COMP 7650 Data Mining and Knowledge Discovery

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 formula on how to calculate these values:

(a) Euclidean distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

0.4690

(b) Cosine distance

$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\|$,

0.9583

(c) Pearson's correlation

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

0.8552

(d) City block distance (Manhattan distance)

$$dist = \sum_{k=1}^{n} | p_k - q_k |$$

0.8000

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

$$\boldsymbol{dist} = \left(\sum_{\boldsymbol{k}=1}^{\boldsymbol{n}} | \boldsymbol{p_k} - \boldsymbol{q_k} |^{\boldsymbol{r}}\right)^{\frac{1}{\boldsymbol{r}}}$$

0.8000 and 0.4690

Note that when r = 1, the Minkowski distance becomes the city block distance. When r = 2, the Minkowski distance becomes the Euclidean distance.

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

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

KL(x_s,x_r) = -0.3544 and KL(x_s,x_r) = 0.9008

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?

Answer:
(1) Nearest neighborhood classifiers
Strengths: non-parametric (no parameter estimation); simple and easy to implement
Weaknesses: require large data storage; high computational cost (calculate similarity with all stored samples)
Assumption: the majority of neighbors can determine the sample's class label

(2) Naïve Bayes classifiers
Strengths: simple parameter estimation procedure; simple and easy to implement
Weaknesses: the conditional independence assumption of attributes may deteriorate the classification performance
Assumption: the attributes are conditionally independent in terms of the class label

(3) Gaussian mixture models
Strengths: a relatively simple parameter estimation procedure based on the expectation maximization algorithm; arbitrary density modeling ability with enough mixture components
Weaknesses: the number of mixture components is difficult to determine; random initialization may affect the classification performance
Assumption: according to a certain metric, the data have different densities in different classes

(4) Hidden Markov models

Strengths: a systematic parameter estimation procedure; arbitrary sequential data modeling with enough hidden states
Weaknesses: the number of hidden states and the meaning of hidden states are not easy to determine; the computational cost may be high if the topology of HMMs is complex
Assumption: the left-right HMMs assume that signals are piecewise stationary.

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 four 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.

Answer: there are 4 combinations of categories for this sequence, so we need to calculate the total probability first as a normalization factor (You can program to achieve this goal).

P_total = [N(0.6; 0,1)*3/4]*[N(0.1; 0,1)*3/4] + [N(0.6; 0,1)*3/4]*[N(0.1; 0.5,1)*1/4] + [N(0.1; 0,1)*3/4]*[N(0.6; 0.5,1)*1/4] + [N(0.6; 0.5,1)*1/4]*[N(0.1; 0.5,1)*1/4]
= 0.1354

P(w1, w1) = [N(0.6; 0,1)*3/4]*[N(0.1; 0,1)*3/4] / P_total
= 0.5494

(b) Repeat for the sequence ω_1, ω_2

P(w1,w2) = 0.1699

(c) Find the sequence having the maximum probability

Because P_total is fixed for all combinations, we only need to compare the numerator. As a result, the best sequence is ω_1, ω_1

4. Suppose we have the left-right HMM with four 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 $P(O|\Omega)$ using the forward-backward algorithm.

Hints: build up the forward variable "alpha" and the backward variable "beta" matrices according to the forward-backward algorithm.

(1) Forward variable: alpha matrix

0.3970	0.0529	0	
0	0.0946	0.0383	0.0115

(2) Backward variable: beta matrix

0.0115	0.0235	0
0	0.0326	0.1173

$P(O|\Omega)$ = 0.0115

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

Hints: build up the "phi" matrix according to the Viterbi algorithm.

The best hidden state sequence is $\omega_1, \omega_2, \omega_3, \omega_3, \omega_4$.