Machine Learning: Foundations	Spring Semester, 2001/2

Homework 1: March 20

Lecturer: Yishay Mansour

Homework number 1.

Theory question I: Suppose you have two dices. A *white dice* that has five sides color white and one black and an *equal dice* that has three sides color white and three color black. You chose a dice, the white dice has probability a third and the equal dice has probability two-thirds. You perform an experiment, you role the dice that you choose once and it comes up white.

1. What is the Maximum Likelihood dice?
2. What is the Maximum A Posteriori dice?
3. What is the probability that this is the white dice?
4. Suppose you role the *same* dice again, what is the probability it will come up white?

Theory question II: Consider a collection of points, $x_1, \cdots x_m$, sampled according to an Exponential Distribution. (Recall that an exponential distribution has a parameter λ and the density of x is $\lambda e^{-\lambda x}$.)

1. What is the Maximum Likelihood estimate of λ.
2. What is the Maximum A Posteriori value of λ given that its prior distribution is exponential with parameter 1.

Programming assignment:

Write a program to implement the Naive Bayes algorithm with Gausian distribution. Run the program on the `iris` data set at the home page of the course.
File format: For each database you have a data file and an information file, that describing the input. In a nutshell, each line in the data file is a record, and the last value in the line is the one you would like to predict.
Methodology: Randomly split the data to two parts, a *training set*, which will be used to train the Naive Bayes algorithm, and a *test set.* Using the training set you build the predictor. For each point in the test set, you compute the prediction, and compare it to the true value. The size of the training set should be a parameter that you can adjust.
Output: Plot the error frequency on the test set as a function of the number of points in the training set. (It is enough to consider training sets of size 20, 50, 100, 200, 400, and 600. For each size average 5 different runs.)

The homework is due in three weeks