Scalable Machine Learning CS281B, STAT241B — January 25, 2012
http://alex.smola.org/teaching/berkeley2012 — due on February 7, 2012
UC Berkeley — Alexander J. Smola

Problem Set 1 — Systems and Statistics

---

1 Incentive Compatible Homework Scoring (10)

This problem both explains how the assignment scoring works and it lets you prove that your optimal policy is to be truthful, i.e. to self-assess your assignments in such a way as to correlate most strongly with the correct answer.

Dapo, our fearless GSI needs to score assignments. However there are too many students to score all problem sets for all students. Hence he decides to put the students in charge of their own grading. Since it is clearly in the students interest to score their problem sets as well as possible he designs the following randomized *mechanism*:

1. Sally Student solves the assignments and submits her solution on the deadline.
2. Dapo puts the a solution set with templates online after collecting the homework.
3. Sally self-grades her solutions (she kept a copy of the assignments for this purpose), using the solution templates. For problem i she picks score $s_i \geq 0$ for all problems $1 \dots p$. She submits the self-graded scores the following week in class.
4. Dapo independently chooses a subset of k problems $I \subset \{1, \dots p\}$ with equal probability and grades them, thereby obtaining the scores d_i.
5. He computes a squared estimate of the score via

$$\hat{S}(s, d, I) = 2\frac{p}{k} \sum_{i \in I} \sqrt{d_i s_i} - \sum_{i=1}^{p} s_i. \tag{1}$$

1.1 Prove that the mechanism is incentive compatible

In other words, prove that it is in Sally's best interest to assess herself using scores that are as close to Dapo's scores as possible. Some hints:

1. Compute the expected value of $\hat{S}(s, d, I)$.
2. Show that in expectation $s = d$ yields the highest scores.

1.2 Variance reduction

1. Compute the variance of the estimate of $\hat{S}(s, d, I)$ for fixed s, d.
2. Can you change the sampler for Dapo such as to reduce the variance of the estimate? How should he change the probability of grading problem i?
3. Assume that Sally is extremely risk averse. She is happy to have a slightly lower score in exchange for a smaller risk. What should she do?
4. Bound the probability of large deviation for an omniscient Sally (she knows d but not I) using e.g. Bennett's inequality or using McDiarmid's theorem.

More work on incentive compatible grading can be found here:

- http://www.kellogg.northwestern.edu/faculty/baliga/htm/nih50.pdf
- http://www.cs.ubc.ca/~kevinlb/teaching/cs532l%20-%202011-12/hw1.pdf

Problem Set 1 — Systems and Statistics

2 Highly available (key, value) storage (10)

Which modifications would you need to apply to a simple replicated (key, value) storage to make it highly available. In particular it needs to satisfy the following conditions:

1. Fault tolerance. That is, describe how your service will handle failures of nodes (consider detection, repair).
2. Scalability. How can you ensure automatic load-balancing as you increase the number of computers?
3. How can you deal with computers of different sizes?
4. Assume we have two groups of computers G and G' of sizes m and m'. Assume that G and G' have capacity k and k' respectively. We decide to use the argmax hashing for load balancing

$$g(x) = \operatorname*{argmax}_{g \in G \cup G'} c_g h(x, g). \tag{2}$$

   How do you need to choose the coefficients c_g in order to ensure that the load distribution occurs according to their capacities. hint — use the cumulative distribution function approach to determine which group receives the keys.
5. How do you need to modify this if you have three (or an arbitrary number of) groups?
6. What happens if one machine is removed? (How) do you need to adjust c_g?

3 Batch gradient using MapReduce (5)

Assume you want to minimize some function

$$f(x) = \sum_{i=1}^{m} f_i(x) \tag{3}$$

using the update procedure

$$g_t = \partial_x f(x_t) \text{ and } x_{t+1} = x_t - \gamma_t g_t. \tag{4}$$

Ignoring issues regarding the actual optimization algorithm, how can you parallelize the algorithm using MapReduce. Describe the associated steps in detail.

4 Optimal Decision (5)

Assume we have a conditional class probability $p(y|x)$ in some classification problem (x denotes the covariates and $y \in \mathcal{Y}$ is drawn from the set of labels). You can assume that $|\mathcal{Y}| = k$.

1. Find (and prove) a rule for $y^*(x)$ such that the probability of choosing the wrong label is minimized.

5 Optimal Rate (5)

Prove that the *rate* for ϵ in terms of the sample size m is tight in Hoeffding's theorem. hint — pick a distribution where you can compute the rate explicitly and show that it is of the same order. You do not need to prove that the dependence on the probability of deviation is optimal.

Problem Set 1 — Systems and Statistics

6 Maximum Entropy Distribution (10)

The entropy of a distribution is given by

$$H[p] = -\int \log p(x) dp(x). \tag{5}$$

Often it is a good idea to find the distribution with largest entropy that satisfies a number of conditions:

1. Assume that we want to find the distribution with the largest entropy which has a given expectation, that is, it satisfies

   $$\mathbf{E}_{x\sim p(x)}[\phi(x)] = \mu. \tag{6}$$

   Derive an expression for $p(x)$ in terms of $\phi(x)$. hint — you need to express the problem as a constrained maximization problem. Use the fact that $H[p]$ is convex.
2. Now we relax condition (6). In particular we only require

   $$\left\|\mathbf{E}_{x\sim p(x)}[\phi(x)] - \mu\right\| \leq \epsilon. \tag{7}$$

   What can you prove in this case? Derive the optimization problem.

Note - in your proofs you can assume that $\phi(x)$ is finite dimensional and moreover, that $\|\cdot\|$ is the Euclidean norm in this space.

7 Benchmarking (15)

You want to find out how good your computer really is without relying on the specification sheet of the manufacturer. Write code that does some of the following:

CPU Determine the clock speed, number of cores, cache size.

Memory Determine bandwidth, access latency, and an estimate of bus width. hint — speeds for reading and writing may differ. Keep in mind that your CPU has a cache. Keep in mind that your operating system is likely to use a swap file.

Disk Determine the read / write speed for sequential and random access. Is there any difference in speed depending on where on the disk you're writing? hint — keep in mind that your disk has a cache. Also keep in mind that it rotates with constant speed. Can you measure the difference?

Which other things can you measure? Part of the assignment is to decide what else could be measured. Some suggestions:

- Downloading benchmarks is not OK.
- For several pieces you will need to be able to write code in a compiled language (i.e. Java, Python, R, Scala are not going to work). If you do not know such a language, do the following:
  - Describe an algorithm in detail (pseudocode) that would suffice for the purpose.
  - Explain why an interpreted code cannot be used.
  - Consider other parts that you *can* measure using an interpreter.
- You need make your code readable such that Dapo can review it. This means that you must comment it and moreover you should include a description of what your code does and why it measures an interesting quantity.
- Having a nice digital format would be good (e.g. LaTeX).
- For many attributes you will only need a short piece of code (1-2 screens).