Problem 1: Linear algebra refresher

a) i Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ $AA^T = I$

$$\Rightarrow \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \begin{matrix} a^2+b^2=1 \\ ac+bd=0 \\ b^2+d^2=1 \end{matrix}$$

By inspection:

If $a^2 = \frac{1}{2}$, $b^2 = \frac{1}{2}$, then $a, b = \frac{1}{\sqrt{2}}$ and $a^2+b^2=1$

Same for $c^2, d^2 = \frac{1}{2}$, but if $c = -\frac{1}{\sqrt{2}}$, then $ac+bd = -\frac{1}{2}+\frac{1}{2}=0$ ✓

$a, b, d = \frac{1}{\sqrt{2}}$ $c = -\frac{1}{\sqrt{2}}$

$$A = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$$Av = \lambda v$$

$$(A-\lambda)v = 0$$

For non-trivial solutions solve $|A - \lambda I| = 0$

$$\Rightarrow \left|\begin{bmatrix} \frac{1}{\sqrt{2}}-\lambda & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}-\lambda \end{bmatrix}\right| = 0$$

$$\left(\tfrac{1}{\sqrt{2}}-\lambda\right)\left(-\tfrac{1}{\sqrt{2}}-\lambda\right) - \tfrac{1}{2} = 0$$

$$-\tfrac{1}{2} - \tfrac{1}{\sqrt{2}}\lambda + \tfrac{1}{\sqrt{2}}\lambda + \lambda^2 - \tfrac{1}{2} = 0$$

$$\lambda^2 = 1$$

$\boxed{\lambda_1 = 1, \lambda_2 = -1}$

$(A-\lambda_1)V_1 = 0$ Let $V_{11} = 1$ $\Rightarrow$ $\frac{1}{\sqrt{2}} - 1 + \frac{1}{\sqrt{2}}V_{12} = 0$ $V_{12} = \sqrt{2}\left(1 - \frac{1}{\sqrt{2}}\right) = \sqrt{2}-1$

$V_1 = \begin{bmatrix} 1 \\ \sqrt{2}-1 \end{bmatrix}$

↳ scale invariance of eigenvectors

$(A-\lambda_2)V_2 = 0$ Let $V_{21} = 1$ $\Rightarrow$ $\frac{1}{\sqrt{2}} + 1 + \frac{1}{\sqrt{2}}V_{22} = 0$ $V_{22} = \sqrt{2}\left(-1 - \frac{1}{\sqrt{2}}\right) = -\sqrt{2}-1$

$V_2 = \begin{bmatrix} 1 \\ -\sqrt{2}-1 \end{bmatrix}$

$V_1 \cdot V_2 = 1 + (\sqrt{2}-1)(-\sqrt{2}-1) = 1 - 2 + \sqrt{2} - \sqrt{2} + 1 = 0$

$V_1 \perp V_2$

$\boxed{\lambda\text{'s are } 1, -1 \text{ and eigenvectors are orthogonal}}$

ii

Eig definition: $Av = \lambda v$

Norm 2 of both sides: $\|Av\|_2 = \|\lambda v\|_2$

$$\Rightarrow \sqrt{(Av)^T Av} = \|\lambda\| \, \|v\| \quad (\lambda \text{ scalar})$$

$(Av)^T Av = v^T A^T A v = v^T I v = v^T v$

$$\Rightarrow \sqrt{v^T v} = \|\lambda\| \sqrt{v^T v} \Rightarrow \boxed{\|\lambda\| = 1}$$

Thus A has λ's where $\|\lambda\| = 1$

iii

Goal: show $v_1^T v_2 = 0$ if $\lambda_1 \neq \lambda_2$

$A v_1 = \lambda_1 v_1 \qquad A v_2 = \lambda_2 v_2$

$$v_1^T v_2 = v_1 A^T A v_2 = \{(A v_1)^T A v_2 = (\lambda_1 v_1)^T \lambda_2 v_2\}$$

$$\Rightarrow v_1^T A^T A v_2 = \lambda_1 \lambda_2 v_1^T v_2$$

($A^T A = I$)

$$(1 - \lambda_1 \lambda_2) v_1^T v_2 = 0$$

Since $\|\lambda\| = 1$, and $\lambda_1 \neq \lambda_2$, then $v_1^T v_2 = 0$,

thus $v_1 \perp v_2$

iv

x can be reflected and rotated, but it cannot be scaled based on the properties derived above

b) i

SVD A: $A = U\Sigma V^T$ where U and V are orthogonal ($U^TU, V^TV = I$)

$$AA^T = U\Sigma V^T (U\Sigma V^T)^T$$
$$= U\Sigma V^T V \Sigma^T U^T$$
$$= U\Sigma\Sigma^T U^T$$

left sv of A is eigV of AA^T

$$A^TA = (U\Sigma V^T)^T U\Sigma V^T$$
$$= V\Sigma^T U^T U\Sigma V^T$$
$$= V\Sigma^T\Sigma V^T$$

right sv of A is eigV of A^TA

Thus the left singular vectors of A are eigenvectors of AA', and right singular vectors of A are eigenvectors of A'A

ii From above $A = U\Sigma V^T$

$$AA^T = U\Sigma\Sigma^T U^T \qquad \Sigma\Sigma^T = \Sigma^T\Sigma$$
$$A^TA = V\Sigma^T\Sigma V^T \qquad = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix}$$

The eigenvalues of A'A and AA' are the squares of the singular values of A.

c) i

False They need not be distinct.

Counter example? $\text{eig}(I_2)$, λ's $= 1, 1$

ii False Counter: for $\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$ $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ ($\lambda=1$), $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ ($\lambda=3$) are eigenvectors for $\lambda = 1, 3$ respectively

$\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, but $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is not an eigenvector of I.

iii

True

$A\vec{v} = \lambda\vec{v}$ eig definition

$x^T A x \geq 0 \quad \forall x$ PSD definition

$v^T A v = v^T \lambda v = v^T v \lambda$

If $v^T A v \geq 0$, and $v^T v \geq 0$ by definition,

then $\lambda \geq 0$

iv True

Rank-Nullity Theorem:

$\text{rank}(A) + \text{nullity}(A) = n$ for

$\text{rank}(A) = \#$ of eigenvalues

But they need not be distinct

For example: $\text{Rank}\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right) = 2 \qquad \lambda_1, \lambda_2 = 1, \lambda_3 = 0$

Rank 2, 1 distinct, non-zero λ

v True

The sum of two vectors is a linear combination.

If a λ has two eigenvectors, then $\dim(E_\lambda) = 2$ (eigenspace; dimension of eigenspace)

$Sp(E_\lambda) = \{v_{1\lambda}, v_{2\lambda}\} \quad \forall\, v_1, v_2$ as eigenvectors for λ

$1 v_{1\lambda} + 1 v_{2\lambda} \in Sp(E_\lambda)$

A sum of two eigenvectors for λ will be in the span of E_λ, so it is still an eigenvector.

Problem 2: Probability refresher

Notation: $t = T$, $h = H$, $h^3 = H, H, H$

a) i $P(H_{50} \mid t)$?? : $P(t \mid H_{50}) = 0.5$ $P(H_{50}) = 0.5$

$$P(t) = \overset{P(H_{50})}{0.5} \cdot \overset{P(t \mid H_{50})}{0.5} + \overset{P(H_{60})}{0.5} \cdot \overset{P(t \mid H_{60})}{0.4} = 0.45$$

$$P(H_{50} \mid t) = \frac{P(t \mid H_{50}) P(H_{50})}{P(t)} = \frac{0.5 \cdot 0.5}{0.45} = \frac{5}{9}$$

$$\boxed{P(H_{50} \mid t) = \frac{5}{9}}$$

ii $P(H_{50} \mid t, h, h, h)$:

$$P(t, h, h, h \mid H_{50}) = 0.5^4 \qquad P(H_{50}) = 0.5$$

$$P(t, h, h, h) = \overset{P(H_{50})}{0.5} \cdot \overset{P(t,h,h,h \mid H_{50})}{0.5^4} + \overset{P(H_{60})}{0.5} \overbrace{(0.4 \cdot 0.6^3)}^{P(t,h,h,h \mid H_{60})}$$

$$P(H_{50} \mid t, h^3) = \frac{P(t, h^3 \mid H_{50}) P(H_{50})}{P(t, h^3)} = \frac{0.5^5}{0.5^5 + 0.5(0.4)(0.6)^3} \approx 0.4197$$

$$\boxed{P(H_{50} \mid t, h^3) \approx 0.4197}$$

iii

$$P(H_{50}), P(H_{55}), P(H_{60}) = \tfrac{1}{3}$$

$$P(h^9 t) = \underbrace{\tfrac{1}{3} \cdot 0.5^{10}}_{P(h^9 t \mid H_{50})} + \underbrace{\tfrac{1}{3} \cdot 0.55^9 \cdot 0.45}_{P(h^9 t \mid H_{55})} + \underbrace{\tfrac{1}{3} \cdot 0.60^9 \cdot 0.4}_{= P(h^9 t \mid H_{60})} \approx 0.0024$$

$$P(H_{50} \mid h^9 t) = \frac{\frac{1}{3} \cdot 0.5^{10}}{P(h^9 t)} \approx \underline{0.1379}$$

$$P(H_{55} \mid h^9 t) = \frac{\frac{1}{3} \cdot 0.55^9 \cdot 0.45}{P(h^9 t)} \approx \underline{0.2927}$$

$$P(H_{60} \mid 9t) = \frac{\frac{1}{3} \cdot 0.6^9 \cdot 0.}{P(9t)} \approx 0.5694$$

$$P(H_{50} \mid 9t) \approx 0.1379$$
$$P(H_{55} \mid 9t) \approx 0.2927$$
$$P(H_{60} \mid 9t) \approx 0.5694$$

b) A = Pregnant T = Positive Test

$P(T \mid A) = 0.99$ $P(T \mid \bar{A}) = 0.10$ $P(A) = 0.01$

$P(A \mid T) = ??$

$$P(T) = \overset{P(T|A)}{0.99} \cdot \overset{P(A)}{0.01} + \overset{P(T|\bar{A})}{0.1} \cdot \overset{P(\bar{A})}{0.99} \approx 0.1089$$

$$P(A \mid T) = \frac{P(T \mid A)P(A)}{P(T)} = \frac{0.99 \cdot 0.01}{P(T)} \approx 0.0901$$

$P(\text{Pregnant} \mid \text{Positive}) = 0.0901$

Note the high false positive rate makes the probability so low...

$$P(+ \mid \bar{A}) = 0.1 > P(A) = 0.01$$

Hence more false positives are generated then there are pregnant women.

c) $\mathbb{E}(Ax+b) = \mathbb{E}[Ax] + \mathbb{E}[b]$ (Linearity of $\mathbb{E}$; $= b$ [b is deterministic])

$= \mathbb{E}[Ax] = A\,\mathbb{E}[x]$ By determinism of A and iid X:

$$\mathbb{E}[Ax] = \mathbb{E}\left[\sum_{i=1}^{n} A_{ij} X_j\right] = \sum_{j=1}^{n} A_{ij}\, \mathbb{E}(X)_j$$

$$= A\,\mathbb{E}[x]$$

Thus $\mathbb{E}[Ax+b] = A\,\mathbb{E}[x] + b$

d) $\text{cov}(Ax+b) = \mathbb{E}\left[(Ax+b-\mathbb{E}[Ax+b])(Ax+b-\mathbb{E}[Ax+b])^T\right)$

$= \mathbb{E}\left[(Ax+b-A\mathbb{E}[x]-b)(Ax+b-A\mathbb{E}[x]-b)^T\right)$ from c) above

$= \mathbb{E}\left[A(x-\mathbb{E}[x])(x-\mathbb{E}[x])^T A^T\right]$

$= A\,\underbrace{\mathbb{E}\left[(x-\mathbb{E}[x])(x-\mathbb{E}[x])^T\right]}_{=\text{cov}(x)}\,A^T$ determinism of A + iid X

$= \boxed{A\,\text{cov}(x)\,A^T}$

Problem 3: Multivariate Derivatives

a) $\nabla_x x^T A y$? $x \in \mathbb{R}^n$, so $\nabla_x x^T A y \in \mathbb{R}^n$

$x^T A y = x_1 a_{11} y_1 + x_2 a_{12} y_1 + \dots x_n a_{nn} y_n$ $\nabla_x x^T A y = a_{11} y_1 + a_{12} y_2 \dots a_{nn} y_n$

$= Ay \in \mathbb{R}^n$ ✓

$$\boxed{\nabla_x x^T A y = Ay}$$

b) $\nabla_y x^T A y$ $\nabla_y x^T A y \in \mathbb{R}^m$

$x^T A y = x_1 a_{11} y_1 + x_2 a_{12} y_1 + \dots x_n a_{nn} y_n$ $\nabla_y x^T A y = x_1 a_{11} + x_2 a_{12} \dots x_n a_{nn}$

$= x^T A$

$\Rightarrow$ To match dimension $(n \times 1)$ need $(x^T A)^T = A^T x \in \mathbb{R}^m$ ✓

$$\boxed{\nabla_y x^T A y = A^T x}$$

c) $\nabla_A x^T A y$ $\in \mathbb{R}^{n \times m}$

$\Rightarrow$

$$\begin{bmatrix} \frac{\partial x^T A y}{\partial a_{11}} = x_1 y_1 & \cdots & \frac{\partial x^T A y}{\partial a_{1m}} = x_n y_1 \\ \vdots & \ddots & \\ \frac{\partial x^T A y}{\partial a_{n1}} = x_1 y_n & \cdots & \frac{\partial x^T A y}{\partial a_{nm}} = x_n y_n \end{bmatrix}$$

$= x y^T \in \mathbb{R}^{n \times m}$ ✓

$$\boxed{\nabla_A x^T A y = x y^T}$$

d) $\nabla_x (x^T A x + b^T x) \quad \in \mathbb{R}^n$

$= \nabla_x (x^T A x) + \nabla_x (b^T x)$ — linearity of ∇'s

$\nabla_x (b^T x) = b \quad \in \mathbb{R}^n$ ✓

$\nabla_x (x^T A x) = Ax + A^T x$ derived in class :

$A = \begin{bmatrix} A_{11} & \cdots \\ \vdots & \ddots \\ \cdots & A_{nn} \end{bmatrix} \qquad f(x) = \sum_i \sum_j a_{ij} x_i x_j$

$= \sum_{j=1}^n a_{1j} x_j + \sum_{i=1}^n a_{i1} x_i$

$= (Ax)_1 \quad + \quad (A^T x)$

$i=j=1 \quad \frac{\partial a_{11} x_1^2}{\partial x_1} = 2 a_{11} x_1 \;,\; i=1, j \neq 1 \quad \sum_{j=2}^n a_{1j} x_1 x_j = \sum_{j=2}^n a_{1j} x_j$

$j=1, i=1 \quad \sum_{i=2}^n a_{j1} x_i x_j \quad \Rightarrow \quad \frac{\partial f(x)}{\partial x_1} = 2 a_{11} x_1 + \sum_{j=2}^n a_{1j} x_j + \sum_{i=2}^n a_{i1} x_i$

$\nabla_x f(x) = (Ax + A^T x) = (A + A^T) x$

$\in \mathbb{R}^n$ ✓

$$\boxed{\nabla_x (x^T A x + b^T x) = Ax + A^T x + b}$$

e) $f = tr(AB) \quad \nabla_A f \quad \in \mathbb{R}^{n \times n}$

$$tr(AB) = tr\left(\begin{bmatrix} a_{11} b_{11} + a_{12} b_{21} \cdots a_{1n} b_{n2} & \cdots & a_{1n} b_{nn} \cdots a_{1n} b_{nn} \\ \vdots & & \\ a_{n1} b_{11} \cdots a_{nn} b_{1n} & \cdots & a_{n1} b_{1n} \cdots a_{nn} b_{nn} \end{bmatrix} \right)$$

$$= \sum_{i=1}^n a_{1i} b_{i1} + \sum_{i=1}^n a_{2i} b_{i2} + \cdots \sum_{i=1}^n a_{ni} b_{in}$$

$$\frac{\partial tr(AB)}{\partial a_{11}} = b_{11} \qquad \frac{\partial tr(AB)}{\partial a_{12}} = b_{21} \qquad \frac{\partial tr(AB)}{\partial a_{1n}} = b_{n1}$$

$$\Rightarrow \quad \frac{\partial tr(AB)}{\partial a_{ij}} = b_{ji}$$

$$\Rightarrow \quad \boxed{\nabla_A tr(AB) = B^T}$$

Problem 4: Deriving LS

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{n} \|y^{(i)} - W x^{(i)}\|^2$$

To minimize: $\frac{\partial \mathcal{L}}{\partial W} = 0 \qquad \nabla_W \mathcal{L} = 0$

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{n} (y^{(i)} - W x^{(i)})^T (y^{(i)} - W x^{(i)})$$

$$= \frac{1}{2}\sum_{i=1}^{n} y^{(i)2} - 2 y^{(i)T} W x^{(i)} + x^{(i)T} W^T W x^{(i)}$$ only need W terms

$$= \sum_{i=1}^{n} -y^{(i)T} W x^{(i)} + \frac{1}{2} x^{(i)T} W^T W x^{(i)}$$ } $\in \mathbb{R}$ so $\mathrm{tr}(\cdot) = (\cdot)$

$$= \sum_{i=1}^{n} -\mathrm{tr}(y^{(i)T} W x^{(i)}) + \frac{1}{2}\mathrm{tr}(x^{(i)T} W^T W x^{(i)})$$ Notation: Dropping '(i)' for readability...

$$= \sum -\mathrm{tr}(W x y^T) + \frac{1}{2}\mathrm{tr}(W x x^T W^T)$$ $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$

$$= -\mathrm{tr}(\underbrace{W x y^T}_{A}) + \frac{1}{2}\mathrm{tr}(\underbrace{W x x^T}_{A} W^T)$$

$\sum x^{(i)} y^{(i)T} = X Y^T$

$\sum x^{(i)} x^{(i)T} = X X^T$

Now $\nabla_W \mathcal{L}$,

$$\nabla_W \mathcal{L} = -Y x^T + \frac{1}{2} W x x^T + \frac{1}{2} W x x^T$$

$$= -Y x^T + W x x^T = 0$$

$\frac{\partial \mathrm{tr}(WA)}{\partial W} = A^T$, $A = X Y^T$

$\frac{\partial \mathrm{tr}(WAW^T)}{\partial W} = W A^T + W A$, $A = X X^T$

$$\Rightarrow W = (Y x^T)(x x^T)^{-1}$$

Optimal W (to min $\mathcal{L}(W)$) is $(Y X^T)(X X^T)^{-1}$

linear_regression

January 11, 2022

0.1 Linear regression workbook: Sunay Bhat - W2022

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 2022, Prof. J.C. Kao, TAs Y. Li, P. Lu, T. Monsoor, T. wang

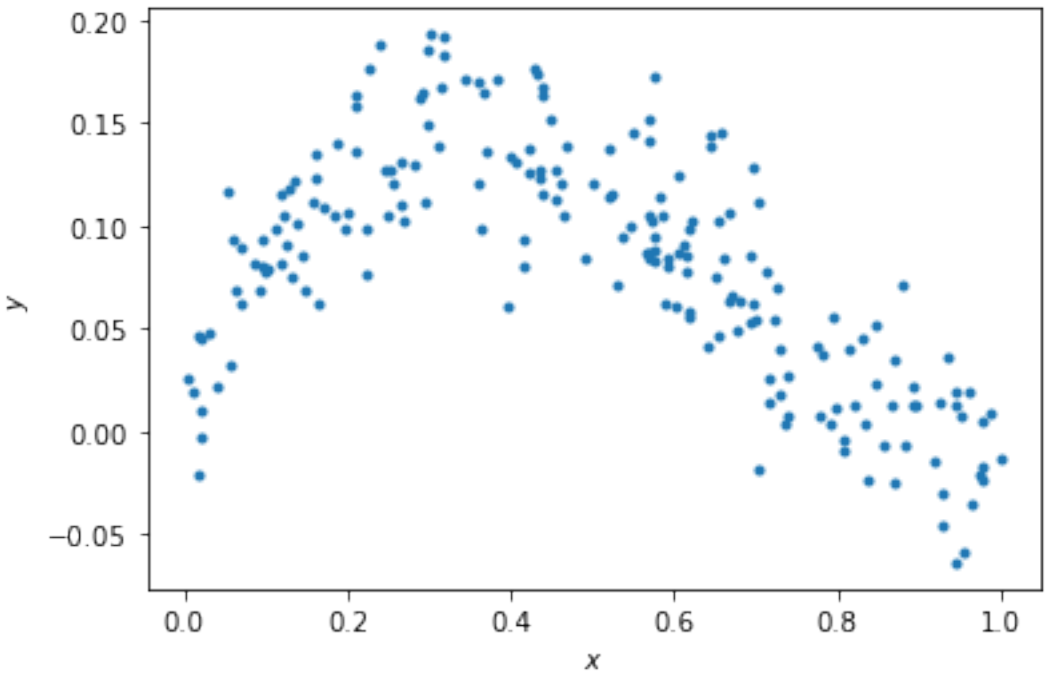
[1]:
```
import numpy as np
import matplotlib.pyplot as plt

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

0.1.1 Data generation

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

[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('$y$')
```

[2]:
```
Text(0, 0.5, '$y$')
```



0.1.2 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?

0.1.3 ANSWERS:

(1) $U(0, 1)$ - Uniform distribution with parameters 0 and 1.

(2) $N(0, 0.3)$ - Normal distirbution with parameters mean $\mu = 0$ and standard deviation $\sigma = 0.3$.

0.1.4 Fitting data to the model (5 points)

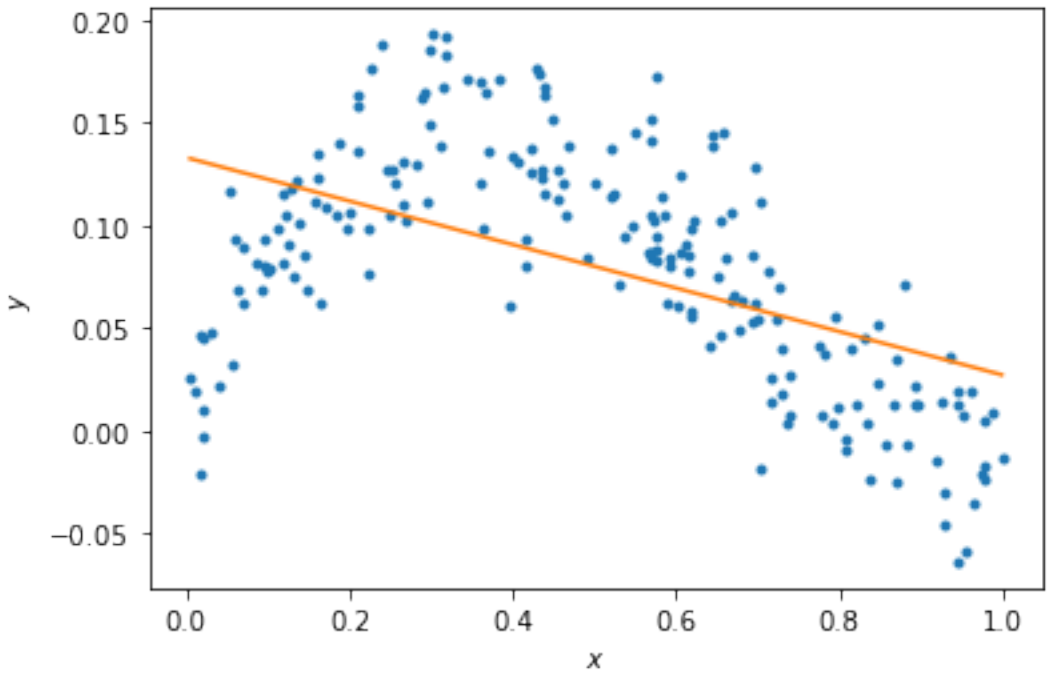
Here, we'll do linear regression to fit the parameters of a model $y = ax + b$.

[3]:
```
# 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]
```

```
# Normal Equation: (X'*X)^-1 * X'*y
theta = np.linalg.inv(xhat.dot(xhat.T)).dot(xhat.dot(y))

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

[4]:
```
# 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))
```

[4]: [<matplotlib.lines.Line2D at 0x7fb9242b5eb0>]



0.1.5 QUESTIONS

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

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

0.1.6 ANSWERS

(1) The linear model **underfits** the data.

(2) We can use a **higher order polynomial** like a 2nd or 3rd order polynomial or many other models with **higher complexity**.

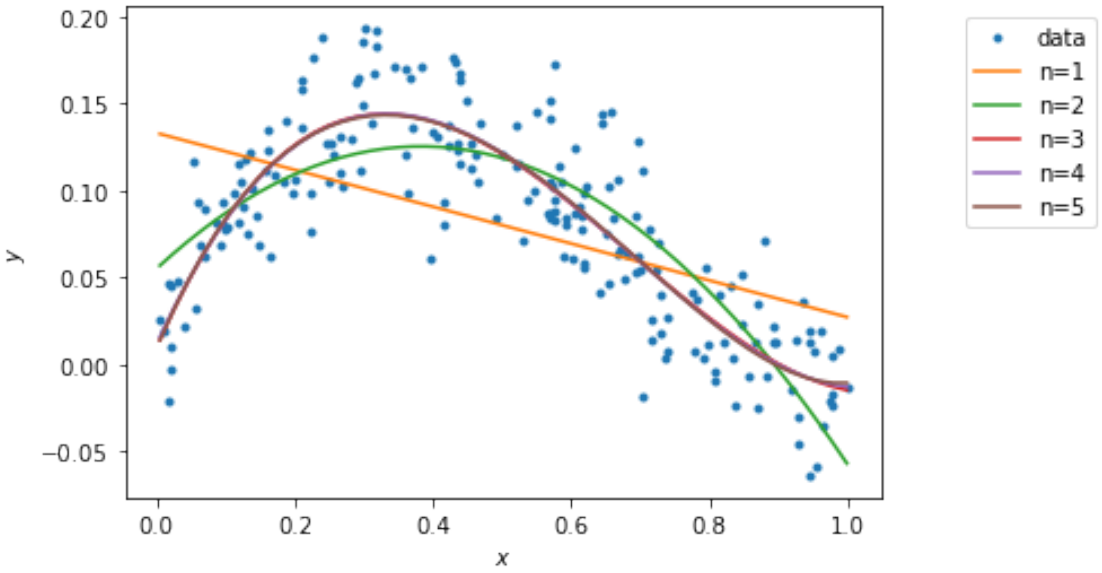
0.1.7 Fitting data to the model (10 points)

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.

[5]:
```
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.

xhats.append(np.vstack((x, np.ones_like(x)))) # Inital xhat = = (x, 1)
for order in np.arange(N):
    if order > 0:
        xhats.append(np.vstack((x**(order+1), xhats[order-1]))) #add next order␣
 ↪features to last xhat and append

    # Normal Equation: (X'*X)^-1 * X'*y
    thetas.append(np.linalg.inv(xhats[order].dot(xhats[order].T)).
 ↪dot(xhats[order].dot(y)))


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

[6]:
```
# 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)
```



0.1.8 Calculating the training error (10 points)

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

$L(\theta) = \frac{1}{2}\sum_j(\hat{y}_j - y_j)^2$

[7]:
```
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␣
 ↪order i+1.

for order in np.arange(N):
    # 1/2 * (theta*xhat - y)' * (theta*xhat - y)'
    errors = thetas[order] @ xhats[order] - y
    training_errors.append(1/2 * (errors.T @ errors))

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

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

```
Training errors are:
 [0.2379961088362701, 0.10924922209268528, 0.08169603801105374,
0.08165353735296982, 0.08161479195525295]
```

0.1.9 QUESTIONS

(1) Which polynomial model has the best training error?

(2) Why is this expected?

0.1.10 ANSWERS

(1) The **5th order** polynomial has the lowest training error.

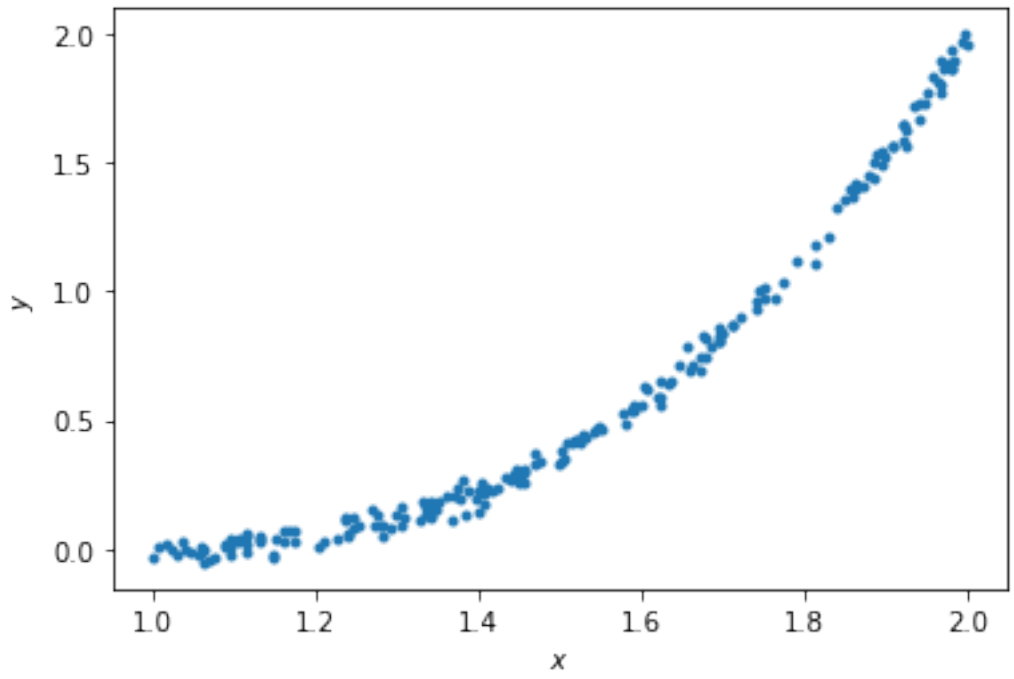
(2) Yes, higher order polynomials will have <= training error than lower order. You can do as well as lower order polynomials by setting higher coefficients to zero (and others equal to each other), but you can generally utilize the added degrees of freedom from higher order terms to better fit the training data.

0.1.11 Generating new samples and validation error (5 points)

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

[8]:
```
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('$y$')
```

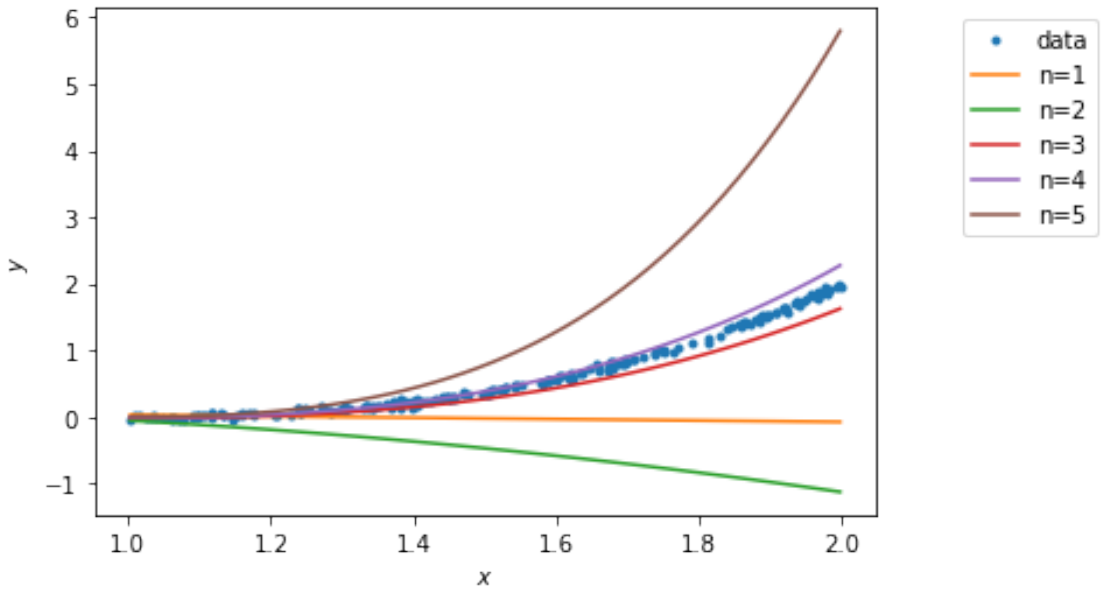
[8]:
```
Text(0, 0.5, '$y$')
```



[9]:
```
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)
```

[10]:
```
# 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)
```



[11]:
```
validation_errors = []

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

# GOAL: create a variable validation_errors, a list of 5 elements,
# where validation_errors[i] are the validation loss for the polynomial fit of␣
 ↪order i+1.

for order in np.arange(N):
    # 1/2 * (theta*xhat - y)' * (theta*xhat - y)'
    errors = thetas[order] @ xhats[order] - y
    validation_errors.append(1/2 * (errors.T @ errors))

# ================== #
```

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

print ('Validation errors are: \n', validation_errors)
```

```
Validation errors are:
 [80.86165184550586, 213.19192445057894, 3.125697108276393, 1.1870765189474703,
214.91021817652626]
```

0.1.12 QUESTIONS

(1) Which polynomial model has the best validation error?

(2) Why does the order-5 polynomial model not generalize well?

0.1.13 ANSWERS

(1) The **4th order** polynomial has the best, or lowest, validation error, with the 3rd order being close.

(2) The 5th order model **overfits the data and does not generalize well when we validate our dataset in the previously unseen region of** $x = [1 : 2]$. In this case, the problem is made acute by validating in an extension of the domain ($x = [1 : 2]$) instead of new points from within the training data domain ($x = [0, 1]$). Specifically, the 4th order coefficient for the 5th order model is too large and rapidly diverges from the 3rd order generative model in the new domain space. The 4th order model's 4th order coefficient is much smaller though, allowing it to perform well even though the generative model is only 3rd order.

[ ]: