CSC 462 Project 1: Supervised Learning

Due: 05-11-2014 1pm.

In this project, you will apply two algorithms to two data sets. The algorithms and the data are included in WEKA. Email your report as a pdf attachment before on due date.

The Data Sets

1. Contact Lens data
    - $WEKA_HOME/data/contact-lenses.arff
2. Iris data
    - $WEKA_HOME/data/iris.arff

The Algorithms

1. K-Nearest Neighbors
    - classifiers/lazy/IBk
2. Decision Trees
    - classifiers/trees/Id3
    - classifiers/trees/J48

Data Sets

What are the differences between the two data sets?

Which algorithm do you expect to perform best on the Contact Lens data? Why?

Which algorithm do you expect to perform best on the Iris data? Why?

KNN on the Contact Lens Data

Run KNN on each data set with 1, 3, 5, 7 and 9 neighbors. Report the results for each run in a confusion matrix and comparisons in a table or graph.

Which K gives the best results? Why?

Holding K constant, try different distance functions on each data set. Which distance function(s) work best for each data set? Why?

KNN on the Iris Data

Run KNN on each data set with 1, 3, 5, 7 and 9 neighbors. Report the results for each run in a confusion matrix and comparisons in a table or graph.

Which K gives the best results? Why?

Holding K constant, try different distance functions on each data set. Which distance function(s) work best for each data set? Why?

Decision Trees on the Contact Lens Data

Based on Weka's vizualizations, which attribute do you expect to be chosen as the split attribute at the root node?

Run each decision tree on the data and report the results for each run in a confusion matrix and comparisons in a table or graph.

How do ID3 and J48 compare in terms of performance?

How does pruning affect test performance and generalization performance? What does that suggest about overfitting?

Decision Trees on the Iris Data

Based on Weka's vizualizations, which attribute do you expect to be chosen as the split attribute at the root node?

Run each decision tree on the data and report the results for each run in a confusion matrix and comparisons in a table or graph.

Why can't you run ID3 on the Iris data?

How does pruning affect test performance and generalization performance? What does that suggest about overfitting?

General

Which algorithm performed best on each data set, for particular definitions of "best?"

Was the (comparative) performance of the algorithms as you expected? Why?

Which data set had the best performance in general across all of the algorithms? Why?

The Report

In your report, answer the questions in your report as they come in the project. Do not use the results to answer the questions; you will have room to analyze the result.