CST 383: Intro to Data Science
Dr. Glenn Bruns

Defining machine learning problems

Purpose. As a data scientist, you'll sometimes be given a data set and be told which variables are predictor variables, and which variable is the target variable. But often, part of your job will be to figure out which variables will be predictors and which variable will be the target. From one data set there can be many different machine learning problems! In this assignment you'll get practice in defining machine learning problems.

Instructions. We've discussed three kinds of machine learning problems: classification, regression, and cluster analysis. Classification and regression are examples of supervised learning problems, while cluster analysis is an example of an unsupervised learning problem. In classification problems, the target variable is categorical; in regression problems, the target variable is quantitative.

Let's look at some example data sets. All of the data sets are available in our class Google drive folder.

iris data set. The data shows various measurements of iris flowers, and in each case shows the species of iris. Download iris.csv from the link above to get the data. Here are the first rows of the data set:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.0	2.0	3.5	1.0	versicolor
6.2	2.2	4.5	1.5	versicolor
6.0	2.2	4.0	1.0	versicolor
6.0	2.2	5.0	1.5	virginica
6.3	2.3	4.4	1.3	versicolor
5.5	2.3	4.0	1.3	versicolor
5.0	2.3	3.3	1.0	versicolor
4.5	2.3	1.3	0.3	setosa
5.5	2.4	3.8	1.1	versicolor
5.5	2.4	3.7	1.0	versicolor
4.9	2.4	3.3	1.0	versicolor
6.7	2.5	5.8	1.8	virginica
6.3	2.5	4.9	1.5	versicolor

Here are some machine learning problems I could define from the data:

classification: predict the value of Species from the values of Sepal.Length and Petal.Length (in other words, we use Sepal.Length as predictors, and Species as the target)

regression: predict the value of Sepal.Length from Petal.Length and Petal.Width (Petal.Length and Petal.Width are the predictors, and Sepal.Length is the target)

cluster analysis: cluster the irises using the values of Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width (in other words, we use features Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width). In the cluster analysis it is normally required that all features are numeric.

Two important things to note:

- In supervised learning problems, you don't need to use all possible features as predictors. In the classification example, I could have also used Sepal.Width and Species as predictors, but chose not to. Similarly, in unsupervised learning problems you can leave out some features.
- In classification problems the target variable (the thing to predict) will be a categorical variable, but sometimes categorical variables are disguised as numbers. For example, in this data set the values of Species could have been provided as the numbers 1,2, and 3.

Your job is to look at some data sets and define possible classification, regression, and cluster analysis problems. For each problem below, list some features in the data set (in the case of a target, or response variable, list just one feature). Provide your answer by editing ml-problems.txt

chickwts data set. Please download chickwts.csv from the link above. This data set shows how feed supplements affected the growth of chickens. Weight is the chick weight; feed is the feed type.

classification problem

1. select one predictor: _____
2. select target: _____

regression problem

3. select one predictor: _____
4. select target: _____

cluster analysis problem

5. select a feature: _____

ToothGrowth data set. Please download ToothGrowth.csv from the link above. This data set shows the effect of Vitamin C on the tooth growth in guinea pigs. 'len' is tooth length, 'supp' is supplement type, and 'dose' is the dose in milligrams.

classification problem

6. select two predictors: _____
7. select target: _____

regression problem

8. select one predictor: _____
9. select target: _____

cluster analysis problem

10. select two features: _____

mtcars data set. Please download mtcars.csv from the link above. This data set is data about cars taken from the 1974 Motor Trend car magazine. See

https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html for details on what the column names mean.

classification problem

11. select two predictors: _____
12. select target: _____

regression problem.

13. select three predictors: _____
14. select target: _____

cluster analysis problem.

15. select three features: _____

Submission. Submit your edited ml-problems.txt on iLearn.

Grading. Each problem is worth 5 points.