Exercise → classification

1. i) morphological classification on whether a word form was derived or inflected

ii) sentiment analysis (positive/negative)

2. similarity measure helps the model to generalize. Objects with somewhat similarity measure can be added to the same class. Measure is used a loss function during classification. We can set a class based the threshold for classification. for example sigmoid in logistic regression.

3. i) position in a sentence
ii) pos distribution of the word in similar places.
iii) ~~sur context~~ surrounding words.

4. cook - sleep

$\overrightarrow{cook} = \langle 98, 1, 1 \rangle$

$\overrightarrow{sleep} = \langle 35, 50, 15 \rangle$

$L_2(cook, sleep) = \sqrt{(98-35)^2 + (1-50)^2 + (1-15)^2}$

$= 81.03$

$\overrightarrow{bake} = \langle 14, 70, 16 \rangle$

$L_2(cook, bake) = \sqrt{(98-14)^2 + (1-70)^2 + (1-16)^2}$

$= 109.74$

$L_2(cook, bake) > L_2(cook, sleep)$

cook is ~~more related~~ closer to bake than sleep. Makes sense!

5.

$P(Diät_1) = \frac{1}{1500} = \frac{1}{1500}$

$P(Diät_2) = \frac{1}{800}$

$P(Diät_3) = \frac{8}{200}$, $P(Diät_4) = \frac{69}{1000}$

$S = \{1, 2, 3, 4\}$

dict kilo rempalin

$f = \langle 1, 1, 17 \rangle \quad n = 3$

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} \; P(s) \prod_{j=1}^{n} P(f_j \mid s)$$

$|w_1| = 3500$

$P(s_1) = \frac{1500}{3500} = 3/7$

$P(s_2) = \frac{800}{3500} = 8/35$

$P(s_3) = 200/3500 = 2/35$

$P(s_4) = 2/7$

$$\hat{s}_1 = 3/7 * \left(\frac{1}{1500} \times \frac{1}{1500} \times \frac{1}{1500}\right)$$

$$= 1.27 \times 10^{-10}$$

$$\hat{s}_2 = 8/35 \times \left(\frac{1}{800} \times \frac{12}{800} \times \frac{85}{800}\right)$$

$$= 4.55 \times 10^{-7}$$

$$\hat{s}_3 = 2/35 \times \left(\frac{8}{200} \times \frac{1}{200} \times \frac{3}{200}\right)$$

$$= 171.43 \times 10^{-9} = 1.7143 \times 10^{-7}$$

$$\hat{s}_4 = 2/7 \times \left(69/1000 \times \frac{95}{1000} \times \frac{1}{1000}\right)$$

$$= 1.87 \times 10^{-6}$$

most probable sense = s_4 → weight
→ relates more to diet and kilo
I'm not sure how it relates to repair.

⑥ Values (Wind) = weak, strong

$C = [9+, 5-]$
$C_{weak} = [6+, 2-]$
$C_{strong} = [3+, 3-]$

$E_C = \sum_{i=1}^{c} -P_i \times \log_2 P_i$

$$E = \sum_{i=1}^{c} -P_i \times \log_2 P_i \quad \text{c = 14}$$

$$= -\frac{9}{14} \times \log_2 \frac{9}{14} + -\frac{5}{14} \log_2 \frac{5}{14}$$

$$= 0.940$$

$$E_{weak} = \frac{-6}{14} \times \log \frac{6}{14} - \frac{2}{14} \log_2 \frac{2}{14}$$

$$= 0.811$$

$E_{strong} =$

7. n-fold cross validation
= 1

→ randomly create a set of train: test partitions from data

→ For each partition
  → train model with train set
  → validate trained model using the test set.

→

→ take the average of validation results for all the partitions.

8. a) gold = {blood, rose}, {blood, poppy}, {ocean, sky}, {rose, poppy}

= 4 pairs

model = {ocean, sky}, {rose, poppy}, {ocean, blood}, {sky, blood}

= 4

gold ∩ model = {ocean, sky}, {rose, poppy}

= 2

(b)

$$\begin{cases} c_b = \{\text{ocean, sky, blood}\} \\ c_r = \{\text{rose, poppy}\} \end{cases}$$

$$\begin{cases} g_b = \{\text{ocean, sky}\} \\ g_r = \{\text{rose, poppy, blood}\} \end{cases}$$

model \ gold	blue	red
blue	2	1
red	1	2

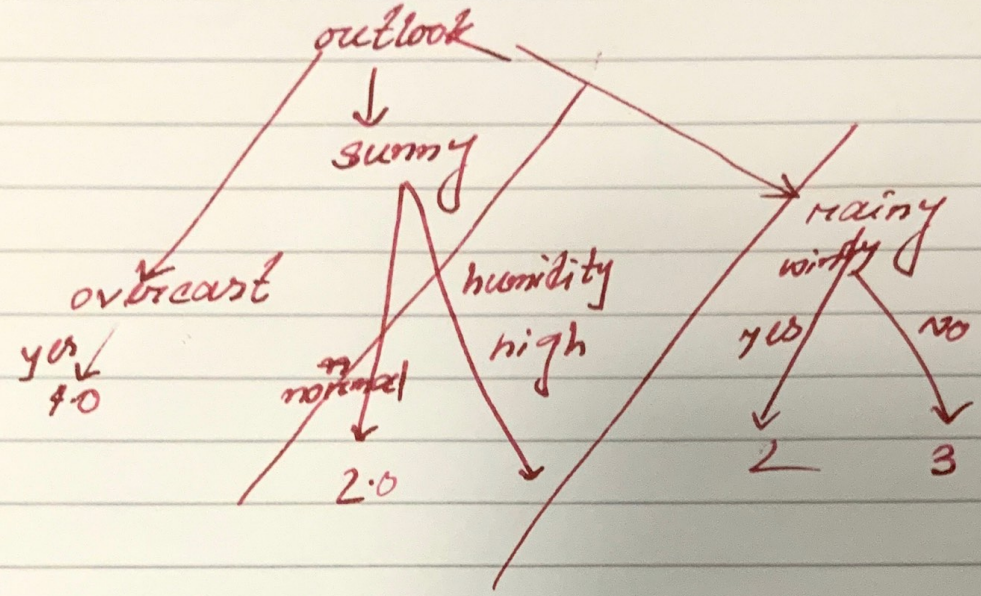
$$\begin{cases} tp = 2 \\ fp = 1 \\ fn = 1 \\ tn = 2 \end{cases}$$

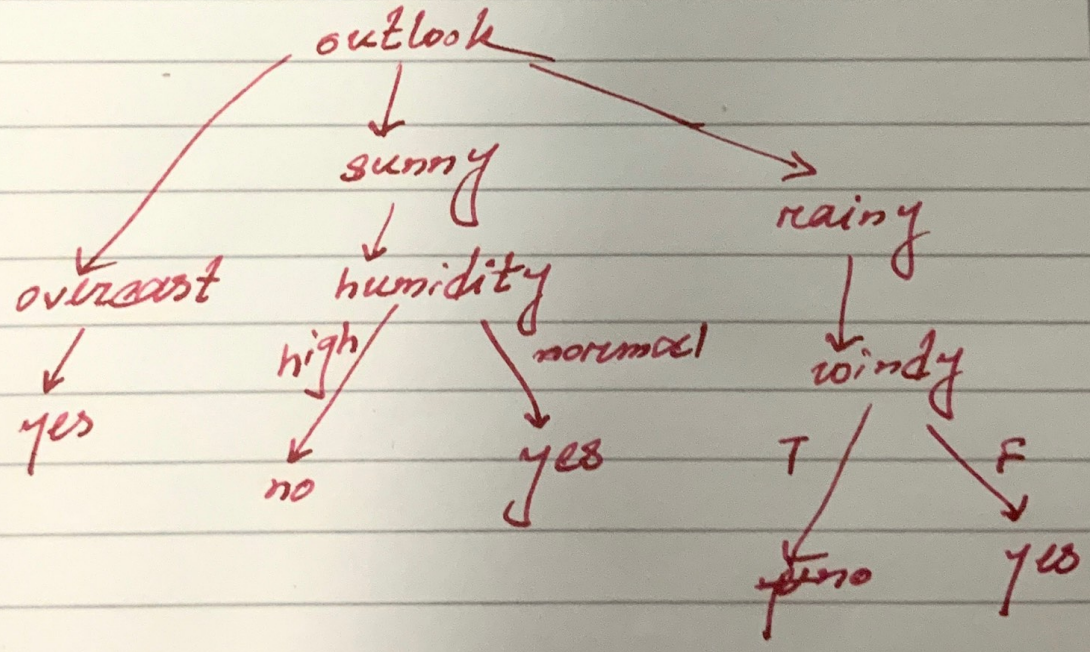
$$\therefore \text{accuracy} = \frac{2+1}{2+1+1+2} = \frac{2+1}{6} = 1/2$$

$$\text{precision} = \frac{tp}{tp+fp} = \frac{2}{2+1} = 2/3$$

$$\text{recall} = \frac{tp}{tp+fn} = \frac{2}{3}, \quad f = \frac{2 \times 2/3 \times 2/3}{2 \times 2/3} = 2/3$$

J48 classifier tree





* can't have more folds than instances for classification

* 3 folds produces the best results.