ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 2

Learning

Classification by kNN

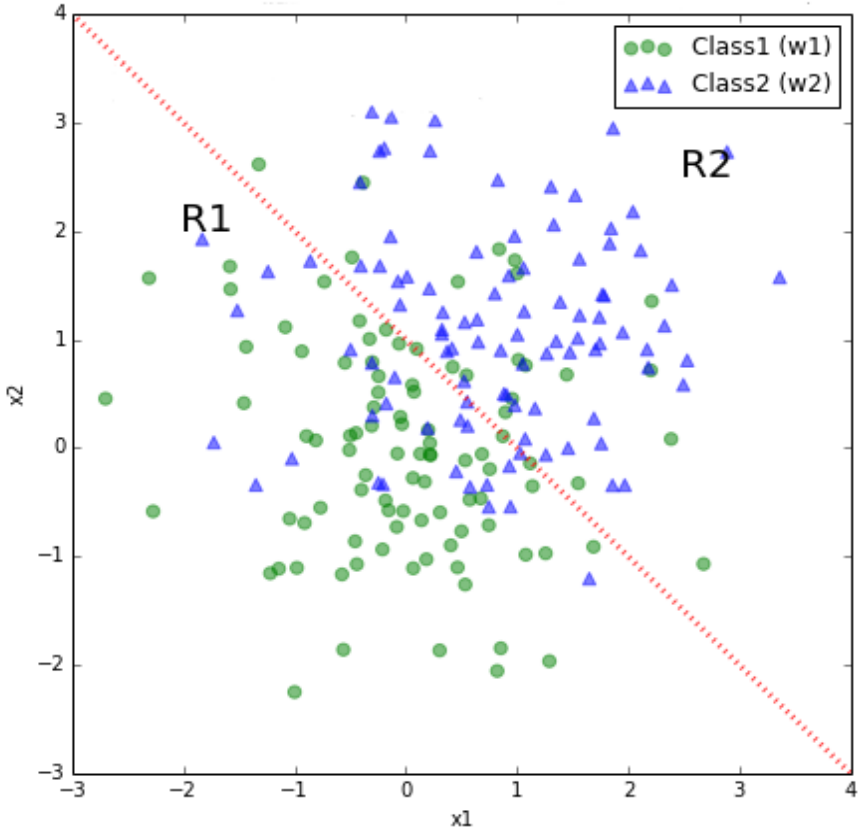
Professor: Dr. Anu Thomas

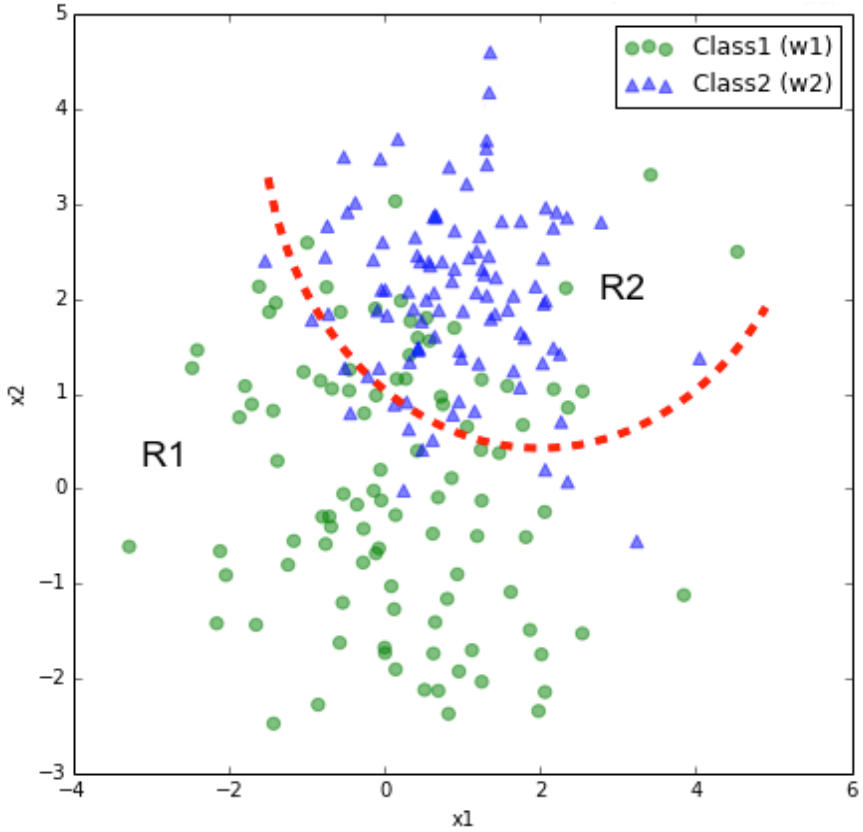
Learning

- Supervised learning – classification, regression
- Unsupervised learning – clustering, outlier detection
- Semi-supervised

Supervised learning: Classification

- Data has class labels
- Based on the labels, classifiers are generated
- New data will be classified based on the generated classifier
- Predicts a **discrete** class label
- Example 1: Cancer dataset – Malignant and benign labels are present for each instance.
- Example 2: Iris dataset – data from 3 types of flowers – every instance has a class label





https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html

Supervised learning: Regression

Regression predicts **continuous** values (numbers) as the output.

Example, housing prices for various houses: 2 bedroom, 3 bedroom, garage size, property size, and the computer must interpolate predictions.

Unsupervised learning

- data has no class labels

- The algorithm tries to identify the objects as being part of some group using a clustering algorithm. Similar instances grouped together to form clusters. (Ex. Insurance: Identifying groups of motor insurance policy holders with a high average claim cost)

- Anomaly detection tries to find those instances which are distinct from the nature of the majority of instances. (Ex. Financial fraud detection)

Semi-supervised learning

- Typically a small amount of labeled data with a large amount of unlabeled data

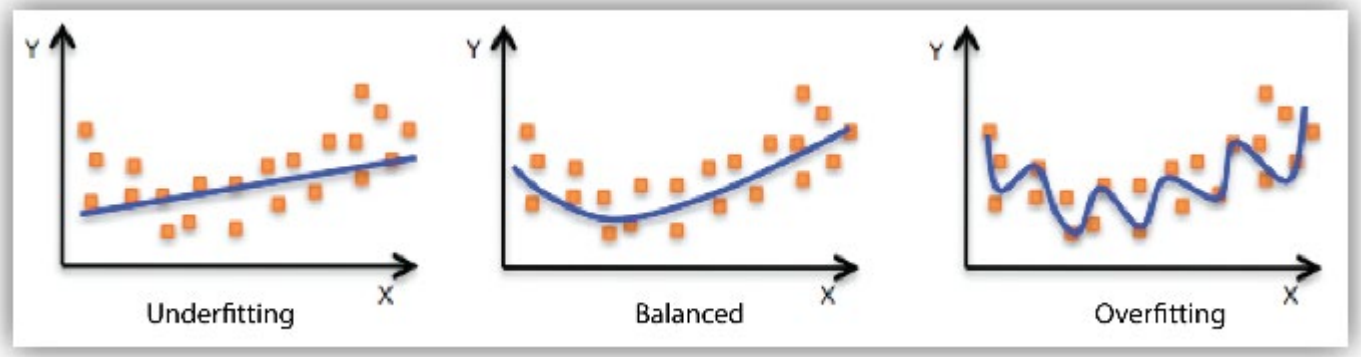
Training & Test set

- To perform learning, you need data.
- Learning will generate a classifier that can perform classification
- In order to test your classifier, you need data which is not used in learning process
- To test the effectiveness of your algorithm, you can split your data into two parts: a training set and a test set.
- The test set should be independent of the training set. It is required to verify the error rate of your algorithm.

Model Overfitting

- Be careful not to over fit. Overfitting is when you are trying to achieve 100% accuracy, even learning from the examples that are wrong. Instead, you want to generalize the data to find the underlying trends.
- Your model is overfitting your training data when you see that the model performs well on the training data but does not perform well on the evaluation data. This is because the model is memorizing the data it has seen and is unable to generalize to unseen examples.

Underfitting vs Overfitting



Your model is *underfitting* the training data when the model performs poorly on the training data. This is because the model is unable to capture the relationship between the input and the target values (often called Y).

Refer to: https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Error Estimation

- Random sampling with repeated holdout. Run once with a random 2/3 training data, 1/3 test data. Then re-run it with a different 1/3 of the data. Continue this process until your error rate stabilizes.
- K-fold cross-validation – partition into K equal groups. K-1 groups are training data, and test on the last remaining group. Repeat this K times, where K is usually 10. Take the average accuracy rate as the overall accuracy.

Accuracy

- A confusion matrix is defined as the possible outcomes:

	Predicted +	Predicted -
Actually +	a	b
Actually -	c	d

Terms

- The ***accuracy*** of your model is the cases you got right: (a+d) / (a+b+c+d)
- The ***precision*** is defined as: a / (a+c)
- The ***recall***, and ***Sensitivity***, both mean: a / (a+b). These are the number of true cases you got right.
- The ***specificity*** is d / (b+d). These are the number of false cases you got right.

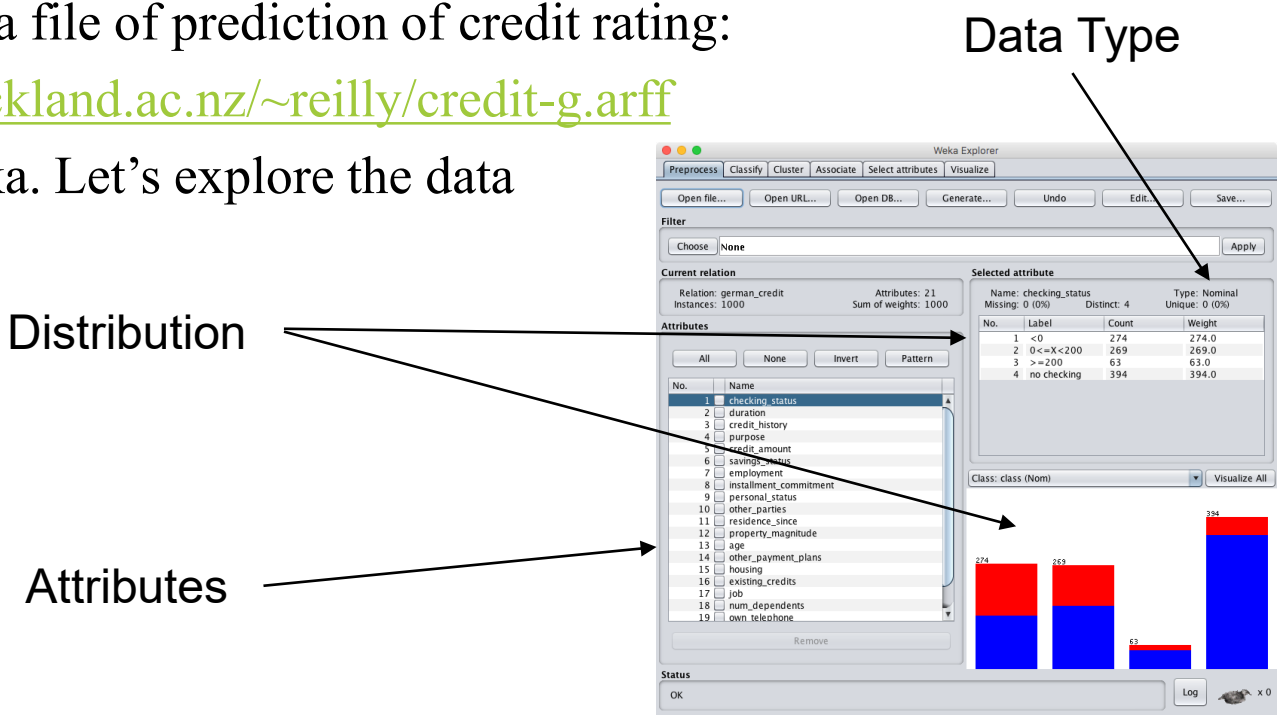
K-Nearest neighbors

- One of the easiest classification algorithms
- Create a plot of the data, and compute which are the K nearest items for your unknown sample.
- From the K-nearest, calculate a simple majority wins estimate for the value you want to predict.
- For predicting final grades, find students with similar final numeric grades, and pick the most popular letter grade.
- For predicting weather, look at previous data for date, temperature, etc. and pick the most popular classification.

Demo in Excel

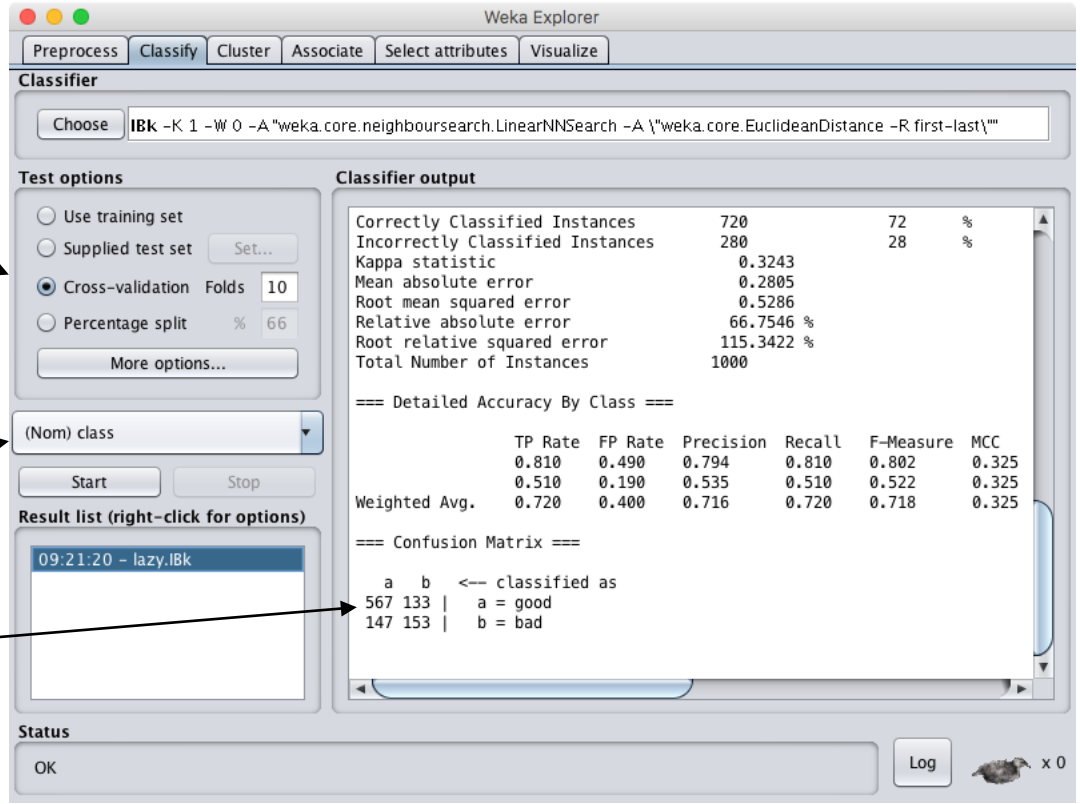
Testing in Weka

- Have a look at a data file of prediction of credit rating:
- https://www.stat.auckland.ac.nz/~reilly/credit-g.arff
- Load the file in Weka. Let's explore the data



Testing

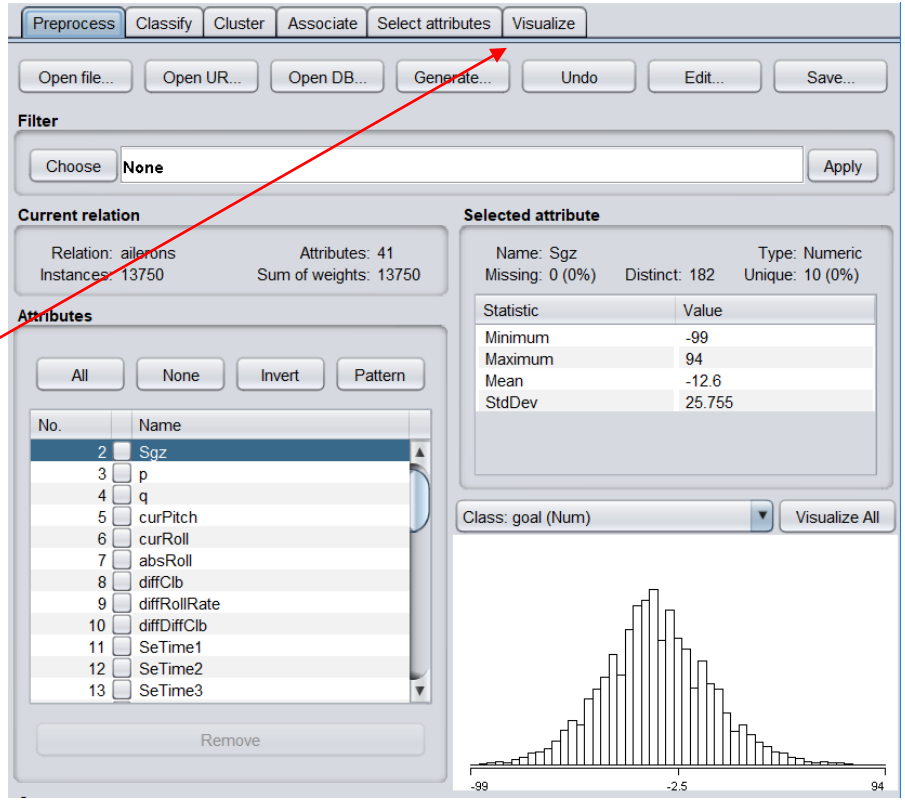
- Validation method
- What are you predicting?
- Confusion matrix



Data Visualization

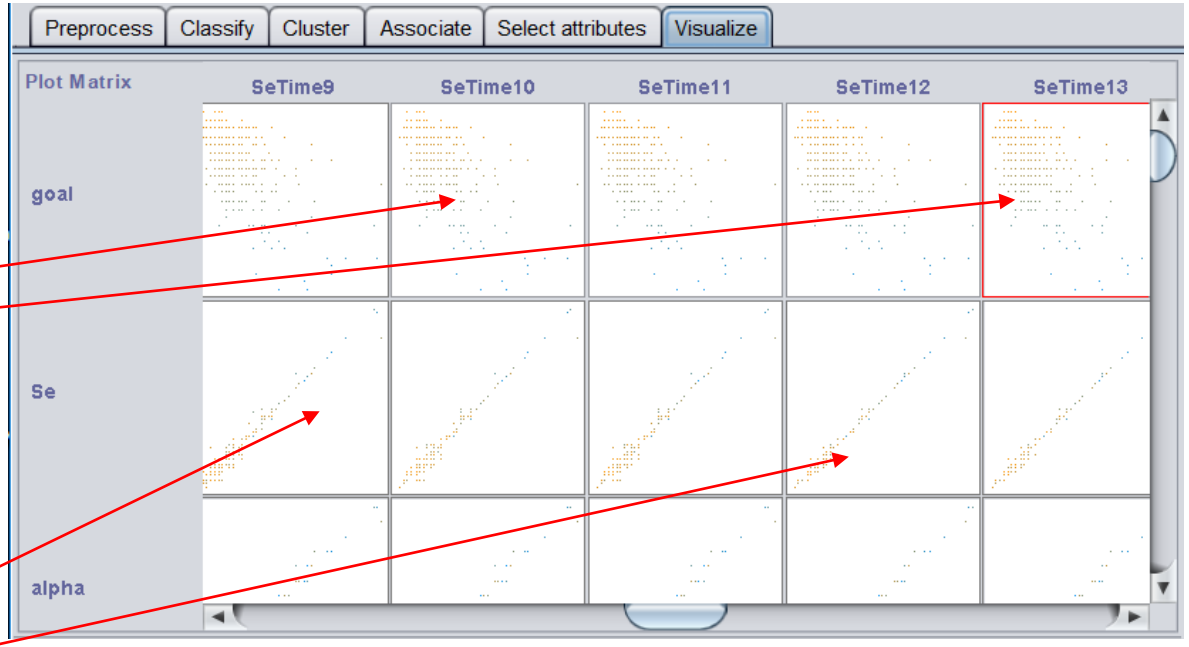
- Looking at regression/ailerons.arff. If the distribution is numeric, it shows the histogram:

- Next click on the “Visualize” tab



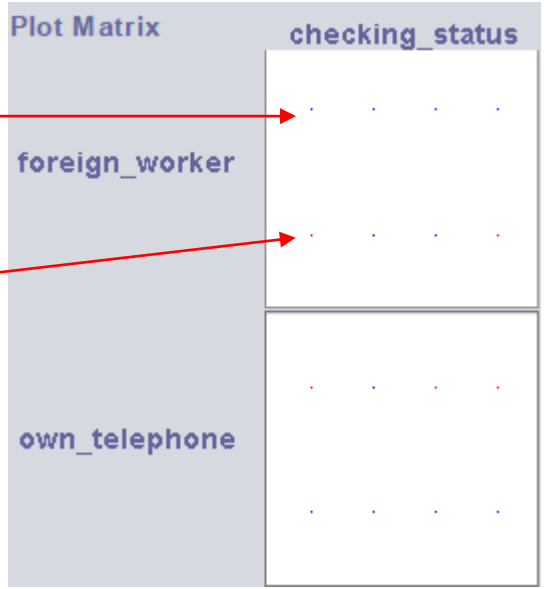
Data Visualization

- You can see the relationship of every variable with each other.
- If there is no relationship, you see a cloud.(No Correlation)
- If there is a relationship, you see a linear pattern. (Correlation)



Data Visualization

- If the data are Categorical, you see clear separation. Here we see yes/no values.
- This is the credit_g.arff file
- Increase the jitter to add some noise. This will give you an idea on how many points are represented by each point.



References

K-Nearest Neighbour:

- http://sens.tistory.com/277
- http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
- https://www.youtube.com/watch?v=SQOdBjjA2y8
- https://www.analyticsvidhya.com/blog/2014/10/introduction-k-neighbours-algorithm-clustering/

Crisp-DM:

- https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
- http://www.sv-europe.com/crisp-dm-methodology/