Assignment 1

Unsupervised Learning

Maneesh Sahani & Yee Whye Teh

Due: Monday 20 Oct, 2008

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 the lecturer. 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. **[10 points] Principles of probability**. 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. **[25 points] Statistics and Distributions**. 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 $2X^2$? [1 point]

   (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 $x_1/3 + 2x_2/3$? [1 point]

   (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? [2 points]

   In the coming weeks we will be making extensive use of the following distributions, which you should know. For each one of these exponential family distributions write down the (i) definition, (ii) mean, (iii) variance (iv) natural parameters (in terms of the conventional parameters), and (v) give the form (and name, if you know it) of the conjugate prior distribution on the parameter(s):

   Binomial, Multinomial, Poisson, Beta, Dirichlet, Gaussian, Gamma

   [3 points each]

3. **[30 points] Models for binary vectors**. Consider a data set of binary (black and white) images. Each image is arranged into a vector of pixels by concatenating the columns of pixels in the image. The data set has N images $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ and each image has D pixels, where D is (number of rows $\times$ number of columns) in the image. For example, image $\mathbf{x}^{(n)}$ is a vector $(x_1^{(n)}, \dots, x_D^{(n)})$ where $x_d^{(n)} \in \{0, 1\}$ for all $n \in \{1, \dots, N\}$ and $d \in \{1, \dots, D\}$.

   (a) Explain why a multivariate Gaussian is not an appropriate model for this data set of images. [3 points]

Assume that the images were modelled as independently and identically distributed samples from a D-dimensional **multivariate Bernoulli distribution** with parameter vector $\mathbf{p} = (p_1, \ldots, p_D)$, which has the form

$$P(\mathbf{x}|\mathbf{p}) = \prod_{d=1}^{D} p_d^{x_d}(1-p_d)^{(1-x_d)}$$

where both $\mathbf{x}$ and $\mathbf{p}$ are D-dimensional vectors

(b) How many bits would it take on average to code this data set? [2 points]

(c) What is the equation for the maximum likelihood (ML) estimate of $\mathbf{p}$? Note that you can solve for $\mathbf{p}$ directly. [5 points]

(d) Assuming independent Beta priors on the parameters p_d

$$P(p_d) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_d^{\alpha-1}(1-p_d)^{\beta-1}$$

and $P(\mathbf{p}) = \prod_d P(p_d)$ What is the maximum a posteriori (MAP) estimate of $\mathbf{p}$? Hint: maximise the log posterior with respect to $\mathbf{p}$. [5 points]

Download the data set `binarydigits.txt` from the course website, which contains $N = 100$ images with $D = 64$ pixels each, in an $N \times D$ matrix. These pixels can be displayed as 8×8 images by rearranging them. View them in Matlab by running `bindigit.m` (almost no Matlab knowledge required to do this).

(e) Write code to learn the ML parameters of a multivariate Bernoulli from this data set and display these paramteres as an 8×8 image. Hand in your code and the learned parameter vector. (Matlab or Octave code is preferred, but C or Java are acceptable). [10 points]

(f) Modify your code to learn MAP parameters with $\alpha = \beta = 3$. What is the new learned parameter vector for this data set? Explain why this might be better or worse than the ML estimate. [5 points]

4. **[10 points] Latent Variable Models**.

(a) Describe a real-world data set which you believe could be modelled using factor analysis. Argue why factor analysis is a sensible model for this data. What do you expect the factors to represent? How many factors do you think there would be? Are the linearity and Gaussianity assumptions reasonable, and if not, how would you modify the model? [5 points]

(b) Describe a real-world data set which you believe could be modelled using a mixture model (do not use the example below). Argue why a mixture model is a sensible model for your real world data set. What do you expect the mixture components to represent? How many components (or clusters) do you think there would be? What parametric form would each component have? [5 points]

5. **[15 points] Principal Components Analysis**.

The conventional latent variable model for Probabilistic Principal Components Analysis has a standard normal latent $\mathbf{y}$ and an arbitrary *loading matrix* Λ.

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{0}, I)$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\Lambda\mathbf{y}, \psi I)\,.$$

An alternative model would be to draw $\mathbf{y}$ from a normal with diagonal covariance (say Υ), and then restrict Λ to be orthogonal:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{0}, \Upsilon); \qquad \Upsilon_{ij} = 0 \text{ for } i \neq j$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\Lambda\mathbf{y}, \psi I); \qquad \Lambda^{\mathsf{T}}\Lambda = I.$$

(a) Show that this alternative model is equivalent to the standard one. [5 points]

(b) Derive the mean and covariance of $p(\mathbf{y}|\mathbf{x})$ within the alternative model in the PCA limit, $\psi \to 0$. [10 points]

6. **[10 points] Linear Latent Models**. We seek to model D-dimensional data vectors using a K-dimensional latent space, with a linear mapping between them.

(a) How many parameters need to be fit for each of two possible models: probabilistic principal components analysis (PPCA) and factor analysis (FA)? [Just count the *raw* parameters, don't worry about accounting for degeneracies.] [4 points]

(b) Recall that the K-dimensional PCA subspace projection can be found by learning the weights of a linear auto-encoder network by gradient descent in the squared reconstruction error. Is the same true of FA if the uniquenesses (another term for the output noise variances Ψ_{dd}) are known? Explain, giving the reconstruction cost function if there is one. [3 points]

(c) What about if the uniquenesses are to be learnt too? [3 points]

BONUS QUESTIONS: please attempt the questions above before answering these.

7. **[Bonus: 10 points] Consistent beliefs**. A friend (perhaps not for much longer) reports that he is willing to bet on the following beliefs about two events A and B, at least one of which must occur (i.e. $P(\neg A \vee \neg B) = 0$):

$$b(A) = 0.5$$
$$b(B|A) = 0.5$$
$$b(A|B) = 0.5$$

(a) Show that these beliefs are inconsistent. [2 points]

(b) Construct a Dutch Book matched to his beliefs. [8 points]

8. **[Bonus: 10 points] Model selection**. In the binary data model above, how would you calculate the (relative) probability of the three different models:

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

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

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