Topics: KNN, Kd-tree, L1 loss, decision tree, k-means clustering, mean-shift

Q1

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
Dist	0.361	0.316	0.141	0.283	0.361	0.849	0.224	0.224	0.567	0.224

KNN-1 Use p_3 Class 1

KNN-3 Pick p_3,(p_7/p_8),p_{10} Class 1, Pick p_3,p_7,p_8 Class 3

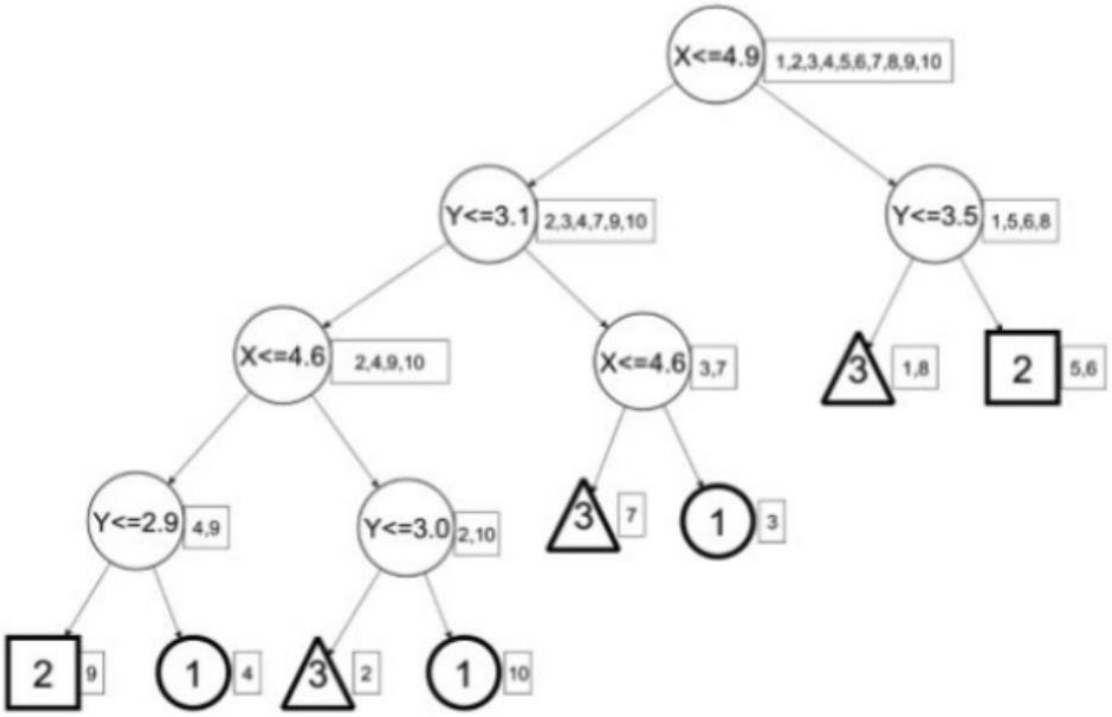
KNN-4 Pick p_3,p_7,p_8,p_{10} Indeterminate (Class 1 or Class 3)

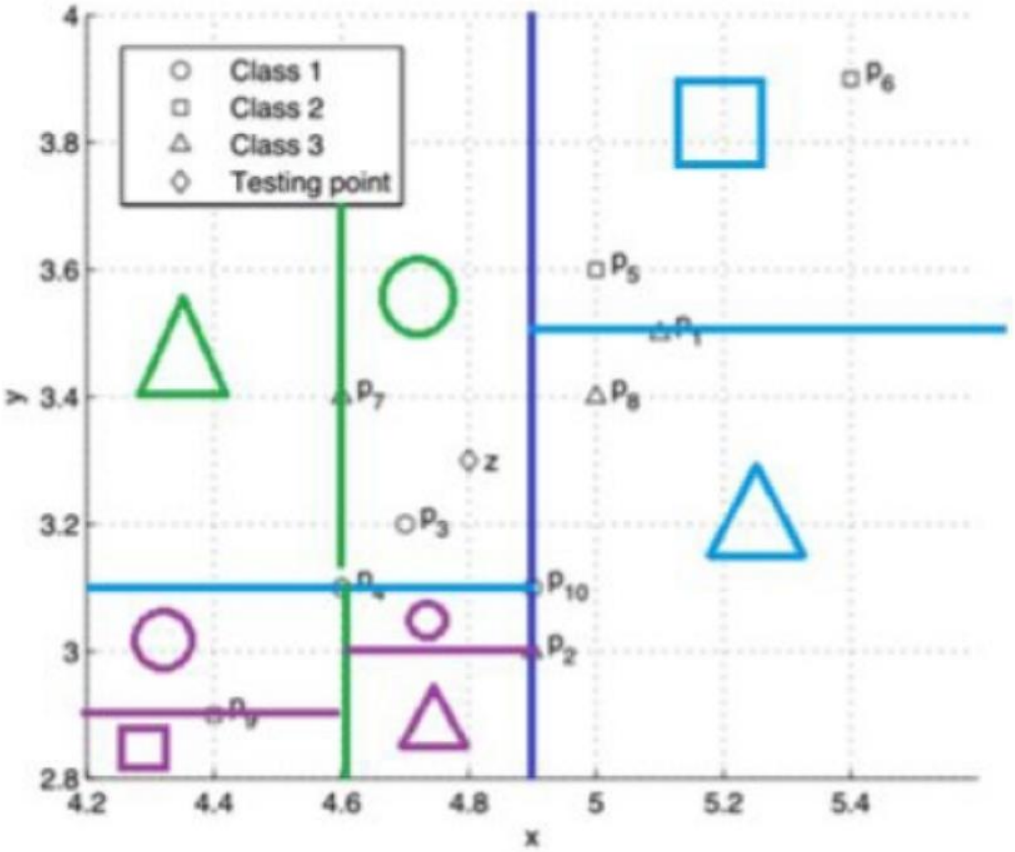
KNN-5 Pick p_3,p_7,p_8,p_{10},p_4 Class 1

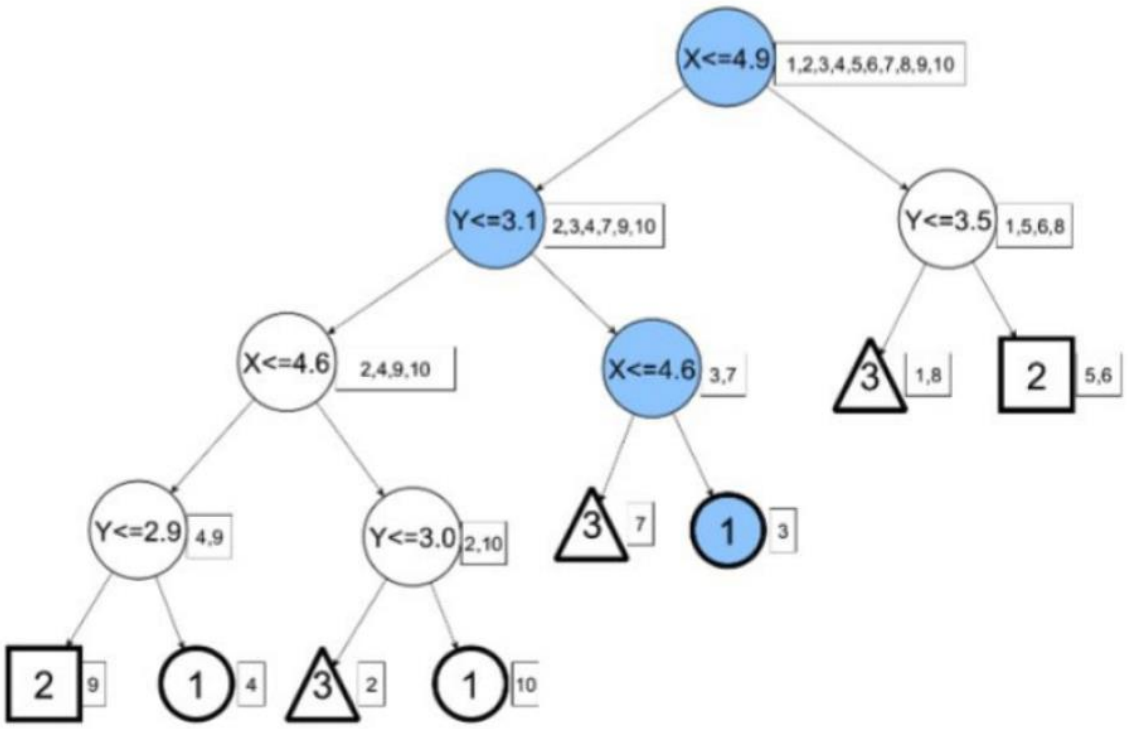
KNN-7 Pick p_3,p_7,p_8,p_{10},p_4,p_2,p_1 Class 3, Pick p_3,p_7,p_8,p_{10},p_4,p_2,p_5 Indeterminate (Class 1 or Class 3)

How you resolve equidistance conflicts is up to you. General approaches include; randomly picking from the subset of conflict points, using the points with the most common class globally, resorting to KNN-1 or using KNN-(k+1).

Q2



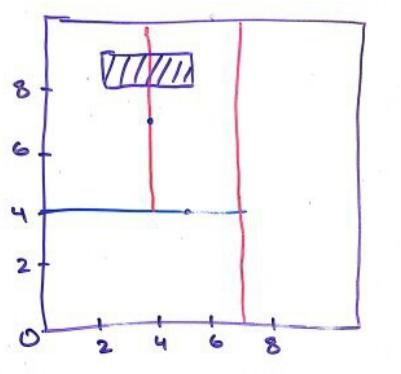

Class 1
Class 2
Class 3
Testing point
p6
p5
p1
p7
p8
z
p3
p4
p10
p2
p9
4
3.8
3.6
3.4
3.2
3
2.8
4.2
4.4
4.6
4.8
5
5.2
5.4
x
y


X<=4.9
1,2,3,4,5,6,7,8,9,10
Y<=3.1
2,3,4,7,9,10
Y<=3.5
1,5,6,8
X<=4.6
2,4,9,10
X<=4.6
3,7
3
1,8
2
5,6
Y<=2.9
4,9
Y<=3.0
2,10
3
7
1
3
2
9
1
4
3
2
1
10

Q3

2-D Range Search with 2-D Trees

To do a 2-D range search, we follow down the tree. Whenever we come across a point whose horizontal/vertical line intersects the 2-D rectangle, we need to "mark" it and ensure we traverse both the left and right subtrees. For example, in the diagram above, if the vertical imaginary line going through (4, 7) intersects the rectangle (even if the point is not in it), we need to search both the left and right subtrees as there may be points in each one that are in the rectangle:



The average run time is $O(R + \lg N)$ where R is the number of points in the rectangle. The worst case is $O(R + \sqrt{N})$ for a balanced tree.

Q4

18.15 The L_1 loss is minimized by the median, in this case 7, and the L_2 loss by the mean, in this case $143/7$.

For the first, suppose we have an odd number $2n + 1$ of elements $y_{-n} < \ldots < y_0 < \ldots < y_n$. For $n = 0$, $\hat{y} = y_0$ is the median and minimizes the loss. Then, observe that the L_1 loss for $n + 1$ is

$$\frac{1}{2n+3}\sum_{i=-(n+1)}^{n+1}|\hat{y}-y_i| = \frac{1}{2n+3}\left(|\hat{y}-y_{n+1}| + |\hat{y}-y_{-(n+1)})|\right) + \frac{1}{2n+3}\sum_{i=-n}^{n}|\hat{y}-y_i|$$

The first term is equal to $|y_{n+1} - y_{-(n+1)}|$ whenever $y_{n+1} \leq \hat{y} \leq y_{-(n+1)}$, e.g. for $\hat{y} = y_0$, and is strictly larger otherwise. But by inductive hypothesis the second term also is minimized by $\hat{y} = y_0$, the median.

For the second, notice that as the L_2 loss of $\hat{y}$ given data $y_1, \ldots, y_n$

$$\frac{1}{n}\sum_i(\hat{y}-y_i)^2$$

is differentiable we can find critical points:

$$0 = \frac{2}{n}\sum_i(\hat{y}-y_i)$$

or $\hat{y} = (1/n)\sum_i y_i$. Taking the second derivative we see this is the unique local minimum, and thus the global minimum as the loss is infinite when $\hat{y}$ tends to either infinity.

Q5

20.21 The main purpose of this exercise is to make concrete the notion of the *capacity* of a function class (in this case, linear halfspaces). It can be hard to internalize this concept, but the examples really help.

a. Three points in general position on a plane form a triangle. Any subset of the points can be separated from the rest by a line, as can be seen from the two examples in Figure S20.4(a).

b. Figure S20.4(b) shows two cases where the positive and negative examples cannot be separated by a line.

c. Four points in general position on a plane form a tetrahedron. Any subset of the points can be separated from the rest by a plane, as can be seen from the two examples in Figure S20.4(c).

d. Figure S20.4(d) shows a case where a negative point is inside the tetrahedron formed by four positive points; clearly no plane can separate the two sets.

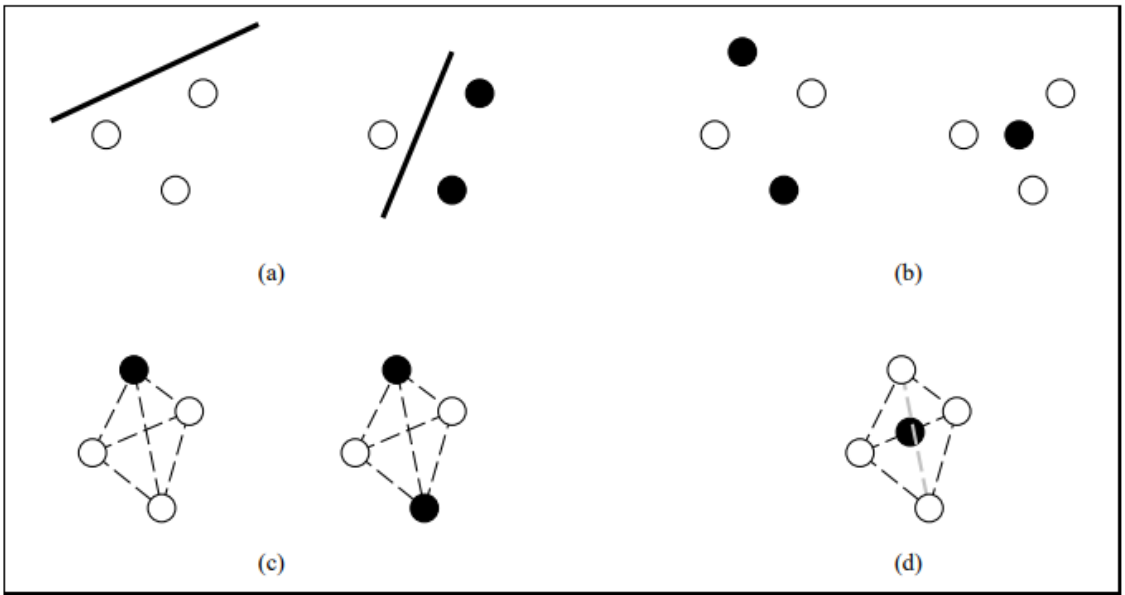


Figure S20.4 Illustrative examples for Ex. 20.21.

Q7

40 red

40 green

Total = 80 H(Total) = 1

Split length by <= 5.5 -> setosa

38 on L>5.5 side, 42 on L<=5.5

a)

H(L > 5.5) = -((37/38) * log(37/38) + (1/38) * log(1/38))

H(L <= 5.5) = -((3/42) * log(3/42) + (39/42) * log(39/42))

IG = 1 - 38/80 * H(L > 5.5) + 42/80 * H(L <= 5.5) = 1 - 0.02510391101 - 0.05866983448 ~= 0.918
Splitting on 5.75 will give the same IG

b)

Same procedure, on L <= 5.5 it would be 3 and on L > 5.5 it would be 3.75

c)

5 splits gives clean separation. 5.5 L -> (3 W -> 4.75 L), (3.75 W -> 6L or anything between 6 and 7.5))

Q8

18.8 This question brings a little bit of mathematics to bear on the analysis of the learning problem, preparing the ground for Chapter 20. Error minimization is a basic technique in both statistics and neural nets. The main thing is to see that the error on a given training set can be written as a mathematical expression and viewed as a function of the hypothesis chosen. Here, the hypothesis in question is a single number $\alpha \in [0, 1]$ returned at the leaf.

a. If α is returned, the absolute error is

$$\begin{aligned} E = p(1-\alpha) + n\alpha = \alpha(n-p) + p &= n \text{ when } \alpha = 1 \\ &= p \text{ when } \alpha = 0 \end{aligned}$$

This is minimized by setting

$$\alpha = 1 \text{ if } p > n$$
$$\alpha = 0 \text{ if } p < n$$

That is, α is the majority value.

b. First calculate the sum of squared errors, and its derivative:

$$E = p(1-\alpha)^2 + n\alpha^2$$
$$\frac{dE}{d\alpha} = 2\alpha n - 2p(1-\alpha) = 2\alpha(p+n) - 2p$$

The fact that the second derivative, $\frac{d^2E}{d\alpha^2} = 2(p+n)$, is greater than zero means that E is minimized (not maximized) where $\frac{dE}{d\alpha} = 0$, i.e., when $\alpha = \frac{p}{p+n}$.

Q9

18.10 This result emphasizes the fact that any statistical fluctuations caused by the random sampling process will result in an apparent information gain.

The easy part is showing that the gain is zero when each subset has the same ratio of positive examples. The gain is defined as

$$I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$$

Since $p = \sum p_i$ and $n = \sum n_i$, if $p_i/(p_i+n_i) = p_j/(p_j+n_j)$ for all i, j then we must have $p_i/(p_i+n_i) = p/(p+n)$ for all i, and also $n_i/(p_i+n_i) = n/(p+n)$. From this, we obtain

$$\begin{aligned} Gain &= I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) \\ &= I\left(\frac{p}{p+n}, \frac{n}{p+n}\right)\left(1 - \frac{\sum_{i=1}^{v} p_i + n_i}{p+n}\right) = 0 \end{aligned}$$