True/False and Multiple Choice (16 points)

1. For the matrix below, indicate whether it is ultrametric, additive, both or neither.

   (a) Ultra-metric
   (b) Additive
   (c) Both
   (d) Neither

\[
\begin{array}{cccc}
  & a & b & c & d \\
 a & 0 & 2 & 8 & 8 \\
b & 2 & 0 & 8 & 8 \\
c & 8 & 8 & 0 & 4 \\
d & 8 & 8 & 4 & 0 \\
\end{array}
\]

2. True / False The minimum number of substitutions for this tree is 5.

3. True / False Given a set of labeled points, the dimension of highest total variance gives the best projection for classification. Justify or give a counter example.

4. True / False SVD will choose the axis indicated below (the dark line) as the first to project upon.

5. True / False When performing protein BLAST with window size \( w \), a matching target and query sequence must have at least one contiguous substring of length \( w \) in common.

6. True / False The \( k \)-means algorithm always globally minimizes the sum of squared errors.
7. **True / False** Hierarchical clustering using average-link, complete-link and single-link always results in the same clustering when the distance matrix is ultrametric.

8. **True / False** EM for motif finding is analogous to Baum-Welch for HMM training, and Gibbs sampling is analogous to Viterbi training.

**Short Answer (34 points)**

9. When scanning for motifs, give an advantage and a disadvantage of using a hash function over a lookup table for storing motif counts.

10. What are the running times for sequence alignment using linear, affine and general gap penalties? What is the justification for using affine gap penalties?

11. Within our framework for sequence alignment, what algorithm (global or local) and costs (gap, match and mismatch) would you use to find:

   (a) Minimum edit distance (i.e. the minimum number of insertions, deletions and substitutions needed to transform one string into the other).

   (b) Length of the longest common substring.
(c) Hamming distance (i.e. number of positions at which two contiguous and equal length strings are different).

12. Give an advantage and a disadvantage of posterior decoding over Viterbi for determining the hidden state at each position.

13. Give an example of two sequences with large hamming distance but small edit distance.

14. Consider the hash function
\[ h(x) = x_A + x_G \pmod{7}, \]
where \( x_i \) is the count of character \( i \) in \( x \). Give 5 DNA sequences of length 6 that have distinct hash values.

15. Give an example of a distance matrix that is not ultrametric, but meets the triangle inequality.
Practical Problems + Design (50 points)

16. Consider one iteration of the EM algorithm for a motif of length 3. Below we have provided you with a set of sequences and a $Z$ matrix ($Z_{ij}$ gives the probability that position $j$ in sequence $i$ is the start of the motif). Compute the next $M$ matrix (a position weight matrix representing the motif at the next iteration). Assume that all pseudocounts are 0 and that the background nucleotide distribution is uniform.

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 1</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>Sequence 2</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Sequence 3</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>G</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$Z$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 1</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Sequence 2</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Sequence 3</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$M$</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

17. Fill in the blanks in the distance matrix below so that the resulting matrix is ultrametric.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

18. Consider the graph below. Find the path where the sum of all the edges along the path is maximized. Indicate the path on the graph.

Hint: This is very similar to the Viterbi algorithm.
19. Consider the following SCFG (in Chomsky normal form) for RNA folding, which models sequences over the simplified alphabet \{A, U\}.

\[
\begin{align*}
S & \rightarrow AS | US | SA | SU | AX | UY | a | u \\
X & \rightarrow SU | u \\
Y & \rightarrow SA | a \\
A & \rightarrow a \\
U & \rightarrow u \\
\end{align*}
\]

Assume that you know the probabilities for each production rule (e.g. \(P(S \rightarrow AX)\), \(P(X \rightarrow u)\), etc.). Also assume that you have run the inside and outside algorithms so that for all \(i \leq j\) and for all nonterminals \(V\), you know \(\alpha(i, j, V)\) and \(\beta(i, j, V)\). Given a pair of positions \(w\) and \(z\), where \(w < z\), determine the posterior probability that the bases in those positions are paired.

Hint: The production \(S \rightarrow AX\) generates \(w \ldots z\) iff \(X\) generates \((w + 1) \ldots z\).

20. We have seen a number of algorithms that use the hamming distance \(D(x, y)\) between two strings \(x\) and \(y\), which treats every mismatch as equally likely, resulting in a fixed penalty/cost. However, in DNA sequences, transitions \((A \leftrightarrow G, C \leftrightarrow T)\) are more likely than transversions (any other mismatch). Design a mapping \(f\) from sequences over \(S = \{A, C, G, T\}\) to sequences over an augmented alphabet \(S' = \{A, C, G, T, +, -\}\), so that the hamming distance \(D(f(x), f(y))\) effectively penalizes transitions with a cost of 1 and transversions with a cost of 2. Show how your mapping would compare sequences GTC and AGC and the resulting score.