BCB4003 / BCB503 Biological and Biomedical Database Mining. A Term / Fall 2011

Exam - October 13, 2011

Prof. Carolina Ruiz
Worcester Polytechnic Institute

NAME: _____

Problem I: _____ **(/25 points)** Data Preprocessing

Problem II: _____ **(/20 points)** Clustering

Problem III: _____ **(/20 points)** Bayesian Methods

Problem IV: _____ **(/25 points)** Markov Models

Problem V: _____ **(/20 points + 5 extra credit points)** Support Vector Machines

TOTAL SCORE: _____ **(/100 points)**

Instructions:

- Show your work and justify your answers
- Use the space provided to write your answers
- Ask in case of doubt

Problem I. Data Preprocessing [25 Points]

1. **[5 Points]** Attribute discretization is a process used to convert a numeric attribute into a nominal attribute. What's the difference between supervised discretization and unsupervised discretization? Explain.

2. Dimensionality reduction. Consider a dataset (similar to that in our Project 3) containing 16 numeric attributes, each corresponding to a dimer count over a set of DNA sequences, plus a Boolean target class. The following table illustrates what the dataset would look like:

AA	**AT**	**AC**	**…**	**GT**	**GC**	**GG**	**class**
64	5	22	…	5	12	77	*true*
53	21	6	…	10	4	8	*false*
…	…	…	…	…	…	…	

1. Assume that we apply Correlation based Feature Selection (CfsSubsetEval in Weka) to this dataset and that the method returns 5 attributes.
    a. **[5 Points]** What property does this set of 5 attributes satisfy?

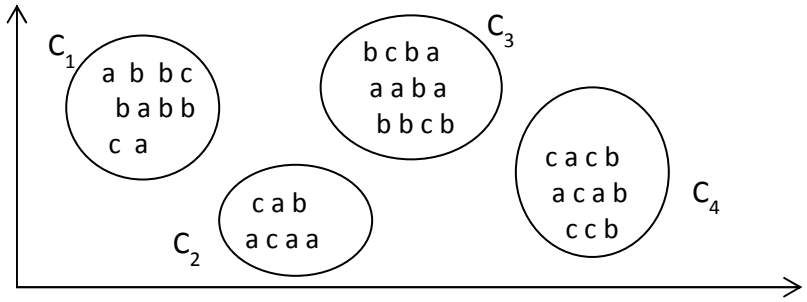
    b. **[5 Points]** Give an example of a possible set of 5 attributes returned by the method (just guess, this is for illustration purposes only).

2. Assume that we apply Principal Components Analysis (or equivalently, Singular Value Decomposition) to this dataset and that the method returns 5 attributes.
    a. **[5 Points]** What property does this set of 5 attributes satisfy?

    b. **[5 Points]** Give an example of a possible set of 5 attributes returned by the method (just guess, this is for illustration purposes only).

Problem II. Clustering [20 Points]

1. Consider the problem of evaluating clusterings. The *purity measure* (defined below, and used in our Project 1) provides an evaluation of a clustering with respect to a target class. Assume that we have a dataset of *n*=40 data instances with a target classification attribute T that has 3 values (*a, b,* and *c*). The figure below depicts a clustering Ω of the dataset into 4 clusters (C_1, C_2, C_3, and C_4).



Purity of this clustering Ω with respect to target T is defined as: purity(Ω,T) = $\frac{1}{n}\sum_{k=1}^{4} m_k$ where m_k is the number of instances in cluster C_k that belong to the majority class in that cluster C_k.

1. **[2 Points]** For example, *b* is the majority class in C_1. So, what's m_1 = ?

2. **[5 Points]** Calculate purity(Ω,T) for the figure above. Show your intermediate results.

3. **[5 Points]** What is the range of values for purity? Is a larger (or smaller) value for purity better?

4. **[3 Points]** What happens to purity when *k,* the number of clusters, increases? Is that good of bad? Explain.

2. **[5 Points]** Given a set of documents, we want to apply clustering to figure out topics and subtopics among those documents. What clustering technique would you use and why?

Problem III. Bayesian Methods [20 Points]

Consider a dataset with attributes A_1, A_2, …, A_n, C where C is the classification target.

1. **[5 Points]** Use Bayes theorem to write the right-hand side of the equation below:

   $P(C \mid A_1, A_2, \ldots, A_n) =$

2. **[5 Points]** In order to construct a classification model for this dataset, we want to find the value of the target attribute C that maximizes (i.e., argmax) the probability above. Describe which terms in the right-hand of the equation above should be kept and which can be ignored. Explain.

   $\underset{\text{values of C}}{\textbf{argmax}} \; P(C \mid A_1, A_2, \ldots, A_n) = \underset{\text{values of C}}{\textbf{argmax}} \; \ldots$

3. **[5 Points]** What's the naïve Bayes assumption? Describe it and also use it to simplify the formula above.

4. **[5 Points]** When constructing a naïve Bayes model over a dataset of documents (each represented as a bag of words). What does the naïve Bayes assumption mean in terms of the words in the documents? Explain.

Problem IV. Markov Models [25 Points]

1. Consider a Markov Chain (the states are observable).

a. **[5 Points]** What is known as the Markov property?

b. **[5 Points]** Given a sequence of states (= emissions) $x_1, x_2, \ldots, x_n$, express the following probability in terms of a product of "simpler" probabilities using the Markov property. Explain.

$P(x_1, x_2, \ldots, x_n) =$

2. Consider now Hidden Markov Models (HMMs). Construct a HMM to model the following over-simplified version of the problem of locating exons and introns in genes. Assume that the emissions are nucleotides ("1-mers").

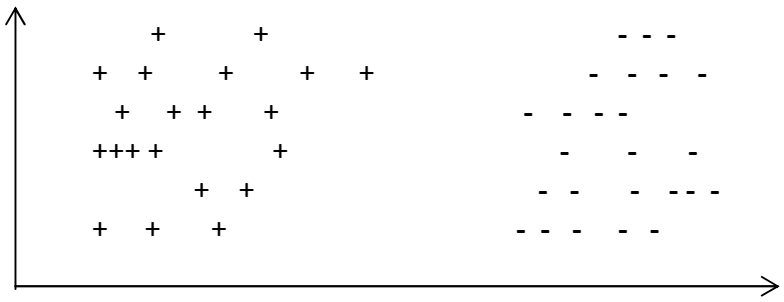
a. **[5 Points]** Depict the topology of a HMM with 2 states (one for exons and one for introns) for this purpose. Include state transitions and emissions explicitly.

b. **[5 Points]** What parameters need to be added to the above topology to complete the construction of the HMM?

c. **[5 Points]** How would you find appropriate values for those parameters? Explain.

Problem V. Support Vector Machines [20 Points + 5 extra credit Points]

1. Consider the dataset depicted in the figure below. The dataset contains two classes +1 (denoted by "+") and -1 (denoted by "-").



   a. **[5 Points]** Draw in the figure above the model that SVM would output for this dataset. Explain.

   b. **[5 Points]** What criterion is maximized in order to construct this model? Explain using words as well as drawings on the figure above.

   c. **[5 extra credit points]** What are the support vectors in the figure? Explain.

2. **[10 Points]** In the non-linear case depicted in the figure below, can you use the same method as above? If not, what do you do? Explain using words as well as drawings on the figure below.