Assignment 1

Unsupervised & Probabilistic Learning

Maneesh Sahani

Due: Monday 17 Oct, 2016

Note: Assignement are due at 11:00 AM (the start of lecture) on the day specified. Please bring them to the lecture hall to hand in. If you are taking this course as COMPGI18, then late assignments may be turned in turned to the CS department following the usual rules, or to Barry Fong in the Gatsby Unit. The usual College late assignments policy will apply.

1. **[28 marks] Statistics and Distributions**. In the coming weeks we will be making extensive use of the following distributions, all of which belong to the exponential family. For each of these distributions, find:

   (a) The standard exponential form, identifying the natural parameters in terms of the conventional parameters given in the table (i.e. the function $\phi(\theta)$), and the sufficient statistic (i.e. $\mathbf{T}(x)$).

   (b) The expected value of the sufficient statistics in terms of the natural or conventional parameters. These expectations are often called the "mean" or "moment" parameters of the distribution.

   The distributions to consider are:

Name	Domain	Symbol	Density or Probability fn
Multivariate Normal	$\mathbb{R}^D$	$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$	$\lvert 2\pi\Sigma \rvert^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$
Binomial	$\mathbb{Z}_{0-N}$	$x \sim \mathsf{Binom}(p)$	$\binom{N}{x} p^x (1-p)^{N-x}$
Multinomial	$[\mathbb{Z}_{0-N}]^D$	$\mathbf{x} \sim \mathsf{Multinom}(\mathbf{p})$	$\frac{N!}{x_1!\, x_2! \dots x_D!} \prod_{d=1}^{D} p_d^{x_d}$
Poisson	$\mathbb{Z}_{0+}$	$x \sim \mathsf{Poisson}(\mu)$	$\mu^x e^{-\mu}/x!$
Beta	$[0,1]$	$x \sim \mathsf{Beta}(\alpha, \beta)$	$\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Gamma	$\mathbb{R}_+$	$x \sim \mathsf{Gamma}(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
Dirichlet	$[0,1]^D$	$\mathbf{x} \sim \mathsf{Dirichlet}(\boldsymbol{\alpha})$	$\frac{\Gamma\left(\sum_{d=1}^D \alpha_d\right)}{\prod_{d=1}^D \Gamma(\alpha_d)} \prod_{d=1}^{D} x_d^{\alpha_d-1}$

   [4 marks each]

2. **[7 marks] ML in the exponential family**.

   Find the maximum-likelihood value of the *mean* parameters (as defined in the question above) of the general exponential family distribution whose natural parameters are θ:

$$p(\mathbf{x}|\theta) = g(\theta) f(\mathbf{x}) e^{\theta^\mathsf{T} \mathbf{T}(\mathbf{x})}$$

   for a data set of iid observations $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$.

3. **[25 marks] 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 would not be an appropriate model for this data set of images. [5 marks]

Assume that the images were modelled as independently and identically distributed samples from a D-dimensional **multivariate Bernoulli distribution** with parameter vector $\mathbf{p} = (p_1, \dots, p_D)$, 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) What is the equation for the maximum likelihood (ML) estimate of $\mathbf{p}$? Note that you can solve for $\mathbf{p}$ directly. [5 marks]

(c) Assuming independent Beta priors on the parameters p_d

$$P(p_d) = \frac{1}{B(\alpha, \beta)} p_d^{\alpha-1}(1 - p_d)^{\beta-1}$$

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 marks]

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).

(d) Write code to learn the ML parameters of a multivariate Bernoulli from this data set and display these parameters as an 8×8 image. Hand in your code and the learned parameter vector. (Matlab or Octave code is preferred, but R or python are acceptable). [5 marks]

(e) 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 marks]

4. **[15 marks] Model selection**. In the binary data model above, write down the expressions needed to 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

Assume the prior probability of all three models is the same. Which model is the most likely given the data in `binarydigits.txt`?

5. **[10 marks] 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 marks]

   (b) Describe a real-world data set which you believe could be modelled using a mixture model. 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 marks]

6. **[15 marks] 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:

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

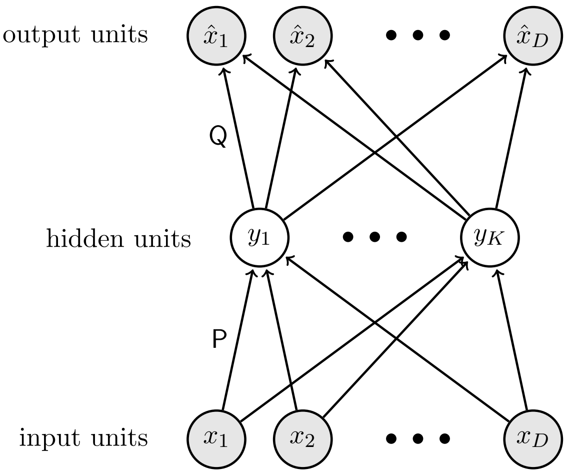
   (a) Show that this alternative model is equivalent to the standard one. [5 marks]

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

BONUS QUESTION: please attempt the questions above before answering this one.

7. **[Bonus: 20 marks] Linear Autoencoders.** A linear autoencoder is a "neural network" implementation of PCA. The network has three "layers" of "units" — each unit represents a single scalar variable, and so each layer represents a vector:

$\hat{x}_j$	$j = 1 \ldots D$	output units
y_k	$k = 1 \ldots K < D$	hidden units
x_i	$i = 1 \ldots D$	input units



   The mappings from input to hidden, and hidden to output layers, are linear:

$$y_k = \sum_i P_{ki}\, x_i$$
$$\hat{x}_j = \sum_k Q_{jk}\, y_k$$

   Given a set of N observed "input" vectors $\{\mathbf{x}_n\}$, the weight matrices P and Q are set to minimise the "autoencoding error" $\sum_n \|\hat{\mathbf{x}}_n - \mathbf{x}_n\|^2$, often using iterative gradient descent. Assume the input vector distribution has zero mean.

(a) Show that after the weights have converged, they obey the identities:

$$P = (Q^\mathsf{T} Q)^{-1} Q^\mathsf{T}$$
$$Q = \Sigma_X P^\mathsf{T} (P \Sigma_X P^\mathsf{T})^{-1}$$

where $\Sigma_X = \sum_n \mathbf{x}_n \mathbf{x}_n^\mathsf{T}$ and we assume that Q is rank K. [5 marks]

(b) It is clear that if P and Q minimize the error, then, for any invertible $K \times K$ matrix C the matrices $P_* = CP$ and $Q_* = QC^{-1}$ also achieve the same minimum. Show that you can always find a matrix C such that $Q_*^\mathsf{T} Q_* = I$ and so $P_* = Q_*^\mathsf{T}$. [2 marks]

(c) Show that the minimisation of the error is equivalent to the maximisation of

$$\mathsf{Tr}\left[Q_* Q_*^\mathsf{T} \Sigma_X\right]$$

and that this maximum is achieved by choosing the columns of Q_* proportional to the first K eigenvectors of Σ_X. (There are many ways to do this, but one option is to start from the argument presented in lecture showing that projection on the leading eigenvector maximises the projected variance.) [7 marks]

This demonstrates the assertion made in lecture that the linear autoencoder will find the subspace of the leading K principal components.

(d) How would you adapt the algorithm to implement factor analysis in the case that the uniquenesses (another term for the output noise variances Ψ_{dd}) are known? [3 marks]

(e) Is it possible to use an autoencoder for FA in the case of unknown uniquenesses? How, if so; or why not, if not? [3 marks]