textual Analysis
- 1929-2009 Climate Visualization and Predictive Analysis
- Wikipedia Link analysis
- Google book N gram analysis
- Analysis of stock market trends
- Hazardous Materials Source Mapping
- Wikipedia Page Traffic Statistics Analysis
- Million Song Dataset Geographical Analysis
- Trending Topics and Page Count Prediction on Wikipedia Traffic Log Data

---

### What is a “good” proposal?

- Completeness, Relevance, Challenging
- Your proposal should demonstrate:
  - Relevance
  - Challenging
  - Feasibility
  - Also, your topic should be appropriate to the academic environment

**Your approach**
- Well structured
- Your approach should lead you to “good results”
  - Then, what is the “good results”?:
    - Interesting knowledge
    - Demonstrating validity of your approach
    - Providing estimable accuracy/performance

---

### Part 1. Large Scale Data Analytics

3. Predictive Analysis

**Linear Regression using MapReduce**
This material is developed based on,


### Linear regression

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math score</td>
<td>85</td>
<td>85</td>
<td>62</td>
<td>67</td>
<td>89</td>
<td>85</td>
</tr>
<tr>
<td>Physics score</td>
<td>89</td>
<td>92</td>
<td>70</td>
<td>85</td>
<td>95</td>
<td>80</td>
</tr>
</tbody>
</table>

- Is there any correlation between a student’s Math score and Physic score?
- If a student’s Math score is known, can we predict his/her Physic score?

**h_θ(x) = θ_0 + θ_1 x_1**

where, the θs are the parameter vectors.

- h_θ(x), (Estimated Physics score) is called the regressand or endogeneous variable
- x_1, (Math score) is called regressor, or exogeneous variable

### Fitting the linear regression model

- The structure

  \[ h_θ(x) = θ_0 + θ_1 x_1 + θ_2 x_2 + θ_3 x_3 + θ_4 x_4 \ldots \]

- How big is the error of the fitted model?
  - We would like to minimize this error.
### Fitting the linear regression model

- **How big is the error** of the fitted model?
  - We would like to minimize this error

- **The model that fits the data best**
  - The model with the minimum sum of errors on the training data
  - e.g. The sum or mean of the squares of the errors
    - Least squares regression

### Squared error

- **Squared error**
  - Strongly penalizes very large errors
  - Drawback
    - Is very sensitive to the data
    - Erroneous or outlying data points can severely skew the resultant linear function

- **We should choose the objective function to optimize**

### Root Mean Squared Error

- Measures the differences between values predicted by model estimator and the values actually observed

\[
\text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (h(x_i) - y_i)^2}
\]

### Linear Regression

- To simplify the notation,
  \[
h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots
\]

### Objective function (Cost function)

- For a given training set, how do we pick, or learn the parameter \( \theta \)?
  - Make \( h(x) \) close to \( y \)
    - Make your prediction close to the real observation
  - We define the objective (cost) function
    - Using Mean Squared Error and multiplying \( \frac{1}{2} \) for convenience
    \[
    J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h(x_i)^2 - y_i)^2
    \]

### Minimization problem

- We have a function \( J(\theta_0, \theta_1) \)
- We want to find \( \min_{\theta} J(\theta_0, \theta_1) \)

- **Goal:** Find parameters to minimize the cost (output of the objective function)

- Outline of our approach:
  - Start with some \( \theta_0, \theta_1 \)
  - Keep changing \( \theta_0, \theta_1 \) to reduce \( J(\theta_0, \theta_1) \) until we end up at a minimum
Part 1. Large Scale Data Analytics
3. Predictive Analysis

Linear Regression: Gradient Descent Algorithm

Concept of Gradient descent algorithm (1/2)

Concept of Gradient descent algorithm (2/2)

Multiple local optimal points

Convex function

Stochastic Gradient descent algorithm

Repeat until convergence {
\[ \theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \]
\[ \theta_0 \leftarrow \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) \]
\[
\text{for } j=0 \text{ and } j=1
\]
Decreasing and increasing $\theta_1$

Decreasing $\theta_1$

$\theta = \theta - \alpha \frac{\partial}{\partial \theta} J(\theta, \theta)$

Increasing $\theta_1$

$\theta = \theta - \alpha \frac{\partial}{\partial \theta} J(\theta, \theta)$

Learning rates

$\theta = \theta - \alpha \frac{\partial}{\partial \theta} J(\theta, \theta)$

Fixed learning rate $\alpha$

$\theta = \theta - \alpha \frac{\partial}{\partial \theta} J(\theta, \theta)$
Simultaneous update

<table>
<thead>
<tr>
<th>Correct: Simultaneous update</th>
<th>Incorrect: Simultaneous update</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta = \theta - \frac{1}{m} \sum_{i=0}^{m} h(x^{(i)}) - y^{(i)}$</td>
<td>$\theta = \theta - \frac{1}{m} \sum_{i=0}^{m} h(x^{(i)}) - y^{(i)}$</td>
</tr>
<tr>
<td>$\theta = \theta - \frac{1}{m} \sum_{i=0}^{m} h(x^{(i)}) - y^{(i)}$</td>
<td>$\theta = \theta - \frac{1}{m} \sum_{i=0}^{m} h(x^{(i)}) - y^{(i)}$</td>
</tr>
</tbody>
</table>

Using Gradient Descent Algorithm for Linear Regression Model

**Gradient descent algorithm**

- Repeat until convergence
  - $\theta_j = \theta_j - \frac{1}{m} \sum_{i=0}^{m} h(x^{(i)}) - y^{(i)}$
- $J(\theta) = \frac{1}{2m} \sum_{i=0}^{m} (h(x^{(i)}) - y^{(i)})^2$

**Linear Regression Model**

- $h(x) = \theta_0 + \theta_1 x$
- $J(\theta) = \frac{1}{2m} \sum_{i=0}^{m} (h(x^{(i)}) - y^{(i)})^2$

<table>
<thead>
<tr>
<th>Case 1: $i = j$</th>
<th>Case 2: $i = j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{\partial}{\partial \theta_i} J(\theta, \alpha) = \frac{1}{m} \sum_{i=0}^{m} (h(x^{(i)}) - y^{(i)}) \theta_i$</td>
<td>$\frac{\partial}{\partial \theta_i} J(\theta, \alpha) = \frac{1}{m} \sum_{i=0}^{m} (h(x^{(i)}) - y^{(i)}) \theta_i$</td>
</tr>
<tr>
<td>$\frac{\partial}{\partial \theta_i} J(\theta, \alpha) = \frac{1}{m} \sum_{i=0}^{m} (h(x^{(i)}) - y^{(i)}) \theta_i$</td>
<td>$\frac{\partial}{\partial \theta_i} J(\theta, \alpha) = \frac{1}{m} \sum_{i=0}^{m} (h(x^{(i)}) - y^{(i)}) \theta_i$</td>
</tr>
</tbody>
</table>

Gradient descent for Linear Regression

- Repeat until convergence
  - $\theta_j = \theta_j - \frac{1}{m} \sum_{i=0}^{m} h(x^{(i)}) - y^{(i)}$
  - $\theta_i = \theta_i - \frac{1}{m} \sum_{i=0}^{m} h(x^{(i)}) - y^{(i)}$

(for $j=0$ and $j=1$)
Fitting $h_\theta(x)$

$h_\theta(x)$

$J(\theta_0, \theta_1)$