Recitation 1
Decision Trees, kNNs

10-701: Introduction to Machine Learning

9/8/2023

1 Decision Trees

1.1 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 $\sum_k \alpha_k = 1$.

We start by writing the entropy, which is the function that we want to maximize, along with the Lagrange multiplier and constraint:

$$f(x) = \sum_k -\alpha_k \log_2 \alpha_k + \lambda \left(\sum_k \alpha_k - 1 \right)$$

We take the derivative with respect to a specific α_k:

$$\begin{aligned} \frac{d}{d\alpha_k} f(x) &= \frac{d}{d\alpha_k} \left[\sum_k -\alpha_k \log_2 \alpha_k + \lambda \left(\sum_k \alpha_k - 1 \right) \right] \\ &= -\log \alpha_k - 1 + \lambda \end{aligned}$$

Setting to 0 and solve for α_k:

$$\begin{aligned} -\log \alpha_k - 1 + \lambda &= 0 \\ -1 + \lambda &= \log \alpha_k \\ \alpha_k^* &= e^{\lambda - 1} \end{aligned}$$

Note that the right hand side does not depend on k, so $\alpha_k = e^{\lambda - 1}, \forall k$. This means that all α_k are equal to each other, and with the additional constraint that they must sum to 1, we conclude that $\alpha_k = \frac{1}{K}, \forall k$.

1.2 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 $\{1, \ldots, n\}$, 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: $x - 1 \geq \log(x)$.

Proving that $D(p||q) \geq 0$ is equivalent to proving $\sum_{i=1}^{n} p(i) \log \frac{q(i)}{p(i)} \leq 0$. Using the given inequality:

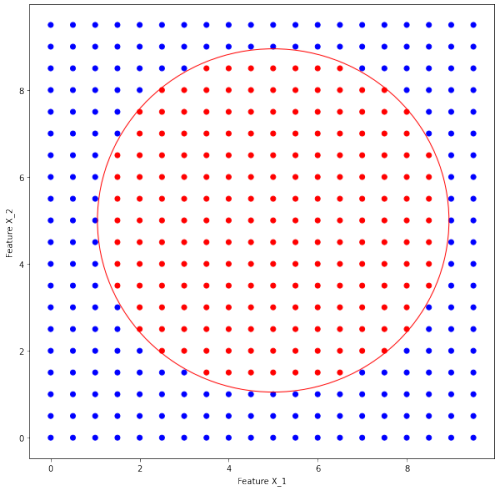
$$\sum_{i=1}^{n} p(i) \log \frac{q(i)}{p(i)} \leq \sum_{i=1}^{n} p(i) \left(\frac{q(i)}{p(i)} - 1 \right) = \sum_{i=1}^{n} q(i) - p(i)$$

Since $p(i)$ and $q(i)$ are normalized, the sum over all events for both is 1. As such:

$$\sum_{i=1}^{n} p(i) \log \frac{q(i)}{p(i)} \leq \sum_{i=1}^{n} q(i) - \sum_{i=1}^{n} p(i) = 0$$

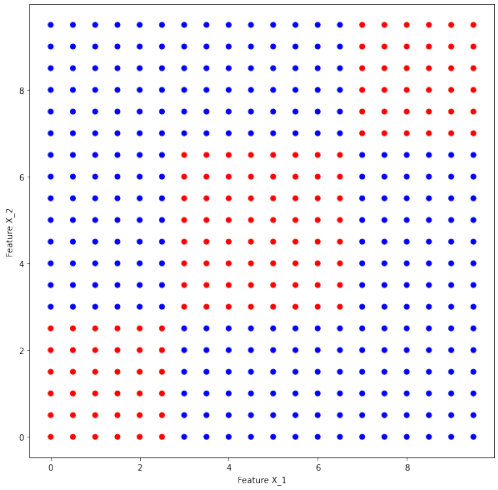
1.3 Decision Tree Decision Boundaries

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



A decision tree can perfectly classify the points above, but the depth of the tree would be huge. This is because the decision boundary of a decision tree must be axis-aligned. This means that we cannot perfectly learn a decision boundary that is described by a circle.

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 $X_1 < 3$?



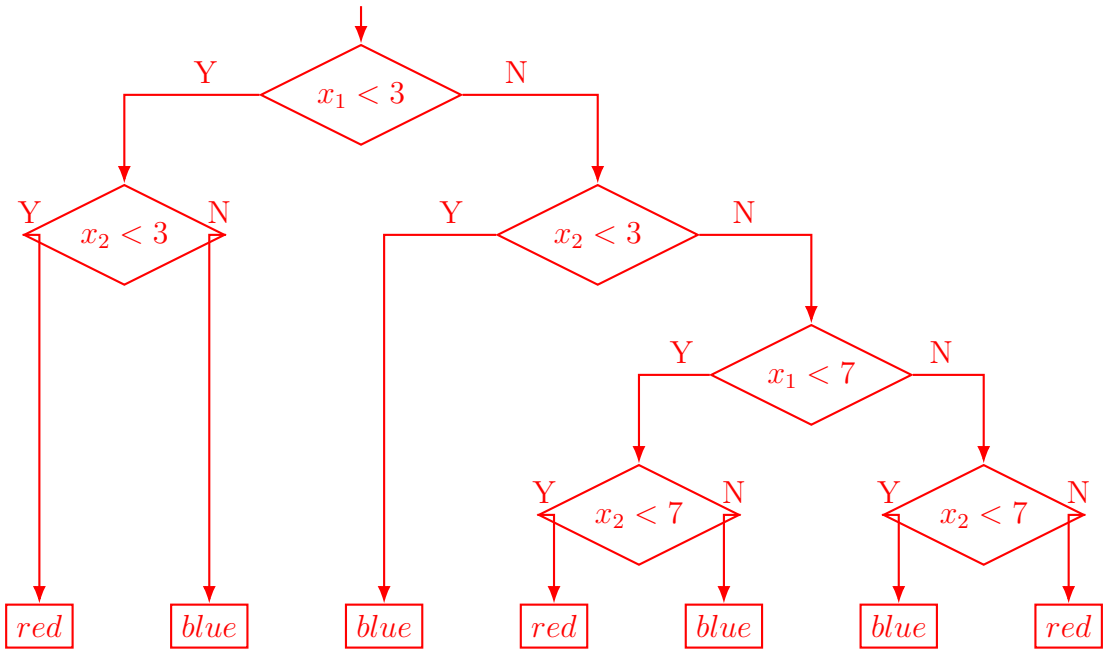
There are 136 red points and $400 - 136 = 264$ blue points. If we let Y denote the observed label distribution, then $H(Y) = -\frac{136}{400}\log_2\frac{136}{400} - \frac{264}{400}\log_2\frac{264}{400} \approx .925$. For points such that $X_1 < 3$, there are 36 red points and 84 blue points for a total of

120 points. So $H(Y|X_1 < 3) = -\frac{36}{120}\log_2\frac{36}{120} - \frac{84}{120}\log_2\frac{84}{120} \approx .881$. For points with $X_1 >= 3$, there are 100 red points and 180 blue points for a total of 280 points. So $H(Y|X_1 \geq 3) = -\frac{100}{280}\log_2\frac{100}{280} - \frac{180}{280}\log_2\frac{180}{280} \approx .94$. The total conditional entropy is $H(Y|X_1) = \frac{120}{400}(.881) + \frac{280}{400}(.94) = .9223$, so the mutual information is $I(Y; X_1) = H(Y) - H(Y|X_1) = .0027$.

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 $X_1 < 3$).

   There are 120 points in this split and from above we know the entropy is .881. Note that the additional split of $X_2 < 3$ perfectly classifies the data in this region. Letting Y' denote the distribution of points with $X_1 < 3$, we have that the information gain is $IG(Y'|X_2) = H(Y') - H(Y'|X_2) = .881 - \frac{36}{120}(0) + \frac{84}{120}(0) = .881$.

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?



We can exactly learn the decision boundary because it is axis-aligned.

2 kNNs in Higher Dimensions

2.1 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: $\int_{x=0}^{1}\int_{y=0}^{1}(x-y)^2 dxdy = \frac{1}{6}$

Let us denote the two points as X and Y, where $X, Y \sim U[0,1]^q$. The distance between them is $d_q(X,Y) = \sqrt{\sum_{i=1}^{q}(X_i - Y_i)^2}$. Squaring both sides and dividing by the number of dimensions, we get:

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

Because we sampled indepdently from the unit hypercube, each dimension is independent of all other dimensions, so by the SLLN the right side converges almost surely to its expectation:

$$\lim_{q\to\infty}\frac{1}{q}\sum_{i=1}^{q}(X_i - Y_i)^2 \xrightarrow{a.s.} \mathbb{E}[(X_i - Y_i)^2] = \int_{x=0}^{1}\int_{y=0}^{1}(x-y)^2 dxdy = \frac{1}{6}$$

Therefore,

$$\lim_{q\to\infty}\frac{d_q(X,Y)^2}{q} = \lim_{q\to\infty}\frac{1}{q}\sum_{i=1}^{q}(X_i - Y_i)^2 = \frac{1}{6}$$

$$\lim_{q\to\infty} d_q(X,Y) = \sqrt{\frac{q}{6}}$$

Counterintuitively, we see that the expected distance between two uniformly sampled points goes to infinity as the number of dimensions increases.

2.2 Sample Complexity

In the previous section we showed that the expected squared distance for two points sampled from a hypercube grows infinitely large with the dimension q. However, for kNN algorithms, we are interested in the distance not to any randomly sampled point, but to the nearest neighbors of a given point.

Consider a point x sampled from one of the edges of a unit hypercube as defined before. Suppose we sample n points $\{x_1, \dots x_n\}$ with replacement from the unit cube. Then, how many points would we need to sample to ensure the probability that the distance from x to its nearest neighbor x_i is at least $\sqrt{d}$ is less than some fixed $\delta > 0$?

HINT: Use the fact that $1 - x < e^{-x}$

Note that the probability that $\min_i \|x - x_i\|$ can be expressed as:

$$\mathbb{P}[\min_i \|x - x_i\| \geq \sqrt{d}] = \prod_{i=1}^{n} \mathbb{P}[\|x - x_i\| \geq \sqrt{d}] = \left(\frac{\sum_{k=d}^{q} \binom{q}{k}}{2^q} \right)^n$$
$$= \left(1 - \frac{\sum_{k=0}^{d-1} \binom{q}{k}}{2^q} \right)^n$$

Using the hint and letting $\sum_{k=0}^{d-1} \binom{q}{k} = s_{d-1}$, we have:

$$\mathbb{P}[\min_i \|x - x_i\| \geq \sqrt{d}] \leq e^{-s_{d-1} n / 2^q}$$

To ensure that this quantity is less than δ, we have:

$$e^{-s_{d-1} n / 2^q} < \delta$$
$$n > \frac{2^q}{s_{d-1}} \log \frac{1}{\delta}$$

Note that $s_{d-1} = O(q^{d-1})$ for d small. As such, for small d, we need a huge number of samples to ensure that the distance to the nearest neighbor of x is less than $\sqrt{d}$