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) Predict the risk of an accident at an intersection, given features such as the time of day and weather.

   (b) Identify cars, bicyclists, and pedestrians in video taken by an autonomous vehicle's cameras.

   (c) Determine the probability that there is a stop sign in an image.

   (d) Generate new road scenarios (generate streets, place stop signs and intersections) for testing autonomous vehicles in a simulation.

2. *Train vs test datasets.* Suppose you are building a classifier that identifies cats and dogs. You have a dataset of 3,000 images containing cats, dogs, or other objects (neither cat nor dog). You randomly split the data into a 2,500 image training set and a 500 image test set.

   (a) Why is it important to "reserve" some images for the test dataset? (Why shouldn't we use all 3,000 images to train the classifier?)

   (b) After training your classifier for a while, you observe it performs well on the training images, but poorly on the test images. What is one possible explanation?

3. *Fitting a known function using samples.* In this problem you will use various nearest neighbor methods to predict $y \in \mathbf{R}$ given $x \in \mathbf{R}$, 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 $\rho = \sqrt{0.0001}$, $\sqrt{0.0003}$, $\sqrt{0.001}$, $\sqrt{0.003}$, $\sqrt{0.01}$. 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.

- `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 $f : \mathbf{R} \to \mathbf{R}$ elementwise to a vector x, use `f.(x)`.

4. *Polynomial embedding.* You are given raw data (u, v) with $u \in \mathbf{R}^3$ and $v \in \mathbf{R}$. 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 $\theta \in \mathbf{R}^d$. Your job is to find an appropriate embedding function $\phi : \mathbf{R}^3 \to \mathbf{R}^d$.

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 $u \in \mathbf{R}^3$ is a linear combination of terms $u_1^p u_2^q u_3^r$, called *monomials*. 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, degree 3 in u_1, degree 2 in u_2, and degree 1 in u_3.

Suggest an appropriate embedding ϕ, based on the expert's advice. *Hint*: $d = 17$.

5. *Confidence set for a probabilistic classifier.* We consider a probabilistic classifier that predicts the probabilities $\pi_1, \ldots, \pi_k$ of K possible outcomes, labeled $k = 1, \ldots, K$. (The probabilities $\pi_1, \ldots, \pi_k$ are nonnegative and sum to one.)

   (a) *Hard classifier.* (This is another term for a non-probabilistic classifier. A probabilistic classifier is sometimes called a *soft classifier.*) Suppose you want a hard classifier that guesses just one of the outcomes. How would you choose the outcome to guess, given the probabilistic classifier output $\pi_1, \ldots, \pi_K$?

   (b) *Classifier confidence set.* The 90% *confidence set* associated with the probabilistic classifier output $\pi_1, \ldots, \pi_K$ is the smallest subset of the possible outcomes $1, \ldots, K$ that has probability at least 90%. Explain how to find the 90% confidence set from the probabilistic classifier output $\pi_1, \ldots, \pi_K$. (A short description is fine.)