K - nearest neighbours

distance metrics

euclidean $d(x, z) = \sqrt{\| x - z \|_2^2} = \sqrt{\sum_{j=1}^{n} (x_j - z_j)^2}$

manhattan $d(x, z) = \| x - z \|_1 = \sum_{j=1}^{n} |x_j - z_j|$

classification

prediction: $\hat{y} = \text{mode}([y_1 \; y_2 \; \dots \; y_k])$

for a datapoint x:

$1\text{-NN} \Rightarrow \hat{y} = +1$

$2\text{-NN} \Rightarrow \hat{y} = ??$ tie

	$d(x, x_i)$	y_i
1	0.05	+1
3	0.25	-1
2	0.13	-1
4	0.8	+1

handling ties:

- weighted K-NN
- random selection
- 1-NN

regression

$\nearrow$ k - nearest points

prediction : $\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$

weighted : $\hat{y} = \dfrac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}$ → normalizes weights

for a datapoint x:

	$d(x, x_i)$	y_i
1	0.5	10
3	2	30
2	1	20

1-NN $\Rightarrow \hat{y} = 10$

2-NN $\Rightarrow \hat{y} = \frac{10 + 20}{2} = 15$

weighted 2-NN $\Rightarrow \hat{y} = \dfrac{\frac{1}{0.5}(10) + \frac{1}{1}(20)}{\frac{1}{0.5} + \frac{1}{1}}$

$$\hat{y} = \frac{20 + 20}{3} = 13.3$$

decision trees

for the sake of time, ill only do first split

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

no = 5/14
yes = 9/14

using entropy & info. gain:

classes

class prob

$$H = - \sum_{i=1}^{k} p_i \log_2 (p_i)$$

leaf nodes

$$\text{Info gain} = H_{root} - \sum_{i=1}^{k} w_i H_i$$

w_i: $\frac{\text{freq. in leaf}}{\text{freq. in root}}$

$$H_{root} = -\left(\frac{5}{14} \log_2 \frac{5}{14} + \frac{9}{14} \log_2 \frac{9}{14} \right) = 0.94$$

outlook ?

sunny	overcast	rain
Y : 2	Y : 4	Y : 3
N : 3	N : 0	N : 2
$w = 5/14$	$w = 4/14$	$w = 5/14$
$H = -\left(\frac{2}{5}\log_2 \frac{2}{5} + \dots\right)$	$H = 0$	$H = 0.97$
$H = 0.97$		

$$\text{info gain (outlook)} = 0.94 - \left(\frac{5}{14}(0.97) + 0 + \frac{5}{14}(0.971)\right)$$

$$= \boxed{0.247}$$

temp

hot	mild	cool
Y : 2	Y : 4	Y : 3
N : 2	N : 2	N : 1
$w = 4/14$	$w = 6/14$	$w = 4/14$
$H = -\left(\frac{2}{4}\log_2 \frac{2}{4} \dots\right)$	$H = 0.918$	$H = 0.811$
$H = 1$		

$$\text{info gain (temp)} = 0.94 - \left(\frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.811)\right)$$

$$= \boxed{0.0291}$$

and so on ...

now, gini index :

\# of classes

$$\text{Gini}(\text{node}) = 1 - \sum_{i=1}^{c} p_i^2$$

→ prob. of class i

$$\text{Gini}_{\text{root}} = 1 - \left(\left(\frac{5}{14}\right)^2 + \left(\frac{9}{14}\right)^2 \right) = 0.459$$

outlook ?

sunny → Y : 2, N : 3, $w = 5/14$

overcast → Y : 4, N : 0, $w = 4/14$

rain → Y : 3, N : 2, $w = 5/14$

$$G = 1 - \left(\frac{2}{5}^2 + \frac{3}{5}^2 \right) = 0.48$$

$$G = 1 - (1 + 0) = 0$$

$$G = 0.48$$

$$\text{info gain (outlook)} = 0.459 - \left(\frac{5}{14}(0.48) + 0 + \frac{5}{14}(0.48) \right) = 0.116$$

and so on ...

tsallis entropy: $S_q(x) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{c} p_i^q\right)$

$q=2 \Rightarrow$ Gini Index $= \frac{1}{2-1}\left(1 - \sum_{i=1}^{c} p_i^2\right) = 1 - \sum_{i=1}^{c} p_i^2$

$q=1 \Rightarrow$ Entropy $= \lim_{q \to 1} \frac{1}{q-1}\left(1 - \sum_{i=1}^{c} p_i\right) = -\sum_{i=1}^{c} p_i \log_2(p_i)$

using L' Hopital