CSCE 590-1: From Data to Decisions with Open Data: A Practical Introduction to AI

Lecture 7: Supervised Machine Learning

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

2ND FEB, 2021

Carolinian Creed: "I will practice personal and academic integrity."

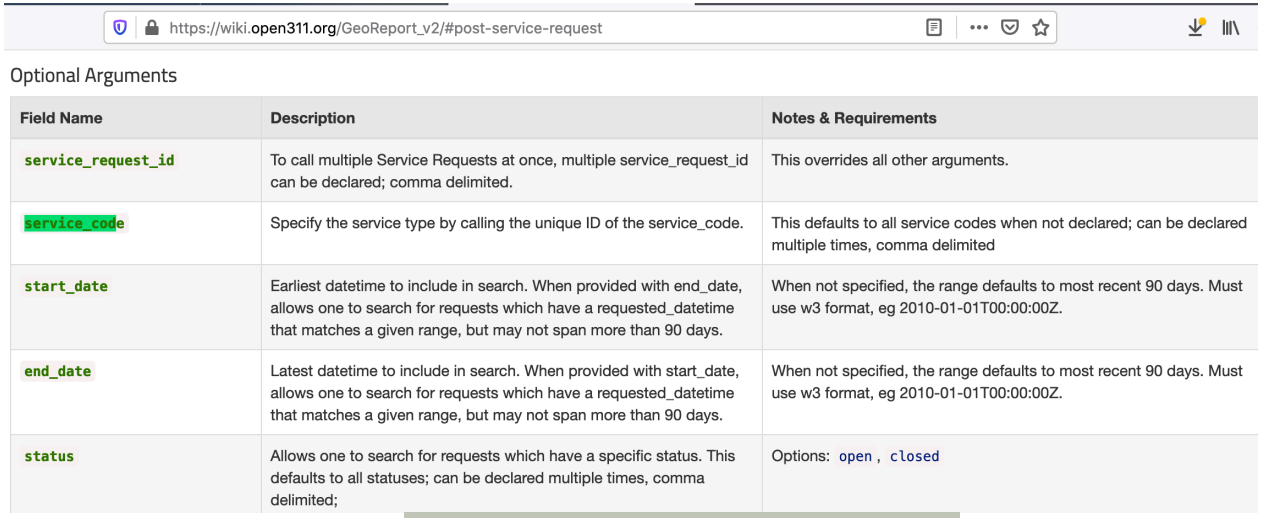
Organization of Lecture 7

- Introduction Segment
  - Recap of Quiz 1
  - Project discussion
  - Adjustment to class schedule – paper readings
  - Recap of Lecture 6
- Main Segment
  - Review datasets
  - Review Weka
  - Decision trees/ random forest
- Concluding Segment
  - About Next Lecture – Lecture 8
  - Ask me anything

Introduction Segment

Recap of Quiz 1

- Q3: Key to answering was looking up details of Open 311 APIs
  - https://wiki.open311.org/GeoReport_v2/



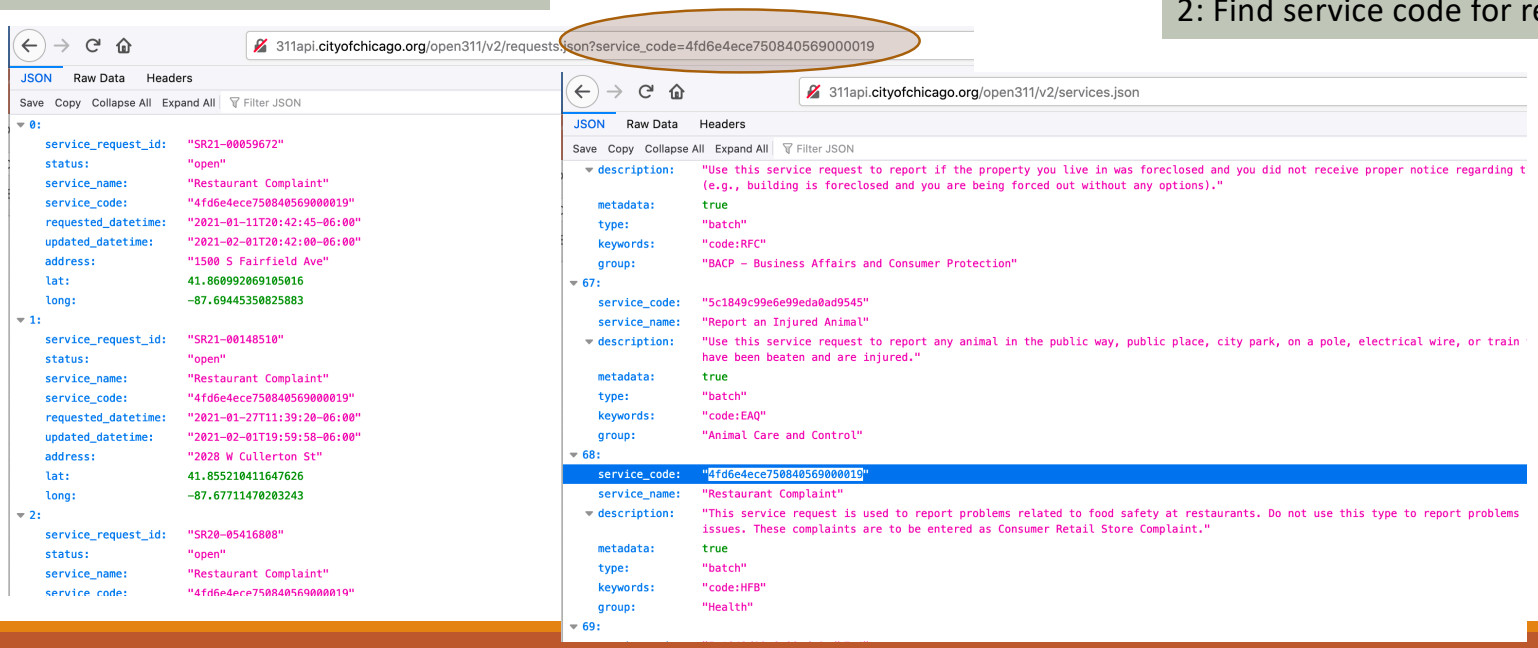
1: service_code can be used to filter

- Q3: Key to answering was looking up details of Open 311 APIs
  - https://wiki.open311.org/GeoReport_v2/

Recap of Quiz 1

3: Filter requests based on service_code

2: Find service code for restaurant



Project Discussion

- Spreadsheet to add details
  - https://docs.google.com/spreadsheets/d/1vNQo_uG0t7lUxVGPSr1yzxt0MWxGmHjeQOE9Txm-r98/edit?usp=sharing

- Create Project Plan and submit by Feb 4, 2021 (next class)
  - Content
    - Identify problem
    - Value of decision: before and after
    - Data-needed
    - Method
    - Evaluation
    - (How can someone) Integrate your work with overall process
  - Milestones

Course Check points
- Project outline: Feb 4
- Mid-sem: March 2
- In-class: March 25
- Class presentations: April 22, 29

Course Project

- (**Undergraduate**) Project: 50% + 10%:
  - Do a significant project of your choice following the framework presented in Lecture 3
  - Deliverable: Project report and code (50%), 1-slide elevator presentation to class (10%)

- **Framework**: Value of decision: before and after, Data-needed, Method, Evaluation, Integrating with overall process

- **Illustration**:
  - Data analysis project
  - Dataset must be from given catalog – Lecture 1 - (US: https://www.data.gov/ or any US state; Text of legislations - LegiScan, https://legiscan.com/; Kaggle datasets: https://www.kaggle.com/datasets; Google datasets search: https://datasetsearch.research.google.com/) OR discussed with instructor
  - Use analytical methods to present new insights
  - Problem (method) to be discussed with instructor
  - Examples: SC traffic deaths, COVID responses of a US states, economic growth and tax rates

Course Project

- (**Graduate**) Project: 50% + 10%:
  - Do a significant project of your choice following the framework presented in Lecture 3
  - Deliverable: Project report and code (50%), 1-slide elevator presentation to class (10%)

- **Framework**: Value of decision: before and after, Data-needed, Method, Evaluation, Integrating with overall process

- **Illustrations**:
  - Project on data analysis (like for undergraduates) but bigger scope
  - Project on creating or exploring new methods (preferred)
  - Problem (method) to be discussed with instructor
  - **Examples**: Estimate crowd from sound, find mask adherence from photos, find algal bloom from images of water

Rubric for Evaluation of Course Project

- Project
  - Project plan
  - Challenging nature of project
  - Actual achievement
  - Report
  - Sharing of code
- Presentation
  - Motivation
  - Coverage of related work
  - Results and significance
  - Handling of questions

Adjustment to Class Schedule

Lecture Number	Date	Topic	Additional Comment
7	Feb 2 (Tu)	Structured: Analysis – Supervised ML	
8	Feb 4 (Th)	Structured: Analysis – Supervised ML and Papers	Paper discussion in class (10 tips) Project Outline due
9	Feb 9 (Tu)	Structured: Discuss papers/ Attend DEEP-DIAL21 workshop	Paper reading in pairs (Which ML to use)
10	Feb 11 (Th)	Structured: Unsupervised	
11	Feb 16 (Tu)	Structured: Unsupervised	
12	Feb 18 (Th)	Structured: Advanced – AutoAI, Explanation	Quiz 2
13	Feb 23 (Tu)	Structured: Time Series Data	

And Schedule Further Down

Lecture Number	Date	Topic	Additional Comment
7	Feb 2 (Tu)	Structured: Analysis – Supervised ML	
8	Feb 4 (Th)	Structured: Analysis – Supervised ML and Papers	Paper discussion in class (10 tips)
9	Feb 9 (Tu)	Structured: Discuss papers/ Attend DEEP-DIAL21 workshop	Paper reading in pairs (Which ML to use)
10	Feb 11 (Th)	Structured: Unsupervised	
11	Feb 16 (Tu)	Structured: Unsupervised	
12	Feb 18 (Th)	Structured: Advanced – AutoAI, Explanation	Quiz 2
13	Feb 23 (Tu)	Structured: Time Series Data	
14	**Feb 25 (Th)**	**Wellness Holiday**	**Wellness Holiday**
15	Mar 2 (Tu)	Invited Guest	Mid-sem Project Review
16	Mar 4 (Th)	Reasoning	Semester - Midpoint
17	Mar 9 (Tu)	Reasoning – Optimization, Uncertainty	
18	Mar 11 (Th)	Text: Data Prep (NLP)	
19	Mar 16 (Tu)	Text: Analysis - Supervised (NLP)	
20	Mar 18 (Th)	Text: Advanced – Summarization, Sentiment	Quiz 3

Course Check points

- Project outline: Feb 4
- Mid-sem: March 2
- In-class: March 25
- Class presentations: April 22, 29

Recap of Lecture 6

- We looked at
  - Supervised learning task
  - Concepts related to data characteristics and quality
  - Evaluation approach
- Also investigated regression method

Papers to Read

- [10 tips] Ten quick tips for machine learning in computational biology
- [Which ML to Use] Data-driven advice for applying machine learning to bioinformatics problems

Main Segment

Machine Learning – Insights from Data

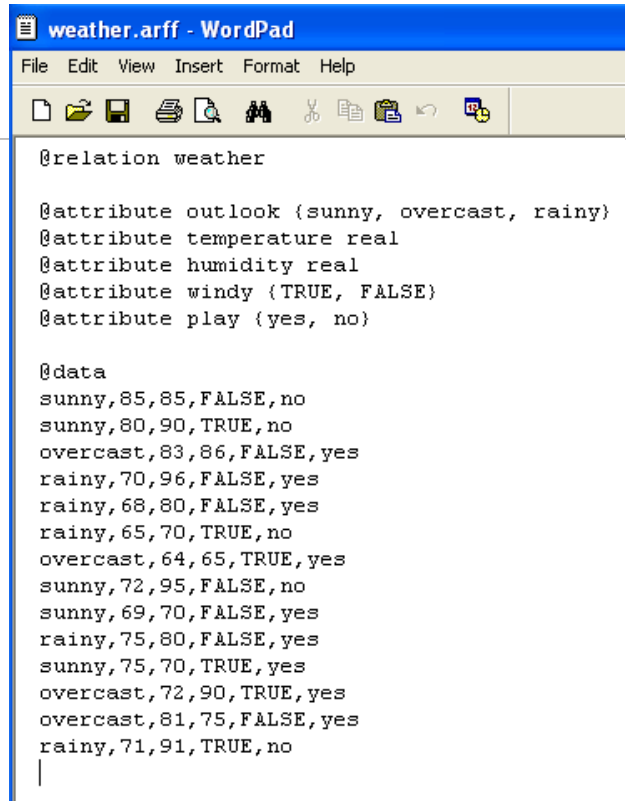
- Descriptive analysis
  - Describe a past phenomenon
  - **Methods**: **classification**, clustering, dimensionality reduction, anomaly detection, neural methods
- Predictive analysis
  - Predict about a new situation
  - **Methods**: time-series, neural networks
- Prescriptive analysis
  - What an agent should do
  - **Methods**: simulation, reinforcement learning, reasoning
- New areas
  - Counterfactual analysis
  - Causal Inferencing
  - Scenario planning

Reference – Data

- UCI Datasets - https://archive.ics.uci.edu/ml/datasets.php
- ARFF format – Used by WEKA

ARFF Data Format

- Attribute-Relation File Format
- Header – describing the attribute types
- Data – (instances, examples) comma-separated list



Slide Courtesy: http://www.cs.iastate.edu/~cs573x/bbsilab.html

Decision Tree

Problem: Classify Weather Data

Input

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
...	...	...	...	...

Class Label

Output (Informal)

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes
```

Slide Adapted From/ Courtesy: Data Mining, Practical Machine Learning Tools and Techniques, Slides for Chapter 6, Trees and Rules of Data Mining by I. H. Witten, E. Frank, M. A. Hall and C. J. Pal

Which Variable to Learn to Create Rules On?

- What do we want?
  - Compact model (e.g., set of rules)
  - High accuracy / low error
- Find the most discriminating variable
  - But how do we measure this
- Corner cases
  - If all the samples are the same, the decision tree is a ?
    - Leaf node with the only class
  - If there are no attributes in the dataset, the decision tree is?
    - A node with most common class

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
...	...	...	...	...

Expected Information/ Entropy

- Concept: Expected Information
  - Let
    - Class label has **m** distinct values (i.e., m distinct classes)
    - s_i be the number of samples of S of Class C_i (i = 1 ..m)
  - $I(s_1, s_2, .., s_m) = - \Sigma_{i=1 \text{ to } m} p_i \log_2 (p_i)$
    - Where P_i is the probability a sample belongs to class Ci; estimated by(s_i /s)

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
...	...	...	...	...

- Entropy / Expected Information after partitioning on Attribute A which has **v** distinct values
  - $E(A) = \Sigma_{j=1 \text{ to } v} (s_{1j} + ... + s_{mj}) / S \quad * (I(s_{1j}, s_{2j}, .., s_{mj})$
  - s_{ij} be the number of samples in S_j of Class C_i (i = 1 ..m)
  - Smaller the entropy, the greater the purity of the subset partitions

Information Gain

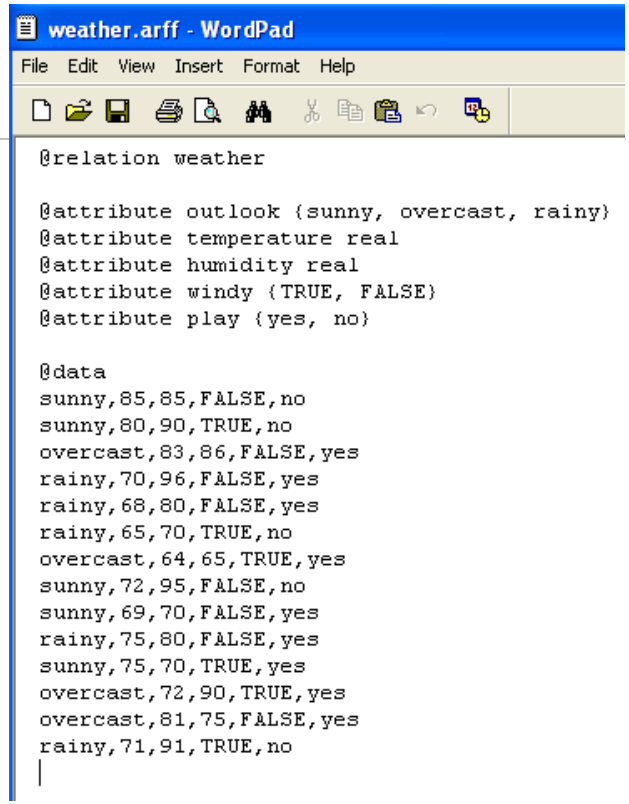
Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
...	...	...	...	...

- Entropy / Expected Information after partitioning on Attribute A which has v distinct values
  - $E(A) = \Sigma_{j=1 \text{ to } v} (s_{1j} + \ldots + s_{mj}) / S \quad * (I(s_{1j}, s_{2j}, .., s_{mj})$
    - s_{ij} be the number of samples in S_j of Class C_i (i = 1 ..m)
- After partition, S_j
  - $I(s_{1j}, s_{2j}, .., s_{mj}) = - \Sigma_{i=1 \text{ to } m} p_{ij} \log_2 (p_{ij})$
  - Where p_{ij} is the probability a sample in S_j belongs to class C_i; estimated by$(s_{ij} / | s_j |)$
- Gain (A) = $I(s_1, s_2, .., s_m)$ - E(A)
  - Is the expected reduction in entropy by knowing the value of Attribute A
- **Method**: Split on the attribute which leads to the highest information gain

Weka Exercise

ARFF Data Format

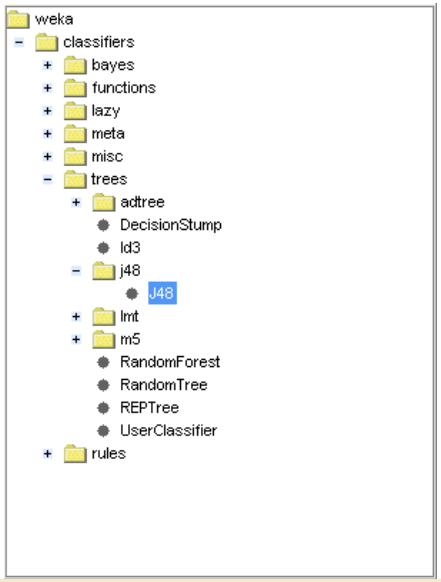
- Data is in ARFF in UCI dataset
- Or Convert
  - File system, CSV → ARFF format
  - Use C45Loader and CSVLoader to convert

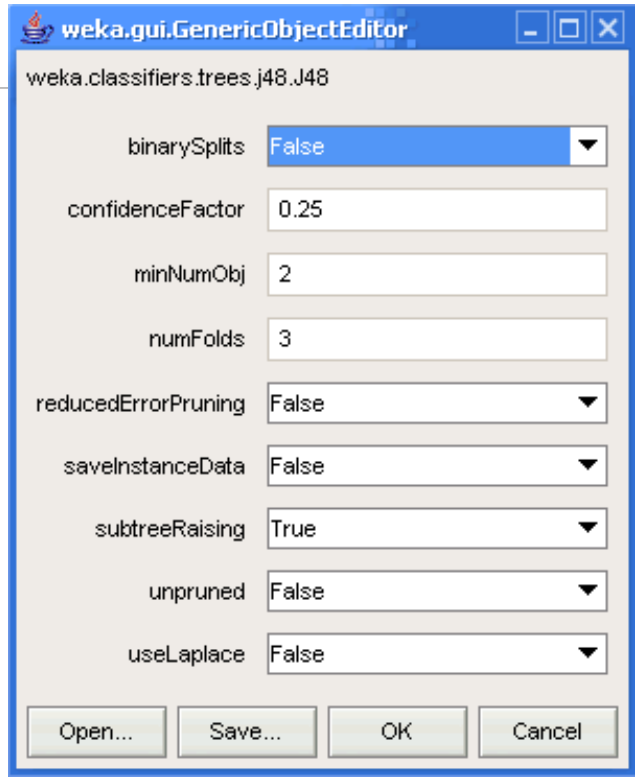


Slide Courtesy: http://www.cs.iastate.edu/~cs573x/bbsilab.html

Weka: weka.classifiers.trees.J48

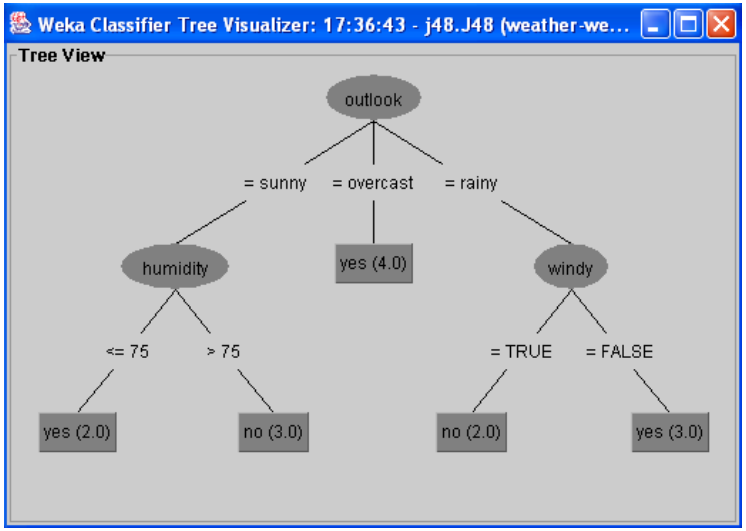
Class for generating an unpruned or a pruned C4.5 decision tree.





Slide Courtesy: http://www.cs.iastate.edu/~cs573x/bbsilab.html

Understanding Output



Slide Courtesy: http://www.cs.iastate.edu/~cs573x/bbsilab.html

Weka: Decision Tree Output

```
J48 pruned tree
------------------

outlook = sunny
|   humidity = high: no (3.0)
|   humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :    5

Size of the tree :        8
```

```
=== Summary ===

Correctly Classified Instances          7               50      %
Incorrectly Classified Instances        7               50      %
Kappa statistic                          -0.0426
Mean absolute error                       0.4167
Root mean squared error                   0.5984
Relative absolute error                  87.5    %
Root relative squared error             121.2987 %
Total Number of Instances                14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.556    0.6        0.625     0.556     0.588      0.633    yes
                 0.4      0.444      0.333     0.4       0.364      0.633    no
Weighted Avg.    0.5      0.544      0.521     0.5       0.508      0.633

=== Confusion Matrix ===

 a b   <-- classified as
 5 4 | a = yes
 3 2 | b = no
```

Slide Courtesy: http://www.cs.iastate.edu/~cs573x/bbsilab.html

Test Options

- Percentage Split (2/3 Training; 1/3 Testing)
- Cross-validation
  - Estimating the generalization error based on resampling when limited data
    - averaged error estimate.
  - Cross-fold validation (10-fold)
  - Leave-one-out (Loo)
  - Stratified

Slide Courtesy: http://www.cs.iastate.edu/~cs573x/bbsilab.html

Random Forest

- An ensemble method
- Credits
  - Ideas introduced by Tin Kam Ho in 1995, https://en.wikipedia.org/wiki/Tin_Kam_Ho
  - Matured by Leo Breiman and Adele Cutler at Berkeley (https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro)
  - History: Khaled Fawagreh, Mohamed Medhat Gaber & Eyad Elyan (2014) Random forests: from early developments to recent advancements, Systems Science & Control Engineering, 2:1, 602-609, DOI: 10.1080/21642583.2014.956265
- Main steps (Input: data, N= number of trees)
  - If the number of cases in the training set is N, sample N cases at random - but *with replacement*, from the original data. This sample will be the training set for growing the tree.
  - If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
  - Each tree is grown to the largest extent possible. There is no pruning.

Slide Courtesy: Leo Breiman and Adele Cutler website

Random Forest in Action

- Code examples:
  - https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l6-l7-supervised-ml/Supervised-RandomForest-Classification.ipynb
- Scikit Library: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Reference - Tool

- Weka - https://www.cs.waikato.ac.nz/ml/weka/
  - Uses data in Weka format
- Demonstration

Metric Types

- **Effectiveness**: what the **user** of a system sees, primarily cares about
- **Efficiency**: what the **executor** in a system sees, primarily cares about

Metrics: Accuracy, Precision, Recall

		Predicted class	
		Class = Yes	Class = No
Actual Class	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Accuracy =
(TP+TN)/
(TP+FP+FN+TN)

Precision =
(TP)/
(TP+FP)

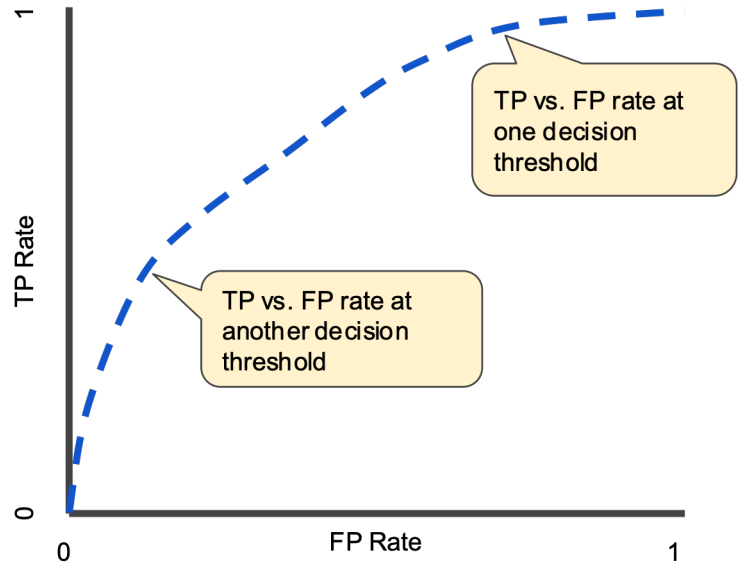
Recall =
(TP)/
(TP+FN)

F1 Score: *Harmonic Mean*

1/F1 = 1/Precision + 1/Recall

F1 = 2*(Recall * Precision) /
(Recall + Precision)

ROC – Receiver Operating Characteristic curve



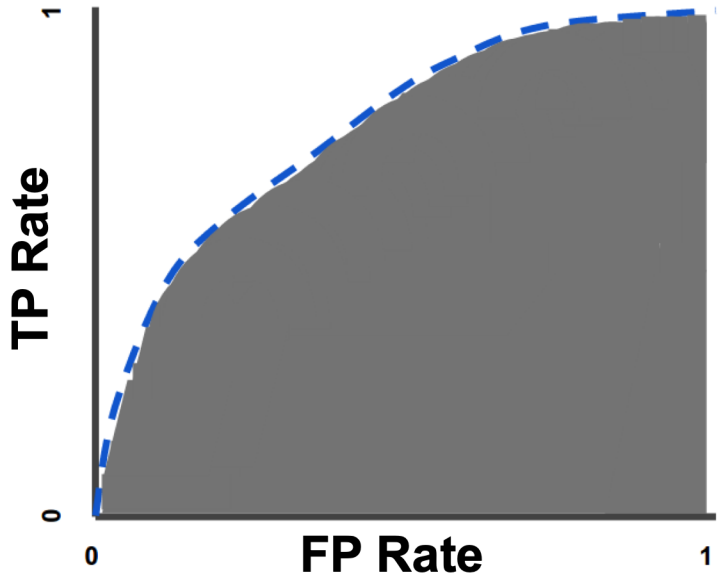
True Positive Rate = Recall = (TP)/ (TP+FN)

False Positive Rate = (FP)/ (FP+TN)

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

AUC – Area Under the ROC Curve



- Aggregate measure of performance across all possible classification thresholds.
- Interpretation: probability that the model ranks a random positive example more highly than a random negative example

Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

References

- Blogs: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/
- Google: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Lecture 7: Concluding Comments

- Reviewed Quiz and Project
- Supervised ML
  - Review datasets
  - Review Weka
  - Decision trees/ random forest

Concluding Segment

About Next Lecture – Lecture 8

Lecture 8: Unsupervised Learning

- Structured Data: Unsupervised Methods
  - Setting and characteristics
- Methods: k-means
- Working with Weka