Data Mining with R

final test structure

Hugh Murrell

test structure

These slides present the anticipated structure for the 2^{nd} and final data mining test.

To prepare for this test, study slide sets 6,7,8,9,10 and read the papers referenced in slides 10.

There will be 4 questions.

general machine learning, (slides 6-10)

a) Describe the *iris* and *wine* datasets and why they are useful for testing machine learning algorithms.

b) Describe why datasets are separated into *training* and *testing* subsets for the purpose of machine learning.

decision trees, (slides 6)

a) what is a decision tree?

b) define *entropy* and indicate how it is computed for a target variable consisting of k classes.

c) derive properties of the entropy function.

d) define *information gain* for a given sample split.

e) what is *recursive partitioning*?

f) why are practical decision tree algorithms *greedy*?

g) how are decision trees implemented in R? (name the package functions)

h) write down a sentence or two describing *random forests*.

the perceptron, (slides 7)

a) understand the role played by the parameters b and $\vec{w}$ in a linear separator.

b) be able to write down the perceptron algorithm for producing a separator for separable data

c) understand the convergence proof for the perceptron algorithm

d) have ideas on how to extend the perceptron algorithm for handling non-separable data.

linear discrimination, (slides 8)

a) given a set of points in feature space, define the *scatter* matrix and its *action* on a unit vector $\vec{w}$.

b) Given a set of feature vectors from two classes, define their *linear Fisher discriminant* and explain how to go about finding it.

c) What is *Mahalonobis separation* and what role does it play in constructing a Fisher discriminant?

d) How is *linear discrimination* implemented in R? (name the functions)

clustering, (slides 9)

a) describe the *k-means* and *k-medoids* algorithms for clustering.

b) discuss two different distance measures on observations

c) how is clustering implemented in R

d) describe *hierarchical clustering* and in particular how distances between clusters are computed when building the hierarchy.

text mining, (slides 10)

a) what is a *corpus*?

b) describe 4 procedures that are commonly used for cleaning a corpus.

c) what is a *term-document matrix* and what is it used for?

d) outline a text mining example where **sentiment** plays a role.

one last machine learning exercise to be completed before you write test 2 ...

- go to the Computer Science admin assistant on your campus
- ask how to complete the COMP717 course evaluation
- grade your instructor.

thank you, goodbye and goodluck