

# RECITATION 1

## DECISION TREES, kNNs

10-701: INTRODUCTION TO MACHINE LEARNING

September 4, 2025

## 1 kNNs in Higher Dimensions

### 1.1 Distance between Points

In this problem we investigate the behavior of the kNN algorithm as the dimension of the datapoints increases. We use the standard Euclidian notion of distance, that is:

$$d_q(X, Y) = \sqrt{\sum_{i=1}^q (X_i - Y_i)^2}$$

Let  $X$  and  $Y$  be two independent samples drawn uniformly from the  $q$ -dimensional unit hypercube. That is, for each dimension  $i$ ,  $X_i, Y_i \sim U[0, 1]$ , and each dimension for a given point is independent of its other dimensions. Show that the expected value of  $d_q(X, Y)$  (the expected distance between the random variables  $X$  and  $Y$ ) approaches infinity as the number of dimensions  $q$  approaches infinity.

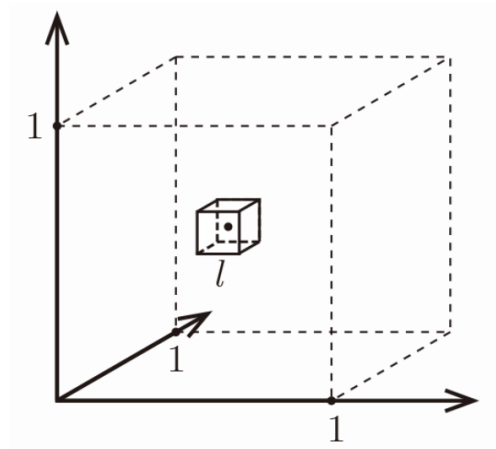
**HINT 1:** Recall the Strong Law of Large Numbers, which tells us that the average of i.i.d random variables converges almost surely to its expectation.

**HINT 2:**  $\int_{x=0}^1 \int_{y=0}^1 (x - y)^2 dx dy = \frac{1}{6}$

## 1.2 The Curse of Dimensionality

Now, consider  $n$  points uniformly spaced throughout a  $q$ -dimensional unit hypercube. Unlike in the previous question, these points are not sampled randomly but rather spaced as a uniform  $q$ -dimensional grid.

For any single given reference point within the unit hypercube, we want to analyze the amount of volume that its  $k$  nearest neighbors occupy as the dimensionality  $q$  increases.



1. Assume some fixed  $k$ ,  $n$  and consider the smallest hypercube of side length  $l$  centered at the test point that contains these neighbors on average. Give an expression of how  $l$  varies with  $q$ . And also compute the limit of the expression as  $q$  approaches infinity.

*Hint: If data points are uniformly distributed in the unit hypercube, what fraction of the total volume must be searched to find  $k$  out of  $n$  total points?*

2. In the previous question, we assumed  $n$  to be fixed. Now, let's flip the question and say that we fix  $l$  to be a small number less than 1. Determine the number of samples required to ensure that a small hypercube with side length  $l$  centered at one point contains the  $k$  closest neighbors to the point. Express your answer as a function of the dimension  $q$ .

## 2 Decision Trees

### 2.1 Entropy

Recall from lecture that the entropy of a distribution for a discrete random variable  $X$  is defined as

$$H(X) = \sum_c -P(X = c) \log_2 P(X = c)$$

Let us define a random variable  $X \in \{0, K - 1\}$ . Define  $\alpha_k = P(X = k)$ . Show that the uniform distribution maximizes the entropy, that is:

$$\max_P H(X) = P(X = k) = \begin{cases} \frac{1}{K} & \text{if } k \in \{0, K - 1\} \\ 0 & \text{else} \end{cases}$$

**HINT:** You will want to introduce a Lagrange multiplier constraint to enforce that the probability terms sum to 1. This constraint is of the form  $\sum_k \alpha_k = 1$ .

**HINT for an alternate approach:** U could use Jensen's inequality to bound the entropy.

## 2.2 KL Divergence

A concept that may be useful for your homework is the KL divergence, a measure of the difference between two probability distributions. For discrete distributions  $p, q$  with support  $\{1, \dots, n\}$ , the KL divergence is:

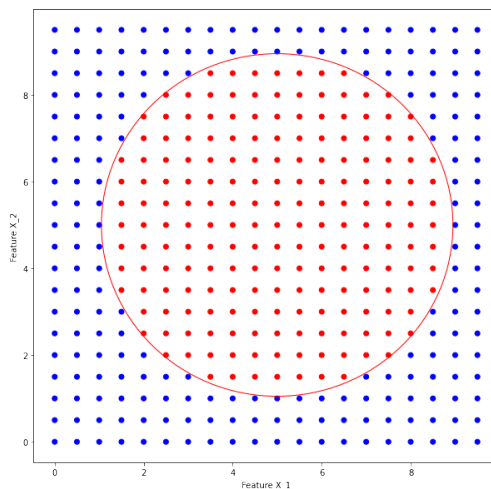
$$D(p||q) = \sum_{i=1}^n p(i) \log \frac{p(i)}{q(i)}$$

Show then, that  $D(p||q) \geq 0$  for all  $p, q$ .

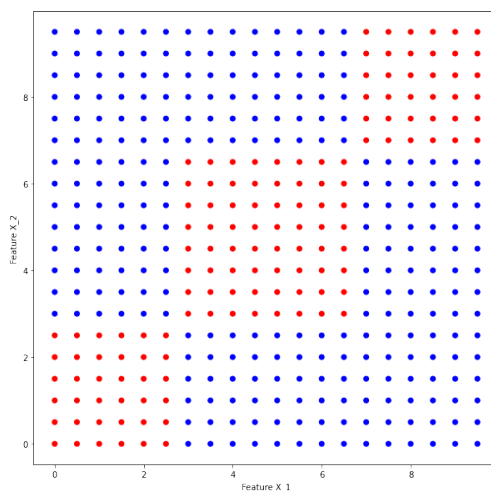
**Hint:** You may use this inequality without proof:  $x - 1 \geq \log(x)$ .

## 2.3 Decision Tree Decision Boundaries

1. Can a learned, finite decision tree using the ID3 algorithm perfectly classify all of the points in the figure below? Can it perfectly learn the decision boundary (portrayed as a circle in red)?



2. Consider the dataset below with 400 total points consisting of three clusters of red points with 36, 64, and 36 points going from bottom left to top right. The mutual information represents how much we gain, in terms of reducing the entropy, from knowing something about an attribute:  $I(Y; X) = H(Y) - H(Y|X)$ , where  $H(Y|X) = \sum_k p(X = k)H(Y|X = k)$  represents the conditional entropy. Starting with the base predictor, what is the information gain (or mutual information) of the split  $X_1 < 3$ ?



3. What is the mutual information of the split  $X_2 < 3$  (we are performing this split AFTER the previous split, i.e. on the points with  $X_1 < 3$ ).

4. Draw out a decision tree that could perfectly classify the points in the figure from Question 2. Why can we exactly learn the decision boundary in this case?