UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

3250 Foundations of Data Science

Module 9: Introduction to Regression and Classification

Course Plan

Module Titles
Module 1 – Introduction to Data Science
Module 2 – Introduction to Python
Module 3 – NumPy
Module 4 – Pandas
Module 5 – Data Collection and Cleaning
Module 6 – Descriptive Statistics and Visualization
Module 7 – Workshop (No Content)
Module 8 – Time Series
Current Focus: Module 9 – Introduction to Regression and Classification
Module 10 – Databases and SQL
Module 11 – Data Privacy and Security
Module 12 – Term Project Presentations (no content)

Learning Outcomes for this Module

- Identify the different types of machine learning algorithms
- Use statsmodels and scikit-learn

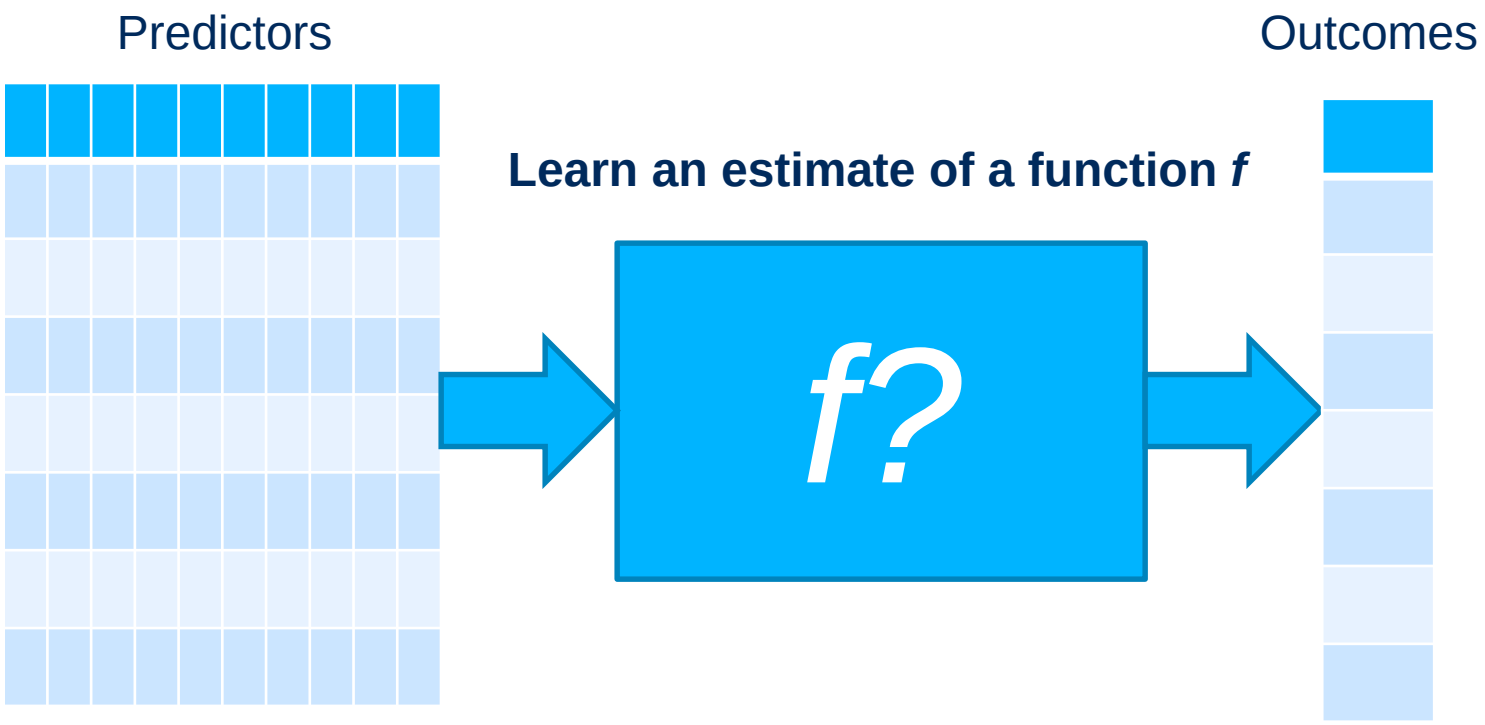
Topics for this Module

- **9.1** Commonly used algorithms and methods
- **9.2** Statsmodels
- **9.3** Scikit-learn
- **9.4** Resources and Homework

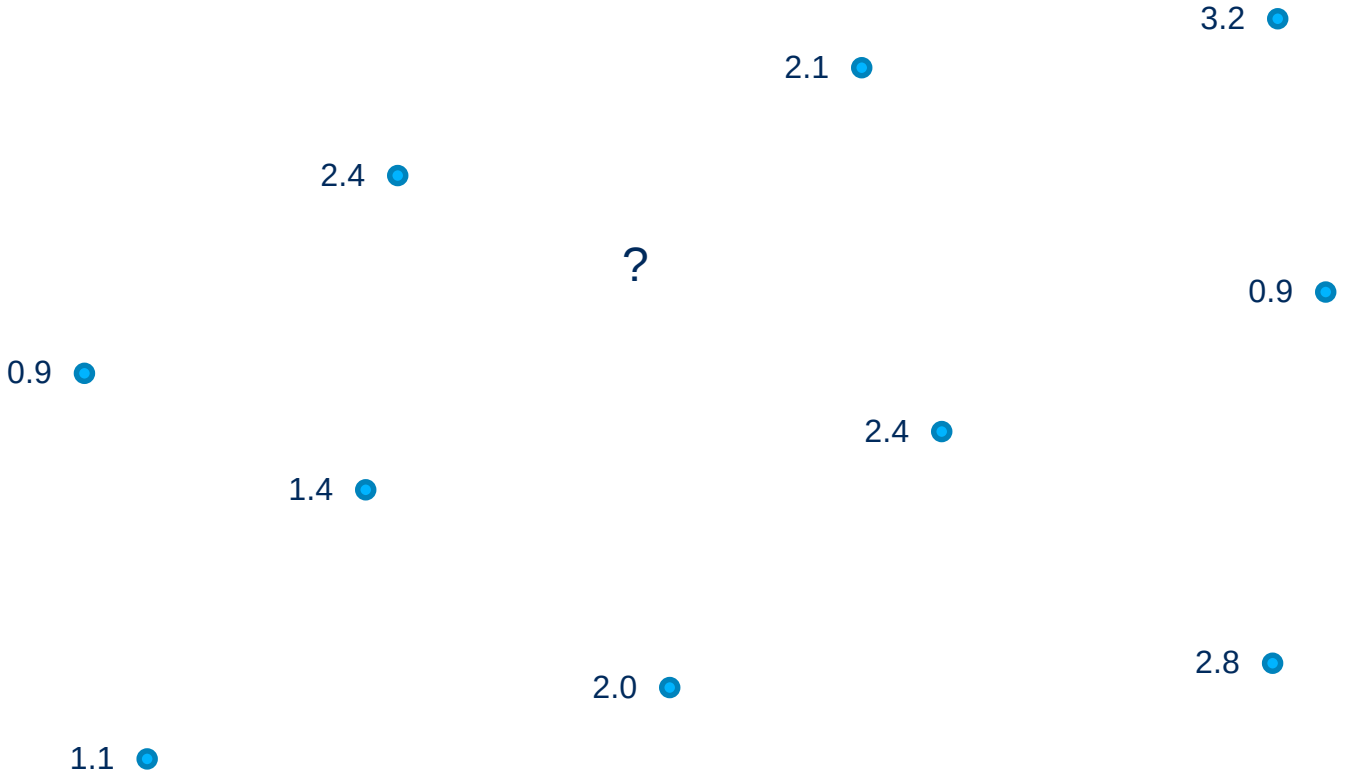
Module 9 – Section 1

Commonly-Used Algorithms and Methods

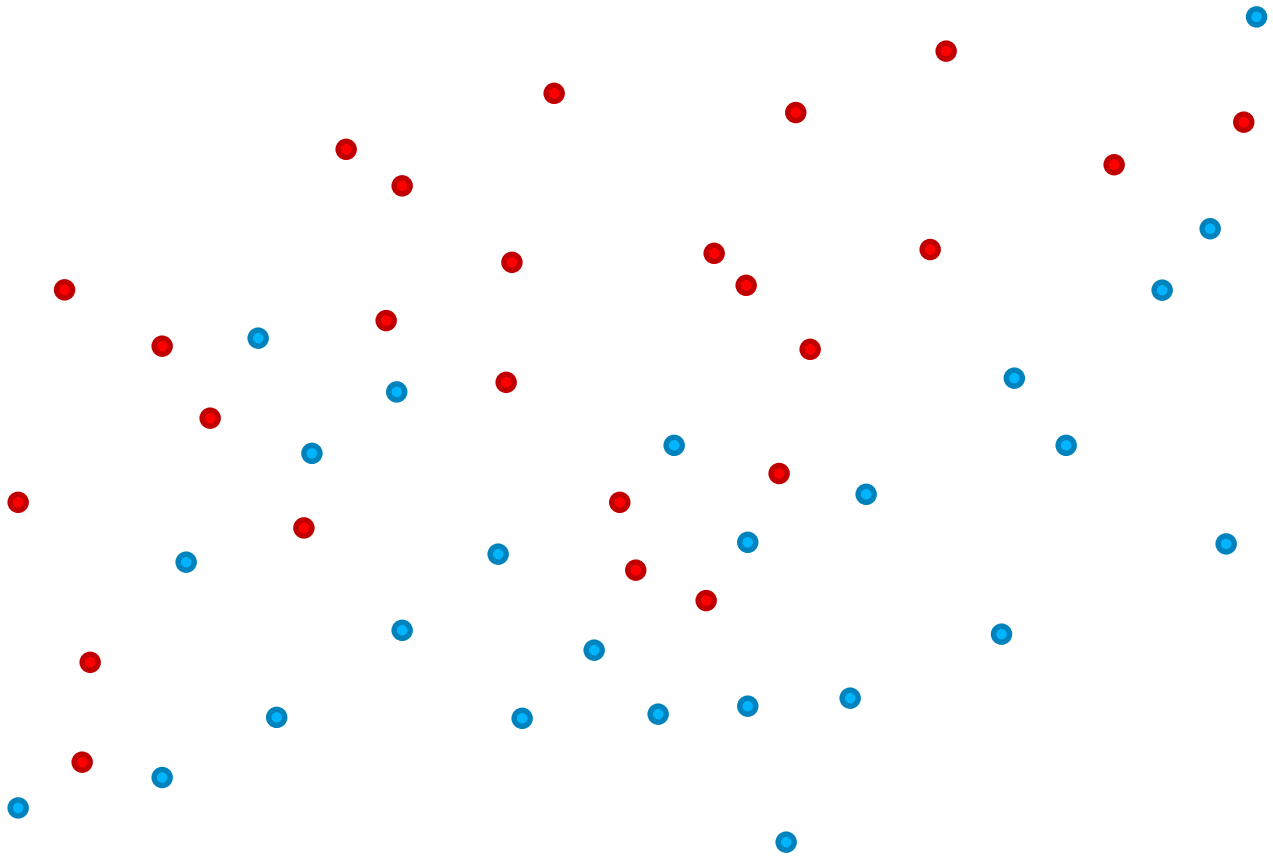
Statistical Machine Learning



Regression (2 Dimensions)



Classification



Regression & Classification

- We can think of the data as an (n-1) dimensional space + the target
- The predictors and target can be:
  - Numeric
  - Categorical
- If the target is numeric we call techniques for understanding the relationship between the predictors and the target (allowing us to make predictions) *Regression*
- If the target is categorical we call it *Classification*

List of Methods

- Over the next slides a brief introduction of important methods and algorithms commonly used in analytics and data science will be given
- More information will be given in subsequent courses
- These are:
  - k-Means
  - Linear Regression
  - Tree-Based Classifiers
  - Support Vector Machines
  - PCA

Clustering - Review

- An *unsupervised* method aiming to create groups of data points
- Data that belong in the same cluster are “close” to each other, while data that belong to different clusters are in general “apart”
- There are plenty of different clustering methods and algorithms
  - K-means algorithm
  - Centroid-based clustering
  - Hierarchical clustering
  - Distribution based clustering
  - Etc.

Clustering - Review (cont'd)

- Each one of these method has advantages and disadvantages and can be more or less appropriate depending on the case
- Appropriate metric (or "distance") needs to be chosen
  - Euclidean $\sum_i (a_i - b_i)^2$
  - Manhattan $\sum_i |a_i - b_i|$
  - Maximum $\max_i |a_i - b_i|$
  - Mahalanobis $\sqrt{(a-b)^T S^{-1} (a-b)}$

K-Means - Review

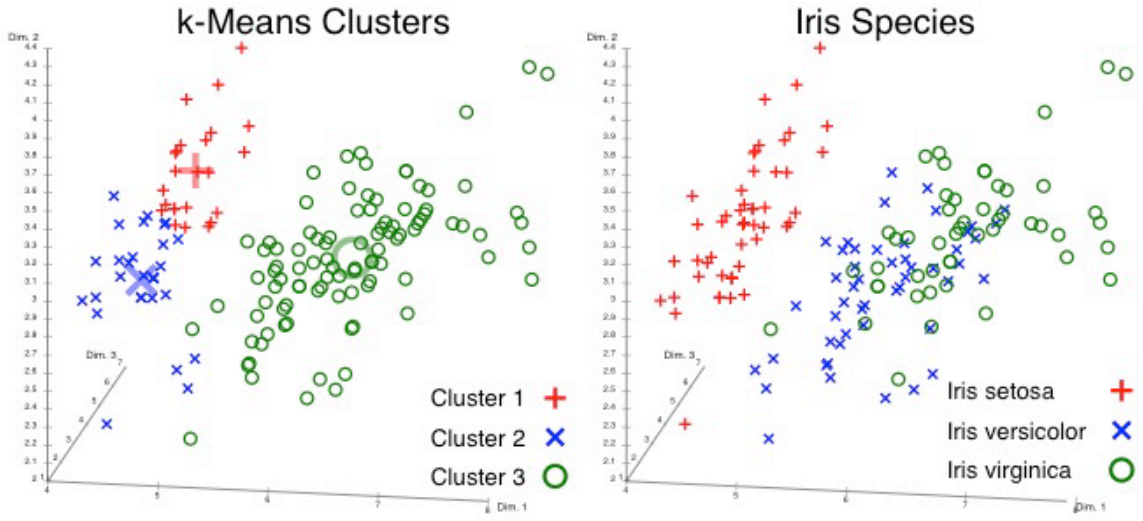
- An algorithm used for clustering
- Given a number k, a set of k clusters is determined so that the sum of the “distance” of each data point from the “centre” of the cluster is minimized
- Minimize the within-cluster sum of squares (WCSS)

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \left\| x - \mu_i \right\|^2$$

- Drawbacks
  - Number of clusters k is assumed known
  - It gives “spherical” clusters
  - It may get stuck at local minima
- Not appropriate if clusters are not spherical, number of clusters are not known, local minima exist, etc.

K-Means - Review (cont'd)

Example of iris data clustered using k-means with k=3



By Chire - Own work, Public Domain,
Source

k-Nearest Neighbours

- Similar to how we assign prices to homes for sale
- An example of a *non-parametric* method
- The distance measure can be in the space of data values
- We pick data points we already have that are close in the space as being representative
- For regression we can average k close examples
- For classification we can count the number of each type that fall in our k close neighbours

Linear Regression

- Regression = numeric prediction
- The idea is to model the relationship between a scalar dependent variable y and one or more independent variables X.
- Multivariate linear regression deals with the prediction of multiple correlated dependent variables, rather than a single scalar dependent variable.
- Nothing is really linear but the models are often useful
- Objective is to find the line (plane, hyperplane, …) that best fits the observations

Linear Regression (cont'd)

- Linear regression models are often fitted using the least squares approach by minimizing the squared error (the L_2 norm).
- They may also be fitted in other ways, e.g. by minimizing the absolute error (the L_1 norm), known as the Least Absolute Error (LAE)
- It assumes a linear relationship between the predictors $\{x_i\}_{i=1}^{n}$ and the response y
- A model is fitted given a set of data, estimating a set of unknown parameters β so that the sum of squares of the differences between the true and predicted responses is minimized

$$\sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2 = \left(y - X\beta\right)^T\left(y - X\beta\right)$$

Ordinary Least Squares Regression

- After we have estimated $\hat{\beta}$, the predicted (or fitted) values from the regression will be $\hat{y} = X\hat{\beta}$
- Method relies on a number of strong assumptions:
  - Linearity
  - Heteroscedasticity (equal variance)
  - Independence of the observations
  - Normality (for hypothesis testing)
- Method is robust, easy to implement and to interpret
- If assumptions do not hold, it does not work well
- There are other more complex alternatives

Naïve Bayesian Classification

- Treats incoming data as having a feature vector of individual, measurable properties of the instance
- Each property is termed a feature (or independent variable), although features may or may not be statistically independent)
- Classification is performed using the feature vector
- Simplifies the calculation of probabilities by assuming that the probability of each feature belonging to a given class value is independent of all other features – a strong assumption but results in a fast and effective method
- The probability of a class value given a value of a feature is called the conditional probability
- By multiplying the conditional probabilities together for each feature of a given class value, we get the probability of an instance belonging to that class
- To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

Naïve Bayesian Classification (cont'd)

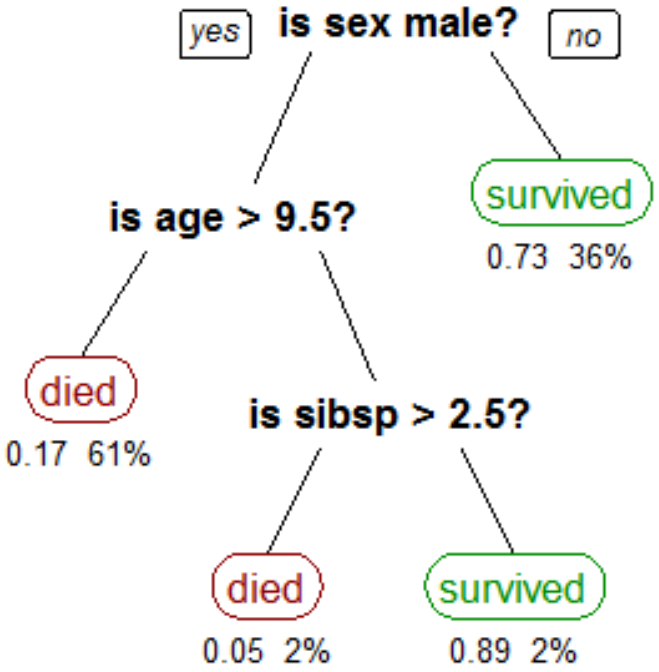
- It doesn't require a large amount of data to get started and can often produce acceptable results for many applications.
- Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam"..
  - The features may be the individual words of the email itself. Spam email will tend to feature sensational advertising terms and words in all caps.
  - By applying a probabilistic score to each feature, the naïve Bayesian classifier can produce a very good sense of the type of the email.
  - If a particular email scores highly on multiple features, it is very likely to be classified as spam.

Tree-based Classifiers

- An algorithm that is used to predict the class of an object based on a number of both categorical and continuous variables
- A tree-like model is built based on existing “training data set”, where each non-terminal node contains conditions related to the variables in the data
- Terminal nodes (leafs) correspond to the predicted classes
- For each new data point follows a path from the top of the tree to some leaf based on its variable values

Tree-based Classifiers (cont'd)

- Tree for classifying passengers of Titanic based on sex, age and number of spouse and siblings on board



By Stephen Milborrow - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=14143467

Trees

- Similar tree models can also be used for regression (i.e. continuous outcome)
- Together they go by the name CART (classification and regression tree) algorithm
- Benefits:
  - Easy to interpret
  - Accepts both continuous and categorical data
  - Ok with large data
- Disadvantages:
  - It can overfit the data (there are methods for *pruning*)
  - It does not work well with some problems (e.g. XOR)
  - Finding the optimal tree may be computationally infeasible (alternative feasible approximations may be suboptimal)

Random Forests

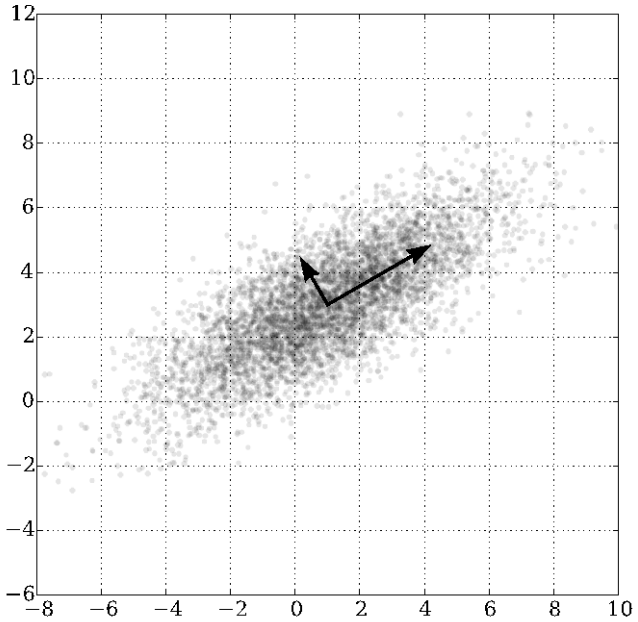
- If we limit a tree's growth it can't overfit
- But neither can it fit a complex function
- What if we create many small trees and have them vote (if a classifier) or average their predictions (if regression)?
- At each split in the tree, we select a random subset of features (and ignore the others)

Principal Component Analysis (PCA)

- PCA is an unsupervised method used to reduce the dimensionality of the data
- The *principal components* are a new set of co-ordinates linearly dependent on the original ones, that are also orthogonal
- The first component in the direction in the “feature space” where the data vary the most, the second component is the second varying direction etc.
- By “projecting” the data to the first few components, we represent them in a system of orthogonal, uncorrelated features that contain as much of the original variation as possible

Principal Component Analysis (Cont'd)

- PCA of a bivariate normal distribution: principal components are shown



By Nicoguaro - Own work, CC BY 4.0, Source

Principal Component Analysis (Cont'd)

- PCA is used for data exploration and understanding of the underline structure and relationship among the original features (covariates)
- Data can be clustered more easily in that new space
- However, PCA is sensitive to rescaling the data
- Also, pc's are only *linear* combinations of the original features
- Methods of non-linear dimensionality reduction exist

Module 9 – Section 2

Statsmodels

Statsmodels

- Regression models
  - Linear
  - Generalized linear
  - Robust
- Discrete choice (classification)
- Many models and functions for time series analysis
- Nonparametric estimators
- A collection of datasets for examples

Statsmodels (cont'd)

- A wide range of statistical tests
- Plays well with NumPy and Pandas
- statsmodels.sourceforge.net

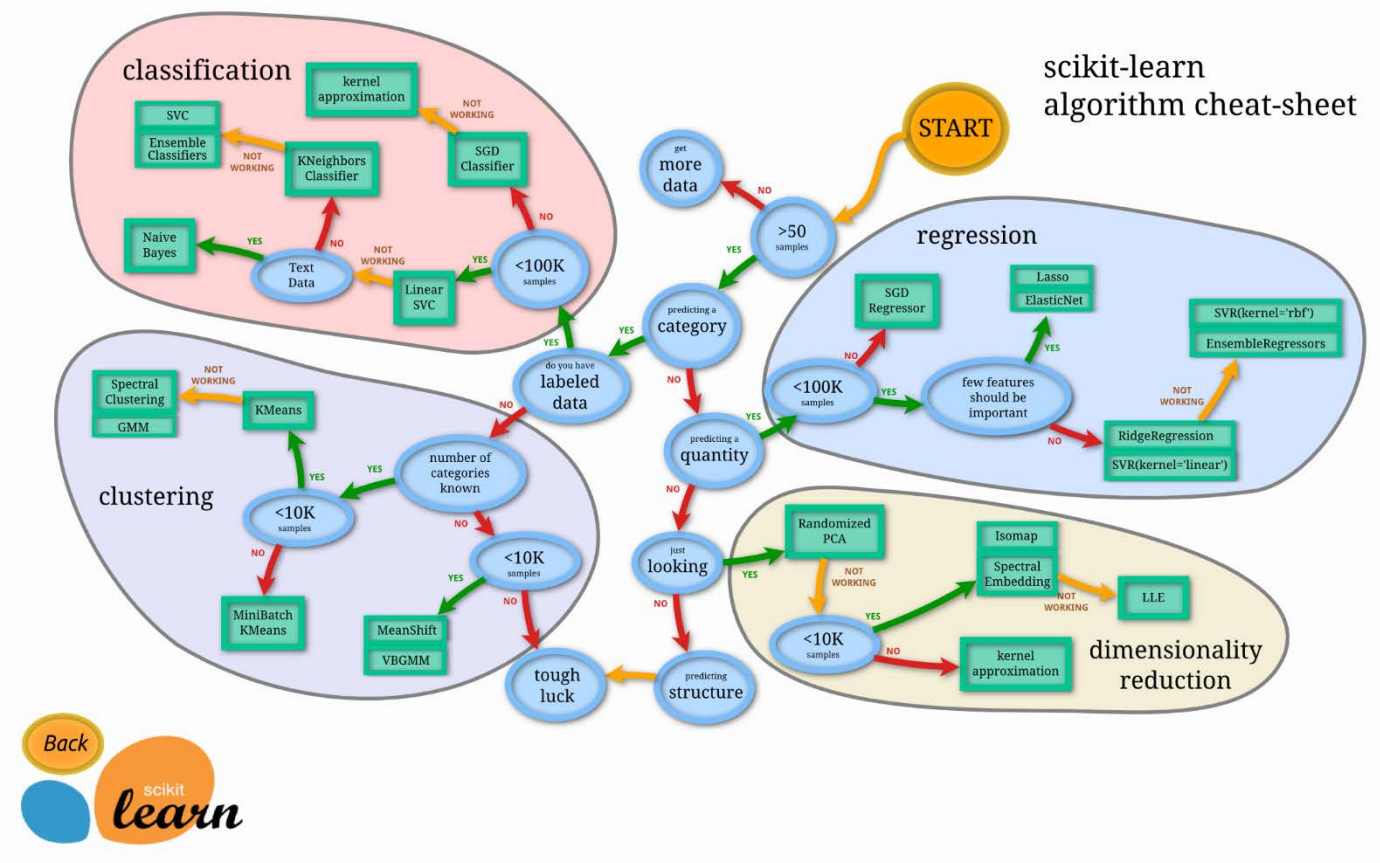
UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 9 – Section 3

Scikit-learn

Scikit-learn

- One of several SciPy toolkits
- Actively supported by INRIA and occasionally Google Summer of Code
- Used by Evernote, Spotify, yhat, booking.com
- Algorithms:


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
regression
predicting a category
YES
NO
do you have labeled data
YES
NO
SGD Regressor
Lasso
ElasticNet
<100K samples
NO
YES
few features should be important
YES
NO
SVR(kernel='rbf')
EnsembleRegressors
NOT WORKING
RidgeRegression
SVR(kernel='linear')
predicting a quantity
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
MiniBatch KMeans
<10K samples
YES
NO
MeanShift
VBGMM
just looking
YES
NO
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
<10K samples
YES
NO
kernel approximation
dimensionality reduction
tough luck
predicting structure
Back
scikit learn

Source

Linear Regression in Scikit-learn

```
from sklearn import linear_model
lm = linear_model.LinearRegression()
lm.fit(X_train, Y_train)
lm.predict(X_test)
lm.coef_
lm.intercept_
lm.score(X_test, Y_test)
```

Consistent Interface

- Training

```
estimator.fit(X_train, Y_train)
```

- Classification, regression, clustering

```
Y_test = estimator.predict(X_test)
```

- Filters, dimension reduction, latent variables

```
X_new = estimator.transform(X_test)
```

- Predictive models, density estimation

```
test_score = estimator.score(X_test, Y_test)
```

Hands-with Statsmodels and Scikit-learn

- Linear Regression with python(refer to "Linear Regression Connor Johnson.ipynb" - class exercise notebook)
- Link to the dataset used in this article

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 9 – Section 4

Resources and Homework

Resources

- statsmodels.sourceforge.net
- scikit-learn.org/stable
- www.ats.ucla.edu
- stats.stackexchange.com

Assignment

- Read Challenger Space Shuttle Disaster
- Go to O-Ring Data set
- Look at the main page and the Data Folder and Data Set Description page (links near top)
- The o-ring-erosion-or-blowby file is on the portal as o-ring-erosion-or-blowby.xlsx

Assignment (cont'd)

- "Blowby" means "leaking"
- Load the file into a pandas DataFrame
- Use statsmodels to do a multiple linear regression
- How many O-rings does the model predict will show erosion or blowby when the temperature is 31 degrees F? (We don't know how much pressure the rings will experience at liftoff so do predictions at 0, 50, 100 and 200 PSI to see what difference it makes.)

Next Class

- Databases & SQL

Follow us on social

Join the conversation with us online:

facebook.com/uoftscs

@uoftscs

linkedin.com/company/university-of-toronto-school-of-continuing-studies

@uoftscs

Any questions?

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Thank You

Thank you for choosing the University of Toronto School of Continuing Studies