Homework 1

1. The different types of machine learning problems. Determine whether the tasks described below involve supervised learning or unsupervised learning. For supervised learning problems, identify them as regression, classification, or probabilistic classification.

   (a) Forecast the daily trading volume of the stock AAPL, given the trading volume of the previous 10 trading days.

   (b) Forecast the trading volume of AAPL over the next 4 trading days, given the trading volume of the past 10 trading days.

   (c) Predict whether or not the trading volume of AAPL on a day will be greater than the 5-day trailing average of trading volume.

   (d) Guess the probability that the trading volume of AAPL on a day will be greater than the 5-day trailing average of trading volume.

   (e) Generate fake or fictitious trading volumes for AAPL for 10 consecutive trading days, with the goal of fooling an expert. (This means the expert can’t distinguish between your generated sequences of volumes, and actual historical ones.)

2. Fitting a known function using samples. In this problem you will use various nearest neighbor methods to predict \( y \in \mathbb{R} \) given \( x \in \mathbb{R} \), for a simple case in which we know the exact relation between \( x \) and \( y \). (This is never the case in practical prediction problems.)

   Consider the function \( f(x) = \sin(10x) \) over \( x \in [0, 1] \).

   (a) Randomly sample 30 points \( x^i \) from \([0, 1]\) using a uniform distribution, and let \( y^i = f(x^i) \). Plot these data points as dots, along with \( f \) as a curve. (To plot \( f \), evaluate it for 500 points uniformly spaced in \([0, 1]\), i.e., \( x = (k-1)/499, k = 1, \ldots, 500. \))

   (b) On eight separate plots, plot the \( k \)-nearest neighbor predictors for \( k = 1, 2, 3 \) and the soft nearest neighbor predictors for \( \rho = 0.0001, 0.0003, 0.001, 0.003, 0.01 \). Include the 30 data points, shown as dots, in these plots.

   (c) RMS error. For each of the eight predictor functions in part (b), evaluate the RMS error on the 500 uniformly spaced points used to plot the functions, given by

   \[
   \left( \frac{1}{500} \sum_{k=1}^{500} (\hat{y}_k - y_k)^2 \right)^{1/2},
   \]

   with \( y_k = f((k-1)/499) \) and \( \hat{y}_k = g((k-1)/499) \), where \( g \) is your predictor.

Julia hints. \texttt{rand(N)} generates \( N \) points from a uniform distribution on \([0, 1]\). To generate a uniformly spaced set of \( N \) values between \( a \) and \( b \) (with \( a < b \)), use
range(a, stop=b, length=N). To apply a function $f: \mathbb{R} \rightarrow \mathbb{R}$ elementwise to a vector $x$, use $f.(x)$.

3. **Polynomial embedding.** You are given raw data $(u, v)$ with $u \in \mathbb{R}^3$ and $v \in \mathbb{R}$. We embed $v$ as $y = v$ and $u$ as $x = \phi(u)$. We will use a linear regression model, $\hat{y} = x^T\theta = \phi(u)^T\theta$, with $\theta \in \mathbb{R}^d$. Your job is to find an appropriate embedding function $\phi: \mathbb{R} \rightarrow \mathbb{R}^d$.

An expert on the data and associated application believes that a polynomial of $u$ will give a good model of $v$. Specifically, she believes that a good prediction model can be found as a polynomial of degree no more than 3, with degree in each component $u_i$ no more than 2. We describe these terms below.

A polynomial of a vector $u \in \mathbb{R}^3$ is a linear combination of terms $u_1^p u_2^q u_3^r$, which are called *monomials*, where $p$, $q$, and $r$ are nonnegative integers, called the degree of the monomial in $u_1$, $u_2$, and $u_3$, respectively. The degree of the monomial $u_1^p u_2^q u_3^r$ is $p+q+r$. The degree of a polynomial of $u$ is the maximum of the degrees of its monomials, and its degree in each $u_i$ is the maximum of the degrees of its monomials in $u_i$. For example, the polynomial $5.7 + u_1^2 u_2 - 3.2 u_1^3 u_2^2 u_3 + 1.3 u_3$ has degree 6 and degree 3 in $u_1$, degree 2 in $u_2$, and degree 1 in $u_3$.

Suggest an appropriate embedding $\phi$, based on the expert’s advice. *Hint: $d = 17$.*