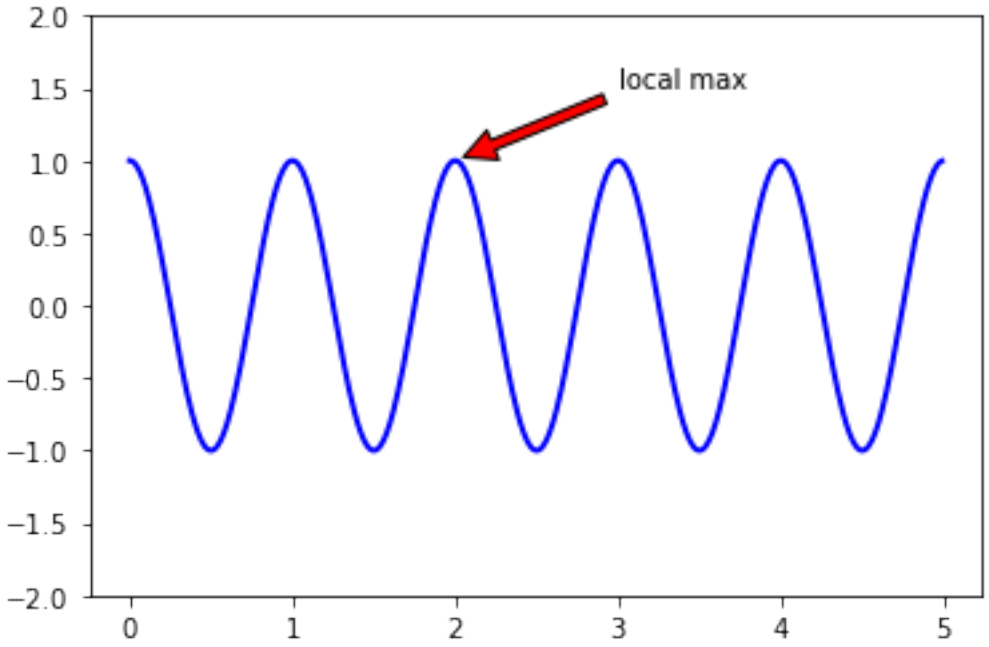
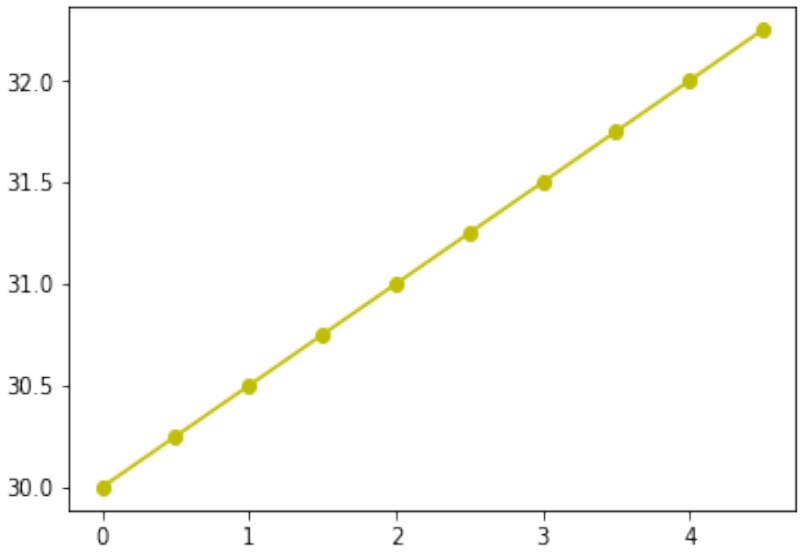
Assignment01

February 15, 2018

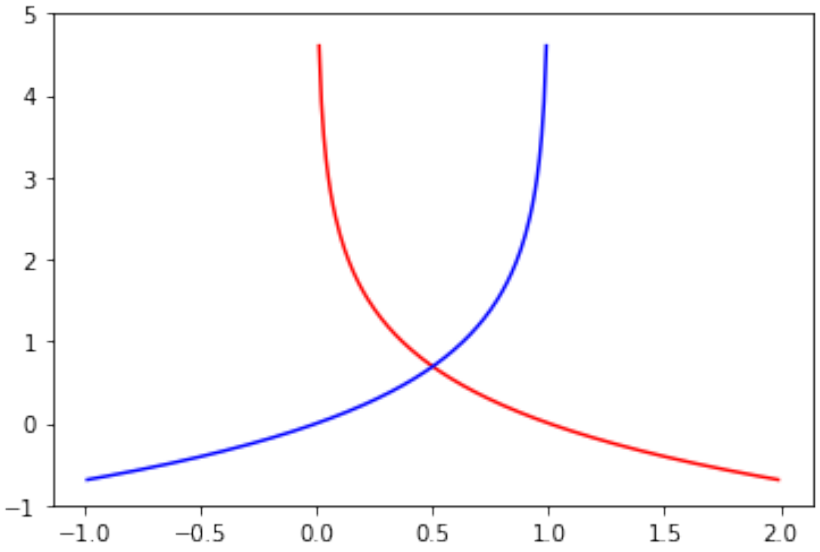
In [2]:
```
# Example
import numpy as np
import matplotlib.pyplot as plt

ax = plt.subplot(111)

t = np.arange(0.0, 5.0, 0.01) #start, stop, step
s = np.cos(2*np.pi*t)

line = plt.plot(t, s, 'b', linewidth=2)

plt.annotate('local max', xy=(2, 1), xytext=(3,1.5), arrowprops=dict(facecolor='red', shrink=0.05))
plt.ylim(-2,2)
plt.show()
```



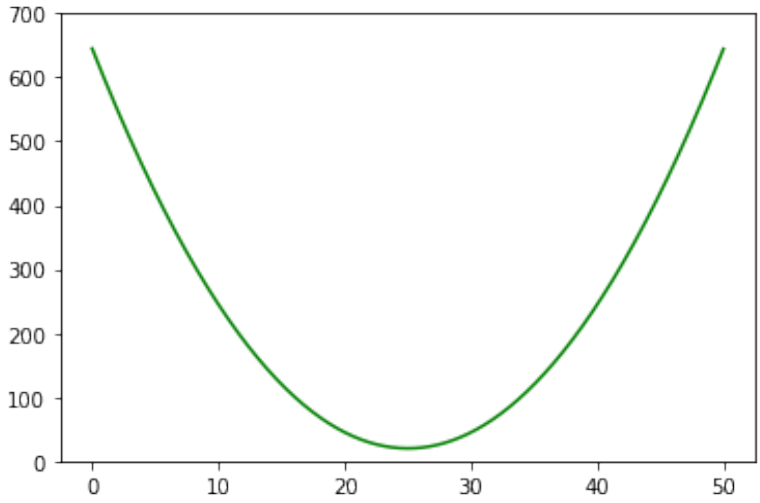
In [3]:
```
# Line function
x = np.arange(0,5,0.5)
y_intercept = 30
slope = .5
y = slope*x+y_intercept
plt.plot(x, y, 'y-o')
plt.show()
```



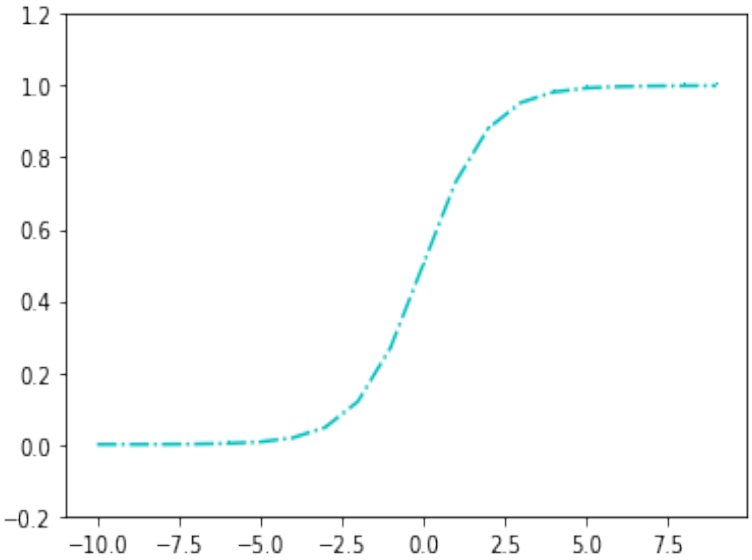
In [4]:
```
# Log function
domain_1 = np.arange(0.01,2.0, 0.01)
domain_2 = np.arange(0.99, -1, -0.01)
equation_1 = -np.log(domain_1)
equation_2 = -np.log(1-domain_2)
# plt.plot(x, equation_1, equation_2, 'r')          # means something else
plt.plot(domain_1, equation_1,'r')
plt.plot(domain_2, equation_2,'b')
plt.ylim(-1,5)
plt.show()
```



In [5]:
```
# Quadratic function
x = np.arange(0,50,.01)
theta_0 = 20
theta_1 = 25
y = (x - theta_1)**2 + theta_0
plt.plot(x, y, 'g')
plt.ylim(0,700)
plt.show()
```



In [6]:
```
# Sigmoid function
x = np.arange(-10,10)
y = 1/(1+np.e**(-x))
plt.plot(x,y, '-.,c')
plt.ylim(-.2,1.2) # first number == ymin; second == ymax in graph
plt.show()
```



1 Resources utilized:

- matplotlib.pyplot.plot() info
- matplotlib.pyplot.plot() tutorial
- numpy.arange() info
- numpy reference
- Jupyter help with install

2.1 (1 Credit) You are given the following tasks, all of which can be solved with a certain type of machine learning algorithms.

Classify a house to be single family or townhouse. A training set is available. Each training sample is provided with size, number of bed rooms, number of bathrooms, and house type (single family or townhouse).

Classify an email to be spam or not. Users already identified some emails to be spam and labeled it for training.

Human tumor Microarray data are provided as a matrix where row correspond to genes and columns to tissue samples. The task is to cluster columns (or samples) to identify disease profiles, i.e. tissues with similar disease should yield similar expression profiles.

Please select one of the following statements that are all correct:

b. i) supervised learning with discrete predictions; ii) supervised learning with discrete predictions; iii) unsupervised learning with discrete results;

3 To classify a fish to be a salmon or sea-bass, we collect a set of training samples, and measure the length of each sample fish. The category label of each training sample is also provided. Which descriptions are correct? (Mark all when applicable)

a. This problem can be solved by supervised learning that uses a single feature as predictor;

d. To solve this problem, we should compare different models, being complex or not, over both training data and testing data before making a selection;

4 *Task*. Please explain briefly the inputs, outputs, and your goal in general.

Data preparation. Please describe how to collect dataset, including training data, validation data, and testing data. Please explain how to get the ground-truth label for all samples;

Solution is on multiple following pages →

4

Discrete Supervised (Multi-class classification problem):
Strategy: Turn the multi-class problem into a binary classification problem

Task:

Goal:

Identify an airplane to be one of four types
Each airplane in training set has four attributes that can be analyzed along side the correct label of the airplane

Input:

- length of airplane from nose to tail
- length of airplane's wing span
- number of engines
- number of wheels

Output:

- the airplane is classified as a Airbus A-380
- the airplane is classified as a Boeing 777
- the airplane is classified as a Boeing 747
- the airplane is classified as a Cessna Skyhawk

Data Preparation:

Collecting dataset:

After granted access to go on airport tarmac make a table of data through recording the following procedures:

- Measure the length of the airplane and wingspan from nose to tail using a really long tape measure.
- Count the number of engines and wheels

4 continued...

Step I. **Preprocessing data** (Airbus A-380 or not):

- manipulate the label of Airbus A-380 to be encoded as **(1)**
- manipulate the label of Boeing 777 and Boeing 747 and Cessna Skyhawk to be **(-1)**, encoding other types of planes with same code, considering them to be one type, a negative type.
- replace the new labels to form a new coding system

Example:

(x1 , x2, x3, x4)

Where:

- x1: length of airplane from nose to tail
- x2: length of airplane's wing span
- x3: number of engines
- x4: number of wheels

Sample1 (x1 , x2, x3, Airbus A-380) → (x1 , x2, x3, **1**)

Sample2 (x1 , x2, x3, Boeing 777) → (x1 , x2, x3, **-1**)

Sample3 (x1 , x2, x3, Boeing 747) → (x1 , x2, x3, **-1**)

Sample4 (x1 , x2, x3, Cessna Skyhawk) → (x1 , x2, x3, **-1**)

Step II. **Looking at threshold** (quantize a distribution)

Look at table with 100 training samples

For each sample we have 4 training features and a label

Question: How do you train your model so that it can classify a flower to be (**1** or **-1**)?

Look into feature 1 (first column) and only look at samples with label (**1**), and check the variance of mean for this specific feature, try to quantize a distribution

Look into feature 1 (first column) and only look at samples with label (**-1**), and check the variance of mean for this specific feature, try to quantize a distribution

4 continued...

Note: Essentially visualizing the different curves
Then try to identify from which boundary we can use to separate the majority of the samples.

For 1 feature we now have a boundary to help make a decision

Repeat for different features to get different boundaries

Step III. For a new sample compare each feature value with each empirical boundary to determine a label

Step IV. Repeat for each individual flower_type

Short Recap for when you are trying to quantize the feature space into few beings:

Training process: for each being you will look into the table and check how many samples have this being's particular length

Testing process: the computer compares each feature value of the sample with the boundary estimation that the training process came up with and **classifies** this sample based off of the boundary decision (inequality/function)

For each feature of a sample, check where it lies in terms of the empirical boundary.

- if all label estimations predicted for each feature are same then classify sample to be that specific label
- if some of the labels do not match then vote based off of potentially adding feature weights

4 continued...

Training data:

Utilize a portion of the collected dataset

Use 50% of dataset to train model(s)

Validation data:

Use 30% of unused the validation test to determine which model function operates with better performance

Testing data:

Test using model that has better performance on the rest of the samples not used in the dataset

Compare the output labels to the actual labels

Obtaining a ground truth label for all samples:

The features should have been recorded with the actual label of the airplane

We can use these feature recordings and label to asses how well the model can predict based off of assigning statistics and percentage error