Yale University
Department of Statistics and Data Science
Quiz 2

STATISTICS 365/565

Issued: 02/24/2021 **Due:** 02/26/2021

Notes: You will have one hour to solve this problem. You cannot discuss this quiz with anybody at any time before the due date. You *can* use notes, online resource, videos, etc... Just nothing adaptive on which you can ask a direct question and get it answered (e.g. no stackoverflow/slack/asking a friend/etc...).

Submission: You will submit this to gradescope as a PDF. Note that there is a time limit, so once you start reading this file your time starts.

Background: K nearest neighbors regression versus classification Suppose we have training data x_i, y_i. For a point x, denote its k-nearest neighborhood as $N_k(x)$. Recall that for kNN regression with the ℓ_2 loss (from HW2) we have

$$f_k(x) = \frac{1}{k} \sum_{i|x_i \in N_k(x)} y_i$$

That is, just the average of the labels of the k closest feature vectors to x.

Suppose that instead of a regression problem we wish to do classification where our labels are $y_i \in \{-1, +1\}$. Note, that the labels are not $\{0, 1\}$.

Problem statement: Suppose you have access to $f_k(x)$, but you do **NOT** have access to the training data. How can you use $f_k(x)$ to build a classifier?