CS199:

Classification of Bioinformatics Research Papers

Shashank Khanna

University of California, Los Angeles

Centre of Excellence for Big Data Computing at UCLA

Department of Bioinformatics

University of California, Los Angeles

Contents

[1 Description of the Problem 1](#_Toc474089588)

[1.1 Background 1](#_Toc474089589)

[1.2 Problem Statement 1](#_Toc474089590)

[2 Part 1 : Algorithms 2](#_Toc474089591)

[2.1 Decision Tree 2](#_Toc474089592)

[2.1.1 Algorithm Summary 2](#_Toc474089593)

[2.1.2 Application to the classification problem 2](#_Toc474089594)

[2.2 SVM (Linear/RBF) 2](#_Toc474089595)

[2.2.1 Algorithm Summary 3](#_Toc474089596)

[2.2.2 Application to the classification problem 3](#_Toc474089597)

[2.3 Nearest Neighbors 3](#_Toc474089598)

[2.3.1 Algorithm Summary 3](#_Toc474089599)

[2.3.2 Application to the classification problem 3](#_Toc474089600)

[2.4 Neural Networks 3](#_Toc474089601)

[2.4.1 Algorithm Summary 3](#_Toc474089602)

[2.4.2 Application to the classification problem 3](#_Toc474089603)

[2.5 Random Forest 4](#_Toc474089604)

[2.5.1 Algorithm Summary 4](#_Toc474089605)

[2.5.2 Application to the classification problem 4](#_Toc474089606)

[2.6 Naïve Bayes 4](#_Toc474089607)

[2.6.1 Algorithm Summary 4](#_Toc474089608)

[2.6.2 Application to the classification problem 4](#_Toc474089609)

[3 Testing algorithms on Iris Training Data 5](#_Toc474089610)

[3.1 Decision Tree 5](#_Toc474089611)

[3.1.1 Algorithm Summary 5](#_Toc474089612)

[3.1.2 Application to the classification problem 5](#_Toc474089613)

[3.2 SVM (Linear/RBF) 5](#_Toc474089614)

[3.2.1 Algorithm Summary 5](#_Toc474089615)

[3.2.2 Application to the classification problem 6](#_Toc474089616)

[3.3 Nearest Neighbors 6](#_Toc474089617)

[3.3.1 Algorithm Summary 6](#_Toc474089618)

[3.3.2 Application to the classification problem 6](#_Toc474089619)

[3.4 Neural Networks 6](#_Toc474089620)

[3.4.1 Algorithm Summary 6](#_Toc474089621)

[3.4.2 Application to the classification problem 6](#_Toc474089622)

[3.5 Random Forest 6](#_Toc474089623)

[3.5.1 Algorithm Summary 7](#_Toc474089624)

[3.5.2 Application to the classification problem 7](#_Toc474089625)

[3.6 Naïve Bayes 7](#_Toc474089626)

[3.6.1 Algorithm Summary 7](#_Toc474089627)

[3.6.2 Application to the classification problem 7](#_Toc474089628)

[4 Feature Identification: Creating The Training Set 8](#_Toc474089629)

[4.1 Decision Tree 8](#_Toc474089630)

[4.1.1 Algorithm Summary 8](#_Toc474089631)

[4.1.2 Application to the classification problem 8](#_Toc474089632)

[4.2 SVM (Linear/RBF) 8](#_Toc474089633)

[4.2.1 Algorithm Summary 8](#_Toc474089634)

[4.2.2 Application to the classification problem 9](#_Toc474089635)

[4.3 Nearest Neighbors 9](#_Toc474089636)

[4.3.1 Algorithm Summary 9](#_Toc474089637)

[4.3.2 Application to the classification problem 9](#_Toc474089638)

[4.4 Neural Networks 9](#_Toc474089639)

[4.4.1 Algorithm Summary 9](#_Toc474089640)

[4.4.2 Application to the classification problem 9](#_Toc474089641)

[4.5 Random Forest 9](#_Toc474089642)

[4.5.1 Algorithm Summary 10](#_Toc474089643)

[4.5.2 Application to the classification problem 10](#_Toc474089644)

[4.6 Naïve Bayes 10](#_Toc474089645)

[4.6.1 Algorithm Summary 10](#_Toc474089646)

[4.6.2 Application to the classification problem 10](#_Toc474089647)

[5 Extracting the Features And Producing Test Data 11](#_Toc474089648)

[5.1 Decision Tree 11](#_Toc474089649)

[5.1.1 Algorithm Summary 11](#_Toc474089650)

[5.1.2 Application to the classification problem 11](#_Toc474089651)

[5.2 SVM (Linear/RBF) 11](#_Toc474089652)

[5.2.1 Algorithm Summary 11](#_Toc474089653)

[5.2.2 Application to the classification problem 12](#_Toc474089654)

[5.3 Nearest Neighbors 12](#_Toc474089655)

[5.3.1 Algorithm Summary 12](#_Toc474089656)

[5.3.2 Application to the classification problem 12](#_Toc474089657)

[5.4 Neural Networks 12](#_Toc474089658)

[5.4.1 Algorithm Summary 12](#_Toc474089659)

[5.4.2 Application to the classification problem 12](#_Toc474089660)

[5.5 Random Forest 12](#_Toc474089661)

[5.5.1 Algorithm Summary 13](#_Toc474089662)

[5.5.2 Application to the classification problem 13](#_Toc474089663)

[5.6 Naïve Bayes 13](#_Toc474089664)

[5.6.1 Algorithm Summary 13](#_Toc474089665)

[5.6.2 Application to the classification problem 13](#_Toc474089666)

# Description of the Problem

## Background

Currently, Bioinformatics software tools are present in silos with no common endpoint which could serve as a one size fits all solution for researchers to access the software they need. This present condition, leads us to create a solution to collate all present bioinformatics software based resources and provide a common access point for them. In order to achieve that, all publications that could potentially contain software related to Bioinformatics, which is henceforth referred to a as a tool in this paper, would need to be inspected, and identified as a tool or a not a tool. Given the scale of this problem, it is physically impossible to individually study every single publication, which leads the problem towards a machine learning solution.

## Problem Statement

The issue of identifying publications as a tool or a non-tool can be looked upon as a traditional classification problem. Given that the publications can be modified to fit the correct consumption format, Python scripts can potentially be used to implement several possible classification algorithms.

# Part 1 : Algorithms

## Decision Tree

Decision Trees classify a pattern through a sequence of questions where the next question depends on the answer to the current question. The questions are asked in a true/false style. (MSU, 2017)

### Algorithm Summary

Classic decision tree works by building a tree and making a decision on a dataset corresponds to a traversal of the tree from root to one of the leaves which corresponds to the decision.

Given a set S, we can organise the test data into a tree where a class is applied to each test. Hence, we can continually split the training set into smaller subsets with the subsets defined by a particular property, until we have reached a node where there is no need to divide the set further and all the data corresponding to that node have the same class label.

### Application to the classification problem

The given problem is

## SVM (Linear/RBF)

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Nearest Neighbors

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Neural Networks

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Random Forest

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Naïve Bayes

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

# Testing algorithms on Iris Training Data

## Decision Tree

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## SVM (Linear/RBF)

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Nearest Neighbors

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Neural Networks

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Random Forest

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Naïve Bayes

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

# Feature Identification: Creating The Training Set

## Decision Tree

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## SVM (Linear/RBF)

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Nearest Neighbors

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Neural Networks

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Random Forest

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Naïve Bayes

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

# Extracting the Features And Producing Test Data

## Decision Tree

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## SVM (Linear/RBF)

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Nearest Neighbors

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Neural Networks

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Random Forest

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.

## Naïve Bayes

We require a bound copy and an electronic copy of your dissertation. .

### Algorithm Summary

We require two canvas bound copies with hard or soft cover. On the spine the lettering should be in the format:

### Application to the classification problem

These should be in Word or PDF format. It is useful to have it on a CD, but if the whole lot can be zipped to under 10MB then you can also email it to your tutor.