Homework 1

Shalin Shah

January 13, 2020

1 Multivariable Calculus

The given equation is:

$$y = zsin(x)e^{-x}$$

If we want to take the partial of y with respect to x, $\frac{\partial y}{\partial x}$, we should differentiate the equation assuming that all variables other than x and y are constants. The partial looks as the following:

$$\frac{\partial y}{\partial x} = z\frac{d}{dx}[sin(x)e^{-x}]$$

$$\frac{\partial y}{\partial x} = z[sin(x)\frac{d}{dx}[e^{-x}] + \frac{d}{dx}[sin(x)]e^{-x}]$$

$$\frac{\partial y}{\partial x} = z[-sin(x)e^{-x} + cos(x)e^{-x}]$$

$$\frac{\partial y}{\partial x} = -zsin(x)e^{-x} + zcos(x)e^{-x}]$$

2 Linear Algebra

2.a Inner product?

$$\mathbf{y}^T = \begin{pmatrix}1 & 3\end{pmatrix} \ \mathbf{z} = \begin{pmatrix}2\\3\end{pmatrix}$$

$$\mathbf{y}^T\mathbf{z} = 2*1+3*3 = 11$$

2.b Product Xy?

$$\mathbf{X} = \begin{pmatrix}2 & 4\\1 & 2\end{pmatrix} \ \mathbf{y}^T = \begin{pmatrix}1\\3\end{pmatrix}$$

$$\mathbf{Xy} = \begin{pmatrix}2 & 4\\1 & 2\end{pmatrix}\begin{pmatrix}1\\3\end{pmatrix} = \begin{pmatrix}2*1+4*3\\1*1+2*3\end{pmatrix}$$

$$\mathbf{Xy} = \begin{pmatrix}14\\10\end{pmatrix}$$

2.c Is X invertible?

$$\mathbf{X} = \begin{pmatrix}2 & 4\\1 & 2\end{pmatrix}$$

If det($\mathbf{X}$) is 0 then the matrix is not invertible.

$$\det(\mathbf{X}) = 2*2-4*1 = 0$$

Since det($\mathbf{X}$) = 0, $\mathbf{X}$ is not invertible.

2.d Rank of X?

Rank is the total number of linearly independent columns in a matrix. Matrix $\mathbf{X}$ has two columns:

$$\begin{pmatrix}2\\1\end{pmatrix}, \begin{pmatrix}4\\2\end{pmatrix}$$

Since column 1 is just column 2 scaled down by a factor of 2, column 1 and column 2 are not linearly independent. Thus, $\mathbf{X}$ has a rank of **1**.

3 Probability and Statistics

Sample $S = (X_1, X_2, X_3, X_4, X_5) = (1,1,0,1,0)$.

3.a Sample mean?

Sample mean = $\frac{1+1+0+1+0}{5}$
Sample mean = $\frac{3}{5}$

3.b Unbiased sample variance?

Variance formula is: $\frac{\sum_{i=1}^{N}(X_i - \bar{x})^2}{N}$
For our problem: $\frac{\sum_{i=1}^{5}(X_i - \bar{x})^2}{5} = \frac{1}{5}(3 * \frac{4}{25} + 2 * \frac{9}{25}) =$
Variance = $\frac{6}{25}$

3.c Probability of observing this data?

Probability of observing data is given by:
$P(X_1 = 1)P(X_2 = 1)P(X_3 = 0)P(X_4 = 1)P(X_5 = 0) = 0.5^5 =$
Probability = $\frac{1}{32}$

3.d What value will maximize probability of sample S?

Let's call probability of heads p. Then probability of tails is $1 - p$. We want to maximize the relation: $P(11010) = P(p) = p^3(1-p)^2$.
$\frac{d}{dp}P(p) = \frac{d}{dp}p^3(1-p)^2 = (x-1)x^2(5x-3)$.
The maximum and minimum points can be found through setting this equal to 0 and solving. When we do this, the roots are $x = 0, x = \frac{3}{5}$. When we check this, $\frac{3}{5}$ does yield the maximum probability value for P, so the answer is $p = \frac{3}{5} = 0.6$.

3.e Probability of $X|Y$?

$P(X = T|Y = b) = \frac{P(XandY)}{P(Y)} = \frac{0.1}{0.25}$
$P(X = T|Y = b) = \frac{2}{5}$

4 Probability axioms

Are the axioms true or false in general?

4.a $P(A \cup B) = P(A \cap (B \cap A^c))$

The intersection will end up yield $P(A \cap B)$.
False

4.b $P(A \cup B) = P(A) + P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
False

4.c $P(A) = P(A \cap B) + P(A^c \cap B)$

$P(A \cap B) + P(A^c \cap B)$ ends up being P(B), not P(A).
False

4.d $P(A|B) = P(B|A)$

$P(A|B) = \frac{P(A)}{P(A \cap B)}$ $P(B|A) = \frac{P(B)}{P(A \cap B)}$ So this is only true when P(A) = P(B).
In general, it's not true.
False

4.e $P(A_1 \cap A_2 \cap A_3) = P(A_3|(A_2 \cap A_1))P(A_2|A_1)P(A_1)$

$P(A_3|(A_2 \cap A_1))P(A_2|A_1)P(A_1) = \frac{P(A_3 \cap (A_2 \cap A_1))}{\cancel{P(A_2 \cap A_1)}} \frac{\cancel{P(A_2 \cap A_1)}}{\cancel{P(A_1)}} \cancel{P(A_1)}$

$$P(A_3 \cap (A_2 \cap A_1)) = P(A_3 \cap A_2 \cap A_1)$$

True

5 Discrete and Continuous Distributions

5.a Guassian

(v)

5.b Exponential

(b)

5.c Uniform

(ii)

5.d Bernoulli

(i)

5.e Binomial

(iii)

6 Mean and Variance

6.a What is the mean and variance of *Bernoulli(p)* random variable?

From 5d, we see that the distribution for a Bernoulli random variable is seen by $p^x(1-p)^{1-x}$. The mean is another way to say the expected value. For Bernoulli, x is either 0 or 1 and so the expected value the probability that we'll get a 1 which is just p. If we set p to 0.6 for example, we're saying that for 100 runs if we add up all of the values of 0 or 1 we'd assume that we get about 60 1's. If we do this for 1 run, however, we assume that we'll get .6 as an expected value since on average the value will be 1 60% of the time.

The variance will be $p(1-p)$. Variance is expected value of X^2 minus the mean squared: $Var[X] = E[X^2] - E[X]^2$. From earlier in this problem, we showed that $E[X] = p.E[X^2] = 1-p$ (from my notes from Stats 100A) and so $Var[X] = p(1-p)$.

6.b Var[2X]? Var[X+3]?

$Var[X] = \sigma^2$
$Var[aX + b] = a^2Var[X]$.
So, $Var[2X] = 2^2Var[X] = 4Var[X], and Var[X + 3] = Var[X]$

$Var[2X] = 4\sigma^2, Var[X + 3] = \sigma^2$

7 Algorithms

7.a Big-O notation

List which of following are true: $f(n) = O(g(n)), g(n) = O(f(n))$, or both.

7.a.i $f(n) = ln(n), g(n) = lg(n)$

Only $g(n) = O(f(n))$ is true. When comparing these two runtimes, we see that $log_e = ln$ is always behind $log_2 = lg$ since $lg(n) > ln(n)$ for large numbers for n. Thus, lg serves as an upper bound for ln and thus g(n) is O(f(n)).

7.a.ii $f(n) = 3^n, g(n) = n^{10}$

Only $f(n) = O(g(n))$ is true. For very large numbers of n, 3^nhas a higher output than does n^{10}. Take n = 100, for example. $3^{100} = 5*10^{47}$ while $100^{10} = 1*10^{20}$. Exponential functions are always $O(n^a)$ if a is a constant.

7.a.iii $f(n) = 3^n, g(n) = 2^n$

Only $f(n) = O(g(n))$ is true. In general, if $a > b$ then $a^n > b^n$ and so f(n) serves as an upper bound for g(n) for large values of n.

7.b Divide and Conquer

Given array with n elements, all that are 0's or + 1's (the array is "sorted" in the sense that all the 0's are front-loaded and the 1's are back-loaded), find the element with last occurance of 0.

This problem could use binary search as a solution. We could start at the middle of the array and check its number. If it's a 0, we check the next element in the array to see if it's a +1. If it's a +1, we check the previous element in the array to see if it's a 0. To account for edge cases we can make sure that if we get to the end of the array, i.e. array[len(array)-1], and the value is 0, then we return -1 and if we get to the beginning of the array, i.e. array[0], and the value is +1, then we return -1. Other than that, if the current is 0 and the next is +1, then return current or if the current is +1 and the previous is 0, then return current-1. Else, recursively check half of the remaining array to see where the transition is. If our current value is a 0 and we get a 0 for the next value, then we know that anything behind our current is also a 0 and thus we only need to check the second half of the array. Similarly, if our current value is a +1 and our previous isn't a 0, we know that anything ahead of our current is also a +1 and thus we only need to check the first half of the array. We keep recursively iterating through the array and either find the location of the transition and return that index, or return -1.

This implementation is correct because it accounts for every case in the array. We don't check every value in the array, but that's because we don't need to. This implementation runs in $O(logn)$ because each iteration of the algorithm cuts the array in half, and we need a maximum of log(n) iterations for us to either find the value or realize the value isn't in the array. Checking the current/next/previous values is an $O(1)$ operation that's done $O(logn)$ number of times, so the run time is unaffected by the checking operation.

8 Probability and Random Variables

8.a Mutual and Conditional Independence

If X and Y are independent random variables, show that E[XY] = E[X]E[Y]. I'll show this for the discrete case, but the same principle applies to the continuous case as well.

The expected value of a random variable X is: $E[X] = \sum_{i=1}^{N} x_i f(x_i)$

For a case where there are two random variables, we need to account for all possible combinations of the first variable and the second variable since their outputs are being multiplied together: $E[XY] = \sum_{j=1}^{N} \sum_{i=1}^{N} x_i y_j f(x_i, y_j)$ where $f(x_i, y_j)$ is the probability distribution.

Now, if X and Y are independent, then there exists no function $f(x_i, y_j)$ that relates X and Y and thus $f(x_i, y_j) = f_x(x_i) f_y(y_i)$. So:
$E[XY] = \sum_{j=1}^{N} \sum_{i=1}^{N} x_i y_j f_x(x_i) f_y(y_j)$
$E[XY] = \sum_{j=1}^{N} y_j f_y(y_j) \sum_{i=1}^{N} x_i f_x(x_i)$
But, $\sum_{j=1}^{N} y_j f_y(y_j) = E[Y]$ and $\sum_{i=1}^{N} x_i f_x(x_i) = E[X]$, so
E[XY] = E[Y]E[X] =E[X]E[Y] ✓✓

8.b Law of Large Numbers and Central Limit Theorem

Provide one line justification.

8.b.i If a fair die is rolled 6000 times, the number of times 3 shows up is close to 1000.

A fair die sees an equal probability of landing any of the 6 numbers and due to the law of large numbers, if the fair die is rolled many many times, then it's likely that each side of the die will be seen roughly an equal amount of times and $\frac{6000}{6} = 1000$.

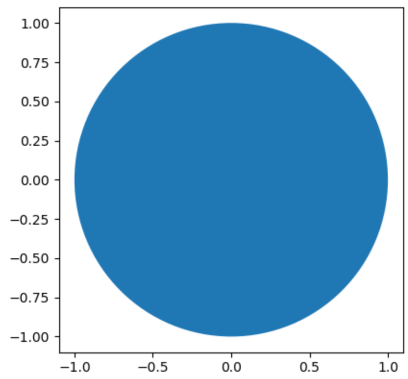
8.b.ii Distribution of $\bar{X}$ will be normal with mean 0 and SD $\frac{1}{2}$ if fair coin is tossed n times and $\bar{X}$ is the distribution of average # of heads

A fair coin has either heads or tails and due to the law of large numbers, if tossed an infinite amount of times will have an average value of $\frac{1}{2}$ [we're subtracting $\frac{1}{2}$ from the distribution so it makes sense our mean is 0] and a variance of $0.5 * 0.5 = \frac{0.25}{n}$ [we're multiplying $\bar{X}$ by $\sqrt{n}$ and $Var(aX + b) = a^2 Var[X]$ so we get $\frac{n}{n} = 1 * 0.25$ as our variance].

9 Linear Algebra

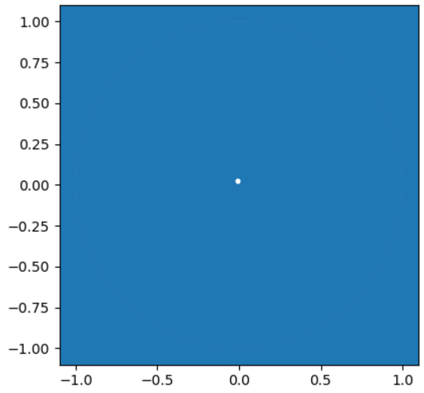
9.a Vector Norms

9.a.i Draw regions corresponding to vectors $\mathbf{x} \in \mathbf{R}^2$such that $||\mathbf{x}||_2 \leq 1$:



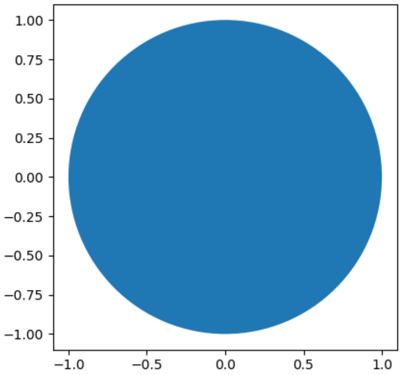
We see here that, since we square the values, that any value above 1 will cause the vector to be above 1 and not return back to 1. Through this means, we see that the unit circle is what gives the relation that we are looking for.

9.a.ii Draw regions corresponding to vectors $\mathbf{x} \in \mathbf{R}^2$such that $||\mathbf{x}||_0 \leq 1$:



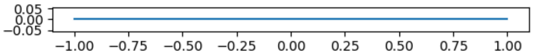
We see here that all of $\mathbf{R}^2$ is included besides the origin, since x_i cannot be 0.

9.a.iii Draw regions corresponding to vectors $\mathbf{x} \in \mathbf{R}^2$such that $||\mathbf{x}||_0 \leq 1$:



We see here that, since we take the absolute of the values, that any value above 1 will cause the vector to be above 1 and not return back to 1. Through this means, we see that the unit circle is what gives the relation that we are looking for.

9.a.iv Draw regions corresponding to vectors $\mathbf{x} \in \mathbf{R}^2$such that $||\mathbf{x}||_\infty \leq 1$:



We see here that it's a straight line of values that we can get since we're choosing the max value (we take the absolute value so it's the same on both sides of the y-axis, only going up to 1).

9.b Matrix Decompositions

9.b.i Define eigenvalues and eigenvectors of square matrix

A square matrix is an $n \times n$ matrix. An eigenvalue is a value that causes the determinant of the matrix subtracted by the value multiplied by the identity matrix to equal 0. Symbolically, for matrix **A**: $\det(A - \lambda I) = 0$. An eigenvector is a vector that equals the zero vector when multiplied by $A - \lambda I$. Symbolically, for matrix **A** and vector **x**: $(A - \lambda I)\mathbf{x} = \mathbf{0}$.

9.b.ii Find eigenvalues and eigenvectors for matrix A

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

$$\lambda I = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

$\det(A - \lambda I) = 0$

$(2 - \lambda)(2 - \lambda) - 1 * 1 = 0$

$3 - 4\lambda + \lambda^2 = 0$

$(\lambda - 3)(\lambda - 1) = 0$

$\lambda = 3$ or $\lambda = 2$

$\lambda = 3$:

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \mathbf{x} = 0$$

$\mathbf{x} = (1, 1)$

$\lambda = 1$:

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \mathbf{x} = 0$$

$\mathbf{x} = (1, -1)$

Eigenvalues: $\lambda_1 = 3, \lambda_2 = 1$

Eigenvectors: $\mathbf{x}_1 = (1, 1), \mathbf{x}_2 = (1, -1)$

9.b.iii Show eigenvalues of $\mathbf{A}^k$ are k^{th} powers of eigenvalues of A and that each eigenvector of A is an eigenvector of $\mathbf{A}^k$.

If λ is an eigenvalue and **x** is an eigenvector, then A**x**=λ**x**.

$\mathbf{A}^k = \mathbf{A}\mathbf{A}\cdots_k \mathbf{A}$

$\mathbf{A}^k\mathbf{x} = \mathbf{A}\mathbf{A}\cdots_k \mathbf{A}\mathbf{x}$, but $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, so $\mathbf{A}^k\mathbf{x} = \mathbf{A}\mathbf{A}\cdots_{k-1} \mathbf{A}[\mathbf{A}\mathbf{x}]$

$\mathbf{A}^k\mathbf{x} = \mathbf{A}\mathbf{A}\cdots_{k-1} \mathbf{A}[\lambda\mathbf{x}]$

$\mathbf{A}^k\mathbf{x} = \lambda\mathbf{A}\mathbf{A}\cdots_{k-2} \mathbf{A}[\lambda\mathbf{x}] = \lambda^2\mathbf{A}\mathbf{A}\cdots_{k-2} \mathbf{A}$

$\vdots$

$\mathbf{A}^k\mathbf{x} = \lambda^k\mathbf{x}$

So, eigenvalues of $\mathbf{A}^k\mathbf{x}$ are k^{th} power of eigenvalue of $\mathbf{A}$. Also, since $\mathbf{x}$ is the same on the left and right side of the equation, an eigenvector of $\mathbf{A}$ is also an eigenvector of $\mathbf{A}^k$.

9.c Vector and Matrix Calculus

Vectors $\mathbf{x}$ and $\mathbf{a}$ and symmetrix matrix $\mathbf{A}$

9.c.i First derivative of $\mathbf{a}^T\mathbf{x}$ with respect to $\mathbf{x}$?

$\frac{\partial \mathbf{a}^T x}{\partial x} = \frac{\partial (x_1 a_1, x_2 a_2, \ldots, x_n a_n)}{\partial x}$
$\frac{\partial x_1}{\partial x} = 1$
So, through this logic, $\frac{\partial \mathbf{a}^T x}{\partial x} = (a_1, a_2, \ldots, a_n) = \mathbf{a}$

9.c.ii First derivative of $\mathbf{x}^T\mathbf{A}\mathbf{x}$ with respect to $\mathbf{x}$? Second derivative?

$\frac{\partial \mathbf{x}^T \mathbf{A}\mathbf{x}}{\partial x}$?
The limit definition of the derivative is as follows: $f^{'}(x) = \lim_{h \to \infty} \frac{f(x+h) - f(x)}{h}$.
Here, $f(x+h) = (\mathbf{x}+h)^T\mathbf{A}[\mathbf{x}+h]$ and $f(x) = \mathbf{x}^T\mathbf{A}\mathbf{x}$.
$f(x+h) - f(x) = (\mathbf{x}+h)^T\mathbf{A}[\mathbf{x}+h] - \mathbf{x}^T\mathbf{A}\mathbf{x}$
$f(x+h) - f(x) = \cancel{\mathbf{x}^T\mathbf{A}\mathbf{x}} + h^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{A}h + h^T\mathbf{A}h \cancel{-\mathbf{x}^T\mathbf{A}\mathbf{x}}$
$f(x+h) - f(x) = h^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{A}h + h^T\mathbf{A}h$
But, since $\mathbf{A}$ is symmetric, we can switch the order of the multiplication to make the difference equal to:
$f(x+h) - f(x) = h^T\mathbf{A}h + \mathbf{x}^T\mathbf{A}h + \mathbf{x}^T\mathbf{A}h = h^T\mathbf{A}h + 2\mathbf{x}^T\mathbf{A}h$
$h^T\mathbf{A}h << 2\mathbf{x}^T\mathbf{A}h$ as $h \to 0$, so we can cancel the first term out when calculating our limit. The derivative is $2\mathbf{x}^T\mathbf{A}$. The second derivative is just $2\mathbf{A} sincethe \mathbf{x} termsaredifferentiatedaway.$

9.d Geometry

9.d.i Show $\mathbf{w}$ is orthogonal to $\mathbf{w}^T\mathbf{x} + b = 0$

$\mathbf{w}^T\mathbf{x} = \mathbf{w}_1 x_{11} + \mathbf{w}_2 x_{12} + b = 0$
$\mathbf{w}^T\mathbf{x} = \mathbf{w}_1 x_{21} + \mathbf{w}_2 x_{22} + b = 0$
$\mathbf{w}_1 x_{11} + \mathbf{w}_2 x_{12} - \mathbf{w}_1 x_{21} - \mathbf{w}_2 x_{22} = \mathbf{w}^T[x_1 - x_2] = 0$
Since this product is equal to 0, the vector $\mathbf{w}$ is orthogonal to the line.

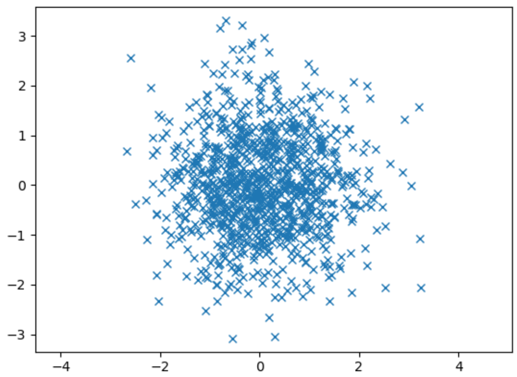
9.d.ii Argue that distance from origin to line is $\frac{b}{||\mathbf{w}||^2}$

Since $\mathbf{w}$ is orthogonal to the line, we know that the shortest distance to the origin has to be a multiple of $\mathbf{w}$ because that is the definition of orthogonality. The distance is formula is $\sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2}$ but x_{12} and x_{22} are both 0, so we end up getting the norm of $\mathbf{w}$ as the distance. It's also added to b so there is a scale factor of b at the top of the equation.

10 Sampling from a Distribution

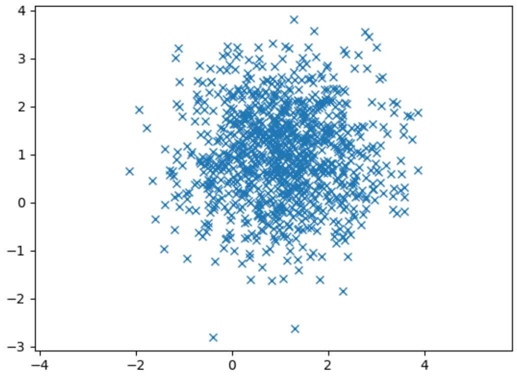
10.a Draw 1000 samples

The plot of 1000 samples when plotting x_1 vs x_2 is shown below.



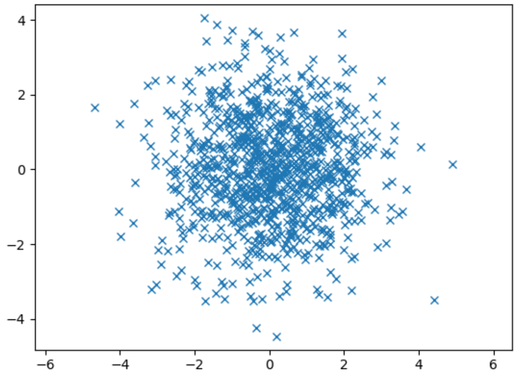
10.b What if mean is (1 1)?

The plot of 1000 samples when plotting x_1 vs x_2 with the mean being (1 1) is shown below.



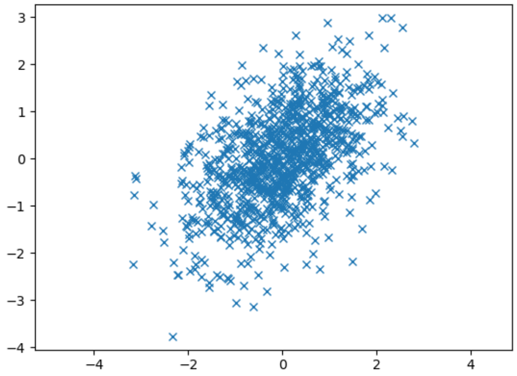
10.c What if we double variance?

The plot of 1000 samples when plotting x_1 vs x_2 with the variance being doubled for both x_1 and x_2 is shown below. The variance being double for both results in a doubling of the covariance since the two are related in the corelation equation by $\frac{1}{Var(x_1)}\frac{1}{Var(x_2)}$



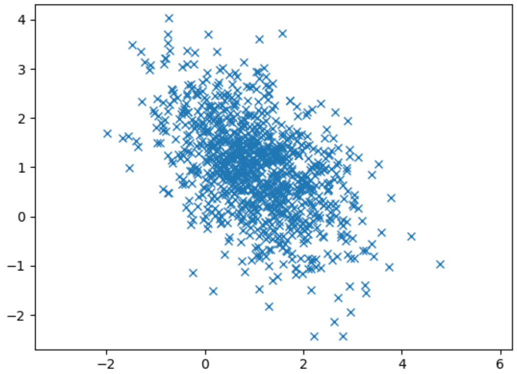
10.d What if we change covariance matrix?

The plot of 1000 samples when plotting x_1 vs x_2 with the covariance matrix being changed is shown below. We see an upward trend in this distribution which introduced by the changed covariance matrix.



10.e What if we change covariance matrix?

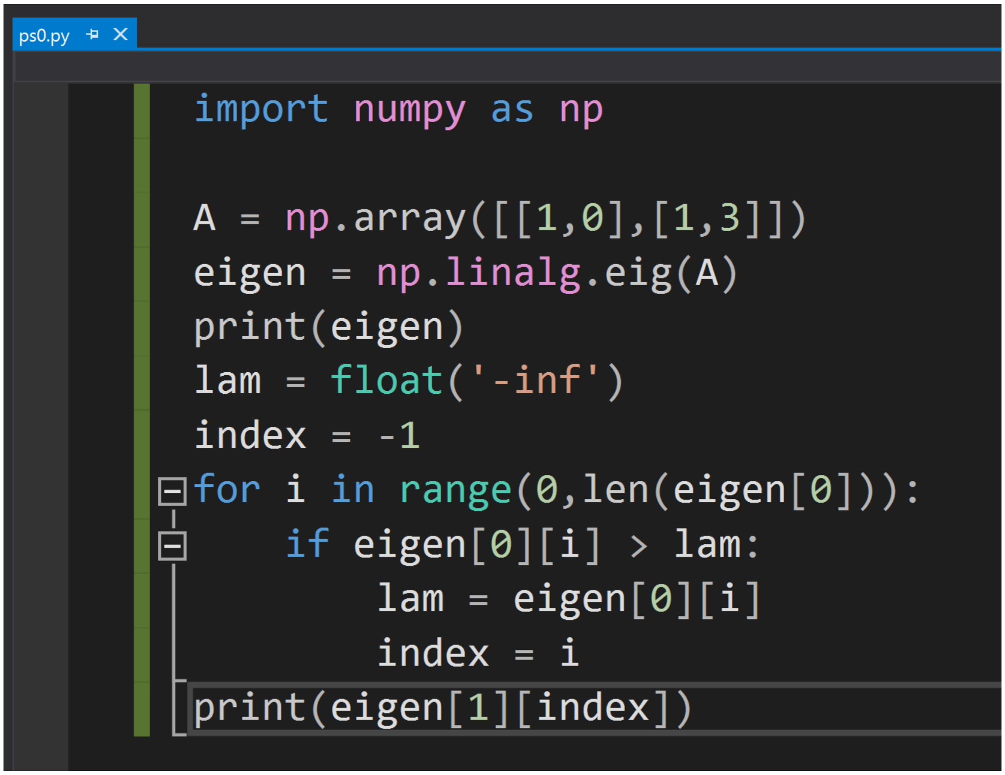
The plot of 1000 samples when plotting x_1 vs x_2 with the covariance matrix being changed is shown below. We see a downward trend in this distribution which introduced by the changed covariance matrix.


4
3
2
1
0
−1
−2
−2
0
2
4
6

11 Eigendecomposition

Write Python program to compute the eigenvector corresponding to the largest eigenvalue of matrix **A**.

The code for the program is included below.



```python
import numpy as np

A = np.array([[1,0],[1,3]])
eigen = np.linalg.eig(A)
print(eigen)
lam = float('-inf')
index = -1
for i in range(0,len(eigen[0])):
    if eigen[0][i] > lam:
        lam = eigen[0][i]
        index = i
print(eigen[1][index])
```

The output of my program is: [0. 0.89442719]

12 Data

Find machine learning dataset that's public, free, supervised. Then provide:

12.a Name

The CIFAR-10 dataset

12.b Where to find

https://www.cs.toronto.edu/ kriz/cifar.html https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf

12.c Brief description

This is a widely used dataset in the computer vision sphere that is a subset of a much larger, 80 million image dataset, and has a collection of 60000 images that computer scientists wish to split into 10 different classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. The model uses 50000 images to fit the data and 10000 to test the model on the data and basically uses different computer vision algorithms to break down each of the pictures into specific features and, using a bag-of-words implementation, group pictures that have similar features and hopefully be able to predict which class a specific image should be classified as.

12.d Number of Examples

There are 6000 examples for each class.

12.e # of features

The features for this data set are less concrete than one may hope for because training the model requires working with different featuers from each image. One implementaiton could be to use SIFT, in which we're looking to have edge detection, whereas another implementation could be to use the Harris-Corner Detector in which we're looking for corners, whereas yet another could just look for the actual RGB pixel value and associate based on that. The purpose of many of these studies is to find out which feature detection algorithm and methodology will work the best and thus there isn't one set feature value or type.