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 $Prob(\mid X \mid \geq 1) = \frac{1}{k^2}$
   See the top answer at this link

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

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?
   From lecture 7, and page 16 of the book –

$$Vol(S) = \left(\frac{1}{2}\right)^{d/4} \times Vol(B(1, d, 4)).$$

To see this, let

$$\begin{aligned} B(R,d,4) &= \{(x_1,x_2,\ldots,x_d)\in\mathcal{R}^d \mid x_1^4+x_2^4+\cdots+x_d^4\le R^4\} \\ &= \left\{R\left(\frac{x_1}{R},\frac{x_2}{R},\ldots,\frac{x_d}{R}\right) \mid \sum_{i=1}^{d}\left(\frac{x_i}{R}\right)^4\le 1\right\} \end{aligned}$$

It follows that $Vol(B(R,d,r)) = R^d Vol(B(1,d,4))$. We have

$$S := \{(x_1,x_2,\ldots,x_d)\in\mathcal{R}^d \mid x_1^4+x_2^4+\cdots+x_d^4\le \frac{1}{2}\} \Rightarrow S = B\left(\frac{1}{\sqrt[4]{2}},d,4\right).$$

That is, $R^4 = \frac{1}{2}$. Therefore

$$Vol(S) = \left(\frac{1}{\sqrt[4]{2}}\right)^d \times Vol(B(1,d,4)) = \left(\frac{1}{2}\right)^{d/4} \times Vol(B(1,d,4)).$$

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

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

Following the procedure of lecture 7, suppose V_{d-1} is the volume of the $(d-1)$-dimension unit ball under the ℓ_4-norm. Following the concept/method of lecture 6, if T denotes the material outside this slab, then

$$\begin{aligned} Vol(T) &\le 2\int_{c/d^{1/4}}^{1}(1-x^4)^{(d-1)/4}V_{d-1}dx \\ &\le 2\int_{c/d^{1/4}}^{\infty\leftarrow\text{see this!}}(1-x^4)^{(d-1)/4}V_{d-1}dx \\ &\le 2\int_{c/d^{1/4}}^{1}e^{x^4(d-1)/4}V_{d-1}dx \\ &\le 2\int_{c/d^{1/4}}^{1}\left(\frac{x}{(c/d^{1/4})}\right)^3 e^{x^4(d-1)/4}V_{d-1}dx \\ &= 2V_{d-1}\times\frac{d^{3/4}}{c^3}\times\frac{1}{d-1}\times\left(-e^{x^4(d-1)/4}\right)\Big|_{x=c/d^{1/4}}^{\infty} \text{ (cf. figure 1)} \\ &\le \frac{3V_{d-1}}{c^3(d-1)^{1/4}}e^{-c^4/4} \text{ for large } d. \end{aligned}$$

Following the same logic as in lecture 7, we find a lower-bound for $Vol(K)$

–

$$
\begin{aligned}
Vol(K) &= 2\int_0^1 (1-x^4)^{(d-1)/4} V_{d-1} dx \\
&\geq 2\int_0^{1/d^{1/4}} (1-x^4)^{(d-1)/4} V_{d-1} dx \\
&\geq \frac{2V_{d-1}}{(d-1)^{1/4}} \left(1 - \frac{1}{d-1}\right)^{(d-1)/4} \\
&\geq \frac{2V_{d-1}}{(d-1)^{1/4}} \left(1 - \frac{1}{\cancel{d-1}} \times \frac{\cancel{d-1}}{4}\right) = \frac{3V_{d-1}}{2(d-1)^{1/4}} \\
&\geq \frac{V_{d-1}}{(d-1)^{1/4}}
\end{aligned}
$$

Therefore, the fraction of the volume of K outside the slab $| x_1 | \leq c/d^{1/4}$ is at most

$$\frac{\frac{3V_{d-1}}{c^3(d-1)^{1/4}} e^{-c^4/4}}{\frac{V_{d-1}}{(d-1)^{1/4}}} = \frac{3}{c^3} e^{-c^4/4}.$$

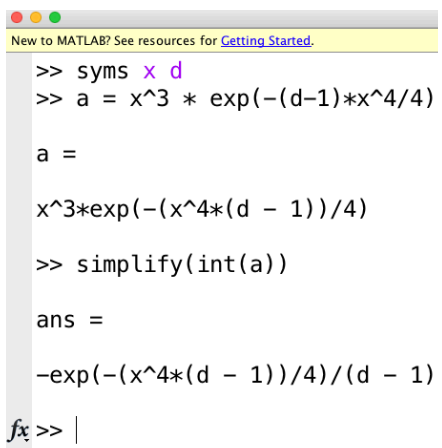

Figure 1: Using MATLAB to compute the integral in the answer to problem 2b.

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 $E\{\mathbf{x}\}$?

   $E\{x_i\} = 0$, therefore $E\{\mathbf{x}\} = \mathbf{0}$.

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

By symmetry across all d-dimensions –

$$E\{\mathbf{x}_i^2\} = \frac{1}{d}\underbrace{E\left\{\sum_{i=1}^{d}\mathbf{x}_i^2\right\}}_{=1} = \frac{1}{d},$$

which means, $var(\mathbf{x}_i) = E\{\mathbf{x}_i^2\} - E\{\mathbf{x}_i\}^2 = \frac{1}{d}$.

(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 $\mathbf{u}^T\mathbf{x}$.

$$\begin{aligned}
var(\mathbf{u}^T\mathbf{x}) &= E\{(\mathbf{u}^T\mathbf{x})^2\} - \underbrace{(E\{\mathbf{u}^T\mathbf{x}\})^2}_{=0} \\
&= E\{(\mathbf{u}^T\mathbf{x})^2\} \\
&= \sum_{i,j} E\{\mathbf{u}_i\mathbf{u}_j\mathbf{x}_i\mathbf{x}_j\} = \frac{1}{d}\overbrace{\sum_i \mathbf{u}_i^2}^{=1} \quad \text{Note: } E\{\mathbf{x}_i\mathbf{x}_j\} = 0, i \neq j,\ E\{\mathbf{x}_i^2\} = 1/d \\
&= \frac{1}{d}.
\end{aligned}$$

The standard deviation of $\mathbf{u}^T\mathbf{x}$ is $1/\sqrt{d}$.

(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.

The ratio of the volume of the above intersection and the volume of each unit ball equals 2 times the fraction of the "northern hemisphere" (see lecture 7) above the plane $x_1 = a/2$, which is at most

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

(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)

If $a = c/\sqrt{d-1}$ and $c \gg 1$, and let $r = 1$ (if not, just scale appropriately), the above fraction is at most

$$\frac{8}{c} e^{-c^2/8}$$

which gets to be very small very quickly. This means we would very very very rarely get a data-point within this intersection. The routine-procedure of computing the distance between a data-point and the two centers will yield a clustering algorithm that will very very very rarely fail.

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 $\mathbf{y}$).

   Since $k < d$, we know that $\mathbf{A}$'s right null-space is non-trivial. That is, we can find two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{R}^d$ such $\mathbf{x} \neq \mathbf{y}$, and $\mathbf{Ax} = \mathbf{Ay} = \mathbf{0}$. That is, the distance between $\mathbf{x}$ and $\mathbf{y}$ is non-zero, but the distance between their images $\mathbf{Ax}$ and $\mathbf{Ay}$ is zero. The distortion will be infinity for this pair.