CS 4780

Problem 1: Version Spaces

(a) $\{mild, medium, hot\} \times \{no meat, chicken, beef\} \times \{yes, no\}$

$$3 \quad \times \quad 3 \times 2 = 18$$

Instance space of size 18

(b) $2^{18} = 262144$ is the size of H

(c) clearly, this hypothesis is not consistent with the table. By definition h is consistent if and only if $h(x) = y$ for all (x,y) in the table. By Bob's hypothesis, h (any entry with hot) should be yes, however the entry for lentil curry contradicts this as the dish is hot but Alice does not like it. Thus Bob's hypothesis is not consistent with table T.

(d) With 5 of the 18 entries determined, the size of the Version space is 2^{13} or $VS(H,T) = 8192$.

(e) 3 possible combinations:

$S + M$

$h_1(S,M) = \begin{cases} yes, & \text{if } S+M > 0 \\ no, & \text{otherwise} \end{cases}$

$M + C$

$h_2(M,C) = \begin{cases} yes, & \text{if } M+C > 0 \\ no, & \text{otherwise} \end{cases}$

$h_1'(S,M) = \begin{cases} no, & \text{if } S+M > 0 \\ yes, & \text{otherwise} \end{cases}$

$h_2'(M,C) = \begin{cases} no, & \text{if } M+C > 0 \\ yes, & \text{otherwise} \end{cases}$

$S + C$ $h_3(S,C) = \begin{cases} yes, & \text{if } S+C > 0 \\ no, & \text{otherwise} \end{cases}$ $h_3'(S,C) = \begin{cases} no, & \text{if } S+C > 0 \\ yes, & \text{otherwise} \end{cases}$

$H' = \{h_1, h_1', h_2, h_2', h_3, h_3'\}$, thus H' has size 6.

(f)

Dish	S	M	C	h_1	h_2	h_3	h_1'	h_2'	h_3'
spicy chicken	1	0	1	yes	yes	yes	no	no	no
pepper steak	0	1	-1	yes	no	no	no	yes	yes
beef stew	-1	1	-1	no	no	no	yes	yes	yes
lentil curry	1	-1	1	no	no	yes	yes	yes	no
chicken soup	0	0	-1	no	no	no	yes	yes	yes

Clearly h_1 is the only function of $VS_{(H', T')}$
as it is consistent for T'. $h_2, h_3,$ and h_1'
are inconsistent due to the incorrect output
for pepper steak whereas h_2' and h_3' are
inconsistent due to incorrect output for
spicy chicken entry. Thus, the size
of $VS_{(H', T')}$ is 1.

2.

Red — *
Blue — •

(a) The boundary can be found by
finding all points where 3 points with
at least one of each color are equidistant.
These points are then connected from these
points, if there are not two lines
that have already connected to
them, a line is extended to
the edge of the rectangle


maintaining equidistance.

(b)   Red → *   Blue → •
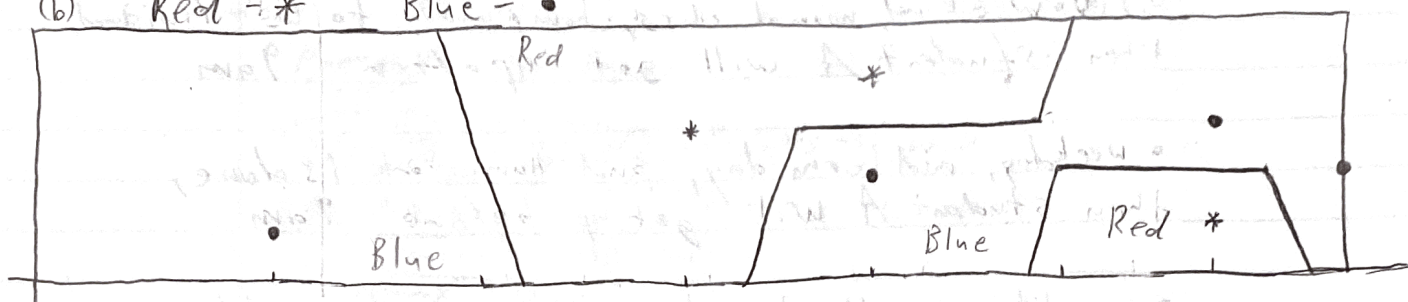
When working with real data this effect
becomes a problem as points in the
original unscaled boundaries and data may
be classified differently than when they
are scaled. For example, the point (0,5) would
be classified as Red, however following scaling,
and the boundaries moving, this now scaled
point (0,5) would be blue. One solution to mitigate
these effects would be to simply scale both
axis equally.

(c) Under the original set of points, (3,4) would
be classified as Red under 3-NN as the
three nearest points are (3,3), (4,4), and (4,2)
of which 2 are Red. To change the classification
of (3,4) under 3-NN, a minimum of 2 points must
be added such that they are now the closest
points to (3,4) such as (3.5,4) and (3,4.5) both
Blue. Thus, the nearest 3 points are (3.5,4),
(3,4.5) and a tie for third between (3,3) and (4,4)
and two are Blue and one is red thus
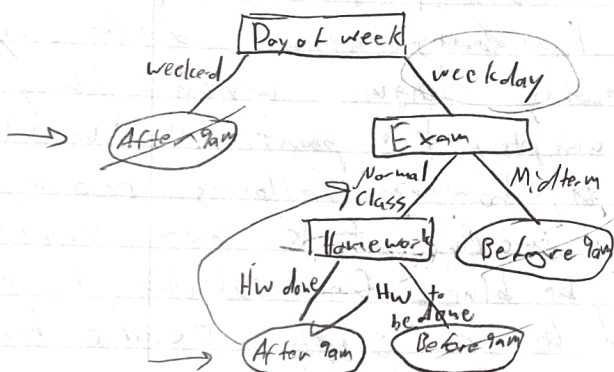classification becomes Blue.

test point (3,4) → ▲   Red → *
(3.5,4) and (3,4.5) → ■   Blue → •
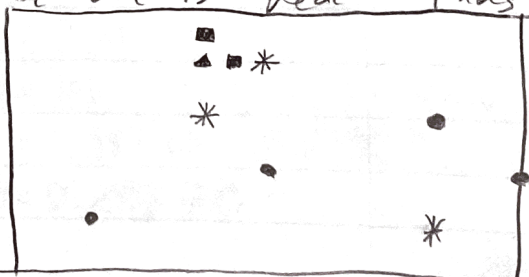
3.) (a) • weekend, normal class, homework to be finished, then student A will get up after 9am
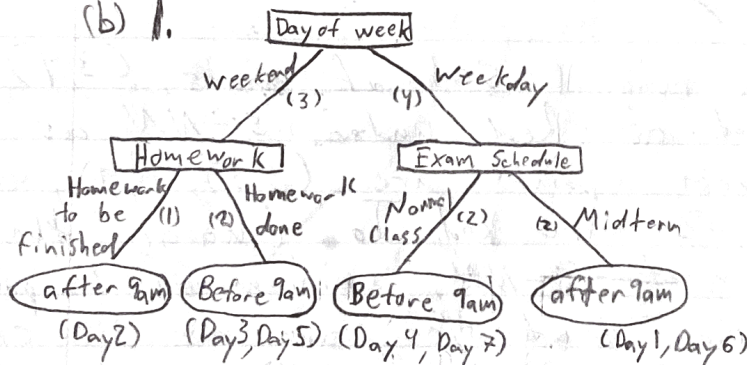
• weekday, midterm day, and homework is done, then student A will get up before 9am

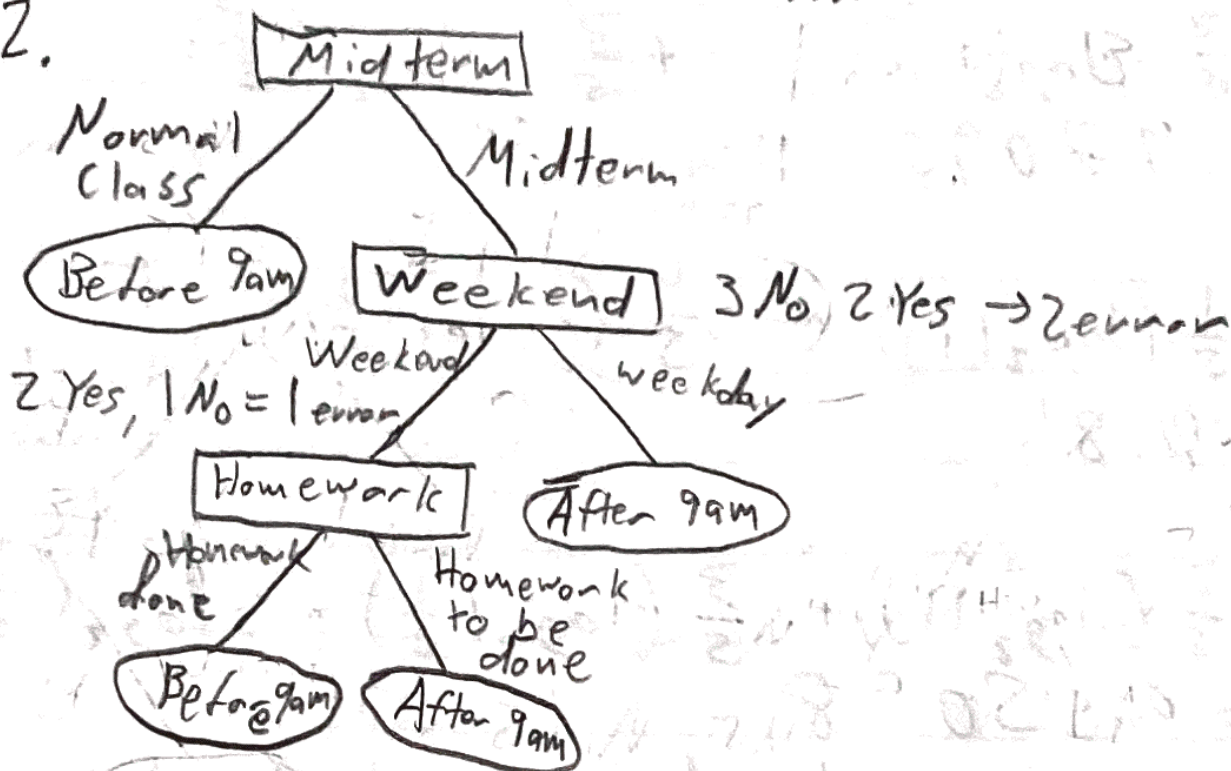• weekday, normal class, homework is done, then student A will get up after 9am

If student A woke up after 9am on a weekday, then it was not a midterm day.

(b) 1.

(b)2.

$$4 \text{ Yes } 3 \text{ No} \rightarrow 3 \text{ error}$$

Midterm

Normal Class / Midterm

Before 9am

Weekend $3 \text{ No, } 2 \text{ Yes} \rightarrow 2 \text{ error}$

2 Yes, 1 No = 1 error   Weekend / weekday

Homework   After 9am

Homework done / Homework to be done

Before 9am   After 9am

(b)3.

$$H(S) = -\sum_{y \in Y} P_s(y) \log_2(P_s(y))$$

$$= -\left(\left(\frac{3}{7}\right) \log_2\left(\frac{3}{7}\right) + \left(\frac{4}{7}\right) \log_2\left(\frac{4}{7}\right)\right)$$

$$= 0.985228$$

$$H(S \mid \text{weekend}) = \sum_{a_i} P_s(S_i) H(S_i)$$

$$= \frac{4}{7} H(S_{yes}) + \frac{3}{7}(H(S_{no}))$$

$$= \frac{4}{7}\left(-\left(\left(\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)\right) + \frac{3}{7}\left(-\left(\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)\right)\right)$$

$$= \frac{4}{7}\left(-\log_2\left(\frac{2}{4}\right)\right)$$

$$= 0.96498$$

$$H(S \mid \text{Homework}) = \frac{2}{7}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right)$$
$$+ \frac{5}{7}\left(-\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)\right)$$
$$= 0.97925$$

$$H(S \mid \text{Midterm}) = \frac{2}{7}\left(-\left(1 \log_2(1)\right)\right)$$
$$+ \frac{5}{7}\left(-\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)\right)$$
$$= 0.69354$$

Maximizing $IG(S) = H(S) - H(S \mid A)$
when using Midterm $IG(S) = H(S) - H(S \mid \text{midterm})$
$$= 0.985228 - 0.69354$$
$$IG(S) = 0.29270$$

$$H = \frac{2}{5} \log_2\left(\frac{2}{5}\right) + \frac{3}{5} \log_2\left(\frac{3}{5}\right)$$
$$= 0.97095$$

$$H_w = \frac{1}{5}\left(\log_2(1)\right) + \frac{4}{5}\left(-\log\left(\frac{2}{4}\right)\right)$$
$$= 0.8$$

$$\text{weekend} = \frac{2}{5}\left(\log_2(1)\right) + \frac{3}{5}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)$$
$$= 0.55098$$

Max is when weekend is next feature

Midterm

Midterm day — Normal class

Weekend — Before 9am

Weekend — weekday

Homework — After 9am

Homework to be done — Homework finished

After 9am — Before 9am