Statistical Learning: Homework 1
Homework due at 11:59pm 16 Esfand 1395

7 Esfand 1395

Instructions: There are 4 questions on this assignment. The last problem involves coding, which could be done in MATLAB (or R or Python).

1. A secret government agency has developed a scanner which determines whether a person is a terrorist. The scanner is fairly reliable; 95% of all scanned terrorists are identified as terrorists, and 95% of all upstanding citizens are identified as such. An informant tells the agency that exactly one passenger of 100 aboard an aeroplane in which you are seated is a terrorist. The agency decide to scan each passenger and the shifty looking man sitting next to you is the first to test positive. What are the chances that this man is a terrorist?

2. Consider the joint distribution of three variables which admit the factorization,

$$p(a, b, c) = p(a \mid b)p(b \mid a)p(a) \tag{1}$$

   where all variables are binary. How many parameters are needed to specify distributions of this form? Compare if you'd like to specify the joint distributions without any available conditional independencies.

3. We say that two random variables are pairwise independent if

$$p(X_2|X_1) = p(X_2) \tag{2}$$

   and hence

$$p(X_2, X_1) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2) \tag{3}$$

   We say that n random variables are mutually independent if

$$p(X_i|X_S) = p(X_i) \quad \forall S \subseteq \{1, \ldots, n\} \setminus i \tag{4}$$

   and hence

$$p(X_{1:n}) = \prod_{i=1}^{n} p(X_i) \tag{5}$$

Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. It suffices to give a counter example.

4. We design and test a classifier for handwritten digit recognition. A subset of the MNIST database is used, which is a benchmark dataset for classification algorithms. See `http://yann.lecun.com/exdb/mnist/` to read more about it. This subset of the data along with the loading code are posted on the course group named as *mnist_all.mat* and *loadMNISTData.m*. This code should be used to load and define matrices of training and test data for the various parts below. In the MNIST database, each training or test example is a 28-by-28 grayscale image. To ease programming of learning algorithms, these images have been converted to vectors of length $28^2 = 784$ by sorting the pixels in raster scan (row-by-row) order. The Matlab reshape command can be used to convert these vectors back to images for visualization. For example, we can plot the third training example of class 1 as follows:

```
>> imshow(reshape(train1(3,:), 28, 28)');
```

To reduce computational complexity and simulation time, we focus on only three of the ten handwritten digits: "1", "2", and "7". We explore the performance of K nearest neighbor (K-NN) classifiers at distinguishing these three digit classes. To determine neighborhoods, we use the Euclidean distance between pairs of vector-encoded digits,

$$d(x_i, x_j) = \|x_i - x_j\|_2 = \left(\sum_{\ell=1}^{784} (x_{i\ell} - x_{j\ell})^2 \right)^{0.5} \tag{6}$$

(a) Implement and submit a function which finds the K nearest neighbors of any given test digit, and classifies it according to a majority vote of their class labels. (This should be your own code, not something extracted from another K-NN implementation.) Construct a training set from the first 200 examples of each class ($N = 600$ total examples). What is the empirical accuracy (fraction of data classified correctly) of 1-NN and 3-NN classifiers on the test examples from these classes?

(b) Display 4 test digits which are correctly classified by the 1-NN classifier, and 4 which are incorrectly classified. Are there any existing patterns?

(c) Implement and submit a function which uses 5-fold cross-validation, on the training dataset from part (a), to estimate the accuracy of a K-NN classifier. Determine a cross-validation accuracy estimate for five candidate classifiers, produced by the use of $K = \{1, 3, 5, 7\}$ nearest neighbors. Create a plot of these accuracy estimates versus K. Which classifier is estimated to be most accurate?