Homework 0

Due: 9/5/19

1. **Formalizing prediction problems.**

In class, we defined a *supervised learning problem*: given a collection of data \((x_i, y_i) \in (\mathcal{X} \times \mathcal{Y})\) for \(i = 1, \ldots, n\), find a mapping \(f: \mathcal{X} \rightarrow \mathcal{Y}\) so that

\[ y_i \approx f(x_i). \]

What might \(\mathcal{X}\), \(\mathcal{Y}\), and \(f: \mathcal{X} \rightarrow \mathcal{Y}\) be in the following cases? What kinds of data (continuous, discrete, nominal, ordinal) do \(\mathcal{X}\) and \(\mathcal{Y}\) contain? Is any of the data likely to be missing? Why?

- **Medical treatment planning:** You receive a patient’s medical record, including a complete history of medical symptoms, diagnostic tests, and treatments. You want to identify which treatment will work best.
- **Electoral campaigning:** Given a voter’s voting history, you want to predict whether a given voter is likely to support your candidate.
- **Time series forecasting:** You’d like to predict how your favorite stock will perform tomorrow.
- **Handwriting recognition:** The post office would like an automated procedure to understand which zip code is written on an envelope.
- **Class placement:** You’re in charge of determining class sections in a middle school. You receive a file on each student with their previous course history and exam results, and want to place them in the appropriate math class.
- **Pick your own problem:** Define \(\mathcal{X}\), \(\mathcal{Y}\), and \(f\) for a big messy prediction problem you’d like to solve.

2. **Coding experience.**

(a) For every student in the class, let \(x \in \mathbb{N}^6\) be a vector describing coding experience. Each entry of \(x\) corresponds to a programming language, and gives the (approximate) number of lines of code that each student has written in each language. Index 1 refers to Julia, 2 to Python, 3 to Matlab, 4 to R, 5 to C or C++, and 6 to Java. Write down your vector \(x_{me}\).
(b) Your TAs have written the following Julia code to process your coding experience vector. What are they trying to do?

```julia
n = 0
for j = 1:6
    if x[j] > 0
        n += 1
    end
end
return n
```

(c) We would like to identify which students have taken a computer science class. To formalize our problem, let’s say that the feature space \( \mathcal{X} = \mathbb{N}^6 \), and the space of outcomes is \( \mathcal{Y} = \{ \text{has taken a CS class}, \text{has not taken a CS class} \} \). Suppose we have found a vector \( w \in \mathbb{R}^6 \) and a number \( b \in \mathbb{R} \) so that \( w^T x > b \) whenever a student with coding experience vector \( x \) has taken a computer science class, and \( w^T x \leq b \) otherwise. Write down a piecewise definition of the function \( f : \mathcal{X} \to \mathcal{Y} \) mapping coding experience vectors to the labels in \( \mathcal{Y} \). (You can also write pseudocode if you prefer.)

(d) Do you think \( w_5 \) is positive, negative, or 0? Why? What about \( w_2 \)? What about \( b \)? Note: this problem does not have a unique right answer...!

(e) Let’s restrict the problem to two dimensions, so we can draw a picture of it. Now \( x \in \mathbb{N}^2 \) and \( w \in \mathbb{R}^2 \) will be vectors in two dimensions. The first coordinate will represent C and C++. The second coordinate will represent Python.

   i. Guess a value for the vector \( w \) and the offset \( b \) that agrees with your reasoning on the previous question.

   ii. On a cartesian grid, draw (your guess of) the vector \( w \).

   iii. On the same grid, draw the line \( w^T x = b \).

   iv. What is the \( x \)-intercept of the line \( w^T x = b \), in terms of \( w \) and \( b \)?

   v. What is the geometric relationship between the vector \( w \) and the line \( w^T x = b \)?

   vi. On the same grid, plot two example coding experience vectors: one that would be classified as a person who has taken a CS class, and one that would not.