Introduction to Data Science:
Homework 2 - Bayesian inference and Supervised learning

Kuo-Shih Tseng

In this homework, you need to write code for Bayesian inference and classification for IMU and laser data.

Compress all of your code and pdf files into a zip file. Then upload it to LMS system on time. Please cite references (e.g., websites or books) if you learn something from them.
NOTICE: DO NOT use any toolbox for this homework. If you use any library, you should make sure it is included correctly. If your code cannot independently run on Prof. Tseng's Matlab. You only got partial points.

1. Bayesian Inference (40%):
There is a Hidden Markov Model in Figure 1. Assume the hidden variables X_i to be boolean ($X_i \in \{1, 0\}$), where i = $0 \sim 10$. Assume the measurement variables Z_i to be boolean ($z_i \in \{1, 0\}$),where i = $1 \sim 10$. Let $P(X_0 = 1) = P(X_0 = 0) = 0.5$. Let the transition matrix $P(X_{t+1}|X_t)$ and sensor matrix $P(Z_t|Xt)$ be given by

$$T = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}, Z = \begin{bmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{bmatrix}$$

, where in the T matrix,
$T_{11} = P(X_{t+1} = 1|X_t = 1); T_{12} = P(X_{t+1} = 0|X_t = 1);$
$T_{21} = P(X_{t+1} = 1|X_t = 0); T_{22} = P(X_{t+1} = 0|X_t = 0);$
and in the Z matrix,
$Z_{11} = P(Z_t = 1|X_t = 1); Z_{12} = P(Z_t = 0|X_t = 1);$
$Z_{21} = P(Z_t = 1|X_t = 0); Z_{22} = P(Z_t = 0|X_t = 0);$
Consider the sequences of measurements $Z_{1:10}$ = {0, 0, 0, 1, 1, 1, 1, 0, 0, 0}
Please answer the following questions.
1) Derive the matrix form of smoothed estimation via aforementioned notations. (20%)
2) Given $Z_{1:10}$, compute the smoothed estimates of X_t, where t=1$\sim$10. (10%)
3) Find the most likely sequence of states $X_{1:10}$ given evidence $Z_{1:10}$. (10%)
Code Delivery:
Matlab code (.m file). The derivation should be electrical form (e.g., pdf files) instead of papers. The two files should be *smooth_seq.m* and *viterbi_seq.m*. Compress all of your code into a zip file. This code should display: The probability or output sequence of each algorithm.

2. Supervised learning for IMU data (20%):
Given three axes acceleration data from a Minibot, the data information is as follows:

There are 200 IMU data points. The format is [a_x, a_y,a_z label]. The a_x, a_y, and a_z represent the measured acceleration data along x-axis, y-axis, and x-axis, respectively. If the label is 0, it means that the Minibot is not inclined. If the label is 1, it means that the Minibot is inclined. The first 100 data is the training data and the other data is the testing data.

The goal is to detect the Minibot is inclined or not. Write four supervised learning algorithms to classify the training and testing data.
*Display the confusion table of the training and test data for each algorithm.
*Plot cost function v.s. iteration for perceptron and logistic regression.
1) Naive Bayes. (4%)
2) Perceptron. (4%)
3) Logistic regression. (4%)
4) Adaboost. (4%)
5) Compare and explain each algorithm's performance. (4%)

3. Supervised learning for laser data (40%):
Given a leg dataset including laser data X and labels Y, the data information is as follows:
There are 720 laser data points at each time step. The laser data was divided into several segments S_t^i, where t denotes the time step and i denote the i-th segments. The data is recorded in 120 seconds. The first 60 data is the training data and the other data is the testing data. The 5 features of data is as follows:
(a). The number of points in each segment:
$n = |S_t^i|$
(b). The standard deviation of each segment:
$\sigma = \sqrt{\frac{1}{n}\sum_j ||\boldsymbol{x_j} - \bar{\boldsymbol{x}}||^2}$, where $\mathbf{x}$ is the x-y data.
(c). The width of each segment:
(d). The circularity (s_c) of each segment:
A circle is represented as $(x-x_c)^2+(y-y_c)^2 = r_c^2$. The unknown variable x' is $[x_c, y_c, (x_c^2+y_c^2-r_c^2)]$. x' can be solved by pseudo inverse $x' = (A^TA)^{-1}A^Tb$, where

$$A = \begin{bmatrix} -2x_1 & -2y_1 & 1 \\ -2x_2 & -2y_2 & 1 \\ \vdots & \vdots & \vdots \\ -2x_n & -2y_n & 1 \end{bmatrix}, b = \begin{bmatrix} -x_1^2 - y_1^2 \\ -x_2^2 - y_2^2 \\ \vdots \\ -x_n^2 - y_n^2 \end{bmatrix},$$

$s_c = \sum_{i=1}^n [r_c - \sqrt{(x_c - x_i)^2 + (y_c - y_i)^2}]^2$
(e). The radius (r_c) of each segment:
The r_c can be computed as (d) mentioned.

The goal is to detect a segment is a leg or not. Write four supervised learning algorithms to classify the training and testing data.
*Display the confusion table of the training and test data for each algorithm.
*Plot cost function v.s. iteration for perceptron and logistic regression.
1) Naive Bayes. (8%)
2) Perceptron. (8%)
3) Logistic regression. (8%)
4) Adaboost. (8%)

5) Compare and explain each algorithm's performance. (8%)

Code Delivery:

Matlab code (.m file). The file name should be *NB.m*, *Perceptron.m*, *logistic.m* and *Adaboost.m*. Please set a variable to switch "IMU_label_data" or "xy_data". Compress all of your code into a zip file.

HW2 Delivery:

Please send a zip file including your paper work (e.g., 1.1 math derivation) in pdf and other matlab code.