# Homework 0, due Monday September 12

COMS 4771 Fall 2016

**Problem 1** (Values)**.** Name one or two of your own personal, academic, or career values, and explain how you hope machine learning can be of service to those values.

**Problem 2** (Stuff you must know). The course website `http://www.cs.columbia.edu/~djhsu/coms4771-f16/` has information about the course prerequisites, course requirements, academic rules of conduct, and other information. You are required to understand this information and abide by the rules of conduct, regardless of whether or not you can solve the following problems.

(a) True or false: I may share my homework write-up or code with another student as long as (1) the write-up only contains solutions for at most half of the problems, (2) the code is at most five lines, and (3) we list each other as discussion partners on the submitted write-up.

(b) True or false: I may use any outside reference material to help me solve the homework problems as long as I appropriately acknowledge these materials in the submitted write-up.

**Problem 3** (More stuff you should know)**.** We'll use the notation $f\colon \mathcal{X} \to \mathcal{Y}$ to declare a function $f$ whose domain is the set $\mathcal{X}$, and whose range is the set $\mathcal{Y}$. For example, $f\colon \mathbb{R} \to \mathbb{R}$ declares a real-valued function over the real line. For a positive integer $d$, the $d$-dimensional vector space called *Euclidean space* is denoted by $\mathbb{R}^d$. For positive integers $m$ and $n$, the space of $m \times n$ matrices over the real field $\mathbb{R}$ is denoted by $\mathbb{R}^{m \times n}$. Every matrix in $\mathbb{R}^{m \times n}$ can be regarded as a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$.

Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{2 \times 2}$ be given by

$$\boldsymbol{A} := \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \quad \boldsymbol{B} := \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}.$$

(":=" is the notation used for "equals by definition".) Also let $\boldsymbol{u} := (2, 1)$ and $\boldsymbol{v} := (1, 2)$, which are vectors in $\mathbb{R}^2$. Note that when we refer to vectors from Euclidean spaces in the context of matrix-vector products, we always regard vectors (like $\boldsymbol{u}$) as *column vectors*, and their *transposes* (like $\boldsymbol{u}^\top$) as *row vectors*:

$$\boldsymbol{u} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad \boldsymbol{u}^\top = \begin{bmatrix} 2 & 1 \end{bmatrix}.$$

(a) What is the rank of $\boldsymbol{A}$?

(b) What is $\boldsymbol{A}\boldsymbol{u} + \boldsymbol{B}\boldsymbol{v}$?

(c) What is $\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{v}$?

(d) The *(Euclidean) norm* (or *length*) of a vector $\boldsymbol{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ is denoted by $\|\boldsymbol{x}\|_2$, and is equal to $\sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}$. What is $\|\boldsymbol{u}\|_2$?

(e) Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be the function defined by

$$f(\boldsymbol{x}) := \boldsymbol{x}^\top (\boldsymbol{A} + \boldsymbol{B})\boldsymbol{x}.$$

The gradient of a real-valued function $g\colon \mathbb{R}^d \to \mathbb{R}$ at a point $\boldsymbol{z} \in \mathbb{R}^d$, denoted by $\nabla g(\boldsymbol{z})$, is the vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_d)$ where

$$\lambda_i := \frac{\partial}{\partial x_i} g(\boldsymbol{x}) \bigg|_{\boldsymbol{x}=\boldsymbol{z}} \quad \text{for all } i = 1, 2, \dots, d.$$

What is $\nabla f(\boldsymbol{v})$?

(f) The *unit circle* in $\mathbb{R}^2$ is the set of vectors in $\mathbb{R}^2$ with unit length, i.e., $\{\boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x}\|_2 = 1\}$. Which vector in the unit circle minimizes $f$ (defined above), and what is the value of $f$ evaluated at this vector? (*Hint*: think about eigenvectors.)

**Problem 4** (Random stuff you should know)**.** A (discrete) probability space is a pair $(\Omega, P)$, where $\Omega$ is a (discrete) set called the *sample space*, and $P\colon \Omega \to \mathbb{R}$ is a real-valued function on $\Omega$ called the *probability distribution*, which must satisfy $P(\omega) \geq 0$ for all $\omega \in \Omega$, and $\sum_{\omega\in\Omega} P(\omega) = 1$. An *event* $A$ is a subset of $\Omega$, and the probability of $A$, denoted by $P(A)$ (somewhat abusing notation), is equal to $\sum_{\omega\in A} P(\omega)$.

(a) A fair coin is tossed three times. Consider the three events:

- $A$: the outcome of the first toss is heads.
- $B$: the outcome of the second toss is tails.
- $C$: the outcomes of all three tosses are the same.
- $D$: exactly one of the outcomes is heads.

Which of the following pairs of events are independent?

- $A$ and $B$.
- $A$ and $C$.
- $A$ and $D$.
- $C$ and $D$.

(b) A student applies to two schools: Trump University and Columbia University. The student has a probability of 0.5 of being accepted to Trump, and a probability of 0.3 of being accepted to Columbia. The probability of being accepted by both is 0.2. What is the probability that the student is accepted to Columbia, given that the student is accepted at Trump?

A *random variable* (*r.v.*) on $(\Omega, P)$ is a real-valued function $X\colon \Omega \to \mathbb{R}$. The notation $X \sim P$ declares the r.v. $X$ and associates it with the probability distribution $P$. (We'll often leave the probability space implicit.) The *expected value* (a.k.a. *expectation* or *mean*) of $X$, written $\mathbb{E}(X)$, is the average value of $X$ under the distribution $P$:

$$\mathbb{E}(X) \;:=\; \sum_{\omega\in\Omega} X(\omega)\cdot P(\omega)\,.$$

An equivalent definition of $\mathbb{E}(X)$ is $\mathbb{E}(X) := \sum_x x \cdot P(X = x)$, where the summation is taken over all $x$ in the range of $X$, and $P(X = x)$ is shorthand for $P(\{\omega \in \Omega : X(\omega) = x\})$.

(c) Consider the sample space $\Omega = \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\}$, and let $P$ be the uniform distribution over $\Omega$, i.e., $P(a, b) = 1/36$ for each $(a, b) \in \Omega$. Let $X$ be the random variable defined by $X(a, b) = \min\{a, b\}$ for each $(a, b) \in \Omega$.

For each $x \in \{1, 2, \ldots, 6\}$, what is $P(X = x)$?

(d) Continuing from (c), what is the expected value of $X$?

(e) A biased coin with $P(\text{heads}) = 1/5$ is tossed repeatedly until heads comes up. What is the expected number of tosses?

(f) You create a random sentence of length $n$ by repeatedly picking words at random from the vocabulary $\{\text{a}, \text{is}, \text{not}, \text{rose}\}$, with each word being equally likely to be picked. What is the expected number of times that the phrase "a rose is a rose" will appear in the sentence? (Note: the appearances may overlap.)

**Problem 5** (More random stuff you should know)**.** We often encounter probability spaces $(\Omega, P)$ where $\Omega$ is not a discrete set. In this class, the only random variables we'll consider on such spaces will either have a discrete image (i.e., $\{X(\omega) : \omega \in \Omega\}$ is a discrete set) or have a *probability density function* $p \colon \mathbb{R} \to \mathbb{R}$, which is a non-negative real-valued function on $\mathbb{R}$ such that, for any open interval $(a, b) = \{x \in \mathbb{R} : a < x < b\} \subseteq \mathbb{R}$,

$$P(X \in (a,b)) \;=\; P(\{\omega \in \Omega : X(\omega) \in (a,b)\}) \;=\; \int_{(a,b)} p(x)\,\mathrm{d}x\,.$$

Random variables with probability density functions will be called *continuous random variables.*

(a) Let $X$ be a continuous random variable with probability density function $p$ given by

$$p(x) \;:=\; \begin{cases} 0 & \text{if } x < 0\,, \\ \lambda e^{-\lambda x} & \text{if } x \geq 0\,. \end{cases}$$

Here, $\lambda$ is a positive number (typically called the rate parameter). If $P(X \leq 1000000) = 0.5$, then what is the value of $\lambda$?

(b) Let $X$ be a *standard normal random variable*, i.e., a continuous random variable whose density is the *standard normal density* $p(x) := e^{-x^2/2}/\sqrt{2\pi}$ for all $x \in \mathbb{R}$. Define the random variable $Y$ on the same probability space as $X$ by $Y := X^2$, i.e., $Y(\omega) := X(\omega)^2$ for all $\omega \in \Omega$. What are $\mathbb{E}(X)$ and $\mathbb{E}(Y)$?

A collection of continuous random variables $X_1, X_2, \ldots, X_d$, all defined on the same probability space, has a *(joint) probability density function* $p \colon \mathbb{R}^d \to \mathbb{R}$ if, for any $A \subseteq \mathbb{R}^d$,

$$P((X_1, X_2, \ldots, X_d) \in A) \;=\; \int_A p(x_1, x_2, \ldots, x_d)\,\mathrm{d}x_1\,\mathrm{d}x_2 \cdots \mathrm{d}x_d\,.$$

We'll often collect several random variables, such as $X_1, X_2, \ldots, X_d$, into a *random vector* $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$. So the equation above can be written as $P(\boldsymbol{X} \in A) = \int_A p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$.

(c) Suppose the pair of random variables $(X_1, X_2)$ has probability density function $p$ given by

$$p(x_1, x_2) \;:=\; \begin{cases} c & \text{if } 0 \leq x_1 \leq 0.5 \text{ and } 0 \leq x_2 \leq 1\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Here, $c$ is a constant (that does not depend on $x_1$ or $x_2$). What should be the value of $c$ so that $p$ is a valid probability density function?

(d) Continuing from (c), what is the probability that $X_2 \geq X_1$?

(e) Continuing from (c), define another random variable $Y$ on the same probability space as $X_1$ and $X_2$ by

$$Y \;:=\; \begin{cases} 1 & \text{if } X_1 > 2X_2\,, \\ -1 & \text{otherwise}\,. \end{cases}$$

Are $X_1$ and $Y$ independent? What is the expected value of $Y$?

(f) Continuing from (c), define yet another random variable $Z$ on the same probability space as $X_1$ and $X_2$ by

$$Z \;:=\; \begin{cases} 1 & \text{if } X_2 > 1/2\,, \\ -1 & \text{otherwise}\,. \end{cases}$$

Are $X_1$ and $Z$ independent? What is the expected value of $X_1 Z$?

Unfortunately the Google Cloud setup is not quite ready yet.

**Problem 6** (Google Cloud; *optional but recommended*)**.** Set up a virtual machine on Google Cloud. Figure out how to install some useful Python packages like `numpy`, `scipy`, `scikit-learn`, etc. Download the OCR image data set `ocr.mat` from Courseworks, and load it into memory:

```
from scipy.io import loadmat
ocr = loadmat('ocr.mat')
```

This file contains four different matrices called `data`, `labels`, `testdata`, and `testlabels`. For example, `data` represents a 60000×784 matrix, which you can verify using the following command:

```
ocr['data'].shape
```

Using the `numpy` and `scipy` libraries, write some code to compute the average squared Euclidean norm of the rows of `data`. The following functions may be useful:

- `numpy.linalg.norm`
- `numpy.mean`

The result should be around 5.7 million. You don't need to submit anything for this problem.