Linear regression workbook

This workbook will walk you through a linear regression example. It will provide familiarity with Jupyter Notebook and Python. Please print (to pdf) a completed version of this workbook for submission with HW #1.

ECE C147/C247, Winter Quarter 2021, Prof. J.C. Kao, TAs: N. Evirgen, A. Ghosh, S. Mathur, T. Monsoor, G. Zhao

In [1]:

```
import numpy as np
import matplotlib.pyplot as plt

#allows matlab plots to be generated in line
%matplotlib inline
```

Data generation

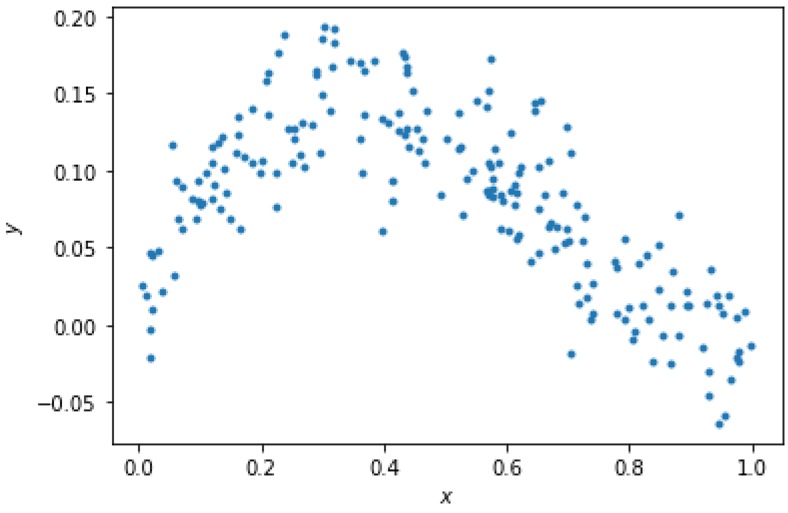
For any example, we first have to generate some appropriate data to use. The following cell generates data according to the model: $y = x - 2x^2 + x^3 + \epsilon$

In [2]:

```
np.random.seed(0)  # Sets the random seed.
num_train = 200     # Number of training data points

# Generate the training data
x = np.random.uniform(low=0, high=1, size=(num_train,))
y = x - 2*x**2 + x**3 + np.random.normal(loc=0, scale=0.03, size=(num_train,))
f = plt.figure()
ax = f.gca()
ax.plot(x, y, '.')
ax.set_xlabel('$x$')
ax.set_ylabel('$y$')
```

Out[2]:

```
Text(0, 0.5, '$y$')
```



QUESTIONS:

Write your answers in the markdown cell below this one:

(1) What is the generating distribution of x?

(2) What is the distribution of the additive noise ϵ?

ANSWERS:

(1) Uniform Distribution.

(2) Normal Distribution.

Fitting data to the model (5 points)

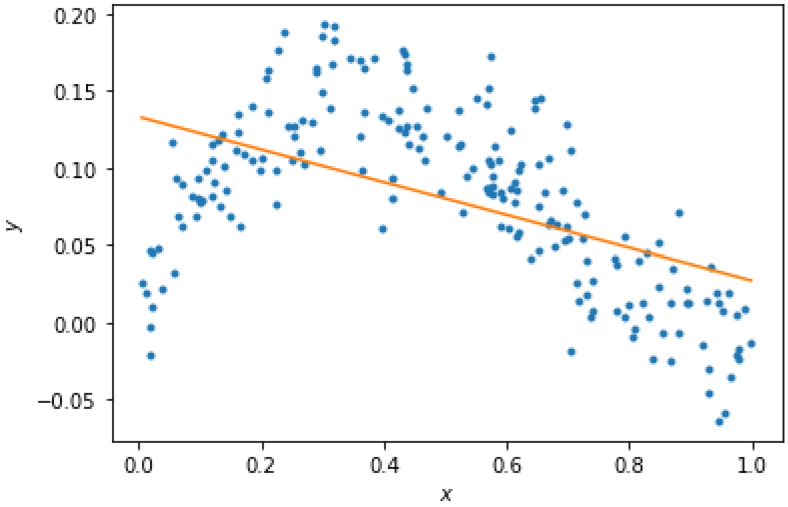
Here, we'll do linear regression to fit the parameters of a model $y = ax + b$.

In [7]:

```
# xhat = (x, 1)
xhat = np.vstack((x, np.ones_like(x)))

# =================== #
# START YOUR CODE HERE #
# =================== #
# GOAL: create a variable theta; theta is a numpy array whose elements are [a, b]

theta =np.linalg.inv(xhat.dot(xhat.T)).dot(xhat.dot(y)) # please modify this line

# ================= #
# END YOUR CODE HERE #
# ================= #
```

```
In [8]:  # Plot the data and your model fit.
         f = plt.figure()
         ax = f.gca()
         ax.plot(x, y, '.')
         ax.set_xlabel('$x$')
         ax.set_ylabel('$y$')

         # Plot the regression line
         xs = np.linspace(min(x), max(x),50)
         xs = np.vstack((xs, np.ones_like(xs)))
         plt.plot(xs[0,:], theta.dot(xs))
```

```
Out[8]:  [<matplotlib.lines.Line2D at 0x29b7a6f1188>]
```



QUESTIONS

(1) Does the linear model under- or overfit the data?

(2) How to change the model to improve the fitting?

ANSWERS

(1) The linear model underfit the data.

(2) Try polynomial regression with an appropriate order.

Fitting data to the model (10 points)

Here, we'll now do regression to polynomial models of orders 1 to 5. Note, the order 1 model is the linear model you prior fit.

In [32]:

```
N = 5
xhats = []
thetas = []

# ==================== #
# START YOUR CODE HERE #
# ==================== #

# GOAL: create a variable thetas.
# thetas is a list, where theta[i] are the model parameters for the polynomial
fit of order i+1.
#   i.e., thetas[0] is equivalent to theta above.
#   i.e., thetas[1] should be a length 3 np.array with the coefficients of the
x^2, x, and 1 respectively.
#   ... etc.

for i in range(N):
    xhat_temp = xhat
    for j in range(i):
        xhat_temp = np.vstack((np.power(x, j+2), xhat_temp))

    xhats.append(xhat_temp)
    theta_temp =np.linalg.inv(xhat_temp.dot(xhat_temp.T)).dot(xhat_temp.dot(y
))
    thetas.append(theta_temp)

# ================== #
# END YOUR CODE HERE #
# ================== #
```

In [33]:

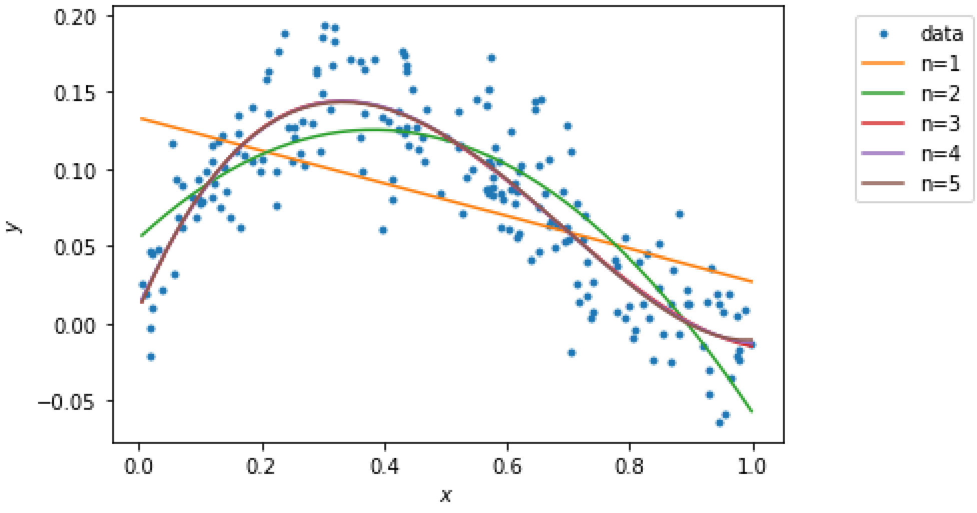
```
thetas[4].shape
```

Out[33]:

```
(6,)
```

In [48]:

```
# Plot the data
f = plt.figure()
ax = f.gca()
ax.plot(x, y, '.')
ax.set_xlabel('$x$')
ax.set_ylabel('$y$')

# Plot the regression lines
plot_xs = []
for i in np.arange(N):
    if i == 0:
        plot_x = np.vstack((np.linspace(min(x), max(x),50), np.ones(50)))
    else:
        plot_x = np.vstack((plot_x[-2]**(i+1), plot_x))
    plot_xs.append(plot_x)

for i in np.arange(N):
    ax.plot(plot_xs[i][-2,:], thetas[i].dot(plot_xs[i]))

labels = ['data']
[labels.append('n={}'.format(i+1)) for i in np.arange(N)]
bbox_to_anchor=(1.3, 1)
lgd = ax.legend(labels, bbox_to_anchor=bbox_to_anchor)
```



Calculating the training error (10 points)

Here, we'll now calculate the training error of polynomial models of orders 1 to 5.

In [65]:

```
training_errors = []

# ==================== #
# START YOUR CODE HERE #
# ==================== #

# GOAL: create a variable training_errors, a list of 5 elements,
# where training_errors[i] are the training loss for the polynomial fit of ord
er i+1.
for i in range(N):
    yhat = thetas[i].dot(xhats[i])
    training_error_temp = np.sum(1/2*(yhat - y)**2)
    training_errors.append(training_error_temp)

# ================== #
# END YOUR CODE HERE #
# ================== #

print ('Training errors are: \n', training_errors)
```

```
Training errors are: 
 [0.23799610883627012, 0.1092492220926853, 0.08169603801105371, 0.08165353735
29698, 0.08161479195525301]
```

QUESTIONS

(1) What polynomial has the best training error?

(2) Why is this expected?

ANSWERS

(1) 5th-order polynomial has the best training error.

(2) A high order polynomial has more freedom to fit data points more closely with more $a_n x^n$ terms.

Generating new samples and testing error (5 points)

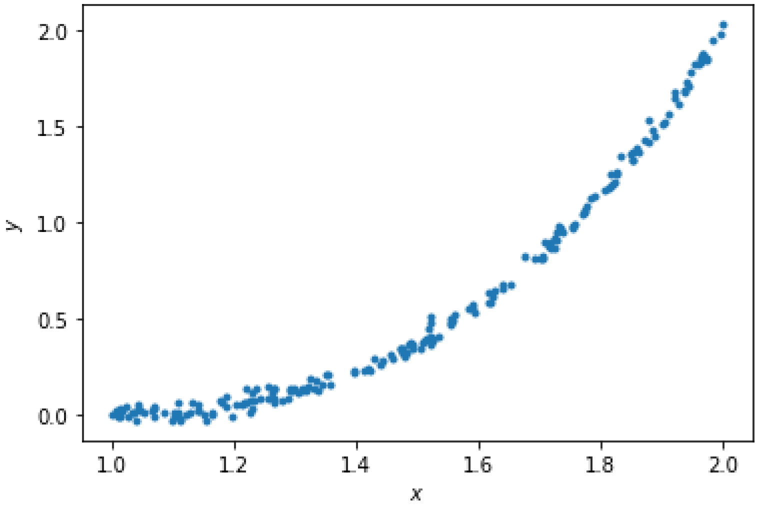
Here, we'll now generate new samples and calculate testing error of polynomial models of orders 1 to 5.

In [78]:

```
x = np.random.uniform(low=1, high=2, size=(num_train,))
y = x - 2*x**2 + x**3 + np.random.normal(loc=0, scale=0.03, size=(num_train,))
f = plt.figure()
ax = f.gca()
ax.plot(x, y, '.')
ax.set_xlabel('$x$')
ax.set_ylabel('$y$')
```

Out[78]:

```
Text(0, 0.5, '$y$')
```

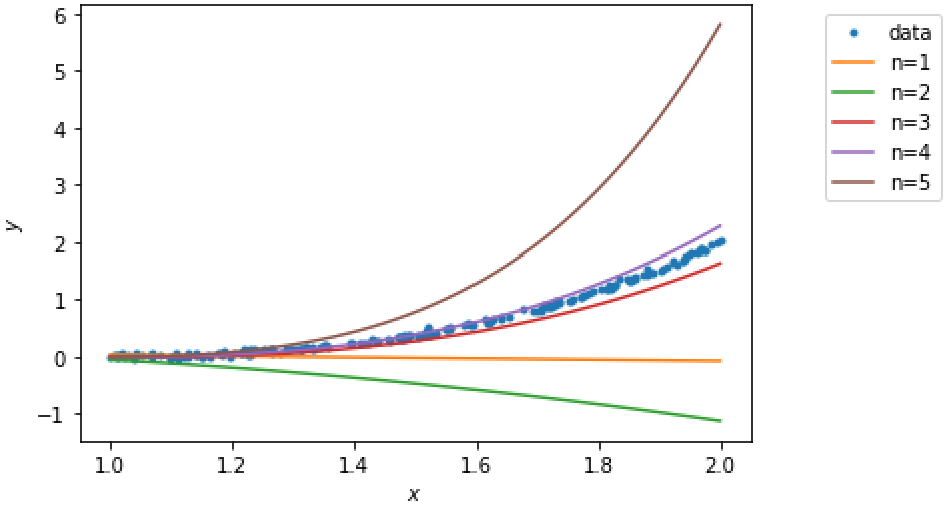


In [79]:

```
xhats = []
for i in np.arange(N):
    if i == 0:
        xhat = np.vstack((x, np.ones_like(x)))
        plot_x = np.vstack((np.linspace(min(x), max(x),50), np.ones(50)))
    else:
        xhat = np.vstack((x**(i+1), xhat))
        plot_x = np.vstack((plot_x[-2]**(i+1), plot_x))

    xhats.append(xhat)
```

In [80]:

```
# Plot the data
f = plt.figure()
ax = f.gca()
ax.plot(x, y, '.')
ax.set_xlabel('$x$')
ax.set_ylabel('$y$')

# Plot the regression lines
plot_xs = []
for i in np.arange(N):
    if i == 0:
        plot_x = np.vstack((np.linspace(min(x), max(x),50), np.ones(50)))
    else:
        plot_x = np.vstack((plot_x[-2]**(i+1), plot_x))
    plot_xs.append(plot_x)

for i in np.arange(N):
    ax.plot(plot_xs[i][-2,:], thetas[i].dot(plot_xs[i]))

labels = ['data']
[labels.append('n={}'.format(i+1)) for i in np.arange(N)]
bbox_to_anchor=(1.3, 1)
lgd = ax.legend(labels, bbox_to_anchor=bbox_to_anchor)
```



In [82]:

```
testing_errors = []

# ==================== #
# START YOUR CODE HERE #
# ==================== #

# GOAL: create a variable testing_errors, a list of 5 elements,
# where testing_errors[i] are the testing loss for the polynomial fit of order
i+1.
for i in range(N):
    yhat = thetas[i].dot(xhats[i])
    testing_error_temp = np.sum(1/2*(yhat - y)**2)
    testing_errors.append(testing_error_temp)

# ================== #
# END YOUR CODE HERE #
# ================== #

print ('Testing errors are: \n', testing_errors)
```

```
Testing errors are: 
 [76.99016367401441, 203.7325309077172, 3.152005458351538, 0.995664253893641
5, 195.75036658114402]
```

QUESTIONS

(1) What polynomial has the best testing error?

(2) Why polynomial models of orders 5 does not generalize well?

ANSWERS

(1) 4th order polynomial has the best testing error.

(2) 5th order polynomial has the best training error since a higher order polynomial has more freedom to approximate the limited training data so that the model tends to learn the details and noise in the training data rather than the overall trend of the data.