IE531: Algorithms for Data Analytics
Spring, 2018
Homework 1: Review of Linear Algebra, Probability & Statistics and Computing
Due Date: March 2, 2018

Instructions

1. You can modify any of the C++ code on Compass to solve these problems, if you want. It might help you with honing your programming skills. If these attempts (at using C++ code) is turning out to to be intense, you can use MATLAB just this once.

2. You will submit a PDF-version of your answers on Compass on-or-before midnight of the due date.

Instructions

1. (*25 points*) **Tightness of the Chebyshev Bound**: This problem is about discovering distributions where the upper-bounds of the Chebyshev Inequality is tight. First, you are going to show (by example) that there is a discrete RV where this bound is tight. Then, you are going to present a cogent argument (no need to be super formal here!) that there can be no continuous RV where the Chebyshev Bound it tight.

   (a) (*5 points*) Show that the Chebyshev Bound is tight for the discrete RV $X \in \{-1, 0, 1\}$, where $Prob(X = -1) = Prob(X = 1) = \frac{1}{2k^2}$. That is, compute $E\{X\}$ and $var(X)$ and plug it into the Chebyshev Bound and arrive at the conclusion that $Prob(\mid X \mid \geq 1) = \frac{1}{k^2}$

   (b) (*20 points*) Show that there can be no continuous distribution over the whole real axis where the Chebyshev Bound is tight.

2. (*25 points*) **Unit-Ball in High Dimensions**: We will use the ℓ_4-norm to define the unit-ball as

$$B(1, d, 4) = \{(x_1, x_2, \ldots, x_d) \in \mathcal{R}^d \mid x_1^4 + x_2^4 + \cdots + x_d^4 \leq 1\}$$

   (a) (*12.5 points*) Suppose we define

$$S := \{(x_1, x_2, \ldots, x_d) \in \mathcal{R}^d \mid x_1^4 + x_2^4 + \cdots + x_d^4 \leq \frac{1}{2}\},$$

   what fraction of the volume of $B(1, d, 4)$ does S occupy?

   (b) (*12.5 points*) For any $c > 0$, prove that the fraction of the volume of $B(1, d, 4)$ outside the slab

$$\mid x_1 \mid \leq \frac{c}{d^{1/4}} \text{ is at most } \frac{1}{c^3} e^{-c^4/4}.$$

3. (*25 points*) **Overlap of Spheres in High-Dimensions**: Let $\mathbf{x}$ be a random sample from the (surface of the) unit sphere in d-dimensions with the origin as center.

   (a) (*5 points*) What is the value of $E\{\mathbf{x}\}$?

   (b) (*5 points*) What is component-wise variance of $\mathbf{x}$? That is, for $i \in \{1, 2, \ldots, d\}$ what is $E\{(\mathbf{x}_i - E\{\mathbf{x}_i\})^2\}$?

   (c) (*5 points*) Show that for any unit length vector $\mathbf{u}$, the variance of the real-valued random variable $\mathbf{u}^T\mathbf{x}$ is $\sum_{i=1}^{d} \mathbf{u}_i^2 E\{\mathbf{x}_i^2\}$. Using this, compute the variance and standard deviation of $\mathbf{u}^T\mathbf{x}$.

   (d) (*5 points*) Given two unit-radius spheres in d-dimensional space whose centers are separated by a distance of a, show that the volume of their intersection is at most

   $$\frac{8e^{-a^2(d-1)/8}}{a\sqrt{d-1}}$$

   times the volume of each sphere.

   (e) (*5 points*) From your solution to problem 3d, present a verbal argument that supports the conclusion that if the inter-center separation of the two spheres of radius r (r is not necessarily unity) is $\Omega(r/\sqrt{d})$, then they share very small mass. From this, make a cogent case for the conclusion that given randomly generated points from the two distributions, one inside each sphere, we can tell "*which sphere contains which point*" (i.e. classify we have a clustering algorithm that separates randomly generated data into two spherical-groups)

4. (*25 points*) **A Counterpoint to the Johnson-Lindenstrauss Lemma**: Prove that for every fixed dimension reduction matrix $\mathbf{A} \in \mathcal{R}^{k \times d}$ with $k < d$, there is a pair of vectors $\mathbf{x}, \mathbf{y} \in \mathcal{R}^d$ such that the distances between their images $\mathbf{Ax}$ and $\mathbf{Ay}$ is hugely distorted (compared to the distance between $\mathbf{x}$ and $\mathbf{y}$).