Problem 1 Constraint Satisfaction

Part A

A.1

The answer is d, vertically and horizontally contiguous spaces.

A.2

The answer is a, the words in the list.

A.3

1. The length of the word must equal the number of spaces. Eg in the H3 space, only Vest, Exam and Nerd can fit.

2. Rows and columns that cross must agree on the letters in the shared space. Eg H1 and V1 must start with the same letter.

Part B

Below, B, G, R, Y stand for the colors. An 0x means circled at step x and an Xy means crossed out at step y.

B.1

Bactracking:

```
B:01   B:02X3 B:05X6 B:09X10
G      G:04   G:06X7 G:011
R      R      R:08   R
Y      Y      Y      Y
```

B.2

Pure constraint propagation (give I = Blue)

```
B:01   B:X2   B:X3   B:X4
G      G      G      G
R      R      R      R
Y      Y      Y      Y
```

B.3

Bactracking with FC

```
B:01   B:X2   B:X3   B:X4
G      G:05   G:X6   G:09
R      R      R:07   R:X8
Y      Y      Y      Y
```

B.4

a. breadth first will find all solutions

b. the space requirements for breadth-first are enormous, BF requires (in the worst case) space exponential in the search depth (number of variables) while DF requires space that is linear in the search depth.

Problem 2

Part A

Excuse the lack of pictures...

The Nearest Neighbor decision boundary is a fairly smooth line separating the two classes. In practice, this separator is piecewise linear (made up of pieces of lines that are equidistant from two points on the boundary between the two classes) but this separator tracks the natural L-shaped boundary between the two classes pretty well.

The ID tree classifier only produces vertical or horizontal lines, since the tests are on one variable at a time. Two lines can be used to separate the classes, one vertical and one horizontal. The vertical one goes all the way down creating a homogeneous set on the left side and an inhomogeneous set on the right, the horizontal line splits the inhomogeneous set (on the right) into two homogoneous subsets.

The single unit perceptron can only draw a linear decision boundary. A line oriented at around 45 degrees produces the fewest misclassified points.

To draw the boundary for a neural net one would need to know the architecture of the net, ie. how many units on which layers.

Part B

Assuming the usual Euclidean metric, there are 499 dimensions of noise and one of signal, all added up. One would not expect reliable classification from nearest neighbor. One would have to modify the distance function to pay attention to only the first dimension.

A perceptron, on the other hand, should work fine. The division between the positives and the negatives is a planar surface in the 500 dimensional space. Another way of saying this is that we can pick weights (a 1 followed by 499 0's) and a threshold (just 0) to perform the classification. The perceptron convergence theorem tells us that we should converge (with enough examples) on such a linear separator.

An ID tree should work fine. The first dimension is tested and there is a value that generates 0 average disorder. So, we get a tree with just one node with perfect performance.

Problem 3 Crazy Frank

Part A

R1 is most likely to get stuck in a local maximum. There is very little diversity – no mutation, everyone descends from "best" parents.

R2 is basically a random search – very high mutation rate means that very little is carried over from one generation to the next.

Part B

```
       Wk 2        Wk 3          Wk 11
     4%    8%    5%     67%     64%    78%
R3:  4/34  8/34  5/101  67/101  64/421 78/421
R4:  1/21  6/21  1/21    6/21    1/21   6/21
```

In R4, the highest scoring individual is always 6 times more likely to be selected than the lowest scoring individual. This is a great deal of selective pressure and is likely to reduce diversity. By contrast in R3, since the scores are improving for all the individuals, the range of probabibilities between high and low scorers drops from approximately 13:1 to 1.2:1 by week 11.

Part C

1. None of these

2. C(A,B) - crossover between fourth and fifth position.

3. M(A) - differs only in second position.

4. Either M(A) - mutation at fifth bit or C(A,B) - crossover between fourth and fifth position.

Part D

P3 is most diverse, every position has both possible values represented.

P1 is least diverse, has 5 positions where all the individuals have the same value. For comparison, P2 has 3 such positions.

Part E

Think of a line as dividing space into two, the 0 side and the 1 side. A single perceptron unit implements such a line (in two dimensions). Two crossing lines divide space into four quadrants (00, 01, 10, 11). Combining the output of two perceptron units allows us to pick which quadrant we want. We can combine such groups of pairs of lines to define more complex shapes.

S1 can be solved by N2. Four lines can divide the A's from the non-A's. Two lines include all of A and a piece of B, two other lines exclude the piece of B sticking into A. N2 implements this.

S2 can be solved by N1. The A region can be separated by two lines, which is what N1 implements.

S3 can be solved by N3. One line separates the A, which is what N1 does.

S4 can be solved by N1. This also needs two (non-crossing lines), which N2 can implement.

Part F

In a feedforward network, a unit cannot point to itself or any other unit in its own layer or a previous layer (that's why they are called feedFORWARD). So, all we need to represent is the upper right hand corner of the connection matrix.

Problem 4 Grading Trees

Part A

Name leads to the lowest disorder, since every name has only one grade and so defines a homogeneous class. Of course, this is useless. In practice, a different measure other than just weighted disorder is used to avoid this problem.

Part B

The answer is 4. This just makes sure that the tests eventually produce homogeneous sets (with minimal disorder).

Part C

Test 1 would be chosen.

Test 1: $16/24 * D(4/16) + 8/24 * D(2/8) = D(1/4) = 1/2$ Test 2: $10/14 * D(2/10) + 4/14 * D(2/4) = 8/14$

Part D

Yes. A single test that produces a homogeneous set is likely to show up in an ID tree.

Part E

```
          F>85
        N/     \Y
      Q1>85    Q1>85
    N/   \Y   N/   \Y
    X  Q1>85 Q2>85  A
       N/\Y  N/\Y
       X  A  X  A
```

No. This tree is unlikely to show up since the first test (and some of the subsequent ones) do not lead to largely homogeneous sets.

Part F

1. If we use the results from last year, we would be ignoring that the grades distribution vary greatly from year to year.

2. We should normalize the data entries (each exam and PS) by subtracting the mean and dividing by the standard deviation.

Part G

All of 2,3,4 and 6 would get Procedure W to learn this concept. The sequences 2 and 3 exploit the ability of changing a set of links at once if the differences are all of the same kind. Sequences 4 and 6 do one change at a time. Sequences 1 and 5 don't start with a positive and are therefore not valid.

Problem 5

Part A

The following is a different (and equivalent) way of defining the tree (and is easier to type):

```
TIGER
 R7
  CARNIVORE
   R3
    MAMMAL
     R1
      GIVES MILK
       QUESTION 1 GIVES MILK? <- YES
       ASSERT GIVES MILK
     ASSERT MAMMAL
    EATS MEAT
     QUESTION 2 EATS MEAT? <- YES
     ASSERT EATS MEAT
   ASSERT CARNIVORE
  FELINE
   R5
    WHISKERS
     QUESTION 3 WHISKERS? <- YES
     ASSERT WHISKERS
    CLAWS
     QUESTION 4 CLAWS? <- YES
     ASSERT CLAWS
   ASSERT FELINE
  STRIPES
   R6
    LIGHT FUR
     QUESTION 5 LIGHT FUR? <- NO
   QUESTION 6 STRIPES? <- YES
   ASSERT STRIPES
 ASSERT TIGER
```

You can read off the order of questions above. Note that since there is a rule for stripes that's tried first, the question is asked only after the rule fails.

Problem 6 Elvis Lives!

Part A

```
P(E=e|T=t) = P(E=e, T=t) =            P(T=t|E=e) P(E=e)
             -----------    --------------------------------------
               P(T=t)        P(T=t|E=e) P(E=e)+ P(T=t|E=/e)P(E=/e)

            =         0.9 * 0.00001
              ----------------------------- = 0.000899
              0.9 * 0.00001 + 0.01 * 0.99999
```

Part B

Given two headlines:

```
P(E=e|T1=t1, T2=t2)
 = P(E=e, T1=t1, T2=t2)
   ---------------------------
   P(T1=t1, T2=t2)

 = P(T=t1|E=e)P(T=t2|E=e)P(E=e)
   ---------------------------------------------------------------
   P(T=t1|E=e)P(T=t2|E=e)P(E=e) + P(T=t1|E=/e)P(T=t2|E=/e)P(E=/e)

 = 0.075
```

For three headlines, the probability is 0.879.

Part C

For n sightings, the formula is:

$$\frac{(0.9)^n * 0.00001}{(0.9)^n * 0.00001 + (0.01)^n * 0.99999}$$

By the time we get to 500 headlines, the probability is indistinguishable from 1.

Part D

The problem is that this model assumes that each headline is independent evidence of Elvis appearing.

Part E

```
                         ---> Tech Headline
Elvis --> Headline       ---> Tech Headline
          chosen at Tech ---> Tech Headline
                         ---> Tech Headline
```

So, all of the headlines make us certain that the headline was chosen by The Tech but not that the sighting is real.