HOMEWORK 1:
BACKGROUND TEST

CMU 10601: MACHINE LEARNING (SPRING 2016)
OUT: Aug. 31, 2016
DUE: 5:30 pm, Sep. 7, 2016
TAs: Ben Cowley, Pradeep Dasigi, Simon Shaolei Du

SOLUTIONS

Guidelines

The goal of this homework is for you to determine whether you have the mathematical background needed to take this class, and to do some background work to fill in any areas in which you may be weak. Although most students find the machine learning class to be very rewarding, it does assume that you have a basic familiarity with several types of math: calculus, matrix and vector algebra, and basic probability. You do not need to be an expert in all these areas, but you will need to be conversant in each, and to understand:

- Basic calculus (at the level of a first undergraduate course). For example, we rely on you being able to take derivatives. During the class you might be asked, for example, to calculate derivatives (gradients) of functions with several variables.
- Linear algebra (at the level of a first undergraduate course). For example, we assume you know how to multiply vectors and matrices, and that you understand matrix inversion.
- Basic probability and statistics (at the level of a first undergraduate course). For example, we assume you know how to find the mean and variance of a set of data, and that you understand basic notions such as conditional probabilities and Bayes rule. During the class, you might be asked to calculate the probability of a data set with respect to a given probability distribution.
- Basic tools concerning analysis and design of algorithms, including the big-O notation for the asymptotic analysis of algorithms.

For each of these mathematical topics, this homework provides (1) a minimum background test, and (2) a medium background test. If you pass the medium background tests, you are in good shape to take the class. If you pass the minimum background, but not the medium background test, then you can still successfully take and pass the class but you should expect to devote some extra time to fill in necessary math background as the course introduces it. If you cannot pass the minimum background test, we suggest you fill in your math background before taking the class. Here are some useful resources for brushing up on, and filling in this background.
Probability

- Lecture notes: http://www.cs.cmu.edu/~aarti/Class/10701/recitation/prob_review.pdf.

Linear Algebra:

- Short video lectures by Prof. Zico Kolter: http://www.cs.cmu.edu/~zkolter/course/linalg/outline.html.
- Handout associated with above video: http://www.cs.cmu.edu/~zkolter/course/linalg/linalg_notes.pdf.
- Book: Gilbert Strang. Linear Algebra and its Applications. HBJ Publishers.

Matlab tutorial

- http://www.math.mtu.edu/~msgocken/intro/intro.pdf.

- http://ubcmatlabguide.github.io/.

Big-O notation:

- http://www.stat.cmu.edu/~cshalizi/uADA/13/lectures/app-b.pdf
- http://www.cs.cmu.edu/~avrim/451f13/recitation/rec0828.pdf
- See ASYMPTOTIC ANALYSIS (Week 1) in the following: https://class.coursera.org/algo-004/lecture/preview

Instructions

- **Homework Submission:** Submit **BOTH** a hard-copy to the hanging folders outside Sandra Winkler's office (GHC 8221) **AND** an electronic version to Gradescope. Both must be submitted on time. For Gradescope, you will need to specify which pages go with which question. Please submit Sections 1 through 4 to Q1, Sections 5 through 6.3 to Q2, and Sections 6.4 to 8 to Q3. We provide a LaTeX template in which you can use to type up your solutions. This template ensures the correct sections start on new pages. Please check Piazza for updates about the homework.
- **Collaboration policy**: For this homework **only**, you are welcome to collaborate on any of the questions with anybody you like. However, you *must* write up your own final solution, and you must list the names of anybody you collaborated with on this assignment. The point of this homework is not really for us to evaluate you, but instead for *you* to determine whether you have the right background for this class, and to fill in any gaps you may have.

Minimum Background Test [80 pts]

1 Vectors and Matrices [20 pts]

Consider the matrix X and the vectors $\mathbf{y}$ and $\mathbf{z}$ below:

$$X = \begin{pmatrix} 9 & 8 \\ 7 & 6 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 9 \\ 8 \end{pmatrix} \qquad \mathbf{z} = \begin{pmatrix} 7 \\ 6 \end{pmatrix}$$

1. What is the inner product of the vectors $\mathbf{y}$ and $\mathbf{z}$? (this is also sometimes called the *dot product*, and is sometimes written as $\mathbf{y}^T\mathbf{z}$)
2. What is the product $X\mathbf{y}$?
3. Is X invertible? If so, give the inverse, and if no, explain why not.
4. What is the rank of X?

1.

$$y^\top z = 9 * 7 + 6 * 8 = 111.$$

2.

$$Xy = \begin{pmatrix} 9 * 9 + 8 * 8 \\ 7 * 9 + 8 * 6 \end{pmatrix} = \begin{pmatrix} 145 \\ 111 \end{pmatrix}.$$

3.

$$X^{-1} = \frac{1}{9 * 6 - 7 * 8} \begin{pmatrix} 6 & -8 \\ -7 & 9 \end{pmatrix} = \begin{pmatrix} -3 & 4 \\ 3.5 & 4.5 \end{pmatrix}.$$

4.

$$2$$

.

2 Calculus [20 pts]

1. If $y = 4x^3 - x^2 + 7$ then what is the derivative of y with respect to x?
2. If $y = \tan(z)x^{6z} - \ln(\frac{7x+z}{x^4})$, what is the partial derivative of y with respect to x?

1.

$$\frac{\partial y}{\partial x} = 12x^2 - 2x.$$

2.

$$\frac{\partial y}{\partial x} = \tan(z)\, 6zx^{6z-1} - \frac{7}{7x+z} + \frac{4}{x}.$$

3 Probability and Statistics [20 pts]

Consider a sample of data $S = \{0, 1, 1, 0, 0, 1, 1\}$ created by flipping a coin x seven times, where 0 denotes that the coin turned up heads and 1 denotes that it turned up tails.

1. What is the sample mean for this data?
2. What is the sample variance for this data?
3. What is the probability of observing this data, assuming it was generated by flipping a biased coin with $p(x = 1) = 0.7, p(x = 0) = 0.3$.
4. Note that the probability of this data sample would be greater if the value of $p(x = 1)$ was not 0.7, but instead some other value. What is the value that maximizes the probability of the sample S? Please justify your answer.
5. Consider the following joint probability table where both A and B are binary random variables:

A	B	$P(A,B)$
0	0	0.1
0	1	0.4
1	0	0.2
1	1	0.3

   (a) What is $P(A = 0, B = 0)$?
   (b) What is $P(A = 1)$?
   (c) What is $P(A = 0|B = 1)$?
   (d) What is $P(A = 0 \vee B = 0)$?

1.
$$4/7.$$

2.
$$12/49.$$

3.
$$L = 0.3 * 0.7 * 0.7 * 0.3 * 0.3 * 0.7 * 0.7 = 0.0064827.$$

4. Let $p = p(x = 1)$, then $L(p) = p^4 (1 - p)^3$. Maximize $L(p)$ we have the optimal p is $4/7$.

5. (a)
$$0.1.$$

   (b)
$$0.5.$$

   (c)
$$4/7.$$

   (d)
$$0.7.$$

4 Big-O Notation [20 pts]

For each pair (f, g) of functions below, list which of the following are true: $f(n) = O(g(n))$, $g(n) = O(f(n))$, both, or neither. Briefly justify your answers.

1. $f(n) = \frac{n}{2}$, $g(n) = \log_2(n)$.
2. $f(n) = \ln(n)$, $g(n) = \log_2(n)$.
3. $f(n) = n^{100}$, $g(n) = 100^n$.

1. $g(n) = O\left(f(n)\right)$.

$$\lim_{n \to \infty} \frac{f(n)}{g(n)} = \frac{\ln 2}{2} \frac{n}{\ln n} = \infty$$

.

2. Both.

$$\frac{f(n)}{g(n)} = \ln 2$$

.

3. $f(n) = O\left(g(n)\right)$.

$$\lim_{n \to \infty} \frac{f(n)}{g(n)} = \lim_{n \to \infty} \frac{100n^{99}}{\ln 100 \times 100^n} = \cdots = \lim_{n \to \infty} \frac{100!}{(\ln 100)^{100} n^{100}} = 0.$$

Medium Background Test [20 pts]

5 Algorithm [5 pts]

Divide and Conquer: Assume that you are given a sorted array with n integers in the range $[-10, +10]$. Note that some integer values may appear multiple times in the array. Additionally, you are told that somewhere in the array the integer 0 appears exactly once. Provide an algorithm to locate the 0 which runs in $O(\log(n))$. Explain your algorithm in words, describe why the algorithm is correct, and justify its running time.

One could do a binary search on the array to find the location of 0. Given the sorted array (A), we start with a random index i (which can be $\frac{n}{2}$, where n is the length of the array). If $A[i] \neq 0$, we repeat the process with $A[i+1:n]$ if $A[i] < 0$ or $A[0:i-1]$ if $A[i] > 0$. At each point, since we potentially split the array in half and repeat the process with one of the halves, the number of steps needed is $O(\log_2(n))$.

6 Probability and Random Variables [5 pts]

6.1 Probability

State true or false. Here Ω denotes the sample space and A^c denotes the complement of the event A.

1. For any $A, B \subseteq \Omega$, $P(A|B)P(B) = P(B|A)P(A)$. True
2. For any $A, B \subseteq \Omega$, $P(A \cup B) = P(A) + P(B) - P(A|B)$. False
3. For any $A, B, C \subseteq \Omega$ such that $P(B \cup C) > 0$, $\frac{P(A \cup B \cup C)}{P(B \cup C)} \geq P(A|B \cup C)P(B \cup C)$. True. Let $D = B \cup C$. By definitions of set operations, $P(A \cup D) \geq P(A \cap D)$, and we know that $P(D) \leq 1$, and thus $\frac{P(A \cup D)}{P(D)} \geq P(A \cap D)$
4. For any $A, B \subseteq \Omega$ such that $P(B) > 0, P(A^c) > 0$, $P(B|A^C) + P(B|A) = 1$. False
5. For any n events $\{A_i\}_{i=1}^n$, if $P(\bigcap_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$, then $\{A_i\}_{i=1}^n$ are mutually independent. False

6.2 Discrete and Continuous Distributions

Match the distribution name to its probability density / mass function. Below, $|\boldsymbol{x}| = k$.

(a) Laplace h

(b) Multinomial i

(c) Poisson l

(d) Dirichlet k

(e) Gamma j

(f) $f(\boldsymbol{x}; \boldsymbol{\Sigma}, \boldsymbol{\mu}) = \frac{1}{\sqrt{(2\pi)^k \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$

(g) $f(x; n, \alpha) = \binom{n}{x} \alpha^x (1-\alpha)^{n-x}$ for $x \in \{0, \ldots, n\}$; 0 otherwise

(h) $f(x; b, \mu) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$

(i) $f(\boldsymbol{x}; n, \boldsymbol{\alpha}) = \frac{n!}{\Pi_{i=1}^k x_i!} \Pi_{i=1}^k \alpha_i^{x_i}$ for $x_i \in \{0, \ldots, n\}$ and $\sum_{i=1}^k x_i = n$; 0 otherwise

(j) $f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x \in (0, +\infty)$; 0 otherwise

(k) $f(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}$ for $x_i \in (0, 1)$ and $\sum_{i=1}^k x_i = 1$; 0 otherwise

(l) $f(x; \lambda) = \lambda^x \frac{e^{-\lambda}}{x!}$ for all $x \in Z^+$; 0 otherwise

6.3 Mean and Variance

1. Consider a random variable which follows a Binomial distribution: $X \sim \text{Binomial}(n, p)$.

   (a) What is the mean of the random variable? np

   (b) What is the variance of the random variable? $np(1-p)$

2. Let X be a random variable and $\mathbb{E}[X] = 1, \text{Var}(X) = 1$. Compute the following values:

(a) $\mathbb{E}[3X]$ 3

(b) $\text{Var}(3X)$ 9

(c) $\text{Var}(X+3)$ 1

6.4 Mutual and Conditional Independence

1. If X and Y are independent random variables, show that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

   $\mathbb{E}[XY] = \int_x \int_y xyp(x,y)dydx = \int_x \int_y xyp(x)p(y)dydx = \int_x p(x)dx \int_y p(y)dy = \mathbb{E}[X]\mathbb{E}[Y]$

2. If X and Y are independent random variables, show that $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.
   Hint: $\text{Var}(X+Y) = \text{Var}(X) + 2\text{Cov}(X,Y) + \text{Var}(Y)$

   $\mathbb{V}[X+Y] = \mathbb{V}[X] + 2\text{Cov}[X,Y] + V[Y]$. Consider $\text{Cov}[X,Y]$:
   $\text{Cov}[X,Y] = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY-\mathbb{E}[X]Y-\mathbb{E}[Y]X+\mathbb{E}[X]\mathbb{E}[Y]] = \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]-\mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. From (1), $X \perp Y \rightarrow \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. This implies $\text{Cov}[X,Y] = 0$. Thus, $\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.

3. If we roll two dice that behave independently of each other, will the result of the first die tell us something about the result of the second die? No, the result of the first die is independent of that of the second die. If, however, the first die's result is a 1, and someone tells you about a third event — that the sum of the two results is even — then given this information is the result of the second die independent of the first die? No, conditioned on the sum being even and the first die's roll is 1, we know the second roll cannot be 2,4, or 6.

6.5 Law of Large Numbers and the Central Limit Theorem

Provide one line justifications.

1. Suppose we simultaneously flip two independent fair coins (i.e., the probability of heads is $1/2$ for each coin) and record the result. After 40,000 repetitions, the number of times the result was two heads is close to 10,000. (Hint: calculate how close.)

   The expected value is 10,000 and the standard deviation is ~85:
   $\mathbb{E}[\sum_{i=1}^{40{,}000} I(c_i^1 = H, c_i^2 = H)] = 10,000$, where I is the indicator function and c_i^j is the ith flip of the jth coin.
   $\mathbb{V}[\sum_{i=1} I(c_i^1 = H, c_i^2 = H)] = \sum_{i=1} 40,000(1-1/4)^2(1/4) + (1/4)^2(3/4) = 7500$, and $\sqrt(7500) \approx 85$.

2. Let $X_i \sim \mathcal{N}(0,1)$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, then the distribution of $\bar{X}$ satisfies

   $$\sqrt{n}\bar{X} \overset{n\rightarrow\infty}{\longrightarrow} \mathcal{N}(0,1)$$

   The sum of Gaussian variables is Gaussian, with expected value and variance:
   $\mathbb{E}[\sqrt{n}\bar{X}] = \mathbb{E}[\sqrt{n}\frac{1}{n}\sum_i X_i] = \frac{1}{sqrtn} \cdot 0 = 0$
   $\mathbb{V}[\sqrt{n}\bar{X}] = \mathbb{V}[\sqrt{n}\frac{1}{n}\sum_i X_i] = \frac{1}{n}\sum_i^n 1 = 1$

   Some useful background material: http://www.cs.cmu.edu/~aarti/Class/10701/recitation/prob_review.pdf

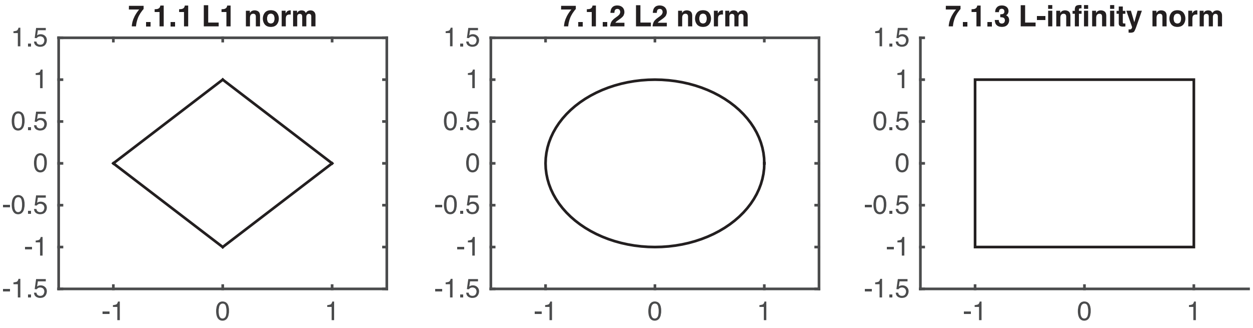
7 Linear algebra [5 pts]

7.1 Norm-enclature

Draw the regions corresponding to vectors $\mathbf{x} \in \mathbb{R}^2$ with the following norms:

1. $||\mathbf{x}||_1 \leq 1$ (Recall that $||\mathbf{x}||_1 = \sum_i |x_i|$)
2. $||\mathbf{x}||_2 \leq 1$ (Recall that $||\mathbf{x}||_2 = \sqrt{\sum_i x_i^2}$)
3. $||\mathbf{x}||_\infty \leq 1$ (Recall that $||\mathbf{x}||_\infty = \max_i |x_i|$)

   Solution:



7.2 Geometry

Prove that these are true or false. Provide all steps.

1. The smallest Euclidean distance from the origin to some point $\mathbf{x}$ in the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ is $\frac{|b|}{||\mathbf{w}||_2}$.
   True. You can formulate this as an optimization problem:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_2, \text{ s.t. } \mathbf{w}^T\mathbf{x} + b = 0$$

and solve this with Lagrangian multipliers. However, there is a more intuitive way.

We want to find a point $\mathbf{x}$ that resides in the hyperplane $\mathbf{w}^T\mathbf{x}+b = 0$ with the smallest distance to the origin.

First, let's find a point in the hyperplane. We know that a point must lie along the vector $\mathbf{w}$, so we can call this point $\lambda\mathbf{w}$. We solve for λ by using the hyperplane equation: $\mathbf{w}^T\lambda\mathbf{w} + b = 0 \rightarrow \lambda = \frac{-b}{\mathbf{w}^T\mathbf{w}}$. So a point on the hyperplane is $\mathbf{x}_0 = \frac{-b}{\mathbf{w}^T\mathbf{w}}\mathbf{w}$. We can now define any point on the hyperplane as $\mathbf{x} = \mathbf{x}_0 + \mathbf{v}$, where $\mathbf{w}^T\mathbf{v} = 0$ (i.e., $\mathbf{v}$ is any vector orthogonal to $\mathbf{w}$).

Second, let's find the point $\mathbf{x}^*$ that is on the hyperplane and closest to the origin. Thus, we want to minimize the Euclidean distance between some point $\mathbf{x}$ and $\mathbf{0}$: $\|\mathbf{x} - \mathbf{0}\|_2$. This Euclidean norm can be broken down more:
$\|\mathbf{x} - \mathbf{0}\|_2 = \|\mathbf{x}_0 + \mathbf{v}\|_2 = \|\frac{-b}{\mathbf{w}^T\mathbf{w}}\mathbf{w} + \mathbf{v}\|_2 = \sqrt{\frac{b^2\mathbf{w}^T\mathbf{w}}{(\mathbf{w}^T\mathbf{w})^2} + \mathbf{v}^T\mathbf{v}}$, where in the last step we used the fact that $\mathbf{w}^T\mathbf{v} = 0$. To minimize this, it is clear any value for $\mathbf{v}$ that is not zero would increase the value, so we set $\mathbf{v} = \mathbf{0}$ and thus the closest point to the origin is $\mathbf{x}^* = \mathbf{x}_0 = \frac{-b}{\mathbf{w}^T\mathbf{w}}$.

This gives us our solution: $\|\mathbf{x}^*\|_2 = \sqrt{(\frac{b^2\mathbf{w}^T\mathbf{w}}{(\mathbf{w}^T\mathbf{w})^2}} = \frac{|b|}{\|\mathbf{w}^T\mathbf{w}\|}$.
Full points will also be given if solved with the Lagrangian method.

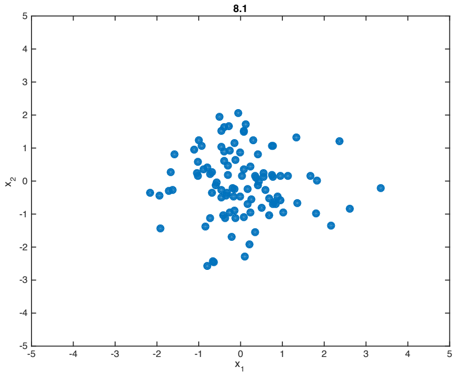
2. The Euclidean distance between two parallel hyperplane $\mathbf{w}^T\mathbf{x}+b_1=0$ and $\mathbf{w}^T\mathbf{x}+b_2=0$ is $\frac{|b_1-b_2|}{||\mathbf{w}||_2}$ (Hint: you can use the result from the last question to help you prove this one).

   True. Consider a point $\mathbf{x}$ on the first hyperplane. Because the two hyperplanes are parallel, the minimum Euclidean distance between $\mathbf{x}$ and some point $\mathbf{y}$ on the second hyperplane is the same for all $\mathbf{x}$. Thus, let $\mathbf{x} = \frac{-b_1}{\mathbf{w}^T\mathbf{w}}\mathbf{w}$, which resides on the first hyperplane. Treat this point as the origin. Then, find some point $\mathbf{y} = \frac{-b_2}{\mathbf{w}^T\mathbf{w}}\mathbf{w} + \mathbf{v}$, which resides on the second hyperplane (where $\mathbf{w}^T\mathbf{v} = 0$). Because we are changing the origin with respect to $\mathbf{x}$, we recenter $\mathbf{y}$ to $\hat{\mathbf{y}} = \frac{-b_2}{\mathbf{w}^T\mathbf{w}}\mathbf{w} + v - \frac{-b_1}{\mathbf{w}^T\mathbf{w}}\mathbf{w}$. The problem is now the same as the previous question (find some point $\hat{\mathbf{y}}$ that minimizes the Euclidean distance to the origin), and we found that $\mathbf{v} = \mathbf{0}$, so the minimum Euclidean distance is $\|\hat{\mathbf{y}}\|_2 = \frac{|b_1-b_2|}{\|\mathbf{w}\|_2}$.

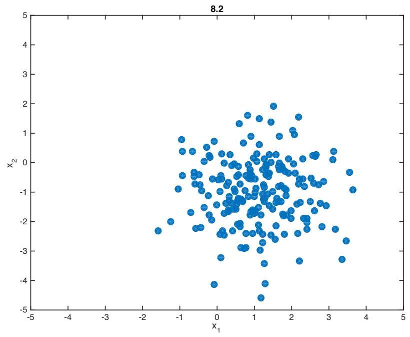
8 Programming Skills - Matlab [5pts]

Sampling from a distribution. For each question, submit a scatter plot (you will have 5 plots in total). Make sure the axes for all plots have the same limits. (Hint: You can save a Matlab figure as a pdf, and then use includegraphics to include the pdf in your latex file.)

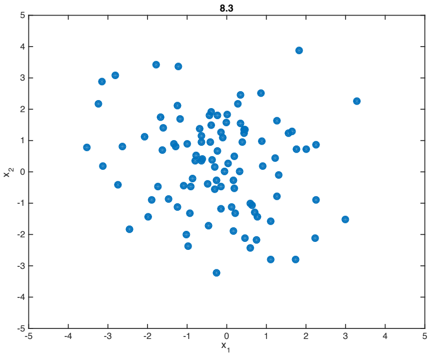
1. Draw 100 samples $\mathbf{x} = [x_1, x_2]^T$ from a 2-dimensional Gaussian distribution with mean $(0,0)^T$ and identity covariance matrix, i.e., $p(\mathbf{x}) = \frac{1}{2\pi}\exp\left(-\frac{||\mathbf{x}||^2}{2}\right)$, and make a scatter plot (x_1 vs. x_2). For each question below, make each change separately to this distribution.



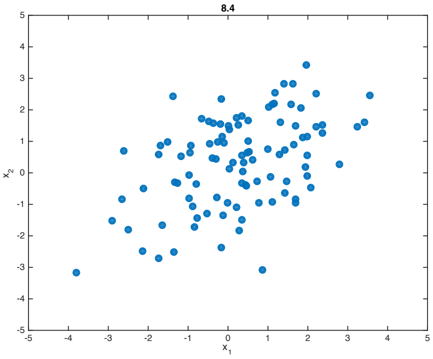
2. Make a scatter plot with a changed mean of $(1,-1)^T$.



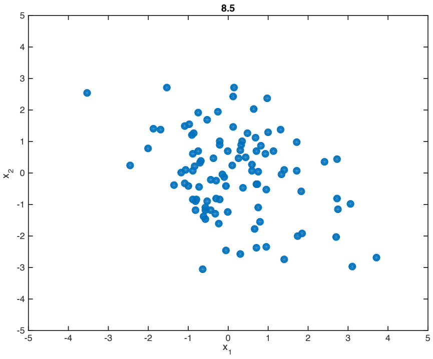
3. Make a scatter plot with a changed covariance matrix of $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$.



4. Make a scatter plot with a changed covariance matrix of $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.



5. Make a scatter plot with a changed covariance matrix of $\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$.


8.5
5
4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5
x_1
x_2