Solution Sheet.

Problem 1

We explore some aspects of the curse of dimensionality.

(a) Let the random variables $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ be i.i.d., each following a $\mathcal{U}([0,1]^p)$ distribution, i.e., they are independently and uniformly distributed within the hypercube $[0,1]^p$. The pairwise Euclidean distance between any two points $X^{(i)}$ and $X^{(j)}$ is defined as:

$$d_{ij} = \|X^{(i)} - X^{(j)}\| = \sqrt{\sum_{k=1}^{p} \left(X_k^{(i)} - X_k^{(j)}\right)^2}.$$

Consider the following:

(1) **Expected Squared Distance**: Show that the expected value of the squared distances between two points $X^{(i)}$ and $X^{(j)}$ is given by $\frac{p}{6}$. That is, demonstrate that:

$$\mathbb{E}\left[\|X^{(i)} - X^{(j)}\|^2\right] = \sum_{k=1}^{p} \mathbb{E}\left[\left(X_k^{(i)} - X_k^{(j)}\right)^2\right] = p \cdot \mathbb{E}\left[(U - U')^2\right] = \frac{p}{6}$$

where U and U' are i.i.d. random variables with $U \sim \mathcal{U}([0,1])$ and $U' \sim \mathcal{U}([0,1])$.

(2) **Standard Deviation of Squared Distances**: Show that the standard deviation of the squared distances between points is approximately $0.2\sqrt{p}$. That is, demonstrate that:

$$\mathrm{Std}\left(\|X^{(i)} - X^{(j)}\|^2\right) = \sqrt{\sum_{k=1}^{p} \mathrm{Var}\left[\left(X_k^{(i)} - X_k^{(j)}\right)^2\right]} \approx 0.2\sqrt{p}$$

(3) **Behavior in High Dimensions**: Use parts (i) and (ii) to discuss the distribution of pairwise distances between points as the dimensionality p becomes very large. Show that distances become increasingly concentrated as $p \to \infty$.

(4) **Implications for Machine Learning**: Explain why this concentration of distances can be problematic in machine learning tasks that rely on distance metrics, such as clustering or nearest-neighbor classification. Discuss how the concept of "nearness" becomes less meaningful in high-dimensional spaces.

(b) Discuss two specific ways in which the curse of dimensionality affects data analysis and machine learning, providing examples of each.

Solution.

(a)(1): Expected Value of the Squared Distance:
The squared distance between two points $X^{(i)}$ and $X^{(j)}$ is given by:

$$\|X^{(i)} - X^{(j)}\|^2 = \sum_{k=1}^{p} \left(X_k^{(i)} - X_k^{(j)}\right)^2$$

where $X_k^{(i)}$ is the k-th coordinate of the point $X^{(i)}$. Since the points are uniformly distributed in $[0,1]^p$, each coordinate $X_k^{(i)}$ is uniformly distributed in $[0,1]$. For two independent uniformly distributed variables U and U' in $[0,1]$, we have:

$$\mathbb{E}\left[(U - U')^2\right] = \mathbb{E}\left[U^2\right] - 2\mathbb{E}[UU'] + \mathbb{E}\left[U'^2\right]$$

Since U and U' are independent and uniformly distributed, we have:

$$\mathbb{E}[U] = \mathbb{E}[U'] = \frac{1}{2} \quad \text{and} \quad \mathbb{E}\left[U^2\right] = \mathbb{E}\left[U'^2\right] = \int_0^1 u^2\,du = \frac{1}{3}$$

$$\mathbb{E}[UU'] = \mathbb{E}[U]\mathbb{E}[U'] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Substituting these values:

$$\mathbb{E}\left[(U-U')^2\right] = \frac{1}{3} - 2 \cdot \frac{1}{4} + \frac{1}{3} = \frac{1}{6}$$

Since the squared distance is the sum of p independent terms, we can write:

$$\mathbb{E}\left[\|X^{(i)} - X^{(j)}\|^2\right] = \sum_{k=1}^{p} \mathbb{E}\left[\left(X_k^{(i)} - X_k^{(j)}\right)^2\right] = p \cdot \frac{1}{6} = \frac{p}{6}$$

(a)(2): Standard Deviation of the Squared Distance
First, we calculate the variance of a single term:
For independent uniformly distributed variables U and U', the variance of $(U-U')^2$ is calculated using:

$$\mathrm{Var}\left((U-U')^2\right) = \mathbb{E}\left[(U-U')^4\right] - \left(\mathbb{E}\left[(U-U')^2\right]\right)^2$$

We already have:

$$\mathbb{E}\left[(U-U')^2\right] = \frac{1}{6}$$

Let's calculate $\mathbb{E}\left[(U-U')^4\right]$:

$$\mathbb{E}\left[(U-U')^4\right] = \int_0^1 \int_0^1 (u-u')^4\,du\,du'$$

Expanding $(u-u')^4$:

$$(u-u')^4 = u^4 - 4u^3u' + 6u^2u'^2 - 4uu'^3 + u'^4$$

Integrating each term over $[0,1]^2$:

$$\int_0^1 \int_0^1 u^4\,du\,du' = \frac{1}{5}, \quad \int_0^1 \int_0^1 u^3u'\,du\,du' = \frac{1}{8}, \quad \int_0^1 \int_0^1 u^2u'^2\,du\,du' = \frac{1}{9}$$

$$\int_0^1 \int_0^1 uu'^3\,du\,du' = \frac{1}{8}, \quad \int_0^1 \int_0^1 u'^4\,du\,du' = \frac{1}{5}$$

Combining these:

$$\mathbb{E}\left[(U-U')^4\right] = \frac{6}{90}$$

Now, calculate the variance:

$$\mathrm{Var}\left((U-U')^2\right) = \frac{6}{90} - \left(\frac{1}{6}\right)^2 = \frac{7}{180}$$

The squared distance is a sum of p independent terms, so the variance of the sum is:

$$\mathrm{Var}\left(\|X^{(i)} - X^{(j)}\|^2\right) = p \cdot \frac{7}{180}$$

The standard deviation is the square root of the variance:

$$\mathrm{Std}\left(\|X^{(i)} - X^{(j)}\|^2\right) = \sqrt{p \cdot \frac{7}{180}} = \approx 0.2\sqrt{p}$$

(a)(3): As the dimensionality p becomes large, the expected squared distance between points, $\mathbb{E}[\|X^{(i)} - X^{(j)}\|^2] = \frac{p}{6}$, grows linearly with p. Meanwhile, the standard deviation of squared distances, $\text{Std}[\|X^{(i)} - X^{(j)}\|^2] \approx 0.2\sqrt{p}$, increases more slowly than the expected value. As a result, the ratio of the standard deviation to the mean distance, $\frac{0.2\sqrt{p}}{p/6} = \frac{1.2}{\sqrt{p}}$, approaches zero as $p \to \infty$. This implies that the distribution of pairwise distances becomes sharply concentrated around the mean distance as p grows, making most distances nearly identical.

(a)(4): In high-dimensional spaces, the concentration of distances means that most points are nearly equidistant from each other. This reduces the effectiveness of distance-based metrics, as algorithms like clustering or nearest-neighbor classification rely on clear distinctions between "near" and "far" points. When distances are almost uniform, it becomes difficult to identify meaningful clusters or neighbors, as the concept of "nearness" loses its relevance, weakening the algorithm's ability to differentiate and group similar data points effectively.

b: See slides of Lecture 3.

Problem 2

We explore some elementary aspects of concentration.

(a) Prove that if Z is a non-negative random variable with expectation $\mathrm{E}[Z]$, then for all $t > 0$, we have $\mathrm{P}[Z \geq t] \leq \mathrm{E}[Z]/t$.

(b) A zero-mean random variable is said to be sub-Gaussian with parameter $\sigma > 0$ if $\mathrm{E}[\exp(tX)] \leq \exp(\sigma^2 t^2/2)$ for all $t \in \mathbb{R}$. Show that $X \sim \mathcal{N}(0, \sigma^2)$ is sub-Gaussian.

(c) Suppose that X is Bernoulli with $\mathrm{P}[X = +1] = \mathrm{P}[X = -1] = 1/2$. Show that X is sub-Gaussian.

(d) Show that any sub-Gaussian random variable X satisfies the two-sided tail bound

$$\mathrm{P}[|X| > t] \leq 2\exp(-\frac{t^2}{2\sigma^2})$$

Solution.

(a) We can calculate that for arbitrary non-negative RVs Z and positive numbers t

$$P[Z \geq t] = E[\mathbb{1}_{\{Z \geq t\}}] = E[\mathbb{1}_{\{\frac{Z}{t} \geq 1\}}] \leq E[\frac{Z}{t}\mathbb{1}_{\{\frac{Z}{t} \geq 1\}}] \leq E[\frac{Z}{t}] = \frac{1}{t}E[Z].$$

(b) We can calculate that for arbitrary positive numbers t it holds that

$$\begin{aligned} E[e^{tX}] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{tx} e^{\frac{-x^2}{2\sigma^2}} dx \\ &\overset{*}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-(x-t\sigma^2)^2}{2\sigma^2} + \frac{t^2\sigma^2}{2}} dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-(x-t\sigma^2)^2}{2\sigma^2}} dx \cdot e^{\frac{t^2\sigma^2}{2}} \overset{**}{=} e^{\frac{t^2\sigma^2}{2}}, \end{aligned}$$

where we used for

$$*) : \frac{-(x-t\sigma^2)^2}{2\sigma^2} + \frac{t^2\sigma^2}{2} = -\frac{x^2 - 2xt\sigma^2 + \sigma^4 t^2}{2\sigma^2} + \frac{t^2\sigma^2}{2} = -\frac{x^2 - 2xt\sigma^2}{2\sigma^2} = -\frac{x^2}{2\sigma^2} + xt.$$

and furthermore we used for $**)$ that $x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-t\sigma^2)^2}{2\sigma^2}}$ is the density function of a $\mathcal{N}(t\sigma^2, \sigma^2)$ distributed random variable - integrating over a density function results in 1 in any case.

(c) We calculate that

$$\begin{aligned}
E[e^{tX}] &= \sum_{k\in\{-1,1\}} P[X=k]e^{tk} \\
&= \frac{1}{2}e^{-t} + \frac{1}{2}e^{t} \\
&= \frac{1}{2}\sum_{j=0}^{\infty}\frac{(-t)^j}{j!} + \frac{1}{2}\sum_{j=0}^{\infty}\frac{t^j}{j!} \\
&= \frac{1}{2}\sum_{j=0}^{\infty}\frac{1}{j!}((-t)^j + t^j) \\
&\overset{*}{=} \frac{1}{2}\sum_{j=0}^{\infty}\frac{1}{(2j)!}2t^{2j} \\
&= \sum_{j=0}^{\infty}\frac{t^{2j}}{(2j)!} \\
&\overset{**}{\leq} \sum_{j=0}^{\infty}\frac{(\frac{t^2}{2})^j}{j!} = e^{\frac{t^2}{2}},
\end{aligned}$$

where we used for $*$) that

$$(-t)^j + t^j = 0 \text{ for } j \text{ odd, and } (-t)^j + t^j = 2t^j \text{ for } j \text{ even.}$$

We then shift the summation index, such that only terms with even j are summed up. For $**$) we used that

$$\frac{t^{2j}}{(2j)!} = (\frac{t^2}{2})^j\frac{2^j}{(2j)!}, \text{ and obviously } \frac{2^j}{(2j)!} \leq \frac{1}{j!}.$$

We can conclude that the RV is sub-gaussian with parameter $\sigma = 1$.

(d) We calculate that

$$P[|X| > t] \leq P[\max\{-X, X\} > t] \leq P[-X > t] + P[X > t] = P[X < -t] + P[X > t].$$

For all positive λ, we get

$$\begin{aligned}
P[X > t] = P[e^{\lambda X} > e^{\lambda t}] &\overset{\text{Markov}}{\leq} \frac{1}{e^{\lambda t}}E[e^{\lambda X}] \\
&\overset{\text{exploit sub-gauss.}}{\leq} e^{\frac{\sigma^2\lambda^2}{2}}e^{-\lambda t} = e^{\frac{\sigma^2\lambda^2}{2}-\lambda t} = e^{-(\lambda t - \frac{\sigma^2\lambda^2}{2})}.
\end{aligned}$$

The term on the right side of the equation chain is minimized (over λ), for the maximum of the quadratic term $q(\lambda) := \lambda t - \frac{\sigma^2\lambda^2}{2}$. We simply calculate the derivative of the quadratic term

$$\frac{dq}{d\lambda}(\lambda) = t - \sigma^2\lambda,$$

and this gets zero for $\lambda = \frac{t}{\sigma^2}$. Hence, we can derive that

$$P[X > t] \leq e^{-(\frac{t}{\sigma^2}t - \sigma^2\frac{t^2}{2\sigma^4})} = e^{\frac{-t^2}{\sigma^2} + \frac{t^2}{2\sigma^2}} = e^{-\frac{t^2}{2\sigma^2}}.$$

We can conduct an analog calculation for $P[X < -t]$, by just taking negative (instead of positive) values for λ into account, and replacing t by $-t$. The result is the same bound, and the final result is proven by taking two times this bound to get our desired tail inequality.

Problem 3

Poisson distribution: Let $X \sim Pois(\lambda)$, that is,

$$\mathrm{P}[X = k] = \frac{\lambda^k \exp(-\lambda)}{k!}.$$

The following limit theorem holds:

Theorem 1.3.4 (Poisson Limit Theorem). *Let $X_{N,i}$, $1 \leq i \leq N$, be independent random variables $X_{N,i} \sim \mathrm{Ber}(p_{N,i})$, and let $S_N = \sum_{i=1}^N X_{N,i}$. Assume that, as $N \to \infty$,*

$$\max_{i \leq N} p_{N,i} \to 0 \quad \textit{and} \quad \mathbb{E}\, S_N = \sum_{i=1}^N p_{N,i} \to \lambda < \infty.$$

Then, as $N \to \infty$,

$$S_N \to \mathrm{Pois}(\lambda) \quad \textit{in distribution.}$$

(a) Show that for any $t > \lambda$,

$$\mathrm{P}[X > t] \leq \exp(-\lambda) \left(\frac{\exp(1)\lambda}{t} \right)^t$$

(b) Show that for $t \in (0, \lambda]$

$$\mathrm{P}[|X - \lambda| > t] \leq 2 \exp\left(-c \frac{t^2}{\lambda} \right)$$

For (b) use the following Chernoff-type inequality for small deviations: For $\delta \in (0, 1]$, we have $\mathrm{P}(|S_N - \mu| \geq \delta\mu) \leq 2\exp(-c\mu\delta^2)$, where $c > 0$ is some constant.

Solution.

(a) We can calculate that

$$\begin{aligned} P[X > t] &= P[X > t] - P[S_n > t] + P[S_n > t] \\ &\leq |P[X > t] - P[S_n > t]| + P[S_n > t] = \\ &\leq |P[X > t] - P[S_n > t]| + e^{-\mu_n} (\frac{e^1 \mu_n}{t})^t \text{ with } \mu_n = E[S_n] \\ &\overset{n\to\infty}{\to} 0 + e^{-\lambda} (e \frac{\lambda}{t})^t. \end{aligned}$$

(b) We can calculate that

$$\begin{aligned} P[|X - \lambda| > t] &\leq |P[|X - \lambda| > t] - P[|S_n - \lambda| > t]| + P[|S_n - \lambda| > t], \text{ and} \\ P[|S_n - \lambda| > t] &= P[|S_n - \lambda| > \frac{t}{\lambda}\lambda] \\ &\overset{\text{Chernoff}}{\leq} 2e^{-c\lambda(\frac{t}{\lambda})^2} \text{ for some constant } c > 0, \\ &= 2e^{-c\frac{t^2}{\lambda}}. \end{aligned}$$

Since $|P[|X - \lambda| > t] - P[|S_n - \lambda| > t]| \overset{n\to\infty}{\to} 0$, the desired result follows.

Problem 4

Let $Z \sim \mathcal{N}(0,1)$ and use the identity $\int_0^\infty (1-3x^{-4})e^{-\frac{x^2}{2}}\,dx = \left(\frac{1}{t} - \frac{1}{t^3}\right)e^{-\frac{t^2}{2}}$.

(a) (Tails of the normal distribution). Then for all $t > 0$, we have

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}} \le \mathrm{P}(Z \ge t) \le \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}.$$

In particular, for $t \ge 1$ the tail is bounded by the density:

$$\mathrm{P}(Z \ge t) \le \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}.$$

(b) Show that for all $t \ge 0$

$$\mathrm{P}(|Z| \ge t) \le 2e^{-\frac{t^2}{2}}.$$

(c) (Truncated normal distribution) Show that for all $t \ge 1$,

$$\mathrm{E}[Z^2 1_{Z>t}] = t\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}} + \mathrm{P}(Z \ge t) \le \left(t + \frac{1}{t}\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}.$$

Solution.

(a) We show first that $P[Z \ge t] \le \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\frac{1}{t}$. To see this, we calculate

$$\begin{aligned} P[Z \ge t] &= \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{\frac{-x^2}{2}}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_0^\infty e^{\frac{-(x+t)^2}{2}}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_0^\infty e^{\frac{-x^2}{2}-xt}dx \cdot e^{\frac{-t^2}{2}} \\ &\overset{e^{-\frac{x^2}{2}}\le 1}{\le} \frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-xt}dx \cdot e^{\frac{-t^2}{2}} \\ &\overset{\int_0^\infty e^{-xt}dx = \frac{1}{t}}{=} \frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}\frac{1}{t}. \end{aligned}$$

We show further, that $(\frac{1}{t} - \frac{1}{t^3})\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}} \le P[Z \ge t]$, by calculating

$$\begin{aligned} P[Z \ge t] &= \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{\frac{-x^2}{2}}dx \\ &\ge \frac{1}{\sqrt{2\pi}}\int_t^\infty (1-3x^-4)e^{\frac{-x^2}{2}}dx \\ &= \frac{1}{\sqrt{2\pi}}(\frac{1}{t} - \frac{1}{t^3})e^{-\frac{t^2}{2}}. \end{aligned}$$

(b) We calculate that

$$\begin{aligned}
P[Z \geq t] &= \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{\frac{-x^2}{2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{\frac{-(x+t)^2}{2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{\frac{-x^2}{2} - xt} dx \cdot e^{\frac{-t^2}{2}} \\
&\overset{xt \geq 0}{\leq} \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{\frac{-x^2}{2}} dx \cdot e^{\frac{-t^2}{2}} \\
&= P[Z > 0] \cdot e^{\frac{-t^2}{2}} = \frac{1}{2} \cdot e^{\frac{-t^2}{2}},
\end{aligned}$$

and the calculation for $P[Z < -t]$ is analogue. Together, they prove the desired result.

(c) We calculate that

$$\begin{aligned}
E[Z^2 \mathbb{1}_{\{Z>t\}}] &= \int_{-\infty}^\infty x^2 \mathbb{1}_{\{x>t\}} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx \\
&= \int_t^\infty (x^2 - 1) \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx + P[Z > t] \\
&= \frac{1}{\sqrt{2\pi}} t e^{\frac{-t^2}{2}} + P[Z \geq t] \\
&\leq \frac{1}{\sqrt{2\pi}} t e^{\frac{-t^2}{2}} + \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}.
\end{aligned}$$

Problem 5

Consider 100 independent coin flips. We wish to find an upper bound on the probability that the number of heads is greater or equal to 75. Use Markov's, Chebyshev's, and Chernoff's inequality and compare.

Solution.

We model the problem by a sum of 100 independent, Bernoulli-distributed RVs, i.e., $X = \sum_{i=1}^{100} X_i$, with $X_i \in \{0,1\}, P[X_i = 0] = 0.5, E[X] = E[\sum_{i=1}^{100} X_i] = \sum_{i=1}^{100} P[X_i = 1] = 0.5 \cdot 100 = 50$.

a) Markov. We calculate that

$$P[X > 75] \leq \frac{E[X]}{75} = \frac{2}{3}.$$

b) Chebyshev. We calculate that

$$P[X > 75] = P[X - 50 > 25] \leq \frac{Var[X]}{25^2} = \frac{25}{25^2} = \frac{1}{25}.$$

c) Chernoff. We calculate that

$$\begin{aligned}
P[X > 75] &\leq e^{-50} (\frac{50e}{75})^{75} \\
&= e^{25} \cdot (\frac{2}{3})^{75} \overset{\text{approximately}}{=} 0.0045.
\end{aligned}$$

Remark: The true probability is approximately 10^{-7}. We still overestimate it by a factor of 40000, even with the Chernoff bound.

PROBLEM 6

Coding problem

(a) Let X be a Poisson random variable with $\lambda = 0.7$. Visually compare the Markov bound, Chernoff bound, and the theoretical probabilities for $x = 1, \dots, 12$. Use Problem 3 (a).

Solution.

We write a program that evaluates all the bounds for $t = 1, ..., 12$.

1. Markov. For this, we simply calculate
$$P[X > t] \leq \frac{E[X]}{t} = \frac{\lambda}{t}.$$
2. Chernoff. For this, we calculate
$$P[X > t] \leq e^{-\lambda}(\frac{e\lambda}{t})^t.$$
3. Theoretical Probabilities. For this, we calculate
$$P[X > t] = 1 - \sum_{k=0}^{t} e^{-\lambda}\frac{\lambda^k}{k!}.$$

The results can be obtained executing the provided python script ex_2_5.py. The Markov bounds are quite loose, while the Chernoff bounds are tighter.