Lecture 17:
• MapReduce naive bayes example
• Looker studio overview

Announcements:
• HW-2 due
• HW-3 out (soon)
• Exam 1 Friday

Map Reduce: Naive Bayes Review

**Naive Bayes** uses probability estimates for instance classification:
- $P(\ell | \bar{x}) = \text{probability a row (feature vector) } \bar{x} \text{ has class label } \ell$
- Goal is to find the label $\ell_i$ that maximizes $P(\ell | \bar{x})$

Naive Bayes uses **Bayes Theorem** to estimate the different $P(\ell_i | \bar{x})$ ...

$$P(\ell_i | \bar{x}) = \frac{P(\bar{x} | \ell_i) \cdot P(\ell_i)}{P(\bar{x})} \approx P(\bar{x} | \ell_i) \cdot P(\ell_i)$$

Where ...
- $P(\ell_i) = \text{proportion of training rows labeled as } \ell_i \quad \ldots \text{Map: } (\ell_i, 1) \text{ pairs}$
- $P(\bar{x} | \ell_i) = \prod_{k=1}^{o} P(x_k = v_k | \ell_i) \quad \ldots \text{Map: } (\ell_i, x_k=v_k) \text{ pairs}$
- for $P(x_k = v_k | \ell_i) = \text{proportion of } \ell_i \text{ rows with } (k\text{-th feature}) x_k = v_k$
Map Reduce: Naive Bayes Review

Sketch of modeling Naive Bayes in MapReduce

- assuming we want to classify instance \([a = 2, b = 2, c = f]\) as either \(n\) or \(y\)
- assume the total number of rows is provided

Alternatively: build up a “table” of probabilities (to classify many instances)

 HW-3 Overview: Looker Studio

Looker Studio: full-featured, cloud native, visualization tool (in GCP)

- primarily for creating interactive reports and dashboards
- similar to Tableau and Power BI
- works over BigQuery plus various other data sources
- automatically converts visualizations to (real-time) SQL queries

Looker Studio Pro:

- the “enterprise”, paid version of Looker Studio
- has additional features, support, SLAs

Looker: underlying platform plus additional UI and support

- developer support including LookML
- API’s for app support and HTML embedding (iFrames)
Basic Terminology used in Looker Platform

Based on Traditional Data Warehouse Concepts ...

**Measures:** values of the “facts” or “observations” in a data warehouse
- e.g., number of items sold and purchase price of each transaction
- the attributes that would go in an SQL aggregate

**Dimensions:** provide context for the facts
- e.g., customer, product, location, time
- attributes that would go in a GROUP BY clause

**Summary – Things to Know**

- Encoding of Naive Bayes in MapReduce
- What Looker Studio is (generally)
- What a dimension and a measure is (in a data warehouse)