COMP 7650 Data Mining and Knowledge Discovery

Assignment 1

September 24th

Due date: October 4th

Instructions:

This assignment is to be submitted online as a single PDF document. Your submission will be confirmed within 5 minutes via Email and online. Submission is to be done via the link associated with Assignment 1 at the Web site

http://www.comp.hkbu.edu.hk/~markus/teaching/comp7650/

Multiple submissions can be made but only the last submitted version will be assessed. Submissions made after the due date and time will not be assessed (no late submissions allowed). The PDF document should have a header containing your full name and your student ID. There is a file size limit of 512KB for submitted material. This means that your PDF file should not exceed 512KB in size. This is an individual assignment! Plagiarism will result in having 0 marks for all students involved.

1. Given two feature vectors, $x_s = (0.1, 0.4, 0.2, 0.5)^T$ and $x_r = (0.3, 0.5, 0.1, 0.9)^T$, calculate the following similarity and distance values and explain the formula on how to calculate these values:

(a) Euclidean distance

(b) Cosine distance

(c) Pearson's correlation

(d) City block distance (Manhattan distance)

(e) Minkowski distance (r = 1 and 2)

(f) Kullback-Leibler divergence (not a metric)

2. Describe the strengths and weaknesses of the following four classifiers: Nearest neighborhood classifiers, naïve Bayes classifiers, Gaussian mixture models and hidden Markov models. What assumptions we made when we use these classifiers?

3. Suppose we have three categories with $P(\omega_1) = 3/4, P(\omega_2) = 1/4$ and the following distributions

- $p(x \mid \omega_1) \sim N(0,1)$
- $p(x \mid \omega_2) \sim N(0.5,1)$

and that we sample the following two points: $x = 0.6, 0.1$

(a) Calculate explicitly the probability that the sequence actually came from ω_1, ω_1. Be careful to consider normalization.

(b) Repeat for the sequence ω_1, ω_2

(c) Find the sequence having the maximum probability

4. Suppose we have the left-right HMM with five hidden states, $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, where ω_1 and ω_4 are non-emitting starting and ending states. The emitting probability of hidden states, ω_2, ω_3, are

- $p(x \mid \omega_2) \sim N(0,1)$
- $p(x \mid \omega_3) \sim N(0.5,1)$

and the transition probabilities are

$$a_{12} = 1, a_{22} = 0.4, a_{23} = 0.6, a_{33} = 0.7, a_{34} = 0.3, a_{44} = 1$$

with other transition probabilities zeros.

We also have a sequence of observations $O = \{0.1, 0.6, 0.3\}$.

(a) The evaluation problem: calculate the conditional probability $P(O \mid \Omega)$ using the forward-backward algorithm.

(b) The decoding problem: find the single "best" state sequence using the Viterbi algorithm.