

Homework 1: Oct 31, 2010

*Lecturer: Yishay Mansour***Homework number 1.**

Theory question I: Consider a collection of points, x_1, \dots, x_m , sampled i.i.d. 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 (ML) estimate of λ .
2. What is the Maximum A Posteriori (MAP) value of λ given that its prior distribution is exponential with parameter 1.
3. Compute the Bayesian posterior distribution for an exponential distribution where λ is distributed according to a Gamma distribution, i.e.,

$$\text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\lambda \beta).$$

Theory question II: Consider a countable hypothesis class H with a distribution $p(h) > 0$ for $h \in H$, and an arbitrary target function f . Show that with probability $1 - \delta$, for every $h \in H$, we have

$$|\widehat{\text{error}}(h) - \text{error}(h)| \leq \sqrt{\frac{-c \log(p(h)\delta)}{m}}$$

for some constant $c > 0$, and where the probability is taken over a sample of size m .

Theory question III: Let $X = [0, 1)$. The class C_k includes concepts that switches k times. Each $c \in C_k$ has k parameters, $0 = z_1 < z_2 < \dots < z_k = 1$. The label of a point x which is in the interval $[z_i, z_{i+1})$ is $c(x) = i \bmod 2$. Consider a set S of m examples (x_i, b_i) where $x_i \in [0, 1)$ and $b_i \in \{0, 1\}$. Compute efficiently a concept $c \in C_k$ which minimizes the error on S . The running time should be polynomial in both m and k . (HINT: Use dynamic programming.)

Programming assignment:

You will need to write the programming assignments in the R (You can download it for free, see <http://www.r-project.org/>) or in MATLAB.

Write a program to implement the Naive Bayes algorithm with Gaussian 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 (example), and the last value in the line is the one you would like to predict.

Methodology: You split the file randomly to two parts. One part is the *training set*, which is used in order to learn the classifier. The other part is the *test set*, which evaluates the classifier you learned.

First, you take the training set and learn a classifier. Then you take the test set and evaluate the error rate of the classifier. The size of the training set should be a parameter that you can adjust.

Output: In your output plot two different error functions as a function of the fraction of points in the training set. Each point should be an average of 10 runs (random splits of training and testing). The error functions are:

1. The error frequency on the test set.
2. The error frequency on the training set.

Put the code you wrote in an open directory, and write the path to access it. Have in the same directory the Iris data set.

The homework is due in two weeks