What's the Deal with Machine Learning?

BERDC Special Topics Talk 14

UNIVERSITY OF NORTH DAKOTA

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Dr. Mark Williamson

Biostatistics, Epidemiology, and Research Design Core

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Opening

Goal: Decode what machine learning is and how to use it in research

- Defining terms
- Machine Learning (ML) Methods
- ML Techniques
- Uses for ML in biological and biomedical research
- Software Tools for running ML
- Worked Examples

Before Moving On:

Pre-test: https://und.qualtrics.com/jfe/form/SV_3asp6ByKwgfrV8W

R code: https://med.und.edu/daccota/_files/docs/berdc_docs/machine_learning_rcode.txt

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Defining Terms

UNIVERSITY OF NORTH DAKOTA

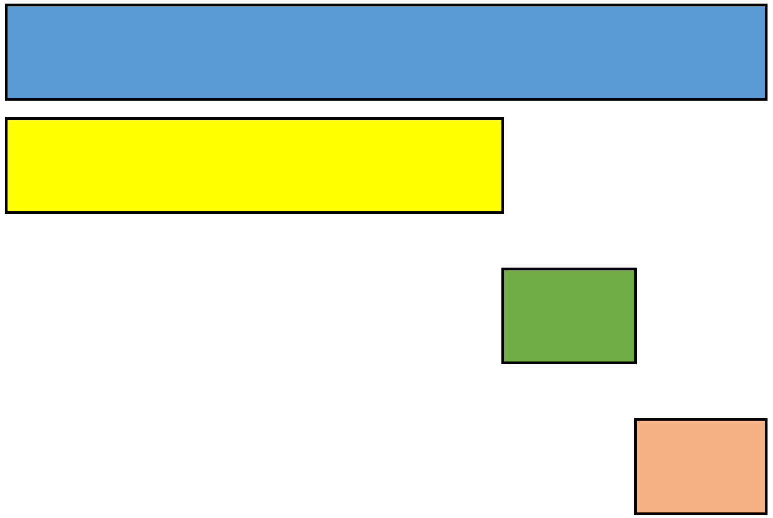
- What is **Machine Learning** (ML)?
  - *Machine learning is branch of AI which focuses on the use of* ***data and algorithms*** *to* ***imitate*** *the way that humans learn, gradually improving its accuracy* [1]
  - *Machine learning (ML) is the process of using* ***mathematical models*** *of data to help a computer learn without direct instruction* [2]
  - *Machine learning is a data analytics technique that teaches computers to do what comes naturally to humans and animals:* ***learn from experience*** [3]
- **Deep Learning**: sub-field of machine learning; less dependent on human intervention to learning [1]; based on neural networks [4]
- **Neural Networks**: sub-field of deep learning; composed of layers, including hidden ones [1]; collection of connected nodes loosely representing neuron connectively in a biological brain [4]

Defining Terms

- **Terms** [5]:
  - Label: dependent variable (statistics)
  - Features: independent variables (statistics)
  - Feature creation: transformation (statistics)
  - Classes: mutually exclusive groups (labels not mutually exclusive)

$$\overset{\textbf{Label}}{y} = \underbrace{x_1 + x_2 + x_3}_{\textbf{Features}}$$

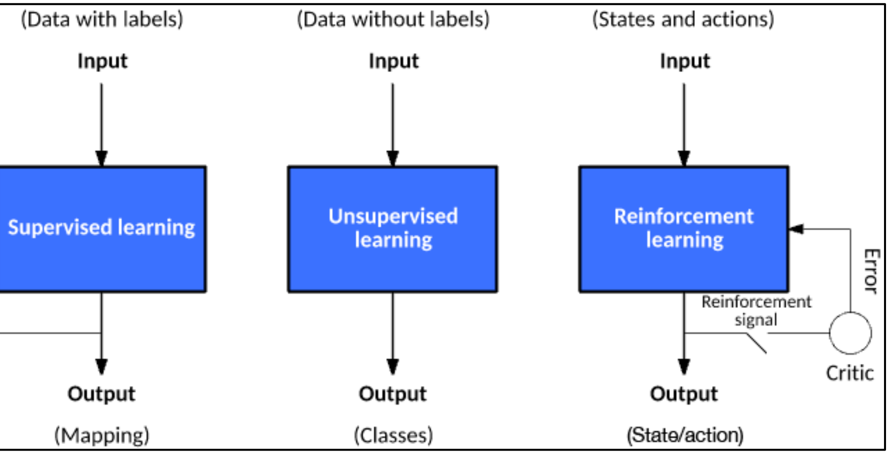
- **Datasets** [4]:
  - Training datasets: used to adjusted parameters of model to improve performance
  - Validation datasets: used to monitor but not influence the training process
  - Test datasets: used for actual research questions



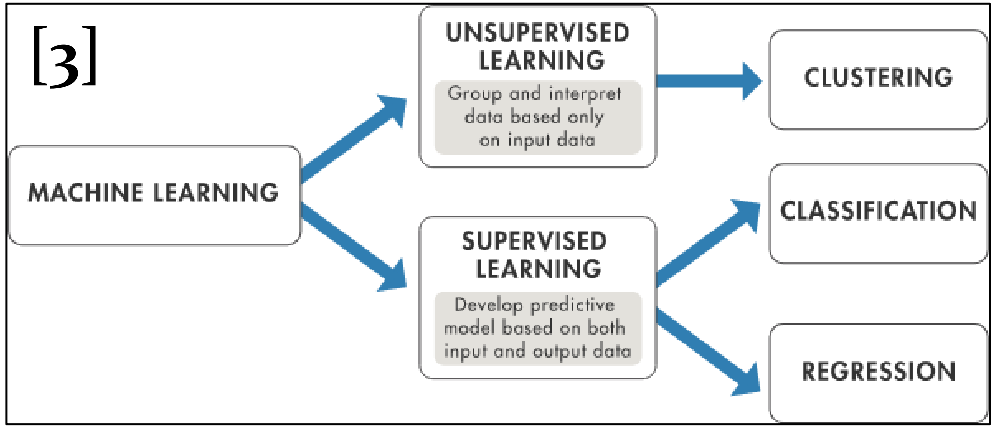
ML Methods

- Supervised, unsupervised, semi-supervised, reinforcement
- **Supervised:** use of labeled datasets to train algorithms to classify data or predict outcomes accurately [6]
- **Unsupervised**: use of ML algorithms to analyze and cluster unlabeled datasets
    - Main difference between the two: labeled data [7]
    - **Semi-supervised**: Happy medium between two; during training, uses a smaller labeled dataset to guide classification and feature extraction from a larger, unlabeled dataset
- **Reinforcement**: Like supervised (receives feedback), but not necessarily for each input or state; ideal algorithm that can learn how to make decisions in an uncertain environment [8]

[8]

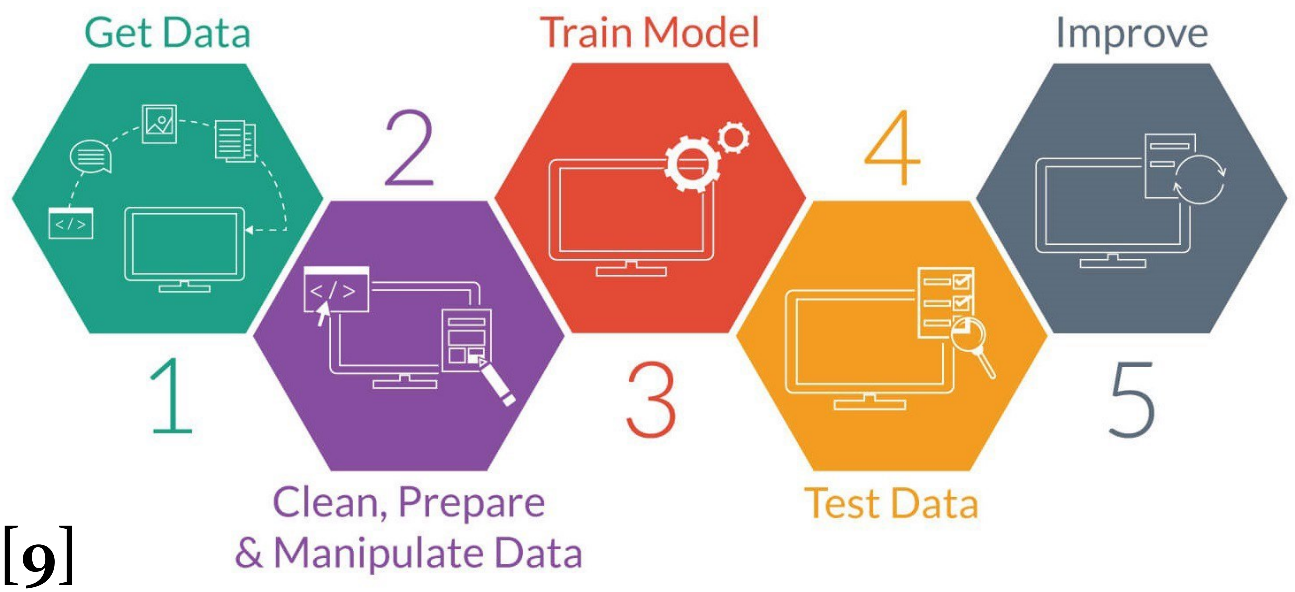


[3]

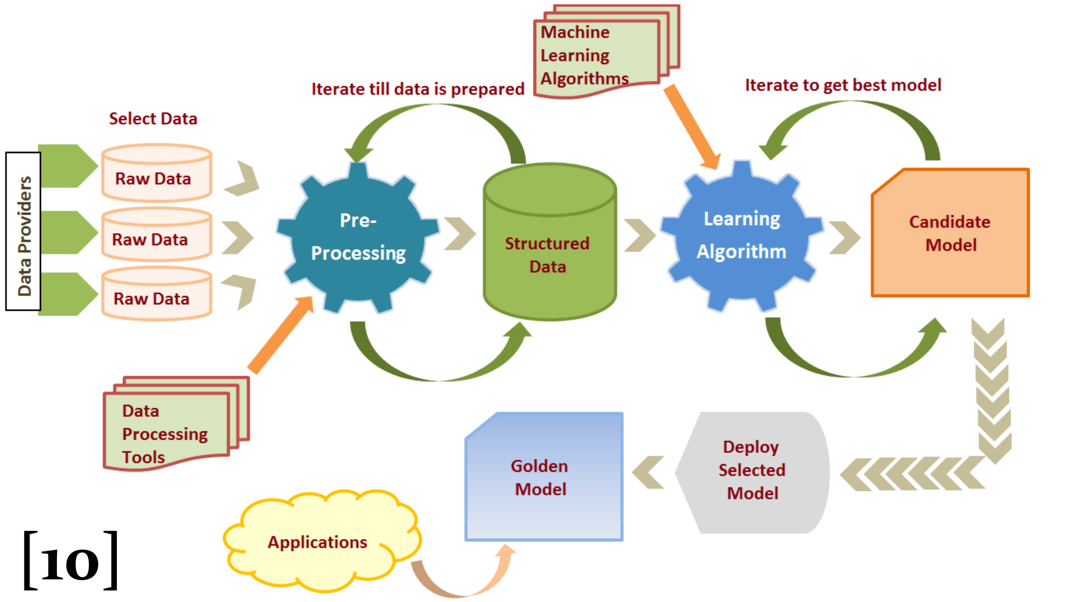


ML Methods

- Overview of process: 1) collect & prepare data, 2) train the model, 3) validate the model, 4) interpret the results [2]



[9]



[10]

ML Techniques

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

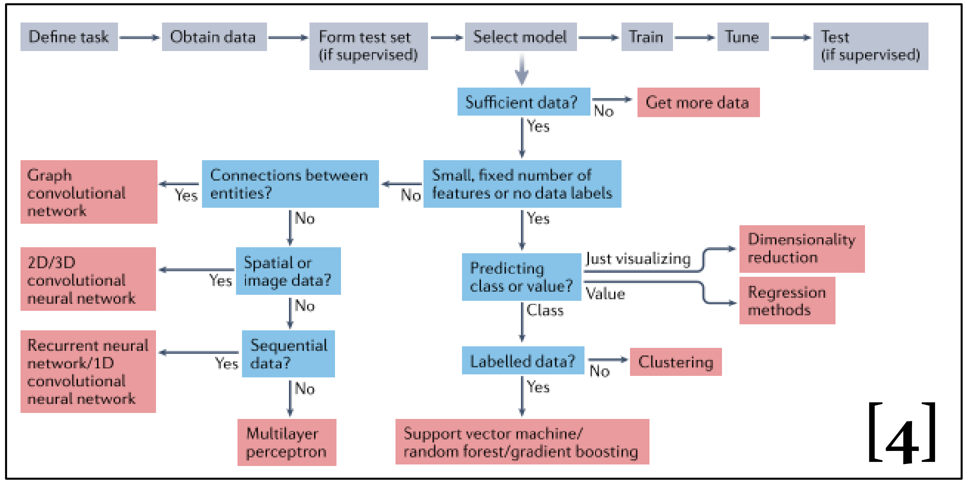
UNIVERSITY OF NORTH DAKOTA

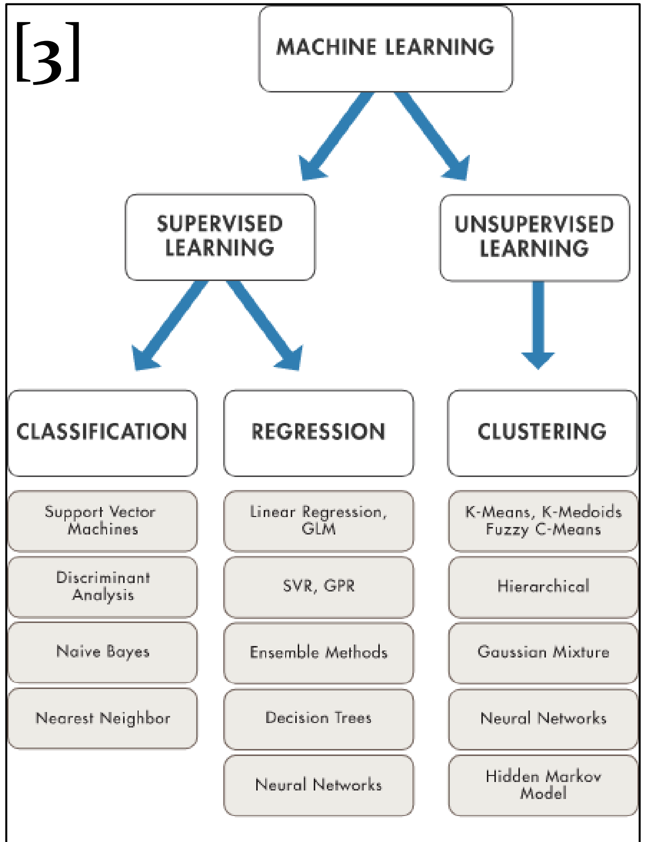
Supervised learning:

- neural networks, genetic algorithm, naïve Bayes, Bayesian networks linear regression, logistic regression, random forests, support vector machines (SVN), decision trees, gradient boosting and bagging, multivariate adaptive regression splines, nearest neighbor

Unsupervised learning:

- neural networks, genetic algorithm, Bayesian networks, PCA and singular value decomposition (SVD), clustering (k-means, probabilistic, etc.), Gaussian mixture models, self-organizing maps, associations and sequence discovery, expectation maximization, Bayesian networks, kernel density estimation, sequential covering rule building

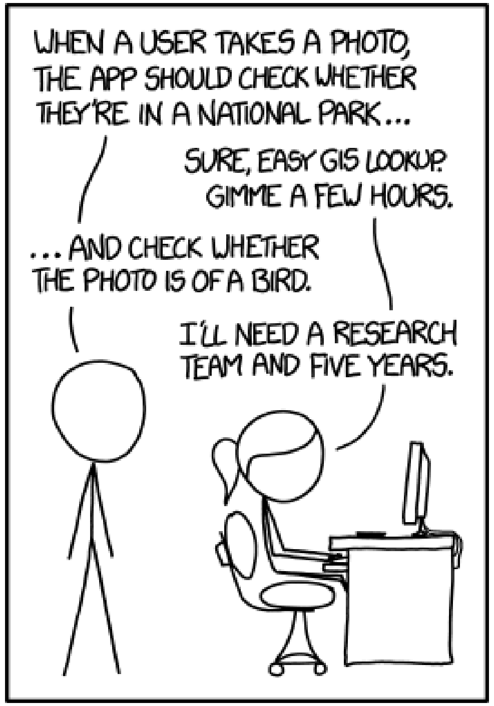




ML Uses

- **Conceptual:** predict values, identify unusual occurrences, find structure, predict categories [2]
- **General:** speech recognition, customer service, computer vision, recommendation engines, automated stock trading [1]
- **Healthcare/Biomedical:**
  - diagnostic tools, patient monitoring, and outbreak predication[2]
  - tumor detection, drug discovery, and DNA sequencing [3]
  - identifying gene coding regions, structure prediction, neural networks (classification of cellular images, genome analysis, drug discovery), AI in healthcare [11]
  - drug manufacturing, personalized medicine, stroke diagnosis [12]

https://xkcd.com/1838/

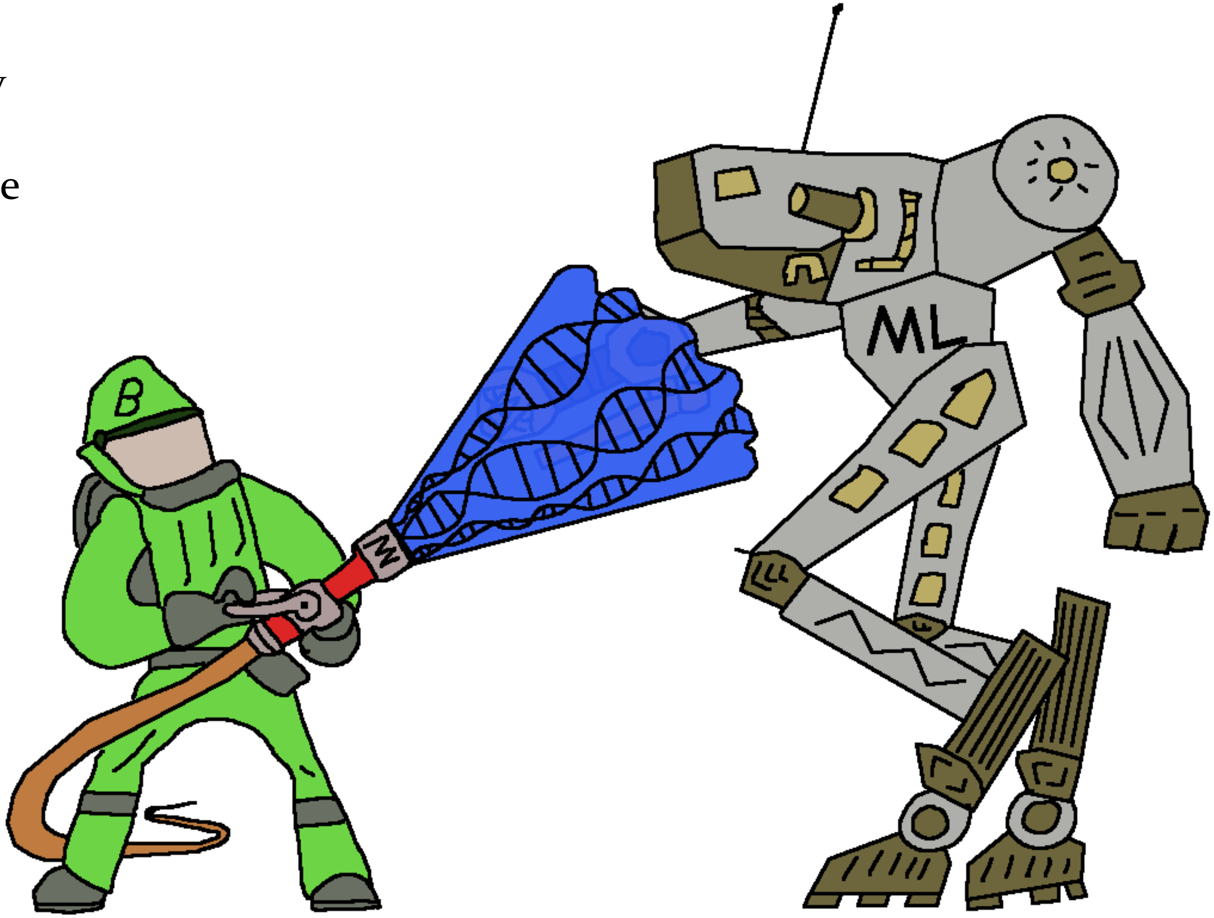


IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

https://xkcd.com/1425/

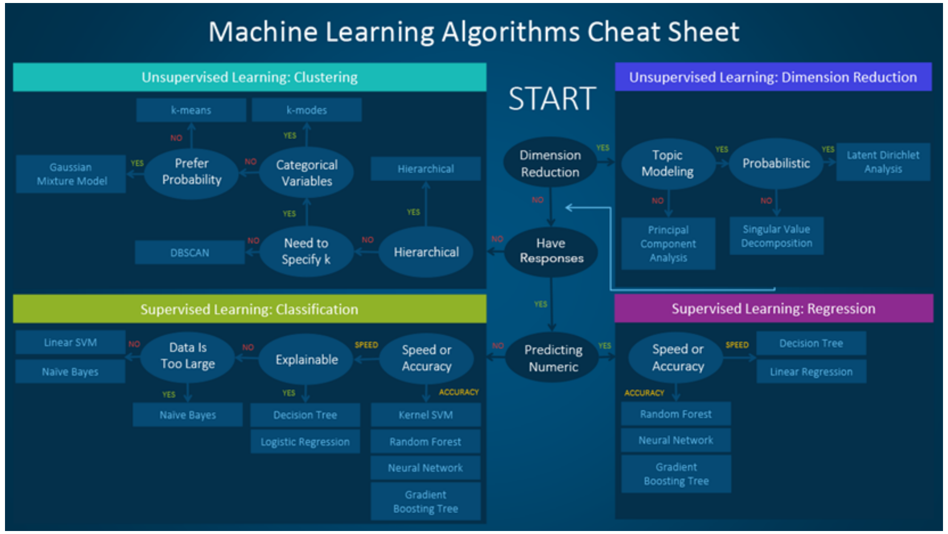
ML Uses

- **Biology** [12]:
  - **Genomics:** regulatory genomics (producing RNA-binding proteins and transcription factors and predicting and classifying gene expression), structural genomics (help classify protein structure), functional genomics (classify mutations and protein subcellular localization), genome sequencing, gene editing, clinical workflow
  - **Proteomics:** mass spectral peaks, protein recognition by sequence database searching
  - **Microarrays:** spotting significant interactions in complex environments, gene classification, clustering, gene analysis (analyze chances in gene patterns), differentiate gene states, predict future gene changes
  - **Systems Biology:** capture interactions between biological components and simulate the whole system's behavior (signal transduction networks, genetic networks, and metabolic pathways), probabilistic graphical modeling, genetic algorithms, Markov chain optimization



ML Tools

- Software Programs [11]:
  - Cell Profiler
  - DeepVarient
  - Atomwise
  - TensorFlow
- Software Languages:
  - SAS [13, 14]
    - Machine learning Algorithms Cheat Sheet
    - Base SAS procedures: ACECLUS, ADAPTIVEREG, CLUSTER, DISCRIM, DISTANCE, FACTOR, FASTCLUS, GLIMMIX, KDE, KRIGE2D, LOGISTIC, MCMC, MDS, MODECLUS, NLIN, PLS, PRINCOMP, REG, ROBUSTREG, VARCLUS



DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

ML Tools

UND UNIVERSITY OF NORTH DAKOTA

- Software Languages (cont.):
  - Python [15]
    - Numpy, Scipy, Scikit-learn, Theano, TensorFlow, Keras, PyTorch, Pandas, Matplotlib
  - Julia [16, 17]
    - MLJ, Scikit Learn, GLM, Decision Tree, Mocha, Knet, Flux, Merlin, MLBase, Strada, TensorFlow
  - R [18, 19, 20]
    - lattice, DataExplorer, Dalex(Descriptive Machine Learning Explanations), dplyr, Esquisse, **caret**, janitor, rpart, data.table, ggplot2, e1071, xgboost, randomforest
    - **caret**: Classification and Regression Training

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

ML Tools

- Other Resources:
  - Online course on Machine Learning [21]
  - Machine Learning in Python (Part 1 of 4) [22]
  - Machine Learning in R (Book) [23]
  - Book list for Machine Learning in R [24]
  - Presentation slides on Machine Learning in SAS [25]

Worked Examples

R

- Example 1: Classification of tumors (Supervised) **[20]**
- Example 2: Regression decision tree for tumors (Supervised) **[26, 27]**
- Example 3: K-means clustering for tumors (Unsupervised) **[28-29]**

Biopsy Data on Breast Cancer Patients

V1: clump thickness
V2: uniformity of cell size
V3: uniformity of cell size
V4: marginal adhesion
V5: single epithelial cell size
V6: bare nuclei
V7: bland chromatin
V8: normal nucleoli
V9: mitosis
Class: 'benign' or 'malignant'

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Example 1.1

UNIVERSITY OF NORTH DAKOTA

Classification of tumors

#Get and check data

```
>data(biopsy)
>head(biopsy)
>sapply(biopsy, class)
```

ID	V1	V2	V3	V4	V5	V6	V7	V8	V9	class
"character"	"integer"	"integer"	"integer"	"integer"	"integer"	"integer"	"integer"	"integer"	"integer"	"factor"

#New dataset (excluding missing data and setting variable to numerical)

```
>biopsy1 <-na.exclude((biopsy[,2:11]))
>biopsy1[,1:9] <- sapply(biopsy1[,1:9],as.numeric)
>sapply(biopsy1, class)
```

V1	V2	V3	V4	V5	V6	V7	V8	V9	class
"numeric"	"numeric"	"numeric"	"numeric"	"numeric"	"numeric"	"numeric"	"numeric"	"numeric"	"factor"

#Validation Dataset

```
>validation_index <- createDataPartition(biopsy1$class, p=0.80, list=FALSE)
  # select 20% of the data for validation
>validation <- biopsy1[-validation_index,]
  # use the remaining 80% of data to training and testing the models
>dataset <- biopsy1[validation_index,]
```

Example 1.2

Classification of tumors

```
# Run algorithms using 10-fold cross validation
>control <- trainControl(method="cv", number=10)
>metric <- "Accuracy“

# Test Five Classification models
>set.seed(1)
>fit.lda <- train(class~., data=dataset, method="lda", metric=metric, trControl=control)
>set.seed(1)
>fit.cart <- train(class~., data=dataset, method="rpart", metric=metric, trControl=control)
>set.seed(1)
>fit.knn <- train(class~., data=dataset, method="knn", metric=metric, trControl=control)
>set.seed(1)
>fit.svm <- train(class~., data=dataset, method="svmRadial", metric=metric, trControl=control)
>set.seed(1)
>fit.rf <- train(class~., data=dataset, method="rf", metric=metric, trControl=control)
```

Example 1.3

Classification of tumors

#Summarize accuracy of models

>results <- resamples(list(lda=fit.lda, cart=fit.cart,
knn=fit.knn, svm=fit.svm,
rf=fit.rf))

>summary(results)

>dotplot(results)

>print(fit.rf)

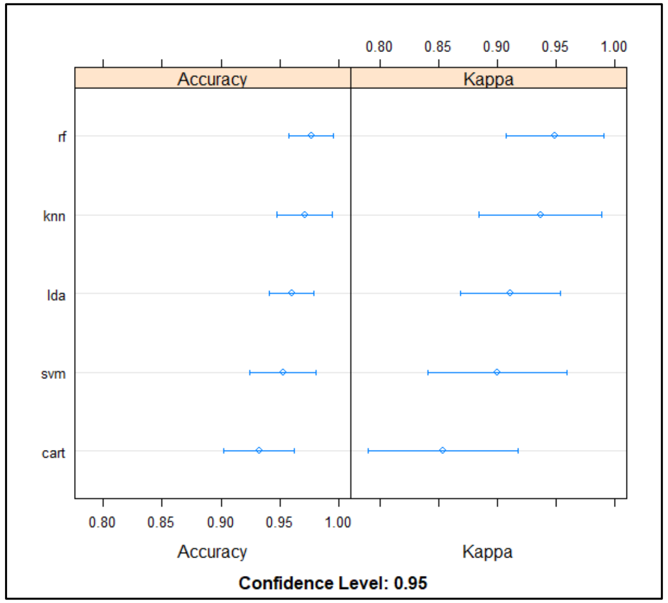
#Make predictions

>predictions <-predict(fit.rf, validation)

>confusionMatrix(predictions, validation$class)

```
Call:
summary.resamples(object = results)

Models: lda, cart, knn, svm, rf
Number of resamples: 10

Accuracy
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
lda  0.9259259 0.9327922 0.9632997 0.9598942 0.9817340 1.0000000    0
cart 0.8571429 0.9078283 0.9363636 0.9325493 0.9634680 0.9818182    0
knn  0.9090909 0.9498316 0.9818182 0.9709716 1.0000000 1.0000000    0
svm  0.8928571 0.9272727 0.9537037 0.9527537 0.9909091 1.0000000    0
rf   0.9285714 0.9544613 0.9818182 0.9764598 1.0000000 1.0000000    0

Kappa
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
lda  0.8335901 0.8508984 0.9182515 0.9109532 0.9591931 1.0000000    0
cart 0.6956522 0.7999152 0.8597709 0.8535073 0.9203543 0.9592894    0
knn  0.8014440 0.8903491 0.9597891 0.9365130 1.0000000 1.0000000    0
svm  0.7717391 0.8467967 0.9015716 0.9000753 0.9803852 1.0000000    0
rf   0.8444444 0.9026131 0.9600850 0.9488491 1.0000000 1.0000000    0
```

```
Random Forest

548 samples
  9 predictor
  2 classes: 'benign', 'malignant'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 494, 493, 493, 493, 492, 494, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  2     0.9764598  0.9488491
  5     0.9672679  0.9285802
  9     0.9654497  0.9246091

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```




```
Confusion Matrix and Statistics

           Reference
Prediction  benign malignant
  benign        86         1
  malignant      2        46

               Accuracy : 0.9778
                 95% CI : (0.9364, 0.9954)
    No Information Rate : 0.6519
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9513

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9773
            Specificity : 0.9787
         Pos Pred Value : 0.9885
         Neg Pred Value : 0.9583
             Prevalence : 0.6519
         Detection Rate : 0.6370
   Detection Prevalence : 0.6444
      Balanced Accuracy : 0.9780

       'Positive' Class : benign
```

Example 2.1

Regression decision tree for tumors

```
#Get and check data
>biopsy2 <- na.exclude((biopsy[,2:11]))
>biopsy2[,1:9] <- sapply(biopsy2[,1:9],as.numeric)
>head(biopsy2)

#Split into training and testing
>sample <- sample(c(TRUE, FALSE), nrow(biopsy2), replace=TRUE,
>prob=c(0.75,0.25))
>train  <- biopsy2[sample, ]
>test   <- biopsy2[!sample, ]
```

Example 2.2

Regression decision tree for tumors

#Modeling

```
>fit.dt <-rpart(class~., data=train, method="class")
>rpart.plot(fit.dt)
>fancyRpartPlot(fit.dt, caption=NULL)
```

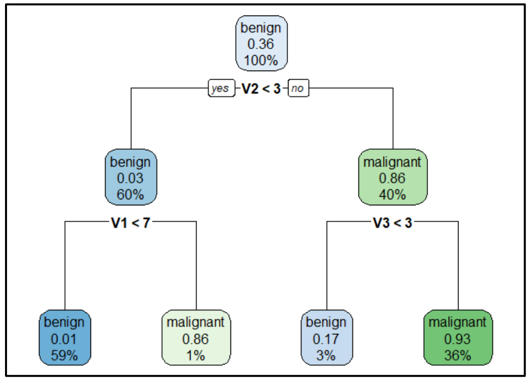
#Feature Importance

```
>varImp(fit.dt)
```

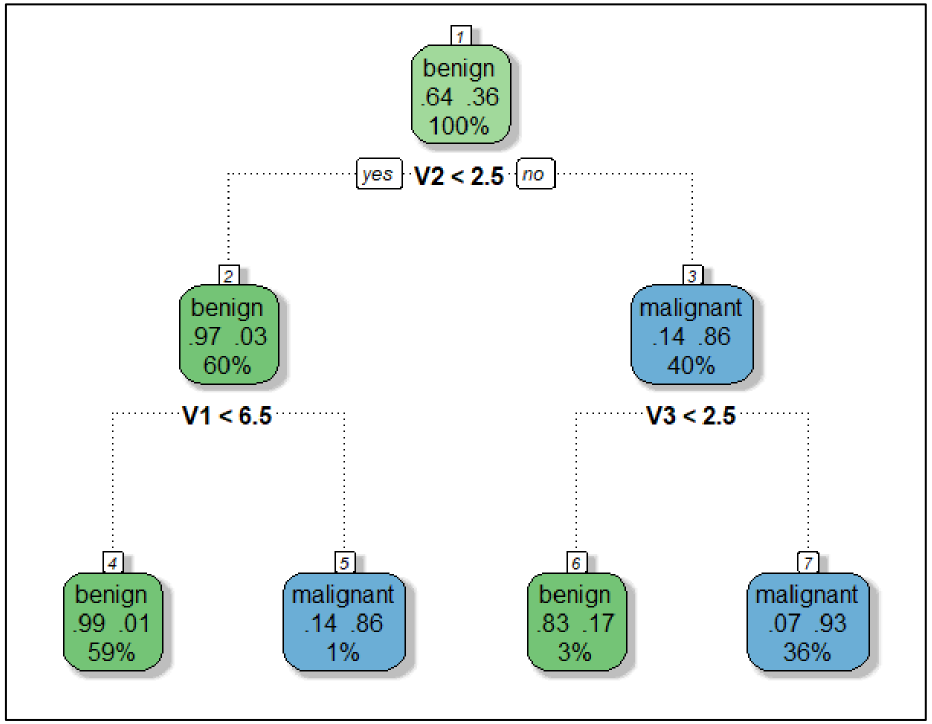
 # Make predictions

```
>predictions2 <-predict(fit.dt, test, type="class")
>head(predictions2)
>confusionMatrix(predictions2, test$class)
```

	Overall
V1	9.906541
V2	188.706682
V3	189.041728
V4	12.363090
V5	145.492667
V6	182.930631
V7	179.969586
V8	6.796046
V9	2.899512




```
Confusion Matrix and Statistics

           Reference
Prediction  benign malignant
  benign       100         5
  malignant      7        47

               Accuracy : 0.9245
                 95% CI : (0.8719, 0.9604)
    No Information Rate : 0.673
    P-Value [Acc > NIR] : 3.302e-14

                  Kappa : 0.8302

 Mcnemar's Test P-Value : 0.7728

            Sensitivity : 0.9346
            Specificity : 0.9038
         Pos Pred Value : 0.9524
         Neg Pred Value : 0.8704
             Prevalence : 0.6730
         Detection Rate : 0.6289
   Detection Prevalence : 0.6604
      Balanced Accuracy : 0.9192

       'Positive' Class : benign
```



DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Example 3

K-means clustering for tumors

#Get data

```
>biopsy3 <- na.exclude((biopsy[,2:11]))
>biopsy3 <- sapply(biopsy3[,1:9],as.numeric)
>head(biopsy3)
```

#Clustering

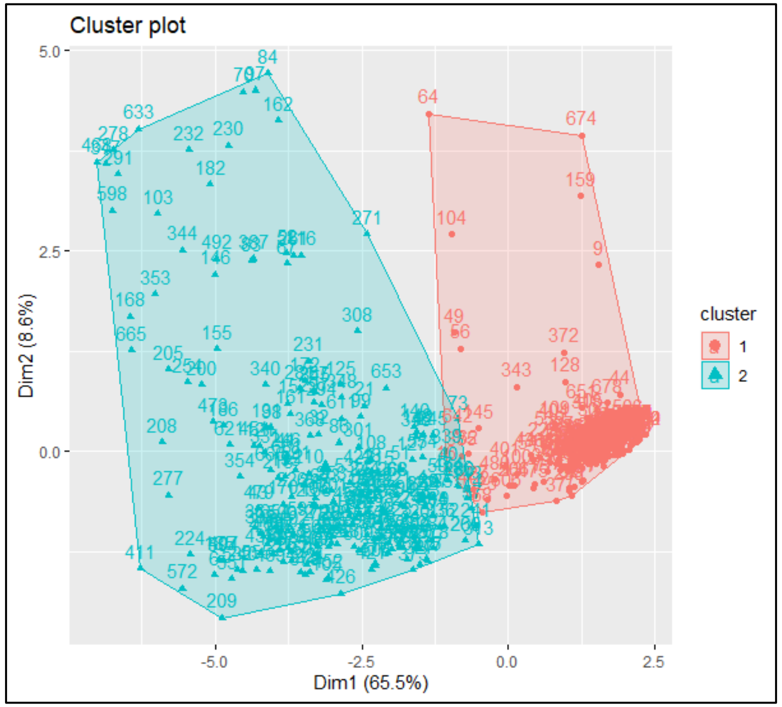
```
>set.seed(1)
>clust.km <-kmeans(biopsy3,2)
>clust.km
```

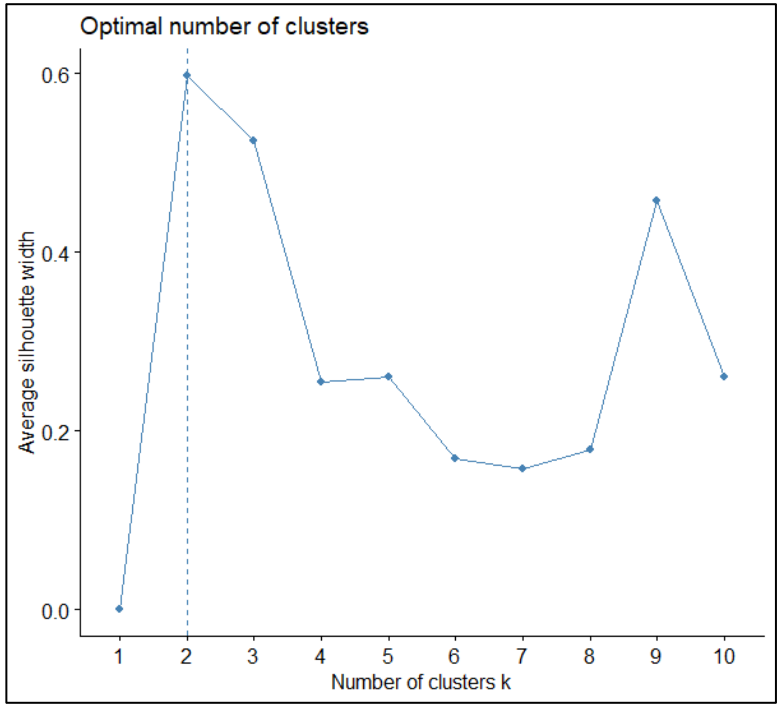
#Graphing

```
>fviz_cluster(clust.km, data=biopsy3)
```

#Optimal clusters

```
>fviz_nbclust(biopsy3[,1:9], kmeans,
              method = "silhouette")
```





Conclusions

- Machine Learning has many applications in biological and biomedical research
- Lots of techniques, which can be run using popular software
- Not as hard as you might think
- Most useful in predictive applications; inferential applications can use standard statistical methods

Please take the post-test and survey:

Post-test: https://und.qualtrics.com/jfe/form/SV_9XLyNZwobmkugU6
Survey: https://und.qualtrics.com/jfe/form/SV_cIokDfUYTfHjwrA

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

References 1

[1] https://www.ibm.com/cloud/learn/machine-learning

[2] https://azure.microsoft.com/en-us/overview/what-is-machine-learning-platform/

[3] https://www.mathworks.com/discovery/machine-learning.html

[4] https://hfenglab.org/NRev21.pdf

[5] https://www.sas.com/en_us/insights/analytics/machine-learning.html

[6] https://www.ibm.com/cloud/learn/machine-learning

[7] https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

[8] https://developer.ibm.com/articles/cc-models-machine-learning/#reinforcement-learning

[9] https://miro.medium.com/max/3056/1*_QGylwpgq831xl54cle_GQ.jpeg

[10] https://cdn.elearningindustry.com/wp-content/uploads/2017/05/73348f2f23b70566eef2d9f10f9fe22c.png

[11] https://www.kolabtree.com/blog/applications-of-machine-learning-in-biology/

[12] https://addepto.com/the-role-of-machine-learning-in-bioinformatics-and-biology/

[13] https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

[14] https://communities.sas.com/t5/SAS-Data-Science/machine-learning-using-base-SAS/td-p/139513

[15] https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning/

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

References 2

[16] https://towardsdatascience.com/machine-learning-in-julia-5bca700e0348

[17] https://www.geeksforgeeks.org/introduction-to-machine-learning-in-julia/

[18] https://www.geeksforgeeks.org/machine-learning-with-r/

[19] https://www.geeksforgeeks.org/7-best-r-packages-for-machine-learning/?ref=rp

[20] https://machinelearningmastery.com/machine-learning-in-r-step-by-step/

[21] https://www.geeksforgeeks.org/machine-learning/

[22] https://pythonforbiologists.com/machine-learning-for-biology-part-one.html

[23] https://edu.kpfu.ru/pluginfile.php/278552/mod_resource/content/1/MachineLearningR__Brett_Lantz.pdf

[24] https://towardsdatascience.com/10-most-brilliant-machine-learning-books-for-r-programmers-9e1780dd21f7

[25] http://www.philasug.org/Presentations/201711/Machine_Learning_for_SAS_Programmers_v3.pdf

[26] https://appsilon.com/r-decision-treees/

[27] https://www.gormanalysis.com/blog/decision-trees-in-r-using-rpart/

[28] https://www.r-bloggers.com/2021/04/cluster-analysis-in-r/

[29] https://data-flair.training/blogs/clustering-in-r-tutorial/

Acknowledgements

The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.

For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications: "***Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)***"