Homework 2: Review of Linear Algebra, Probability Statistics and Computing

Zhenye Na(zna2)
IE531: Algorithms for Data Analytics

March 2, 2018

1. **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) 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($|x| \geq 1) \geq \frac{1}{k^2}$.
   Solution:
   Computing $E\{X\} = \frac{1}{2k^2} + 0 - \frac{1}{2k^2} = 0$
   Computing *var(X)* $= E\{X^2\} = \frac{1}{k^2}$
   Then we plug them into Chebyshev's Inequality, $P(|x - \mu| \geq a) = \frac{\sigma^2}{a^2}$, let a = 1, then we can find this satisfy Chebyshev's Inequality.

   (b) Show that there can be no continuous distribution over the whole real axis where the Chebyshev Bound is tight.
   Source:
   `https://stats.stackexchange.com/questions/235524/random-variables-for-which-markov-chebyshev-inequalities-are-tight`

   Interpretation: The hypothesis fail because it does not have finite variance. Suppose $P(|X| > x) = \frac{1}{x^2}$. From this distribution we can create a continuous distribution: CDF $= 1 - \frac{1}{x^2}$, and PDF $= \frac{2}{x^3}$ if we take the derivative of CDF. The variance and expectation value should be finite in this case, however, $\frac{1}{x^3}$ leads to an undefined expectation value. So, there can be no continuous distribution over the whole real axis where the Chebyshev Bound is tight.

2. **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) 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?

Solution: $(\frac{1}{2})^{\frac{d}{4}}$ of volume of B(1, d, 4) will S occupy.

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

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

Solution:

Integrate an incremental volume that is a disk of width dx_1 and whose face is a ball of dimension $d-1$ and radius $\sqrt{1-x^4}$ because we use ℓ_4-norm here.
Let K denote the portion of the ball with $x_1 \geq c/d^{1/4}$.
let H denote the upper hemisphere.
Let V_{d-1} denote the volume of the unit ball under ℓ_4 norm in (d-1) dimension.
We first compute upper bound.

$$\begin{aligned}
Vol(K) &= \int_{c/d^{1/4}}^{1} (1-x^4)^{(d-1)/4} V_{d-1} dx \\
&\leq \int_{c/d^{1/4}}^{+\infty} (1-x^4)^{(d-1)/4} V_{d-1} dx \qquad \text{(we use } 1-x \leq e^{-x}\text{)} \\
&\leq V_{d-1} \int_{c/d^{1/4}}^{+\infty} \frac{x^3}{(c/d^{1/4})^3} \exp(-x^4(d-1)/4) dx \\
&= V_{d-1} \cdot \frac{d^{3/4}}{c^3} \cdot \frac{1}{d-1} \cdot \big(-\exp(-x^4(d-1)/4)\big)\Big|_{c/d^{1/4}}^{+\infty} \\
&\leq \frac{V_{d-1}}{c^3(d-1)^{1/4}} \exp(-c^4/4) \qquad \text{(for large d's)}
\end{aligned}$$

Now we compute the lower bound.

$$\begin{aligned}
Vol(H) &= \int_0^1 (1-x^4)^{(d-1)/4} V_{d-1} dx \\
&\geq V_{d-1} \int_0^{1/(d-1)^{1/4}} (1-x^4)^{(d-1)/4} dx \qquad ((1-x)^a \geq 1-ax \text{ for } a \geq 1) \\
&\geq \frac{V_{d-1}}{(d-1)^{1/4}} \left(1 - \frac{1}{d-1}\right)^{(d-1)/4} dx \\
&\geq \frac{V_{d-1}}{(d-1)^{1/4}}
\end{aligned}$$

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

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

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

   (a) What is the value of E$\{\mathbf{x}\}$?
   Solution:
   $\mathbf{E}[x_i] = 0$, therefore $\mathbf{E}[x] = 0$

   (b) What is component-wise variance of x? That is, for i $\in$ {1,2,...,d} what is $E\{(x_i - E\{x_i\})^2\}$?
   Solution:
   By symmetry we have

$$\mathbf{E}[x_i^2] = \frac{1}{d}\mathbf{E}[\sum_{i=1}^{d} x_i^2] = 1/d$$

   Therefore

$$\mathbf{Var}[x_i] = \mathbf{E}[x_i^2] - \mathbf{E}[x_i]^2 = 1/d$$

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

$$\begin{aligned}
\mathbf{Var}[\mathbf{u}^\intercal\mathbf{x}] &= \mathbf{E}[(\mathbf{u}^\intercal\mathbf{x})^2] - \mathbf{E}[\mathbf{u}^\intercal\mathbf{x}]^2 \\
&= \mathbf{E}[(\mathbf{u}^\intercal\mathbf{x})^2] \\
&= \sum_{i,j} \mathbf{E}[u_i u_j x_i x_j] \\
&= \frac{1}{d}\sum_{i} u_i^2 \qquad \text{(since } \mathbf{E}[x_i x_j] = 0 \text{ when } i \neq j \text{ and } \mathbf{E}[x_i^2] = 1/d\text{)} \\
&= \frac{1}{d}
\end{aligned}$$

   So the standard deviation of $\mathbf{u}^\intercal\mathbf{x}$ is $\sqrt{\mathbf{Var}[\mathbf{u}^\intercal\mathbf{x}]} = 1/\sqrt{d}$

   (d) 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.

Solution:

The ratio between the volume of intersection and the volume of each unit ball equals 2 times the fraction of the hemisphere above the plane $x_1 = a/2$ (of a unit ball centered at origin), according to text book *Lemma 2.2*, is at most

$$2 \cdot \frac{2}{\sqrt{d-1} \cdot a/2} \exp\left(- \frac{(a/2)^2(d-1)}{2} \right) = \frac{8}{a\sqrt{d-1}} \exp\left(- \frac{a^2(d-1)}{2} \right)$$

(e) 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)

Solution:

For $a = c/\sqrt{d-1}$ (think of $c \gg 1$ and note that we assume that the radius $r = 1$), the fraction of the intersection is at most $\frac{8}{c} \exp\left(- \frac{c^2}{8} \right)$, which is exponentially small in c.

4. **A Counterpoint to the Johnson-Lindenstrauss Lemma:** Prove that for every fixed dimension reduction matrix 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 $\mathbf{y}$).

Solution:

Since $k < d$, we know that A has a non-trivial null space. Because the rank of $\mathbf{A}$ is at most k, and if first k entries in $\mathbf{x}$ and $\mathbf{y}$ are same, then $A\mathbf{x} = A\mathbf{y}$. That means that there exist two vectors $\mathbf{x} \neq \mathbf{y}$ such that $\mathbf{Ax} = \mathbf{Ay}$. Now we have $||\mathbf{x} - \mathbf{y}|| \neq 0$ and $||\mathbf{Ax} - \mathbf{Ay}|| = 0$, which implies the distance can be hugely distorted.