CS 541 Artificial Intelligence: Homework 1

Instructor: Jie Shen

Due: Oct 2, 2020, 8:00 pm EDT

Instruction: See Lecture 1.

Notation: $\|\boldsymbol{x}\|$ denotes the ℓ_2-norm of the vector $\boldsymbol{x}$.

Problem 1: Random Projection for Nearest Neighbor Search

Background. The promise of random projection is that for any data points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, it is possible to construct a mapping matrix $\boldsymbol{A}$, such that

$$\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \approx \|\boldsymbol{A}\boldsymbol{x}_1 - \boldsymbol{A}\boldsymbol{x}_2\|. \tag{1}$$

Therefore, if we think of $\boldsymbol{A}\boldsymbol{x}_1$ (respectively $\boldsymbol{A}\boldsymbol{x}_2$) as the new representation of $\boldsymbol{x}_1$ (respectively $\boldsymbol{x}_2$), their ℓ_2 distance is almost preserved. That means we can always perform a pre-processing step to reduce the dimension of the data, without sacrificing the performance of the algorithm for some specific applications. Observe that Eq. (1) is equivalent to verifying that for any vector $\boldsymbol{x}$, $\|\boldsymbol{x}\| \approx \|\boldsymbol{A}\boldsymbol{x}\|$.

Experiments. Let $d = 1000$. Write a Python or Matlab program to randomly generate a d-dimensional vector $\boldsymbol{x}$, where each entry of $\boldsymbol{x}$ follows a uniform distribution in $[-100, 100]$. Then *for each*

$$k \in \{10, 30, 50, 80, 100, 150, 200, 300, 400, 500, 600, 800, 1000\},$$

Step 1: randomly generate a matrix $\boldsymbol{A} \in \mathbb{R}^{k \times d}$, where each element in $\boldsymbol{A}$ is an i.i.d. draw from the normal distribution $N(0, 1/k)$;

Step 2: plot figures to compare $\|\boldsymbol{x}\|$ and $\|\boldsymbol{A}\boldsymbol{x}\|$.

Conclusion. Based on the figures, summarize your findings.

Problem 2: Reliable Data Annotation

Background. Most of the AI applications require high-quality data. While raw data are easy to obtain, it is expensive to gather correct labels even in the presence of crowdsourcing platforms. In particular, consider we have a data point $\boldsymbol{x}$ (e.g. you can think of it as an image), and we want to get the correct label of whether it is a digit or not. Suppose that the groundtruth is $+1$, i.e. it is a digit. Furthermore, suppose that a crowd worker i will return $+1$ with probability 0.6, and return -1 otherwise. That is,

$$\Pr(y_i = 1) = 0.6, \quad \Pr(y_i = -1) = 0.4. \tag{2}$$

Our goal is to obtain the correct label for this image with a confidence as high as 0.99, by querying multiple workers and following the majority vote.

1. Give a condition phrased in terms of $y_1, \ldots, y_n$ under which the majority vote returns correct label;
2. Use Chebyshev's inequality to estimate n;
3. Use Chernoff bound to estimate n;
4. Compare the estimates given by the above two inequalities, and summarize your findings.

Appendix

Chernoff bound. Let $Z_1, Z_2, \ldots, Z_n$ be n independent random variables that take value in $\{0, 1\}$. Let $Z = \sum_{i=1}^n Z_i$. For each Z_i, suppose that $\Pr(Z_i = 1) \leq \eta$. Then for any $\alpha \in [0, 1]$

$$\Pr(Z \geq (1+\alpha)\eta n) \leq e^{-\frac{\alpha^2 \eta n}{3}}.$$

When $\Pr(Z_i = 1) \geq \eta$, for any $\alpha \in [0, 1]$

$$\Pr(Z \leq (1-\alpha)\eta n) \leq e^{-\frac{\alpha^2 \eta n}{2}}.$$