**CSCI B609 – Foundations of Data Science** **Fall 2017**

# Homework 1: September 11

*Name: YOUR NAME HERE* *Due: September 24, 11:59pm EST*

**Problem 1.1 (Exercise 2.4)** *The Chebyshev inequality states that* $Pr[|X - \mathbb{E}[X]| \geq k\sqrt{Var[X]}] \leq \frac{1}{k^2}$ *for* $k \geq 1$. *Give a probability distribution and a value* $k$ *for which:*

1. *The inequality is tight.*
2. *The inequality is not tight.*

**Problem 1.2 (Exercise 2.21)** *What is the volume of the largest* $d$*-dimensional hypercube that can be placed entirely inside a unit radius* $d$*-dimensional ball? Argue that no larger cube can be placed.*

**Problem 1.3 (Exercise 2.47)** *Let* $x_1, x_2, \ldots, x_n$ *be independent samples of a random variable* $\mathbf{x}$ *with mean* $\mu$ *and variance* $\sigma^2$. *Let* $m_s = \frac{1}{n}\sum_{i=1}^n x_i$ *be the sample mean. Suppose one estimates the variance using the sample mean rather than the true mean, that is,*

$$\sigma_s^2 = \frac{1}{n}\sum_{i=1}^n (x_i - m_s)^2.$$

*Prove that* $\mathbb{E}[\sigma_s^2] = \frac{n-1}{n}\sigma^2$ *and thus one should have divided by* $n-1$ *rather than* $n$.

***Hint****: First calculate the variance of the sample mean and show that* $Var(m_s) = \frac{1}{n}Var(\mathbf{x})$. *Then calculate* $\mathbb{E}[\sigma_s^2] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n (x_i - m_s)^2]$ *by replacing* $x_i - m_s$ *with* $(x_i - \mu) + (\mu - m_s)$.

**Problem 1.4 (Exercise 2.49, Part 1)** *Suppose you want to estimate the unknown center of a Gaussian in* $d$*-dimensional space which has variance one in each direction. Show that* $O(\log d/\epsilon^2)$ *random samples from the Gaussian are sufficient to get an estimate* $\mathbf{m}$ *of the true center* $\mu$, *so that with probability at least* 99/100,

$$\max_i \left[|\mu_i - \boldsymbol{m}_i|\right] \leq \epsilon.$$