Purpose:

- Review essentials of probability
- Review theory of linear regression
- Use Python to solve a linear regression problem

Directions: This homework is to be done individually. Please upload a set of solutions containing your name and `@ucsc.edu` e-mail address to Canvas. Typeset (e.g. TeX) solutions are preferred, but scans or photographs of hand-written solutions are acceptable *provided that they are neat and legible.* The TA may deduct points for poorly organized or illegible solutions.

Question:	1	2	3	4	5	Total
Points:	18	27	30	24	6	105
Bonus Points:	0	0	0	5	5	10
Score:						

Questions: (ordered roughly by increasing difficulty)

1. **Conditional Probability Review:** There are 52 cards in a standard deck (excluding Jokers), with 13 cards per suit. The suits are Hearts, Spades, Diamonds, and Clubs, where Hearts and Diamonds are *red* and Spades and Clubs are *black*. Each suit contains 3 *face* cards (Jack, Queen, King), and 10 additional cards (Ace, Two, ..., Ten). Suppose we remove both *red* Kings and draw a card uniformly at random from the rest of the deck. *Give your answer as a fraction. Correct answers need not show work.*
   (a) (3 points) What is the probability that the card you draw is a *face* card? $P(F)$
   (b) (3 points) What is the probability that the card you draw is *black*? $P(B)$
   (c) (3 points) What is the probability that the card you draw is a *black face* card? $P(BF)$
   (d) (3 points) Given that the card you draw is black, what is the probability that it is a *face* card? $P(F|B)$
   (e) (3 points) Given that the card you draw is a *face* card, what is the probability it is *black*? $P(B|F)$
   (f) (*Ungraded*) *Verify that the five probabilities just calculated are consistent with the definition of conditional probability and Bayes's Rule* (i.e. $P(BF) = P(F|B)P(B) = P(B|F)P(F)$)
   (g) (3 points) Given that the card you draw is *not* a Two, what is the probability it is a Heart?

2. **Linear Regression:** Consider the vector equation $X\theta = \mathbf{y}'$ where
   - X is an $(n \times d)$ matrix
   - θ is a $(d \times 1)$ column vector
   - $\mathbf{y}'$ is an $(n \times 1)$ column vector
   - $1 < d < n$ and the underlying field is $\mathbb{R}$

   We will consider each row X_i as a feature vector corresponding to a calculated label y_i' and a true label y_i. The first element of each row of X is 1, and the remaining $d-1$ values in each row are given by the features of our data set. θ is chosen to minimize the average squared error between the components of $\mathbf{y}' = X\theta$ and $\mathbf{y}$.
   (a) (3 points) What is the maximum dimension of the hyperplane defined by $X\theta$?

(b) (3 points) What is the dimension of the vector space in which $\mathbf{y}$ is a point?

(c) (9 points) What point in the $X\theta$ hyperplane is closest (Euclidean distance) to $\mathbf{y}$? *Justify your answer.*

(d) (12 points) Consider that there exists some vector $\mathbf{v}$ such that

$$\mathbf{y} = \mathbf{y}' + \mathbf{v}$$

Show that

$$X^T\mathbf{v} = 0$$

Hint: Use proof by contradiction and the identity $\langle A\mathbf{w}, \mathbf{u}\rangle = \langle \mathbf{w}, A^T\mathbf{u}\rangle$.
Also note that this result derives the normal equation:

$$\left(X^T\mathbf{v} = 0\right) \implies \left(X^T\mathbf{y}' = X^T\mathbf{y}\right) \implies \left(X^TX\theta = X^T\mathbf{y}\right) \implies \left(\theta = (X^TX)^{-1}X^T\mathbf{y}\right)$$

3. **Linear Regression in Python:**

For this question, we will consider artificial data. Let us first start with the equation

$$X\theta' = \mathbf{z}'$$

where we fix X and give an arbitary vector θ from which to determine the value of $\mathbf{y}$'.

$$X = \begin{bmatrix} 1 & 3 & 9 & 2 \\ 1 & 6 & 9 & 1 \\ 1 & 7 & 7 & 7 \\ 1 & 8 & 6 & 4 \\ 1 & 1 & 0 & 8 \end{bmatrix} \quad ; \quad \theta' = \begin{bmatrix} 3 \\ 0 \\ 2 \\ -1 \end{bmatrix} \quad \implies \quad \mathbf{z}' = \begin{bmatrix} 19 \\ 20 \\ 10 \\ 11 \\ -5 \end{bmatrix}$$

To $\mathbf{z}'$, we add a small amount of noise $\mathbf{v}$ (which is *not* necessarily orthogonal to the $X\theta$ hyperplane!) and consider the resultant vector

$$\mathbf{y} = \mathbf{z}' + \mathbf{v} = \begin{bmatrix} 19 \\ 19 \\ 10 \\ 11 \\ -3 \end{bmatrix}$$

We now consider the linear regression problem of finding θ in

$$X\theta = \mathbf{y}'$$

such that the mean squared error between the components of and $\mathbf{y}'$ and $\mathbf{y}$ is minimized. For the purpose of a checking the reasonableness of our answer, we have generated our data in this way so θ and θ' will be close(*ish*).

We will be using Python for this problem, for which we now adopt the language of machine learning:

```
data =
[[3, 9, 2, 19],
 [6, 9, 1, 19],
 [7, 7, 7, 10],
 [8, 6, 4, 11],
 [1, 0, 8, -3]]
```

Our *training* data is a list of examples (or *instances*), where each example has been written on its own line in a row of four values. The first three values of each row are *features* of the data (corresponding to the values of X without the column of ones). The last entry in each line is the *label* (or *target*) of the instance and is the corresponding component of $\mathbf{y}$ (*not* $\mathbf{y}'$).

Our training data consists of 5 instances. We may label the first three features of the data with the variables x_1, x_2, and x_3 (considering $x_0 = 1$ for the omitted column of X). The first instance therefore has features $x_1 = 3, x_2 = 9, x_3 = 2$ and target $z = 19$.

(a) (3 points) Tell us which machine learning tool (or library, such as Scikit-learn) in Python you will use to solve this linear regression problem (this is your preference).

(b) (12 points) Run a linear regression algorithm on the full training set. The input features should be a 4-dimension vector $\mathbf{x} = (x_0, x_1, x_2, x_3)$, where $x_0 = 1$ is a constant, and x_1, x_2, x_3 correspond to the first, second, and third column in data. Report the model and "root mean squared error". The root mean squared error is defined as $RMSE = \sqrt{\sum_i (f(\mathbf{x}^{(i)}) - y^{(i)})^2 / N}$, where $f(\mathbf{x}^{(i)})$ is the prediction of instance-i, and N is the number of samples. Note in some tools, such as scikit-learn, the constant feature x_0 is added by default thus the input of your feature should simply be (x_1, x_2, x_3).

(c) (9 points) Suppose you had an *unlabeled* instance $\mathbf{x} = [3, 3, 5]$. What prediction for the label would the model from part (b) give?

(d) (6 points) If the examples are re-ordered (so the rows of X and elements of $\mathbf{y}$ are permuted), what happens to the learned θ vector and why?

4. **More Probability Review:** Suppose there are n students in CSE 142 this quarter. Each student is assigned a random integer sampled uniformly (with replacement) from the set $\{1, 2, ..., x\}$, where $x \geq n$.

(a) (6 points) What is the probability that $k \in \{0, 1, ..., n\}$ students are assigned the integer i?

(b) (9 points) What is the probability that *at least* two students are assigned the same random number?

(c) (9 points) Suppose k students are assigned the integer i. Because integers are sampled uniformly, the expectation value for k is $\bar{k} = (n/x)$. What is the variance $\mathbb{E}\left[\left(k - \bar{k}\right)^2\right]$? *You may appeal to computer algebraic tools (e.g.* Wolfram Alpha*) to simplify expressions, but the final answer is expected to be given in analytic form as a function of x and n.*

(d) (5 points (bonus)) Let σ^2 be your answer to part (c). What is the probability that greater than $\big((n/x) + \sigma\big)$ students are assigned the value i, as $n \to \infty$? *Hint: this answer does not actually depend on your answer to part (c). As an aside, note that if this probability is less than 0.05, and we were to observe such a finding, we could publish with evidence that the sampling was likely non-uniform!*

5. **A Continuous Variable plus Bayes's Rule:** Suppose it will rain $g(x) = \cot(x\pi/2)$ cm tomorrow, where $x \in (0, 1]$ is some unknown parameter. In this part of the world, $(1 - x)$ is precisely the probability of hearing thunder before sunset. This morning you assigned probability density $(n+1)x^n$ to each value of $x \in (0, 1]$ (your prior belief) where $n \geq 1$.

(a) (6 points) What is your expected value for x at midday?

(b) (5 points (bonus)) You hear no thunder before sunset. What probability density do you now assign to each value of x (your posterior), following Bayes's Rule? *Give your answer as a function $f(x)$ depending on n. Hint: constant factors may be ignored until the end, when we only need to ensure $\int_0^1 f(x)dx = 1$.*