DATA 133 - Introduction to Data Science I

Instructor: Renzhi Cao
Computer Science Department
Pacific Lutheran University

Announcement

Course evaluations

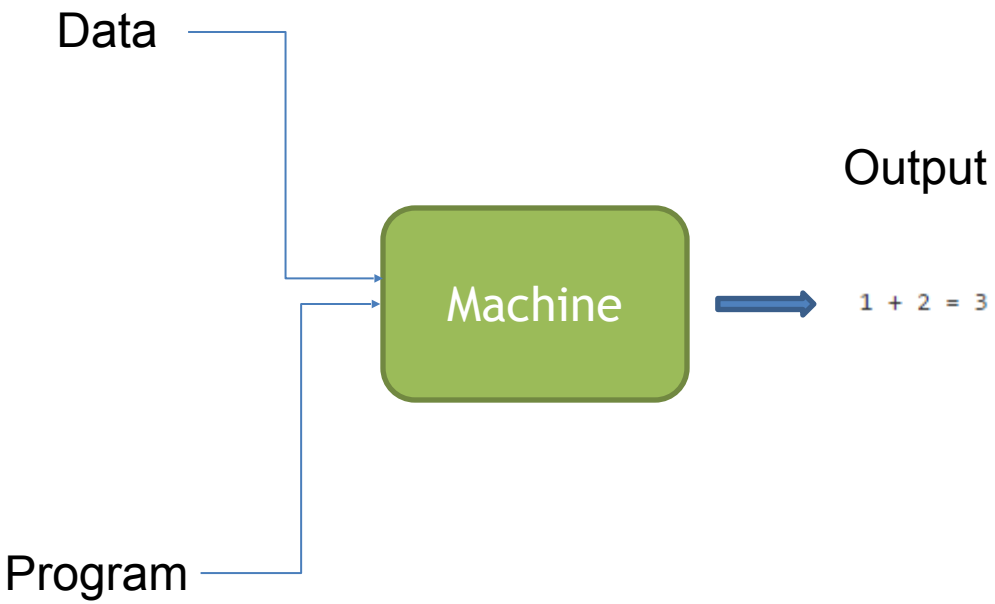
	Title	Response rate
DATA133	Intro to Data Science	2/23 = **8.7%**
CS270L02	Intro to Computer Science	0/11 = **0%**
CS330	Intro to Artificial Intelligence	7/24 = **29.17%**

- How is the project?
- Project labs next week

Machine learning - Neural Network

Traditional Programming

```
Please give two numbers:
1   2
```

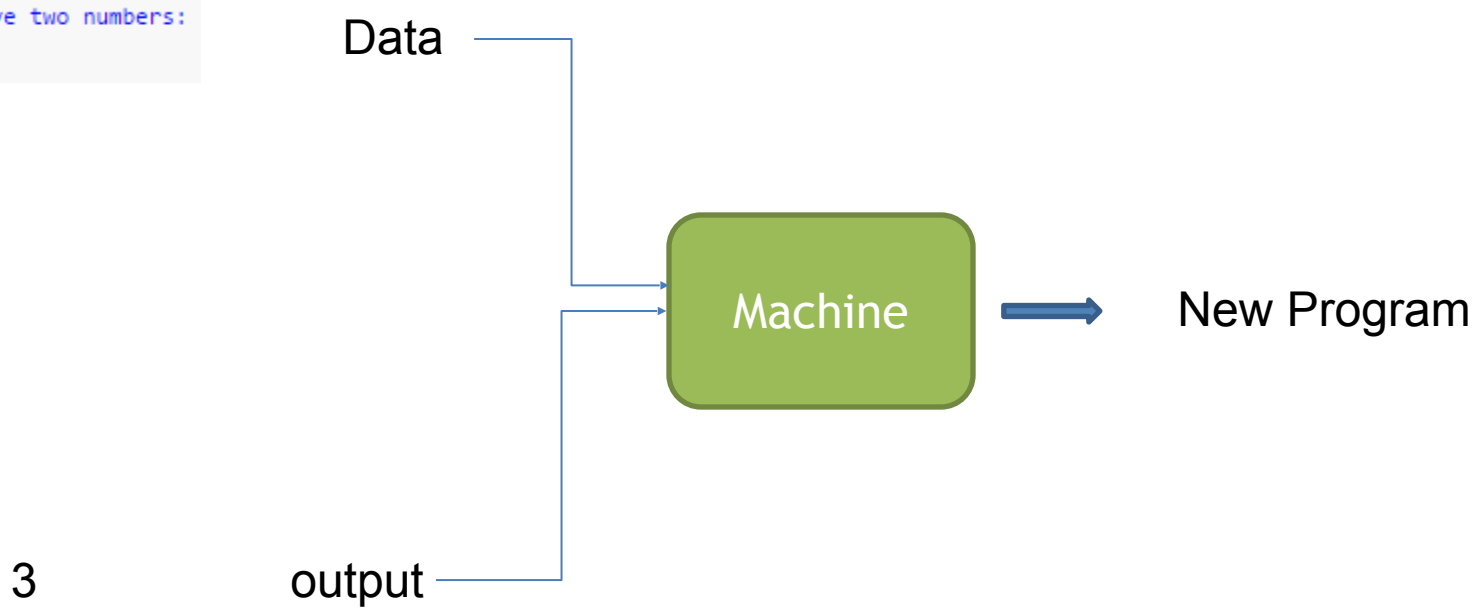


Source on Save

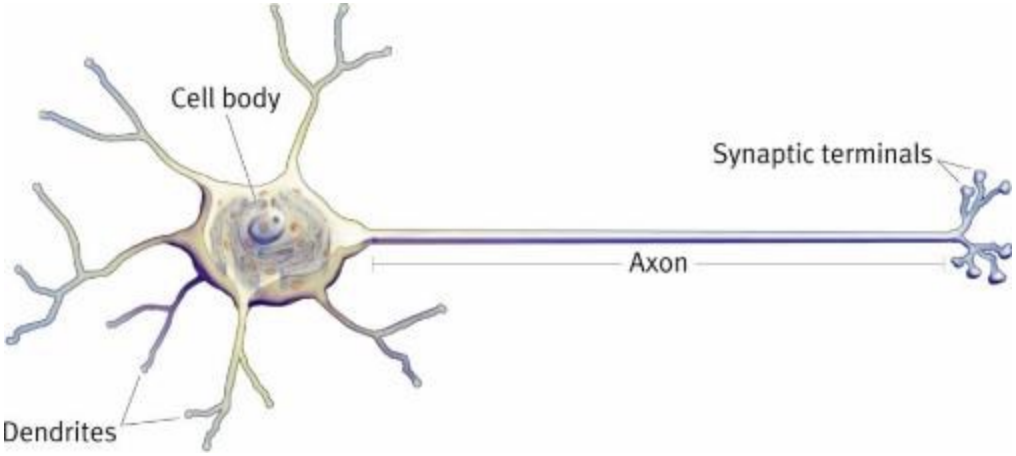
```
fAdd <- function(x,y){
  z <- x+y
  z
}
```

What is Machine learning?

```
Please give two numbers:
1   2
```



Neural Network



$$Y = w * x$$

output Weight Input

Input : 2
Output: 8

$$0 = w * x - Y$$

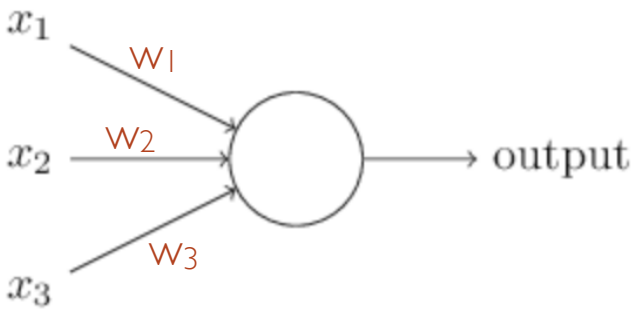
$$8 = w * 2$$

$$w = 8 \div 2$$

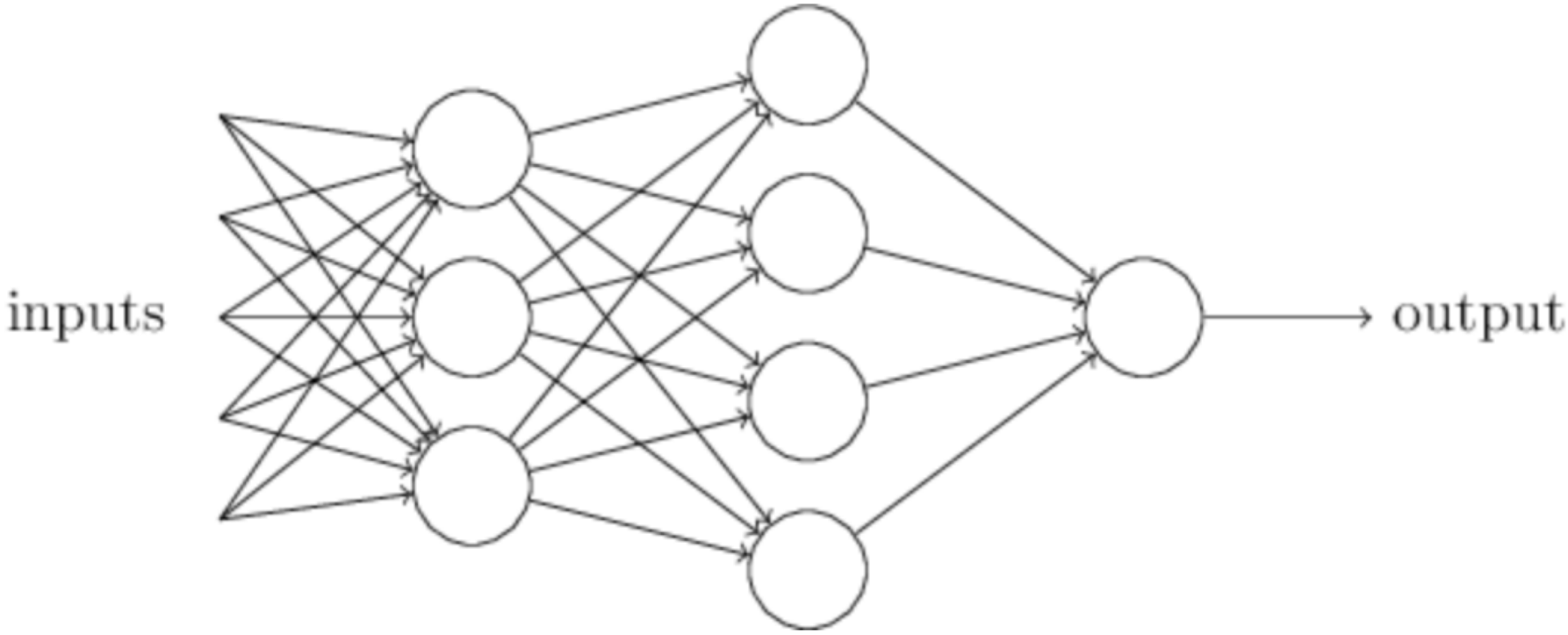
$$Error = |w * x - Y|$$

Feature units

decision units



Learned weight


inputs
output

Python: Pytorch

- https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html
- Open Run in Google Colab and try it.

Data preparation

ISLR's built in College Data Set which has several features of a college and a categorical column indicating whether or not the School is Public or Private.

```
#install.packages('ISLR')
library(ISLR)

print(head(College,2))
```

source: http://www.kdnuggets.com/2016/08/begineers-guide-neural-networks-r.html

Data processing

It is important to normalize data before training a neural network on it!

We use build-in scale() function to do that.

```
# Create Vector of Column Max and Min Values. apply(data, 1 for row, 2 for column, fun)
maxs <- apply(College[,2:18], 2, max)
mins <- apply(College[,2:18], 2, min)

# Use scale() and convert the resulting matrix to a data frame
scaled.data <- as.data.frame(scale(College[,2:18],center = mins, scale = maxs - mins))

# Check out results
print(head(scaled.data,2))
```

Train and Test Split

Training and testing dataset.

```
# Convert Private column from Yes/No to 1/0
Private = as.numeric(College$Private)-1
data = cbind(Private,scaled.data)

library(caTools)
set.seed(101)

# Create Split (any column is fine)
split = sample.split(data$Private, SplitRatio = 0.70)

# Split based off of split Boolean Vector
train = subset(data, split == TRUE)
test = subset(data, split == FALSE)
```

Neural Network Function

Before we actually call the neuralnetwork() function we need to create a formula to insert into the machine learning model

```
feats <- names(scaled.data)

# Concatenate strings
f <- paste(feats,collapse=' + ')
f <- paste('Private ~',f)

# Convert to formula
f <- as.formula(f)

f
```

Neural Network training

```
#install.packages('neuralnet')
library(neuralnet)
nn <- neuralnet(f,train,hidden=c(10,10,10),linear.output=FALSE)
```

```
# save your model and load it back for future usage
saveRDS(nn,"./nnModel.rds")

…

nn <- readRDS(“./nnModel.rds")
```

Predictions and Evaluations

We use the compute() function with the test data (jsut the features) to create predicted values.

```
# Compute Predictions off Test Set
predicted.nn.values <- compute(nn,test[2:18])

# Check out net.result
print(head(predicted.nn.values$net.result))
```

Predictions and Evaluations

Notice we still have results between 0 and 1 that are more like probabilities of belonging to each class.

```
predicted.nn.values$net.result <- sapply(predicted.nn.values$net.result,round,digits=0)
```

Now let's create a simple confusion matrix:

```
table(test$Private,predicted.nn.values$net.result)
```

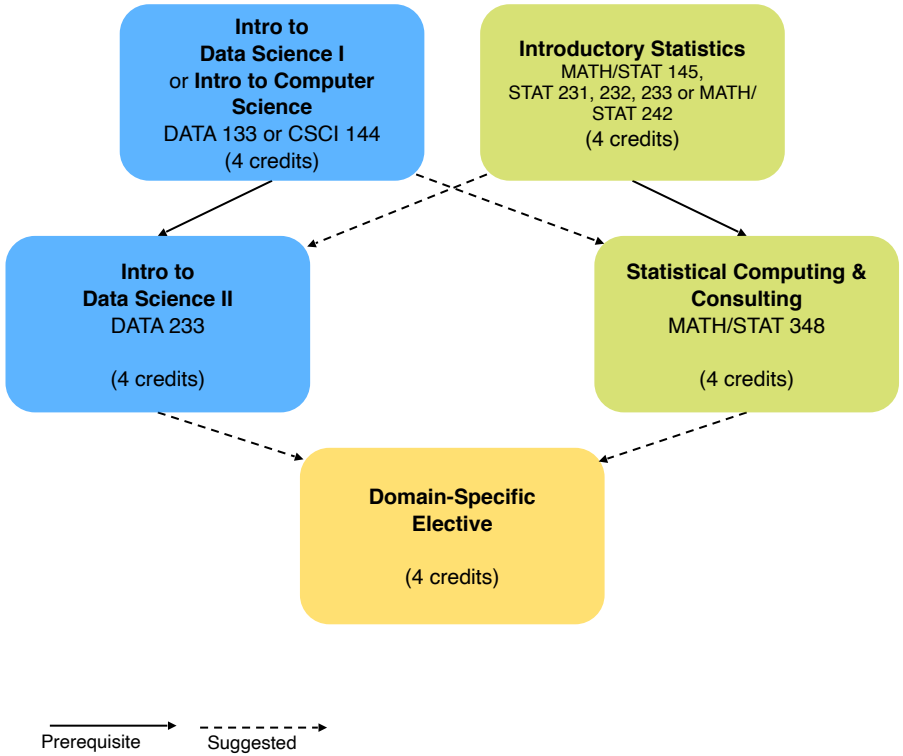
Visualizing the Neural Net

We can visualize the Neural Network by using the plot(nn) command.

Break

- Practice

Minor in Data Science



Requirements - *20 semester hours*

Computational and Data Science Foundations
8 semester hours

Statistical Foundations
8 semester hours

Domain-Specific Elective
4 semester hours

Prerequisite: Math 140 Precalculus or equivalent

https://www.plu.edu/computer-science/data-science/

DS 233: Intro to Data Science II (Spring 2020)

Learning objectives

- Learn how to get different type of data in Python.
- Learn how to process data in Python.
- Learn hands-on skills on data mining and machine learning techniques, and be able to build machine learning models in Python.
- Develop skills to analyze data from different fields, such as business field, bioinformatics field, etc.
- Develop skills to work on interdisciplinary projects.
- Develop teamwork skills using tools like Github.
- Learn hands-on skills to write SQL for storing, manipulating and retrieving data in databases.

Tentative - subject to change as course progresses

Date	Description	Weeks
2/8/2017, 2/10/2017	Introduction, basic Python programming and comparison with JAVA	Week 1
2/13/2017, 2/15/2017, 2/17/2017	Advanced Python programming and getting data from different sources	Week 2
2/20/2017, 2/22/2017, 2/24/2017	Working with data, cleaning and manipulating data **No classes, President's Day on 2/20**	Week 3
2/27/2017, 3/01/2017, 3/03/2017	Machine learning and k-Nearest Neighbor	Week 4
3/06/2017, 3/08/2017, 3/10/2017	Naive Bayes and Simple Linear Regression	Week 5
3/13/2017, 3/15/2017, 3/17/2017	Multiple regression	Week 6
3/20/2017, 3/22/2017, 3/24/2017	Logistic Regression	Week 7
3/27-3/31/2017	Invited talk for applications of data science in different fields, such as Business and biology. Mid-term exam	Week 8
4/3/2017, 4/5/2017, 4/7/2017	**Spring break**	Week 9
4/10/2017, 4/12/2017, 4/14/2017	Review mid term Decision Tree **Easter Break for 4/14**	Week 10
4/17/2017, 4/19/2017, 4/21/2017	Neural Networks and clustering	Week 11
4/24/2017, 4/26/2017, 4/28/2017	Natural Language Processing, Network analysis and recommender system	Week 12
5/1/2017, 5/3/2017, 5/5/2017	Databases and SQL	Week 13
5/8/2017, 5/10/2017, 5/12/2017	Topics of Data science on social networks, finance data, text mining, bio-tech analysis, web data.	Week 14
5/15/2017, 5/17/2017, 5/19/2017	Data science in Business, Biology, Geoscience, etc. Final project presentation (Report due on May 24th) Final Exams Review	Week 15
Final exam week	Final Exams on ??? Wednesday, May 24th, 2:00pm - 3:50pm	Week 16

Any suggestions?

Announcements

Extra credit for course evaluation!

>60 % - 10 bonus homework points

>70 % - 15 bonus homework points

>80 % - 20 bonus homework points

>90 % - 5 bonus points on final exam

>95 % - 7 bonus points on final exam

100 % - 10 bonus points on final exam!!!