Assignment 1: Statistical Foundations

Unsupervised Learning

Maneesh Sahani

Due: Thursday 12 Oct, 2006

Note: all assignments for this course are to be handed in to the Gatsby Unit, **not** to the CS department. Please hand in all assignments at the beginning of lecture on the due date to Maneesh. Late assignments will be penalised. If you are unable to come to class, you can also hand in assignments to Rachel Howes in the Alexandra House 4th floor reception.

1. Read "Nuances of Probability Theory" by Tom Minka: http://research.microsoft.com/~minka/papers/nuances.html. Pick one of the topics and write a short paragraph discussing it (do you agree or disagree, can you think of another example, etc).

2. Read the preface and chapters 1 and 2 of E.T Jaynes *Probability Theory: The Logic of Science*, available on the web at http://omega.math.albany.edu:8008/JaynesBook.html. Pay particular attention to and try to understand: the desiderata (p 112-114) and the derivations of the sum and product rules. Write a short paragraph paraphrasing the arguments in your own words.

3. You will need to be familiar with the following terms from statistics.

   expected value, unbiased estimator, sufficient statistics, exponential family

   Find definitions in a textbook or on the web. Answer the following questions:

   (a) Let X be a Gaussian random variable with mean μ and variance σ^2. What is the expected value of $2X^2$?

   (b) Let $x_1 \dots x_n$ be samples from a Gaussian random variable with mean μ and variance σ^2. Is x_1 an unbiased estimator for μ? What about $x_1/3 + 2x_2/3$?

   (c) Let $x_1 \dots x_n$ be samples from a Gaussian random variable with mean μ and variance σ^2. What are the sufficient statistics for μ? What are the sufficient statistics for σ?

   (c) What is the definition for an exponential family distribution? Is the Binomial distribution in the exponential family? What are the natural parameters of a Gaussian distribution?

   In the coming weeks we will be making extensive use of the following distributions, which you should know. For each one of these write down the definition, the mean, and the variance:

   Bernoulli, Binomial, Multinomial, Beta, Dirichlet, Gaussian, Gamma

4. Assume you have a data set of independent and identically distributed **binary** vectors $X = \{\mathbf{x_1}, \dots, \mathbf{x_N}\}$ each of which is D-dimensional. Describe a simple statistical model for your data. What is the expression for the likelihood of the data given the parameters of your model? Write down the equations for how you would estimate the parameters of your model from the data.

Bonus In an example as the above, how would you calculate the (relative) probability of the three different models:

   (a) all D components are generated from a Bernoulli distribution with $q = 0.5$

   (b) all D components are generated from Bernoulli distributions with unknown (but identical) q

   (c) each component is Bernoulli distributed with separate, unknown q_d