Unit 2 Review: Topics in Big Data

March 11, 2021

Exercise 1

A known limitation of histograms is handling categorical data. You will consider a few different proposals for such histograms.

The "Hash-to-gram"

Let's suppose that we have a dataset of N items, each of which falls into D categories. Let's further suppose that we have a hash function h that hashes each of those categories to one of $k < D$ buckets. We construct a hash-to-gram as follows:

```
Hgrm[:] = 0

for i=0...N
    j = h(data[i])
    Hgrm[j] += 1/N
```

Derive an estimator to estimate the frequency of any desired category $i \in \{0, ..., D-1\}$

Is this proposal sensible? Why or Why Not?

Exercise 2

The "Sort-to-gram"

Let's suppose that we have a dataset of N items, each of which falls into D categories. Let's further suppose that we have a known ranking function r that ranks categories by their expected frequency (0 means more likely, 1 means less likely). We use these ranks to construct an equi-width histogram (equal size buckets, dividing the D categories into k buckets).

Compare and contrast this proposal to the previous one.

Assume the categories follow a geometric distribution with a probability parameter of 0.2**, what is the expected bias?**