CSRC Data Science Workshop

Dates: May 28-30, 2019

Description:

The workshop will include an introduction to data science, statistical and machine learning, and related topics. The two main methods used are random forests and neural networks. Both supervised and unsupervised learning will be covered. Classification and regression problems, as well as clustering for the two methods will be presented.

The format will be a series of short lectures, followed by corresponding labs. The labs will provide participants with hands-on experience implementing random forests and neural network methods.

Prerequisites:

A short review of the structure of linear and logistic regression will be provided to give some insight into the neuron model used in modern neural nets. During the workshop, we will cover some basic topics which are important to accuracy and generalization such as under and over fitting.

Knowledge of MATLAB well enough to follow code snippets and run numerical experiments is assumed. R at an introductory level is assumed.

Day 1:

1. Introduction to Data Science

- Objectives of the analyses: prediction, classify, cluster, dimension reduction, identify variables, establish relationships between variables, etc.
- Supervised and unsupervised learning
- Model assessment
- Model selection

2. Random Forests:

- trees and growing trees
- nonparametric bootstrap
- ensemble learning method and predictions
- tuning parameters
- variable importance

Applications: classification, regression

Day 2:

3. Neural Networks:

The workshop will cover the basics of modern neural networks. Emphasis will be given to supervised learning methods, starting from pre-trained models and moving to transfer learning. An applied perspective with examples in computer vision problems will guide the presentation. If time permits, we will cover some topics outside of computer vision, such as audio or sequence models.

Building blocks:

- fully connected layer, convolution layer, skip connections, weights, biases,
- ReLu, Sigmoids, pooling, stride, padding,
- drop-out, batch-norm, regularization, learning rate,
- stochastic gradient descent, mini batches, etc.

Architectures:

- LeNet, VGG, GoogleLeNet*, ResNet, DenseNet*,
- Autoencoders*

Applications:

- Classification, regression, segmentation*

Techniques:

- Visualize weights, dreaming, learning curves, etc.

* if time permits

Day 3:

4. Unsupervised Learning

- Introduction: distance or proximity matrix
- Algorithms

5. Clustering with Random Forests

6. Clustering with Neural Networks

Day 3: Afternoon

7. Analyze dataset of your choice

8. Short presentations of analyses and results