1 Question 1

Part 1 Without knowledge, the expected utilities of the candidates for Party A are:

- $U(NewMexico) = 0.6 * 90 + 0.4 * 70 = 82$
- $U(Indiana) = 0.6 * 100 + 0.4 * 30 = 72$
- $U(Virginia) = 0.6 * 20 + 0.4 * 80 = 44$

Party A would therefore choose the candidate from New Mexico and the expected utility is 82.

With knowledge of the candidate of party B, the utility for Party A is 100 if party B chooses the candidate from Connecticut, and 80 if party B chooses the candidate from Massachusetts. Thus the expected utility is $0.6 * 100 + 0.4 * 80 = 92$.

The expected value of knowing which candidate party B is going to select is $92 - 82 = 10$.

Part 2 The candidate from New Mexico. See above.

Part 3 If party B chooses the candidate from Connecticut, party A would choose the candidate from Indiana.

If party B chooses the candidate from Massachusetts, party A would choose the candidate from Virginia.

2 Question 2

$A : \{1, 2, 3\}$ $B : \{2, 3, 4\}$ $C : \{2, 3, 4\}$

3 Question 3

Part (a)

- $P(A) = 0.112 + 0.448 + 0.028 + 0.012 = 0.6$
- $P(B) = 0.112 + 0.448 + 0.048 + 0.192 = 0.8$
- $P(C) = 0.112 + 0.028 + 0.048 + 0.112 = 0.3$

Part (b)

- $P(A, B) = 0.112 + 0.448 = 0.56$ but $P(A)P(B) = 0.48$. A and B are not independent.
- $P(B, C) = 0.112 + 0.048 = 0.16$ but $P(B)P(C) = 0.24$. B and C are not independent.
- $P(A, C) = 0.112 + 0.028 = 0.14$ but $P(A)P(C) = 0.18$. A and C are not independent.

Part (c)

- $P(A, B|C) = P(A, B, C)/P(C) = 0.112/0.3 = 0.373$, but $P(A|C)P(B|C) = (P(A, C)/P(C)) * (P(B, C)/P(C)) = (0.14/0.3) * (0.16/0.3) = 0.249$.
- $P(B, C|A) = P(A, B, C)/P(A) = 0.112/0.6 = 0.187$, but $P(B|A)P(C|A) = (P(A, B)/P(A)) * (P(A, C)/P(A)) = (0.56/0.6) * (0.14/0.6) = 0.218$.
- $P(A, C|B) = P(A, B, C)/P(B) = 0.112/0.8 = 0.14$, and $P(A|B)P(C|B) = (P(A, B)/P(B)) * (P(B, C)/P(B)) = (0.56/0.8) * (0.16/0.8) = 0.14$.

A and C are independent given B.

Part (d) In graphs 1, 3, 6, 7, 8, 9, 10, A and C are not independent given B. In graph 4, A and C are independent. Only graphs 2 and 5 are consistent with our observations above.

4 Question 4

Let $w_1, ..., w_n$ be the words of a sentence and $t_1, ..., t_n$ be its part-of-speech tags. We assume that there are T kinds of tags, and that $P(t_1, ..., t_n | w_1, ..., w_n) = \prod_{i=1}^{n} P(t_i | w_i, t_{i-1})$.

The probabilities may be estimated by Maximum Likelihood: $P(t_i | w_i, t_{i-1}) = \frac{count(w_i, t_{i-1}, t_i)}{count(w_i, t_{i-1})}$

A dynamic programming algorithm can be used to find the best tag sequence, with the a table where $\pi[i, j]$ is defined to be the maximum probability of a partial tag sequence ending at the i^{th} tag being assigned to t_j, $i = 1, ..., n$ and $j = 1, ...T$.

For each of the nT entries in the table, the algorithm needs to consider each of the T possible tags for the previous word. The time complexity is thus nT^2.

5 Question 5

Part (1a) Time complexity is equivalent to doing two breadth-first searches on a tree of height $\frac{d}{2}$ with branching factor b, plus the constant time needed for testing intersection. The time complexity is thus $O(2b^{\frac{d}{2}})$.

Part (1b) Space complexity is also equivalent to doing two breadth-first searches on a tree of height $\frac{d}{2}$ with branching factor b. The space complexity is thus $O(2b^{\frac{d}{2}})$.

Part (2) If one direction is performed using a breadth-first search, we can use depth-first search for the other direction, since we know that the depth of the tree is finite, and that we can immediately return the path when we reach the frontier of the search from the other direction.

Part (3) Yes. The successor function for the opposite direction is straightforward: for state n, the successor is $\frac{n}{2}$ if n is even and $\frac{n-1}{2}$ is n is odd. Thus, the two searches can be conducted simultaneously, e.g., using depth-first search until their frontiers meet, as before.

6 Question 6

Part (1) True. Assuming each node has only two children, a decision tree of height one is equivalent drawing a straight line to separate data points.

Part (2) False. Consider a feature space with two data points with different labels. It is separable by a decision of height 1 but not separable by 1-nearest neighbor.

Part (3) True. Given enough partitions, a decision tree can separate any feature space, except when there is inconsistent data, which would also not be separable by a perceptron.

Part (4) Duplicated.

Part (5) False. Consider a feature space with a cluster of 3 $+$ data points and a cluster of 2 $-$ data points that are far from one another. This space is separable by 1-nearest neighbor but not by 2-nearest neighbor.

Part (6) True. If each leave in the tree of height k contains only data points from one label, it will also contain only data points from the same label when it is further split.

7 Question 7

The output is $w = (8)(-1)\langle 1,2,3\rangle + (3)(-1)\langle 1,4,1\rangle + (2)(+1)\langle 1,7,4\rangle + (3)(+1)\langle 1,5,4\rangle = \langle -6,1,-7\rangle$.

8 Question 8

Part (a) The "commute" feature. The ratio $R_{commute}(\geq 1hr, TB)/R_{commute}(\geq 1hr, noTB) = \frac{8}{16}/\frac{20}{87}$ is greater than the corresponding ratios of all other features.

Part (b) Given that the patient does not live in an overcrowded condition, the new $S(1)$ is $0.004 * R_{crowded}(0,1) = 0.004 * \frac{11}{11+6} = 0.00259$, and the new $S(0)$ is $0.01 * R_{crowded}(0,0) = 0.01 * \frac{75}{50+75} = 0.006$.

Part (c) No, since $R_{crowded}(0,1) = \frac{11}{11+6} > \frac{75}{50+75} = R_{crowded}(0,0)$.

Part (d) Please ignore the requirement that the numerical entries be more consistent with the intuition.

Values along the lines of the following would tip the balance in favor of a prediction of TB in the case in part (b).

	crowded	not crowded
TB	1	9
not TB	9	1

Part (e) Multiply $P_0(TB)$ to $S(0)$ and $P_0(no\_TB)$ to $S(1)$.