

CS434 - HW 1

Statistical Estimation

1. Maximum likelihood estimation of λ

- a. Log-likelihood function

$$\log P(D|\lambda) = \sum_{i=1}^N \log \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right)$$

- b. Derivative

$$\frac{d}{d\lambda} \log P(D|\lambda) = \frac{\sum_{i=1}^N x_i}{\lambda} - N = 0$$

$$\lambda = \frac{1}{N} \sum_{i=1}^N x_i$$

2. Maximum a posteriori estimate of λ with a gamma prior

?

3. Deriving the posterior of a poisson-gamma model

?

k-Nearest Neighbor

4. Encodings and distance

The issue with assigning categorical variables as ordinal is that those categories with higher assigned numbers will have a higher weight for the Euclidean distance in kNN. Giving each category its own boolean value will allow for those weights to not only be the same, but also tuned to one's liking.

5. Looking at data

$$1967 / 8000 = 0.246 = 24.6\%$$

Around 25% have an income >50k. This might mean that there is a higher likelihood of labeling something as having a lower income. 70% accuracy wouldn't be great, but it could be better. There are 85 dimensions for each data point.

6. Norms and distances

$$\|x - z\| = \sqrt{\sum_{i=1}^d x_i^2 - z_i^2}$$

We normalize the vectors so that they are on the same plane and then subtract them from each other, giving us the distance between the two vectors.

7. Implement kNN classifier

See code.

8. Implement k-fold cross validation

See code.

9. Hyperparameter search

See code.

10. Kaggle submission

See kaggle.

Debriefing

1. Approximately how many hours did you spend on this assignment?

Around 5 hours

2. Would you rate it as easy, moderate, or difficult?

Overall, moderate. The coding was fairly easy, but I do not entirely understand the math still.

3. Did you work mostly alone or did you discuss the problems with others?

I worked entirely alone.

4. How deeply do you feel you understand the material it covers?

I feel that I have a very good understanding of the coding section. I don't feel that I understand much of the math. I understand how to apply the math to our algorithms, but that's about it.

5. Any other comments?

Are there any supplemental videos I can watch that might help me understand the math sections more?