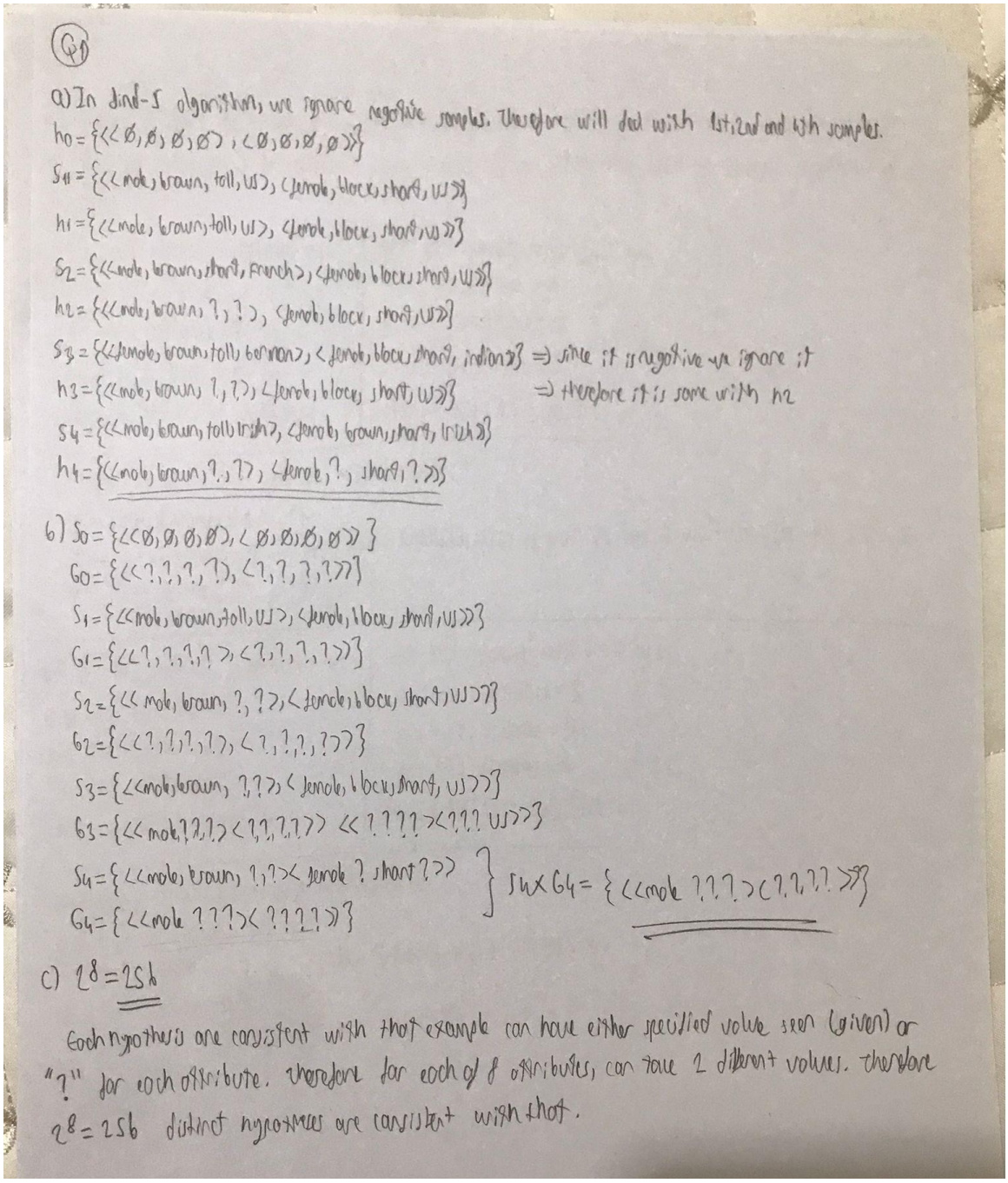
Group Members:
1- Mert Tazeoğlu - b21946606



Q1

a) In find-S algorithm, we ignore negative samples. Therefore will deal with 1st, 2nd and 4th samples.

$h_0 = \{\langle\langle \emptyset, \emptyset, \emptyset, \emptyset\rangle, \langle \emptyset, \emptyset, \emptyset, \emptyset\rangle\rangle\}$

$S_1 = \{\langle\langle$ male, brown, tall, US$\rangle, \langle$ female, black, short, US$\rangle\rangle\}$

$h_1 = \{\langle\langle$ male, brown, tall, US$\rangle, \langle$ female, black, short, US$\rangle\rangle\}$

$S_2 = \{\langle\langle$ male, brown, short, French$\rangle, \langle$ female, black, short, US$\rangle\rangle\}$

$h_2 = \{\langle\langle$ male, brown, ?, ?$\rangle, \langle$ female, black, short, US$\rangle\rangle\}$

$S_3 = \{\langle\langle$ female, brown, tall, German$\rangle, \langle$ female, black, short, Indian$\rangle\rangle\}$ ⇒ since it is negative we ignore it

$h_3 = \{\langle\langle$ male, brown, ?, ?$\rangle, \langle$ female, black, short, US$\rangle\rangle\}$ ⇒ therefore it is same with h_2

$S_4 = \{\langle\langle$ male, brown, tall, Irish$\rangle, \langle$ female, brown, short, Irish$\rangle\rangle\}$

$h_4 = \{\langle\langle$ male, brown, ?, ?$\rangle, \langle$ female, ?, short, ?$\rangle\rangle\}$

b) $S_0 = \{\langle\langle \emptyset, \emptyset, \emptyset, \emptyset\rangle, \langle \emptyset, \emptyset, \emptyset, \emptyset\rangle\rangle\}$

$G_0 = \{\langle\langle ?, ?, ?, ?\rangle, \langle ?, ?, ?, ?\rangle\rangle\}$

$S_1 = \{\langle\langle$ male, brown, tall, US$\rangle, \langle$ female, black, short, US$\rangle\rangle\}$

$G_1 = \{\langle\langle ?, ?, ?, ?\rangle, \langle ?, ?, ?, ?\rangle\rangle\}$

$S_2 = \{\langle\langle$ male, brown, ?, ?$\rangle, \langle$ female, black, short, US$\rangle\rangle\}$

$G_2 = \{\langle\langle ?, ?, ?, ?\rangle, \langle ?, ?, ?, ?\rangle\rangle\}$

$S_3 = \{\langle\langle$ male, brown, ?, ?$\rangle, \langle$ female, black, short, US$\rangle\rangle\}$

$G_3 = \{\langle\langle$ male, ?, ?, ?$\rangle, \langle ?, ?, ?, ?\rangle\rangle, \langle\langle ?, ?, ?, ?\rangle, \langle ?, ?, ?, US\rangle\rangle\}$

$S_4 = \{\langle\langle$ male, brown, ?, ?$\rangle, \langle$ female, ?, short, ?$\rangle\rangle\}$

$G_4 = \{\langle\langle$ male, ?, ?, ?$\rangle, \langle ?, ?, ?, ?\rangle\rangle\}$

$S_4 \times G_4 = \{\langle\langle$ male, ?, ?, ?$\rangle, \langle ?, ?, ?, ?\rangle\rangle\}$

c) $2^8 = 256$

Each hypothesis are consistent with that example can have either specified value seen (given) or "?" for each attribute. Therefore for each of 8 attributes, can take 2 different values. Therefore $2^8 = 256$ distinct hypotheses are consistent with that.

(Q2)

a) Label is same of point x with 1-NN algorithm. Since 1-NN chooses only one nearest point, then there can't be another nearer point to x other than born of nearest neighbours of S1 and S2.

Let S1's nearest neighbours as:
- X → (+), distance = 1
- y → (−), distance = 2
- t → (−), distance = 5

Then, 1-NN(S1) is positive.

Let S2's nearest neighbours as:
- X → (+), distance = 1
- t → (−), distance = 2
- q → (−), distance = 3

Then, 1-NN(S2) is positive.

Therefore (S1 ∪ S2)'s nearest neighbours are:
- X → (+), distance = 1
- y → (−), distance = 2
- t → (−), distance = 2

} Then 1-NN(S1 ∪ S2) is positive and it will consider only point X.

---

b) Let S1's nearest neighbours are:
- a → (+), distance = 1
- b → (−), distance = 2
- c → (+), distance = 4

Then, 3-NN(S1) = positive.

Let S2's nearest neighbours are:
- d → (−), distance = 3
- e → (+), distance = 5
- f → (+), distance = 6

Then, 3-NN(S2) = positive.

Therefore, (S1 ∪ S2)'s nearest neighbours are:
- a → (+), distance = 1
- b → (−), distance = 2
- d → (−), distance = 3

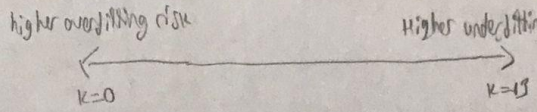
} Then 3-NN(S1 ∪ S2) is negative.

(G1)

a) A point can be its own neighbor. Therefore $k=0$ minimizes training set error.
If $k=0$, then training error is 0.

---

b) Too big k values (for example $k=13$) misclassifies every datapoint (using leave one out cross validation). On the other hand, too small k value (for example $k=0$) leads to overfitting.



---

c)

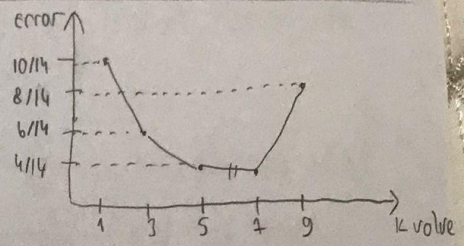
for $k=1$ error is $10/14$

for $k=3$ error is $6/14$

for $k=5$ error is $4/14$

for $k=7$ error is $4/14$

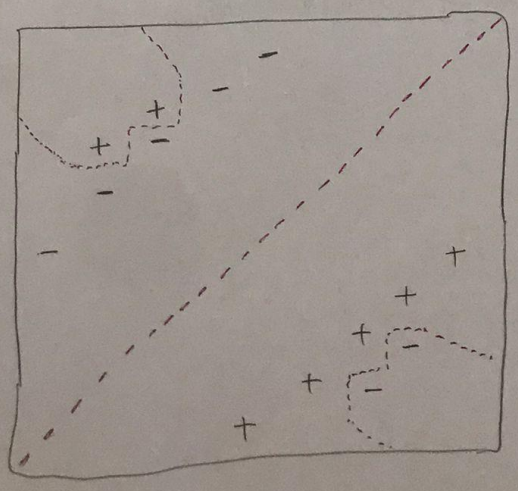
for $k=9$ error is $8/14$



★ $k=5$ and $k=7$ minimizes the leave-one-out cross validation error. Error is $4/14$.

---

d)



---- ⇒ 1 nearest neighbour decision boundary

(Q4)

PREDICTED \ ACTUAL	positive	negative
positive	TP	FP
negative	FN	TN

In 1000 instances;
i) Since 950 are healthy, TP+FN must be 950.
ii) Since 50 are patient, FP+TN must be 50.

a)

	+	−
+	TP = 950	FP = 50
−	FN = 0	TN = 0

In this example;

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} = \frac{950}{1000} = \%95$$

According to accuracy value, this model is very good. However, it is quite bad since it classifies every person as positive. (In other words, it doesn't detect negative)

b)

	+	−
+	TP = 50	FP = 1
−	FN = 900	TN = 49

In this example;

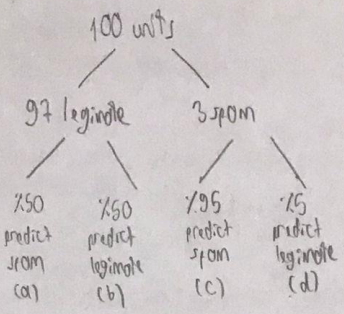
$$\text{Recall} = \frac{TP}{TP+FN} = \frac{50}{950} = 0.052 \text{ (so bad value)}$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{50}{51} = 0.98 \text{ (good value)}$$

In this example, model has a good precision value, but a bad recall value. In this example, using only precision value not enough and both of them should be used. If we use only precision value, although the model is bad it will be evaluated as good.

(Q5)

Solution-1:



a- (97 x 50) / 100 = %48.5
b- (97 x 50) / 100 = %48.5
c- (3 x 95) / 100 = %2.85
d- (3 x 5) / 100 = %0,15

Predict AS spam => a+c = %51.35
Actually spam (predicted as spam) => %2.85

$\frac{2,85}{51,35} = \%5.55$

Solution-2:

This can be solved using Bayes theorem, which says:
$P(A|B) = P(B|A) \times P(A) / P(B)$, where A = email is spam and B = algorithm predict as spam

$P(A|B) = P(B|A) \times P(A) / P(B)$

$P(B|A) = 95/100 = 0.95$

$P(A) = 3/100 = 0.03$

$P(B) = \frac{(0.97 \times 0.5) + (0.03 \times 0.95)}{1 \times 100} = \frac{0.4850 + 0.0285}{1} = 0.5135$

$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} = \frac{(0.95) \times (0.03)}{(0.5135)} = \frac{0.0285}{0.5135} \simeq 0.555 \simeq \%5.55$

(b) (PART-1)

Recursion depth: 1 (root)

Grouped records: {Age: {young: 5, middle: 4, senior: 5}, sex: {M: 7, F: 7}, BP: {High: 4, normal: 6, low: 4}, cholesterol: {normal: 8, high: 6}}

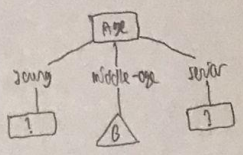
system entropy: 0.940

Feature entropies: {Age: [0.97, 0, 0.71], sex: [0.98, 0.99], BP: [1, 0.65, 0.81], cholesterol: [0.81, 0.91]}

Information gains: {Age: 0.25, sex: 0.1, BP: 0.6, cholesterol: 0.005}

Most informative: Age (young > senior > middle)

↳ in middle age drug B is common



Recursion depth: 2 (young people)

Grouped records: {sex: {F: 3, M: 4}, BP: {high: 2, normal: 2, low: 0}, cholesterol: {normal: 2, high: 2}}

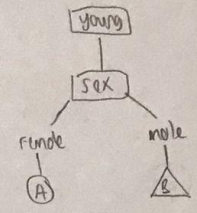
system entropy: 0.940

Feature entropies: {sex: [0, 0], BP: [0, 1], cholesterol: [0, 1]}

Information gains: {sex: 0.94, BP: 0.94, cholesterol: 0.44}

most informative: sex (female > male)

↳ All of young males use drug B.
also all of young females use drug A.



Recursion depth: 2 (senior people)

Grouped records: {sex: {F: 2, M: 3}, BP: {normal: 3, low: 3}, cholesterol: {normal: 3, high: 2}}

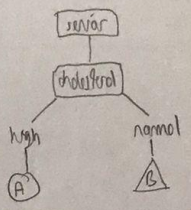
system entropy: 0.864

Feature entropy: {sex: [0, 1], BP: [0, 0.92], cholesterol: [0, 0.92]}

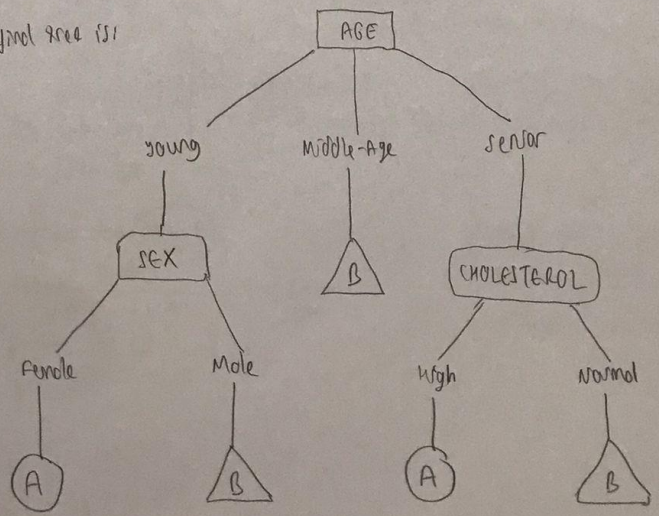
Information gains: {sex: 0.19, BP: 0.40, cholesterol: 0.40} → no exact match in BP but cholesterol provides it therefore we select it

most informative: cholesterol (high > low)

↳ where is a exact match that normal cholesterol = drug B; otherwise drug A



At the end, our final tree is:



(P15) ⇒ since P15 is middle aged, then he/she should use DRUG-B!

b) $P(drug=A) = 5/14$, $P(drug=B) = 9/14$ | $P(cholesterol=normal) = 8/14$, $P(cholesterol=high) = 6/14$ | $P(BP=low) = 4/14$, $P(BP=normal) = 4/14$, $P(BP=high) = 4/14$ | $P(sex=F) = 9/14$, $P(sex=M) = 7/14$ | $P(age=young) = 5/14$, $P(age=middle) = 4/14$, $P(age=senior) = 5/14$

Let's compute $P(X|c_i)$ for each class:

- $P(age='middle' \mid drug='A') = 0/4 = 0$
- $P(age='middle' \mid drug='B') = 4/4 = 1$
- $P(sex='female' \mid drug='A') = 3/4$
- $P(sex='female' \mid drug='B') = 3/7$
- $P(BP='low' \mid drug='A') = 1/4$
- $P(BP='low' \mid drug='B') = 3/4$
- $P(cholesterol='normal' \mid drug='A') = 2/8$
- $P(cholesterol='normal' \mid drug='B') = 6/8$

$X = (age='middle', sex='female', BP='low', cholesterol='normal')$

$P(X|c_i)$: $P(X \mid drug='A') = 0 \times 3/4 \times 1/4 \times 2/8 = 0.000$

$P(X \mid drug='B') = 1 \times 3/4 \times 3/4 \times 6/8 = 0.241$

$P(X|c_i) \times P(c_i) \propto P(X \mid drug='A') \times P(drug='A') = 0.000$

$P(X \mid drug='B') \times P(drug='B') = 0.154$

→ therefore P15 belongs to class (drug=B).

confidence of classification $= \frac{0.154}{0.154 + 0.000} = 1 = \%100$

BOTH OF MODELS SAY THAT P15 SHOULD USE DRUG-B!