NAME: AVIRAT BELEKAR
CWID: 10454332

CS 541-A Artificial Intelligence: Mid-Term Exam

Instructor: Jie Shen

10/20/2020, 18:30 – 21:00 EST

Instructions:

- Open book exam, feel free to use any resource;
- Discussion is not permitted;
- Always give your answer and explain it (guaranteed 5 point for nonempty answer);
- 20 points per problem, totally 110 points (20 * 5 + 10).

0. Write down your name. (10 pts)

1. Choose a topic from the course that you are most interested in, and talk about your understanding.

2. Let X be a random variable. Suppose that $\mathrm{E}[X] = 1$. What are the practical implications? If we further know that $\mathrm{Var}[X] = 10$, what are the practical implications?

3. Given an image, suppose that an expert will present you the correct label with probability at least 0.99, and he charges 100 dollars. On the other side, it is possible to distribute the image to a pool of non-experts, each of which charges you 1 dollar but the probability that he returns the correct label is as low as p ($p > 0.5$). Therefore, when going with the second option we have to hire many workers and take majority vote. Give a sufficient condition on p such that the following two are fulfilled simultaneously:

- the quality of the label from majority voting is as good as the one from the expert;
- it costs less to hire these non-experts.

Note: The desired condition is like $f(p) \geq 0$ for some function f. You need to show what f is, but you do not need to calculate the value of p.

4. Random projection (RP) is a widely used tool for dimension reduction.

- What are the major advantages of RP?

- Give an example where RP fails to boost computational efficiency;
- What do you think will happen if we use different types of random matrices, for example, discrete matrix for random projection? What are possible benefits and what are potential issues?

5. State the main idea of collaborative filtering. What are the possible drawbacks of its current formulation, and how to improve them?

Question 1:

It is very rare for an observation to deviate from its expected value. Both markov's and chebyshev's inequality place this intuition on a firm mathematical background. Markov's inequality relates probabilities to expectation and provide bounds for the cumulative distribution function of a random variable. On the other hand chebyshev's inequality relates variance to probabilities for the distribution function of a random variable.

Let us discuss about Markov's Inequality

Markov's inequality states that for a positive random variable X with finite mean,

$$P(X \geq t) \leq \frac{E(x)}{t}$$

This means that if the mean Of a positive random variable is small then it is more likely to be too large often which implies that probability that it is large is too small.

Proof of Markov's inequality:

Suppose x is a discrete random variable,

$$E[X] = \sum_{x} x \cdot Pr(X = x)$$

$$\geq \sum_{x \geq t} x \cdot Pr(X = x$$

$$\geq t \cdot \sum Pr(X = x)$$

$$= t \cdot Pr(X \geq t)$$

Rearranging we get,

$$P(X \geq t) < \frac{E[X]}{t}.$$

Question 2:

If we were to sample from the distribution of which X is a part of, the sample mean would converge toward 1 as the number of samples is increased.

Now that we know that the variance is 10 ie being 10 times the mean, this implies that the distribution where X comes from is quite spread out.

From an estimation standpoint, the fact that the variance is high also implies that any estimate that we calculate for X or a function of X would have a higher varianc if the number of sample were low. For example, the variance of

the same mean estimate is $Var(x)/n$
Unless n is really high we cannot have sufficient confidence on the estimate given by the sample mean.

Question 4

a. Advantages of random Projection

With very high dimensions if speed is an issue, then consider that on matrix of size $n \times k$, PCA takes $O(k^2 \times n + k)$ time where as random projection takes $O(nkd)$.

Random Projection on sparse matrix is even faster

Random projection does not assume data to be in linear subspace like PCA does.

Random projections are quite fast for reducing the dimensions of a mixture of gaussians.

If the data is very large we do not need to hold memory for random projection.

b.

Random projection fails to provide computational efficiency when it comes to certain clustering algorithms. Since random projection projects and selects features randomly certain clustering algorithm will not provide the desired result. This is proven in a paper named "Random Projection for High Dimensional Data Clustering" by Xiaoli Zhang Fern and Carla Brodley. The paper also provide results as to why random projection is unstable.

Question 5:

Collaborative filtering filters information by using interactions and data collected by the system from other users.
It is based on the main idea that people who agreed in their evaluation of certain items are more likely to agree again in the future.
Most collaborative systems apply the so-called similarity index based technique. They focus on the relationship between the users and items. The similarity of items is determined by the similarity of those items by the users who have rated both items.

Drawbacks of collaborative filtering:

→ Cannot handle fresh items:

The prediction model for given (user, item) pair is the dot product of the corresponding embeddings. So if an item is not seen during training, the model can't create an embedding for it and cant query with this item. This issue is often called the cold-start problem.

To solve this problem we have two approaches

i) Projection in WALS: Given a new item not seen in training if the system has few interactions with user, then the system can easily compute an embedding V_{i0} for this item without having to retrain the whole model.

ii) Heuristics to generate embedding of fresh items:
If the system does not have interactions, the system can approximate its embeddings by averaging the embeddings of items from the same category.

Question 2:
Consider we hire n non experts. To make the cost of hiring n non-experts less than hiring an expert which is \$100 a sufficient condition on $n = 99$ non-experts.

Now since, a vote is bernoulli variable with parameter p, the majority vote would be a Binomial random variable (as summation of Bernoulli is Binomial)

For $n = 99$ non experts' atleast 50 have to be chosen who vote correctly with probability atleast p.

Therefore the probability of a majority vote can be chosen in the following way such that is atleast as good as the probability with which expert can guess correctly.

$$\sum_{i=50}^{99} {}^{99}C_i \, p^i (1-p)^{99-i} \geq 0.99$$

Probability that expert judges correctly.

Binomial distribution of majority votes

$$\sum_{i=50}^{99} {}^{99}C_i \, p^i (1-p)^{99-i} - 0.99 \geq 0$$

$$f(p) = \sum_{i=50}^{99} {}^{99}C_i \, p^i (1-p)^{99-i} - 0.99$$