FBDA (CS/DS 812)

Module 1 (Randomized Algorithms & Probabilistic Methods) Test

Prof.G.Srinivasaraghavan

Date Posted: Feb 23, 2016	**Submit By**: March 3, 2016, Midnight	**Max. Marks**: 100

Q-1: A Vertex Cover for a graph is a subset of vertices such that each edge of the graph is incident on at least one vertex in the subset. Formally for a graph $G = (V, E)$, $C \subseteq V$ is a vertex cover if:

$$\forall (u, v) \in E, (u \in C) \vee (v \in C)$$

We are often interested in finding the smallest such cover for a graph. Imagine the 'hubs' in a social network. Derive a non-trivial upper bound $K(|V|, |E|)$ as a function of the number of vertices and the number of edges, such that for any graph there always exists a vertex cover of size at most $K(|V|, |E|)$. How would this generalize to regular-hypergraphs? A hypergraph is a 'graph' in which each 'hyperedge' connects some subset of vertices - for a normal graph, each edge just connects 2 vertices. A regular hypergraph is one where all hyperedges connect the same number (say r) of vertices each. **Hint**: Lovasz Local Lemma. **15**

Q-2:

a. Show using induction that if a function $f(x)$ is convex then for any $x_1, \ldots, x_n$ and $\lambda_1, \ldots, \lambda_n$ such that $\sum_{i=1}^{n} \lambda_i = 1$, $0 \leq \lambda \leq 1$, then:

$$f\left(\sum_{i=1}^{n} \lambda_i . x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

b. Use this to prove a restricted form of Jensen's Inequality: that assuming the random variable X takes only discrete values, for any convex function f

$$E(f(X)) \leq f(E(X))$$

5

Q-3: For any two random variables X and Y show that

$$E[Y.E[X|Y]] = E[XY]$$

5

Q-4: Consider a random variable X that takes real values and we know from the characteristics of the source from which X was measured that the standard deviation of X is $E[X]/10$. For some intuition about such random variables, imagine a device for which higher the output levels, higher the variance. For instance a typical (no so high-end) audio amplifier will exhibit more noise (variability in the output, that is not in the source) as the volume is stepped up. Now we wish to estimate $E[X]$. We take n samples $X_1, \ldots, X_n$ and use the 'average' of the n samples as an estimator of the real $E[X]$. How many samples should we collect to ensure that the probability of our estimate being within 1% of the actual expected value is at least 99%. **15**

Q-5: The city police wants to estimate the effectiveness of the traffic police deployment on the city roads. They want to convince their political bosses (who obviously know nothing about Chernoff bounds!!) that the effectiveness of the deployment has increased compared to last year. The effectiveness measure they have in mind is "if an offender is driving past a police picket for a 'unit of traffic density', what is the probability that the offender will be caught and fined". They need to prove to their bosses that this probability is indeed getting better. The experiment they conduct to estimate this probability is this. The experiment is conducted on the more 'permanent' offences — missing number plate, no seat belt, not meeting emission standards, etc.. Each traffic police picket stops those they identify as offenders, and do the following:

1. Say the i^{th} person is caught.
2. Ask where the person is driving from. From the deployment chart, they know the pickets that have been deployed on the way from where this person started. Suppose the traffic density (in some normalized units of traffic density) is ρ_{ij} at picket j along the way. Record $\sum_{j=1}^{n_i}(1/\rho_{ij})$, where n_i is the number of pickets on the way. In other words an offender slipping past a picket when the traffic density is ρ, is considered equivalent to slipping past ρ pickets with unit traffic density.
3. Increment the number of offenders caught by one.

Finally compute the estimated effectiveness as $p = (n/\sum_{i=1}^{n}\sum_{j=1}^{n_i}(1/\rho_{ij}))$ where n is the total number of offenders booked during the period of the experiment. Use your expertise in Chernoff bounds to prove that this is indeed an effective strategy. How would you compute the minimum recommended n for which this strategy can be shown to yield an estimate with the required confidence? **10**

Q-6: Turn the answer to question 5 into an algorithm for an optimal deployment of traffic police across the city. Assume you have computed the effectiveness figure (the probability as in Question 5). Formulate the optimal deployment of the available traffic police across the city as a linear programming problem. The measure we are trying minimize is "the expected distance (weighted by the traffic density) travelled by an offender before being caught". **10**

Q-7: We have made measurements of the values of the attributes of objects all of which are supposed to lie on a (hyper)sphere of (unknown) radius R. We need to estimate R. Suppose each data point we observe is a d-tuple of the form $(y_1, \ldots, y_d)$ where each y_i is a noisy form of the actual attribute value x_i. Assume $y_i = x_i + \epsilon$ where ϵ is a random variable that takes values ± 1 with probability $1/2$ each. We can assume that the noise terms for the attributes are all independent of each other and the noise is independent of the attribute value itself. Also assume that for our problem, $\sum_{j=1}^{d} x_i \leq L$ — the sum of the attributes is bounded by a fixed constant L. Our dataset consists of n i.i.d data points $p_i = (y_{i1}, \ldots, y_{id})$, $1 \leq i \leq n$. We use $\left(\sum_{i=1}^{n} r_i^2\right)/n$ as an estimate of R^2, where $r_i^2 = \sum_{j=1}^{d} y_{ij}^2$. Derive a Chernoff bound for how much our estimate of R^2 can deviate from the actual value. In particular prove that

$$Pr\left(\left|\frac{\sum_{i=1}^{n} r_i^2}{n} - R^2\right| > \delta\right) \leq 2e^{-O\left(n(d-\delta)^2\right)}$$

20

Q-8: Suppose I want to predict the transaction volume (in terms of the number of tradeable instruments — equity, bonds, futures, etc. — transacted) during some period. From historical records I know the average number that is transacted daily. Typically the regulatory authorities will

impose a cap on the number that can be transacted in a day — this is typically done to ensure there are uncontrolled bull/bear runs in the market, and there is the required sanity in the market to ensure continued investor confidence. Finally unlike the stock prices, it seems not unreasonable to assume that the daily transaction volumes are independent of each other. Show that it is then possible for me to predict the total transaction volume during a season, with reasonable accuracy and confidence (inverse exponential in the length of the season, and square of the ratio of the daily average to the volume cap). **Hint**: Formulate this as a martingale over the sequence of cumulative transaction volumes, in a slightly modified form. **15**

Note: 5 Marks grace for presentation, clarity, precision.