BUSINESS ANALYTICS CLUB

Workshop Series 9.12

Principles of Data Mining
Applications in Consulting

Janet Ye
Shantanu Joshi

BAC

Set Ups

- Download the zip file from bit.ly/bacdata "Weka Consulting" folder. In the zip file, you'll find:
  - Weka installer
    - PC: weka-3-6-13 jre.exe
    - Mac: Weka-3-6-12-oracle- jvm.app
  - Installation Guide
  - Data set

BAC

Learning Objective

1. Understand Data Mining and learn the use cases in various industries
2. Present the problem, and approach the consulting case as a data consultant
3. Introduce most common algorithms:
   - J48 Trees
   - Logistic Regression
   - k Nearest Neighbors (kNN)
4. Formulate a proposal to the case

BAC

Data Science? Data Mining?

Data science: set of fundamental principles that guide the extraction of knowledge from data

Data mining: the act of extracting knowledge from data, via technologies that incorporate these principles

Data Science and data mining have nothing to do with acquisition of the data!

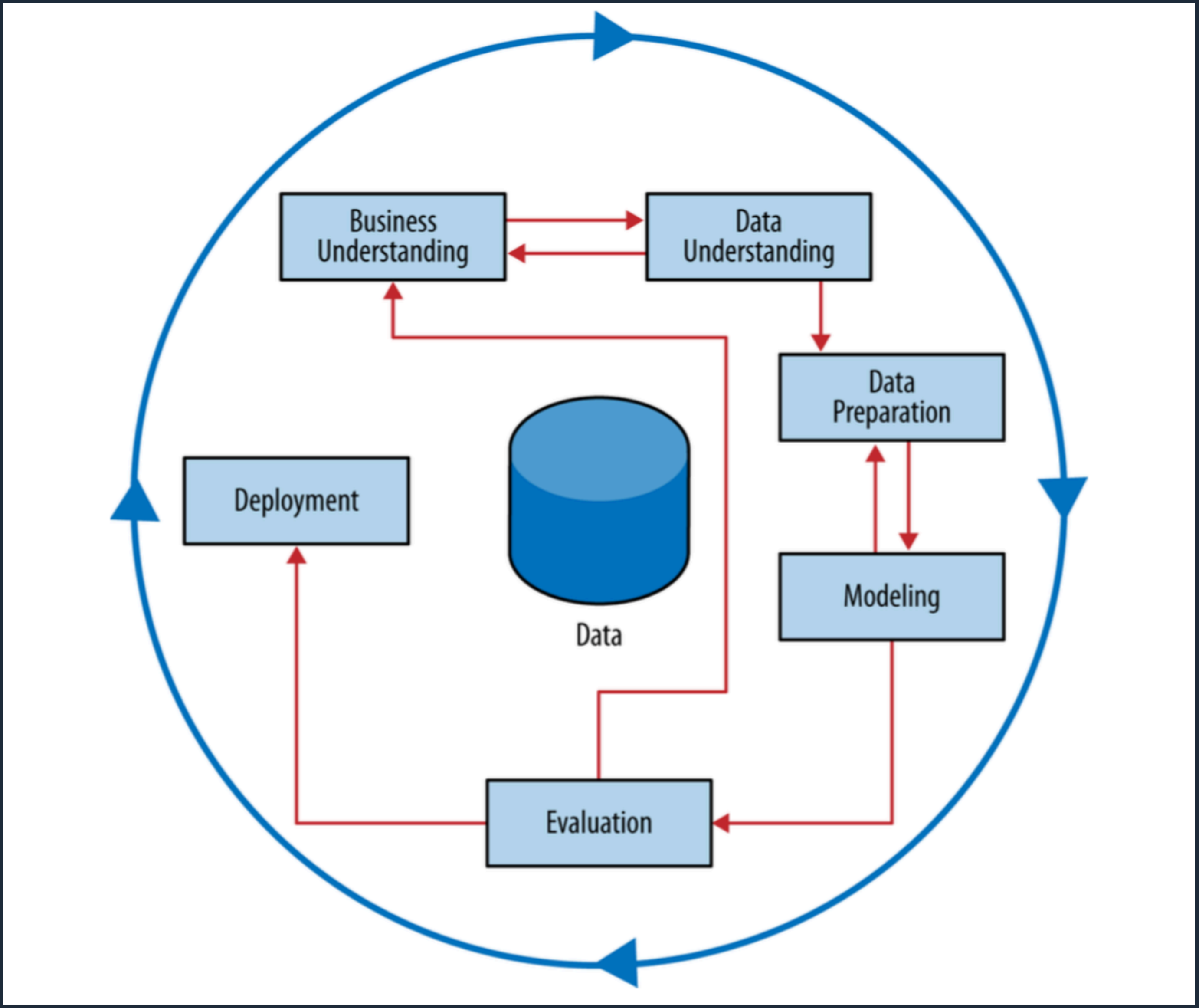
BAC

Solving Business Problems

- NYU Langone – benign or cancerous tumor?
- Apple – increase profit from the iPad Pro?
- Macy's –bundling items?
- Verizon – cell phone usage profiling?
- Google – natural language processor for searching?
- Facebook / LinkedIn – friends you may know?
- Amazon / Netflix / Spotify – product recommendations?

BAC

Data Mining Cycle



BAC

Business Understanding

Your client is an auto dealer that buys used cars from auctions and repairs them for resale. Recently, the firm has been purchasing a record amount of cars. A lot of these cars turned out to be lemons, and the firm's profits have been going down.

What can you do for the firm?

BAC

Data Understanding

- The firm says it has a large cache of data on each car, and which cars have been lemons
- An average consultant would do something like:
  - Segment the data by car type / make / model
  - Might use Excel, plot some graphs, and make assumptive statements on probabilities of certain segments
- How do we do better?
  - As a data consultant, we can use data mining techniques to build a classification model (good cars vs. lemons)

Data Understanding (ctd)

- Example variables:
  - auction info
  - vehicle year
  - vehicle age
  - make
  - color
  - transmission
  - wheel type
  - etc.

Data Preparation

- Weka accepts .arff files
- See appendix for converting a .csv to .arff file

Modeling

- Target variable (y)
- Features (x)
- Instance: feature + target (x, y)

- Supervised
  - Develop a model using an example dataset with both features and target variable, called training set
  - Example: have dataset on cars and whether they were good or lemons. How to better classify?

- Unsupervised
  - Build a model using a dataset with only the features, no target variable
  - Example: find patterns in data with no objectives

BAC

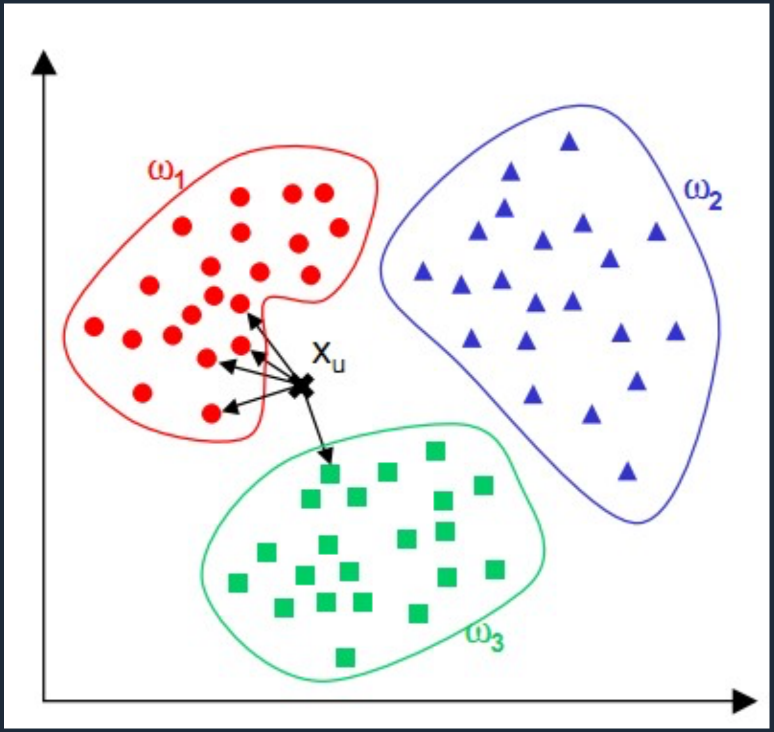
Today we focus on three models.

J48 (Decision Tree)

- Divide the data using the most informative attribute into two sets/branches
- Subdivide the sets using other variables as many times as we would like
- Finally, group the instances into good cars vs. lemons using a chosen metric

kNN (k Nearest Neighbor)

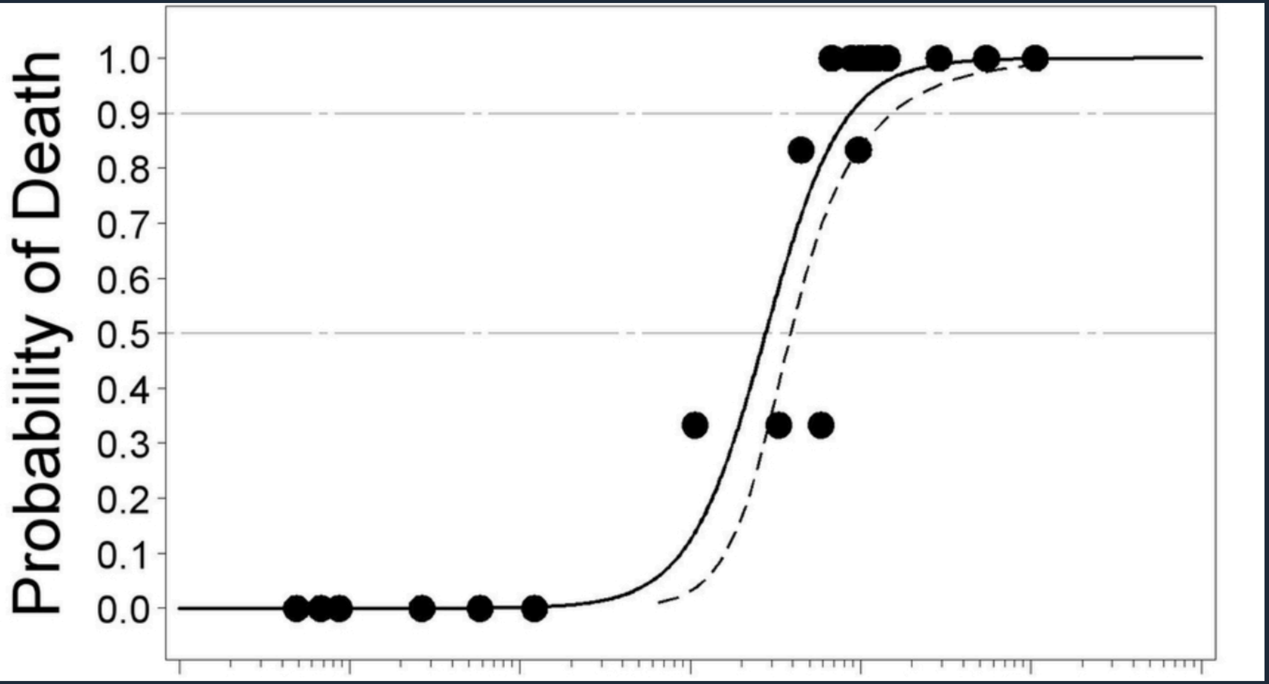
- Form groups of k instances/neighbors
- To predict a new instance, we find its “nearest” neighbors



BAC

Logistic Regression

- Fits a “sigmoid” curve to a special data set with a threshold separating the two cases in a binary outcome
- Examples
  - young vs. old smokers having cancer
  - good cars vs. lemons



BAC

Finding the Best Model

- The metric we use to define "best" is AUC –area under receiver operating characteristic (ROC) curve
- Target variable is assigned as:
  - good car = 0, lemon = 1
- AUC area is the probability that a randomly chosen lemon ranks higher than a randomly chosen good car
  - Perfect model = 1
  - Random model = 0.5
  - 0 < bad model < 0.5

BAC

Finding the Best Moel

- To generate AUC, we adjust the complexity parameter for each model (how simple or complicated a model is)
- Tree – minimum number of objects in a leaf (minNumObj)
  - Simpler model = fewer branches
  - If minNumObj is a large number, we have a lot of instances in one leaf, thus few branches
- Logistic Regression – Ridge parameter
  - Simpler model = larger ridge
- kNN – adjust k
  - Simpler model = larger k
  - If k large, group a lot of instances in one neighborhood, fewer clusters

J48 Complexity

- Click [Explore]
- Open [File]
- Choose [car_data.arff]
- Change to [Classify] tab
- Click [Choose] under classifier
- Under [Trees] look for [J48] and click
- Click the text that says [J48 –C 0.26 –M 2]
- A dialogue box should open with options
- Here we can adjust settings of model
  - Change [Unpruned] to [True]
  - [minNumObj] to 2

BAC

J48 Complexity

- Under [Test Options], select [Percentage Split] and make sure it is set to [66%]
- Make sure the dropdown menu under [Test Options] has [(Nom) IsBaBuy]
- Click [Start]
- In the output, scroll down to [Detailed Accuracy by Class]
- Under wich you will find the AUC for this model (in this case 0.636)
- Save this value in a table
- Increase complexity by factors of 2 (2^0, 2^1, 2^2, …, 2^12)
  - You can stop when AUC plateaus or hits 0.5

Word of Caution

- Note on [Test Options], select [Percentage Split] set to [66%]
- The biggest pitfall in data mining is overfitting
  - Building a model that fits well to the training data but fails to generalize
- Solution: 66% split
  - 66% of data, as training set, used for building a model
  - 33% of data, as test set, used for testing the model
  - Model is judged on how well it performs on test set

Complexity by Models

numMinObj	J48	Ridge	Logistic	Complexity	kNN
2	0.636	1e-4	0.761	4	0.691
4	0.660	1e-2	0.761	8	0.713
8	0.702	0.1	0.761	16	0.732
16	0.714	1	0.761	32	0.739
32	0.719	10	0.762	64	0.741
64	0.733	100	0.762	128	0.735
128	0.729	1,000	0.757	256	0.740
256	0.735	10,000	0.738	512	0.737
512	0.721	100,000	0.726	1024	0.734
1024	0.719	1,000,000	0.724	2048	0.726
2048	0.661			4096	0.706
4096	0.500			8192	0.500
8192	0.500				

Best model is Logistic Regression, Ridge = 100

Now what?

Working with Best Model

- Go back to weka [Explore], click [Choose] under [Classifier] tab
- Under [Functions], look for [logistic]
- Click the text [Logistic –R 1.0E-8 –M -1]
- Select [Percentage Split], make sure [66%] is the test method
- Click [More Options…] under [percentage split]
- Select the box for [Output Predictions]
- Now if you scroll up in [Classifier Output], you'll see probabilities for each instance of being a 0 or 1

Working with Best Model

- We can copy and paste these into Excel and start analyzing them
- We did this for you in "car_costs_and_probabilities.xls"
- The methodology to do this in the Appendix if you are curious

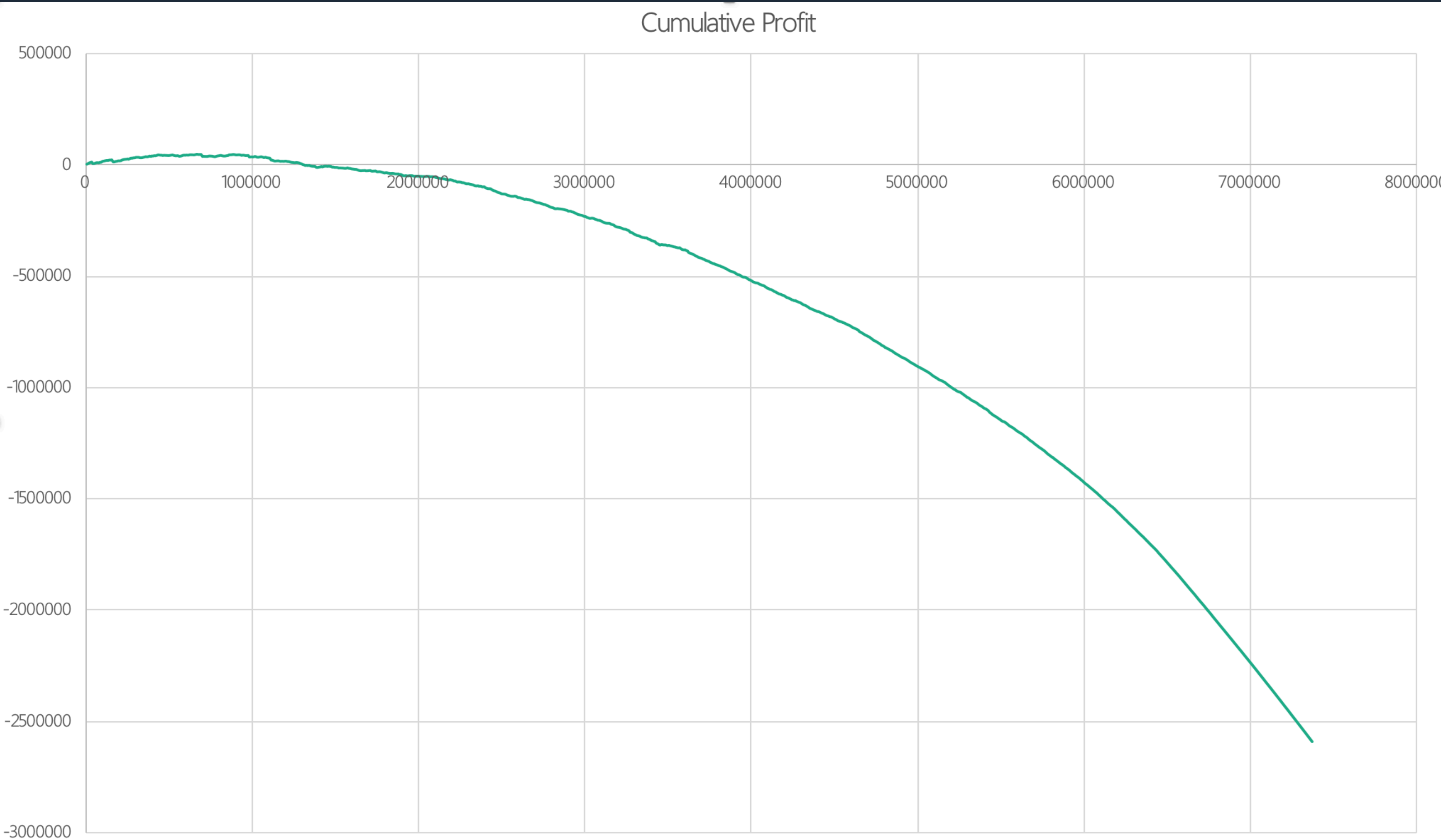
Excel Analysis

- Excel sheet includes costs for a certain car, and prices on the market if not a lemon
- Paste in output probabilities from Weka into Excel sheet
- Sort by P (not lemon) descending
- Calculate expected profit for each instance
  - Expected profit = Expected Retail Price * P(not a lemon) – cost to buy at auction
- Next, calculate cumulative costs
  - For instance 1, this is just cost at auction
  - For instance 2, this is cumulative costs in instance 1 + cost to buy instance 2 at auction
  - Apply second instance formula to rest of the data

Excel Analysis

- Apply the second instance formula to the rest of the data
- Calculate cumulative profit
  - For instance 1, this is just the expected profit
  - For instance 2, this is cumulative profit at instance 1 + expected profit of instance 2
- The graph should automatically adjust

Profit Curve



Profit Curve

- Let's interpret this graph…
- If we buy 0 cars, we spend $0 and make $0
- If we buy around 172 cars, we spend $1,305707, and make around $1408
- Above 172 cars, we start losing money
  - Purchase too many lemons, cut into profits

How many cars should we buy?

Profit Curve

- Ideally, we buy until our profit is maximized
- According to our model…
  - Max profit = $47,858
  - Car #88

- Recommendation: ideal budget is to buy 88 cars, which would cost us around $668,082.
- Other consideration: Should take client's financials into account

Deployment

- Let’s take this to our Clients
- Interpret
  - Ask client to extend engagement
  - Collect data on next 88 purchases and re-evaluate model

Useful Resources

- Free Stanford Machine Learning on Coursera
  - Link: https://www.coursera.org/learn/machine-learning
  - Blog with detailed write-ups
    http://www.holehouse.org/mlclass/index.html

Acknowledgement

Jessica Clark
PhD Candidate, Information,
Operations and Management Sciences

Appendix – Data Prep

- For data preparation we remove extraneous elements in excel
- Export excel to csv
- Use http://ikuz.eu/csv2arff/ to convert csv to arff
    - Make sure online converter correctly identifies numerical, categorical, and binary variables
- Edit the arff file in a text editor to make sure it confines to weka standard