EE 511 – Assignment 1

Due: 16 Jan '19

Part I: written Assignment (10 points)

1. Let s_0 and s_1 be two d-length vectors of Boolean variables. In other words, $s_{k,i} \in \{0, 1\}$ for $i = 1, \ldots, d$. You observe the vector $X = Y \oplus N$, where $y \in \{s_0, s_1\}$, $P_Y(s_i) = 0.5$, and N is a d-length vector of independent and identical Bernoulli variables with $P(N_i = 1) = q < 0.5$. ($\oplus$ indicates mod 2 addition.) Find the MAP decision rule for determining which signal was sent given an observation x. (You should get a result that looks like a minimum Hamming distance.) How would you choose $\{s_i\}$ to minimize the expected error?

2. The number of occurrences x_i of word i in a news article on topic t is modeled as a Poisson random variable with parameter $\lambda_{i,t}$. Assume there are 5 different topics that you are interested in detecting and you use a vocabulary of V words. (Actually, the vocabulary represents $V - 1$ words, and the V-th token counts the number of out-of-vocabulary items observed.) Each topic is equally likely.

   (a) Find the MAP decision rule for determining which topic a document with word count vector X corresponds to, assuming the document must correspond to one of the 5 topics. If all errors have equal cost, how would you characterize performance of the system?

   (b) Now assume that half the documents you see will not correspond to any of these topics. You don't have any examples of documents with other topics. Assume that the cost of missing an in-topic document is R times that of making an error of assigning a document to an incorrect topic. How would you change the decision function? How would you characterize system performance? [Note: This is an open-ended question. There are multiple reasonable answers.]

3. The Erlang distribution is given by

$$f(x) = \frac{\alpha^n x^{(n-1)} e^{-\alpha x}}{(n-1)!} \quad x \in [0, \infty)$$

   with parameters $\alpha > 0$ and integer $n > 0$. (This distribution results when you sum n independent exponential variables with parameter α.) Assuming n is known, find the ML estimate of α.

4. You are given the statistic s for a training set $\mathcal{X} = \{x_1, \ldots, x_n\}$,

$$s = \frac{1}{n} \sum_{i=1}^{n} x_i$$

   where x_i are i.i.d. observations generated by a exponential random variable. You see only s, and not the specifics of $\mathcal{X}$ other than the size n.

(a) Find the maximum likelihood estimate of the exponential parameter λ in terms of s, assuming that λ is fixed but unknown.

(b) Now find the MAP estimate of the unknown parameter λ from s assuming that λ is random and can be characterized by an Erlang distribution with known parameters α and $n = 3$.

Part II: Computer Assignment (10 points)

Download the zipcode data from the **data** page on `web.stanford.edu/ hastie/ElemStatLearn`. Implement and compare the performance of two classifiers: i) a k-nearest-neighbor classifier, and ii) a Gaussian classifier. (This should be easy to implement in Matlab.) Use 10% of the training data as validation data, i.e. to determine k and any choices you make related to the covariance structure. Describe the methods you used to make these choices, and report the performance you obtain on the held out validation data and the test data.