Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning

Instructor: Purushottam Kar
Date: January 31, 2018
Total: 60 marks

ASSIGNMENT 1

Problem 1.1 (The search for fairness ends here)**.** Alice needs fair coin tosses to complete a programming assignment in her randomized algorithms course. She has access to K coins $C_1, C_2, \ldots, C_K$ with biases $p_1, p_2, \ldots, p_K$. She has no idea about the biases of these coins but knows that at least one of these coins (maybe more) is fair i.e. for at least one $i \in [K], p_i = 0.5$. Can you help Alice obtain fair (or almost fair) coin tosses using these K coins?

More specifically, design an algorithm called SAMPLE. Whenever SAMPLE is invoked, it is allowed to toss any/all of the K coins any number of times, and use all their outcomes to produce a single heads or tails verdict. Analyze your algorithm to give a guarantee on $\mathbb{P}[\text{SAMPLE}() = H]$. Remember, what we desire is that $\mathbb{P}[\text{SAMPLE}() = H] \approx 0.5$.

Your can (also) have an algorithm called PREPROCESS using which Alice can play around with these K coins to test them a bit before SAMPLE starts getting invoked. Is it possible to ensure $\mathbb{P}[\text{SAMPLE} = H] = 0.5$? Solutions that are less expensive in terms of coin tosses and produce SAMPLE results that are more fair will get more credit. (10+10 marks)

Solution. Several solutions exist for this problem. Some require variable number of tosses at sampling time, others require a fixed number of tosses. A few are outlined below.

1. Solution 1: During PREPROCESS, choose a coin i which has bias $p_i \notin \{0, 1\}$ i.e. it is not a totally loaded coin. This can be found out with high confidence after tossing each of the coins a few times. At SAMPLE time, toss the chosen coin twice in succession. If the outcome is HT, predict H else if the outcome is TH, predict T, else reject and invoke SAMPLE again. To avoid high rejection rates, during the PREPROCESS, we may want to choose a coin with bias $p_i \approx 0.5$.

2. Solution 2: No PREPROCESS required. At SAMPLE time, toss each of the K coins once each (independently) and count the number of heads. If the number is even, predict H else predict T. This strategy always gives an unbiased coin toss and that too after a fixed, K number of tosses. The key is the presence of an absolutely fair coin, say i^* in the set of coins. Suppose the other coins $j \neq i^*$ resulted in an even number of heads during SAMPLE, then C_{i^*} will cause the final tally to be even with probability 0.5 and odd with probability 0.5. The same holds if the other coins result in an odd number of heads.

Problem 1.2 (Why adding two random variables makes sense)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Show that the set of all real-valued random variables X on this probability space that take finite values almost surely, forms a vector space (over the set $\mathbb{R}$ as scalars). Be careful to verify all properties that a random variable (see Lecture 1, Definition 1.6 and footnotes 1 and 2), and a vector space must satisfy while giving your solution. (10 marks)

Solution. We define two vector space operations, an additive identity, and additive inverse

1. (Vector addition) For any two random variables X, Y that are finite almost surely, we define $X + Y$ as $(X + Y)(\omega) := X(\omega) + Y(\omega)$ for any $\omega \in \Omega$.

2. (Scalar multiplication) For any random variable X that is finite almost surely and a real number $c \in \mathbb{R}$, we define $c \cdot X$ as $(c \cdot X)(\omega) := c \cdot X(\omega)$ for any $\omega \in \Omega$.

3. (Identity) Define the random variable $\mathbf{0}$ as $\mathbf{0}(\omega) = 0$ for any $\omega \in \Omega$.

4. (Inverse) For any random variable X that is finite almost surely, define the random variable $-X$ as $(-X)(\omega) = -X(\omega)$ for any $\omega \in \Omega$.

Note that the identity and inverse random variables are trivially shown to be finite almost surely. Also, for any $c \in \mathbb{R}$, we have $\{\omega : |c \cdot X(\omega)| < \infty\} = \{\omega : |X(\omega)| < \infty\}$ and thus, the random variable $c \cdot X$ is finite almost surely as well. Showing the sum of two almost surely finite random variables to be almost surely finite will take some more work. We have

$$\begin{aligned}
\mathbb{P}\left[\omega : |X(\omega) + Y(\omega)| < \infty\right] &\geq \mathbb{P}\left[\{\omega : |X(\omega)| < \infty\} \cap \{\omega : |Y(\omega)| < \infty\}\right] \\
&\geq \mathbb{P}\left[\omega : |X(\omega)| < \infty\right] + \mathbb{P}\left[\omega : |Y(\omega)| < \infty\right] - 1 \\
&= 1
\end{aligned}$$

where the first step uses the triangle inequality, the second step uses the Fréchet inequality, and the final step holds since the events $\{\omega : |X(\omega)| < \infty\}$ and $\{\omega : |Y(\omega)| < \infty\}$ are almost sure. Note the two operations we defined above inherit several properties from the set of reals

1. (Identity and inverse) $X + \mathbf{0} = X$ and $X + (-X) = \mathbf{0}$

2. (Associativity and commutativity) $(X + Y) + Z = X + (Y + Z)$ and $X + Y = Y + X$

3. (Distributivity) $c \cdot (X + Y) = c \cdot X + c \cdot Y$ and $(c + d) \cdot X = c \cdot X + d \cdot X$

Thus, all we are left to show is that the new random variables we defined in the vector space operations and additive inverse, and the additive identity, are all measurable maps.

1. (Identity) For any $t < 0$, we have $\{\omega : \mathbf{0}(\omega) \leq t\} = \emptyset$ and for any $t \geq 0$, we have $\{\omega : \mathbf{0}(\omega) \leq t\} = \Omega$. Since $\emptyset, \Omega \in \mathcal{F}$, we conclude that $\mathbf{0}$ is measurable.

2. (Inverse) We have $\{-X \leq t\} \in \mathcal{F}$ iff $\{-X > t\} \in \mathcal{F}$ since $\mathcal{F}$ is closed under the complement operation. Also $-X > t$ iff $X < -t$ iff there exists a rational number $r \in \mathbb{Q}$ such that $r < -t$ and $X \leq r$. Thus we have $\{X < -t\} = \cup_{r \in \mathbb{Q}, r < -t} \{X \leq r\} \in \mathcal{F}$ since it is a countable union of measurable sets.

3. (Scalar multiplication) We have $\{\omega : c \cdot X(\omega) \leq t\} = \{\omega : \text{sign}(c) \cdot X(\omega) \leq t/|c|\} \in \mathcal{F}$ for any $t \in \mathbb{R}$ since both X and $-X$ are both measurable.

4. (Vector addition) We have $\{X + Y \leq t\} \in \mathcal{F}$ iff $\{X > t - Y\} \in \mathcal{F}$ by closure under complements. Now $X > t - Y$ iff there exists a rational r such that $X > r > t - Y$. Thus, $\{X > t - Y\} = \bigcup_{r \in \mathbb{Q}} \{X > r\} \cap \{Y > t - r\} = \bigcup_{r \in \mathbb{Q}} \overline{\{X \leq r\} \cup \{Y \leq t - r\}} \in \mathcal{F}$ since it is a countable union of the complements of unions of measurable sets.

There are slightly different proofs that utilize the fact that countable intersections of measurable sets are measurable too.

Problem 1.3 (An overcomplicated definition of independence)**.** Show that two events are independent if and only if the σ-subalgebras they generate are independent (see Lecture 1, Definition 1.2 and Remark 1.10). (10 marks)

Solution. Notice that for any event E, we have $\sigma(E) = \left\{E, \bar{E}, \emptyset, \Omega\right\}$. Thus, one direction follows from the definition of independent algebras itself. For the other direction, we need to show that for two independent events A, B, the pairs of events $A, \bar{B}$, $\bar{A}, B$, and $\bar{A}, \bar{B}$ are independent too. We have $\mathbb{P}\left[A \cap \bar{B}\right] = \mathbb{P}\left[A - (A \cap B)\right] = \mathbb{P}\left[A\right] - \mathbb{P}\left[A \cap B\right]$ by union bound since $A \cap B$ and $A - B$ are disjoint. Since A, B are independent, we get $\mathbb{P}\left[A \cap \bar{B}\right] = \mathbb{P}\left[A\right] - \mathbb{P}\left[A\right]\mathbb{P}\left[B\right] = \mathbb{P}\left[A\right]\mathbb{P}\left[\bar{B}\right]$.

$\mathbb{P}\left[\bar{A} \cap B\right] = \mathbb{P}\left[\bar{A}\right]\mathbb{P}\left[B\right]$ follows similarly. For $\mathbb{P}\left[\bar{A} \cap \bar{B}\right]$ notice that $\bar{A} - \bar{A} \cap B$ and $\bar{A} \cap B$ are disjoint to get $\mathbb{P}\left[\bar{A} \cap \bar{B}\right] = \mathbb{P}\left[\bar{A} - \bar{A} \cap B\right] = \mathbb{P}\left[\bar{A}\right] - \mathbb{P}\left[\bar{A}\right]\mathbb{P}\left[B\right] = \mathbb{P}\left[\bar{A}\right]\mathbb{P}\left[\bar{B}\right]$.

Problem 1.4 (Social Network Analysis)**.** A company that manages a large social messaging network of N users, wishes to find out how many messages do an average pair of users send each other. For any two users $i, j \in [N]$, the company has a count c_{ij} of how many messages have they sent each other. Assume $0 \leq c_{ij} \leq K$ for $i \neq j$ and $c_{ii} = 0$ for sake of simplicity. Thus, the company wishes to find out $\mu := \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j>i} c_{ij}$. Finding this quantity explicitly will take $\mathcal{O}\left(N^2\right)$ time which is prohibitive since N is large.

A SALT alumni who is employed in that company suggested that the company estimate μ by first sampling $n \ll N$ users with replacement, say $I_1, I_2, \ldots, I_n$, and then estimating μ using $\hat{\mu}_S := \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{k>j} c_{I_j I_k}$, calculating which takes only $\mathcal{O}\left(n^2\right)$ time.

1. Would you instead like to sample S without replacement?
2. Would you like to stick with the given definition of $\hat{\mu}_S$ or would you like to tweak it a bit?
3. For the sampling strategy chosen by you and the definition of $\hat{\mu}_S$ chosen by you, show that $\hat{\mu}_S \approx \mu$ with high probability.

Give a proper quantified version of your result (in the form Chernoff, Hoeffding inequalities are presented). You may leave small constants such as $2, 32, 15$ etc unspecified but dependence on accuracy/tolerance and confidence parameters must be exactly presented.

While analyzing the above problem, be careful to note that the random variables $c_{I_1 I_2}$ and $c_{I_1 I_3}$ are not independent even if I_1, I_2, I_3 were independently sampled since they contain I_1 in common that couples them. (20 marks)

Solution. There are several solutions to this problem. Some are given below. We will stick to uniform sampling with replacement. Versions of Chernoff-Hoeffding bounds exist for sampling without replacement but the analyses are more involved. Analyzing stability of estimators also becomes more challenging without replacement. However, sampling without replacement usually gives faster convergence since there is no repetition of samples.

Solution 1

Notice that $\mathbb{E}\left[c_{I_1 I_2}\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[c_{I_1 I_2} \mid I_1 = i\right] = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{N} \sum_{j=1}^{N} c_{ij} = \frac{2}{N^2} \sum_{i=1}^{N} \sum_{j>i} c_{ij}$ since $c_{ii} = 0$ for all i. Thus, I would want to redefine my estimator as

$$\hat{\mu}_S = \frac{1}{n(n-1)} \frac{N}{N-1} \sum_{j=1}^{n} \sum_{k>j} c_{I_j I_k}$$

Using linearity of expectation it is easy to see that $\mathbb{E}\left[\hat{\mu}_S\right] = \mu$. Now use the fact that the estimator $\hat{\mu}_S$ is C-stable for $C = \frac{KN}{n(N-1)} \leq \frac{2K}{n}$ for $N > 1$ and the observation that McDiarmid's inequality pays no heed to the fact that the random variables $c_{I_1 I_2}$ and $c_{I_1 I_3}$ are coupled, to get

$$\mathbb{P}\left[|\hat{\mu}_S - \mu| > \epsilon\right] \leq 2 \exp\left(\frac{-n\epsilon^2}{8K^2}\right)$$

Solution 2

Change the estimator to remove coupling (assume n is a multiple of 2)

$$\hat{\mu}_S = \frac{N}{n(N-1)} \sum_{i=1}^{n/2} c_{I_i I_{n/2+i}}$$

Since there is no coupling now and each term $c_{I_i I_{n/2+i}}$ is a bounded random variable, Hoeffding's inequality may be directly applied.

Solution 3

It is possible to remove coupling without changing the estimator. Notice that solution 2 is not very efficient since it uses only $n/2$ pairs to estimate μ whereas solution 1 uses $n(n-1)/2$ pairs. Removing coupling is such an interesting topic that there is a textbook devoted to the problem itself (Decoupling: From Dependence to Independence, Victor de la Peña, and Evarist Giné, Springer 1999.). There are generic techniques such as Hoeffding decomposition and Serfling's representation which can perform such decoupling. If you are interested, take a look at the following paper (Stéphan Clémençon, Gábor Lugosi and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. Annals of Statistics, 36:844–874, 2008.)