Mitigating the Effect of Sampling Bias on Recommender Systems through Quantization

Fengyu Li, Sarah Dean

Abstract

Placleholder.

1 Introduction

2 Draft Area

(Consider using 2-dimensional Gaussian?)
Assume the (flattened) ground truth ratings is a standard Gaussian with differential entropy

$$\begin{aligned}\mathbb{H}(R \sim N(0,1)) &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \log \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \\ &= \frac{1}{2}\log 2\pi e \\ &= 2.05\end{aligned}$$

We can show that quantization (discretization) loses information. Consider a simple binary quantization that splits the ratings around the mean. Denote the quantized r.v. as Q_1.

$$\mathbb{H}(Q_1) = -2 \cdot \frac{1}{2}\log\frac{1}{2} = 1$$

Consider another quantized r.v. Q_2 that divides the distribution four-fold based on the four quantiles.

$$\mathbb{H}(Q_2) = -4 \cdot \frac{1}{4}\log\frac{1}{4} = 2$$

We can see that different quantization schemes loses different amount of information.

$$\mathbb{H}(R) > \mathbb{H}(Q_2) > \mathbb{H}(Q_1)$$

Next, we can prove that selection bias causes information loss. Assuming the selection bias causes ratings to have unequal probability to be sampled, and higher ratings are more likely to be sampled. The distribution of the observed ratings would then be right-skewed. We prove that, in the binary case, this reduces the information loss. Let p be the probability that the sampled rating is positive.

$$\mathbb{H} = p \log p + (1-p) \log(1-p)$$

$$\frac{d\mathbb{H}}{dp} = 1 + \log p + \frac{1-p}{p-1} - \log(1-p) = 0$$

which is optimized when $p = 0.5$. This result extends to all n-quantizations (can be proved using Lagrange multipliers.) See `http://pillowlab.princeton.edu/teaching/statneuro2018/slides/notes08_infotheory.pdf`.