Assignment 1

Unsupervised & Probabilistic Learning

Maneesh Sahani & Peter Orbanz

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 (i.e. $\langle \mathbf{T}(x) \rangle_{p(x|\theta)}$). These expectations are often called the "mean" or "moment" parameters of the distribution. [Note: show your derivation of the expectations; don't just look them up.]

   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	$[\mathbb{Z}_{0-N}]^D$	$\mathbf{x} \sim \mathsf{Multinom}(\mathbf{p})$	$\frac{N!}{x_1!\, x_2! \ldots x_D!} \prod_{d=1}^{D} p_d^{x_d}$
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**.

   Express the maximum-likelihood value of the *mean* parameters (as defined in the question above) of the general exponential family distribution

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

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

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 $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ 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 $(x_1^{(n)}, \ldots, x_D^{(n)})$ where $x_d^{(n)} \in \{0,1\}$ for all $n \in \{1, \ldots, N\}$ and $d \in \{1, \ldots, D\}$.

(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 $\mathbf{p} = (p_1, \ldots, p_D)$, 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) 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` or in Python by running `bindigit.py`.

(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. Include a listing of your code within your submission, and a visualisation of the learned parameter vector as an image. (You may use Matlab, Octave or Python) [5 marks]

(e) Modify your code to learn MAP parameters with $\alpha = \beta = 3$. Show the new learned parameter vector for this data set as an image. 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, find 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 that all three models are equally likely *a priori*, and take the prior distributions for any unknown probablities to be uniform. Calculate the posterior probabilities of each of the three models having generated the data in `binarydigits.txt`.

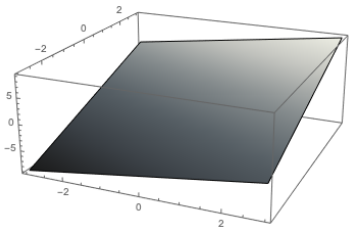
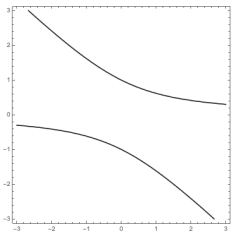
5. **[5 marks] Basic spectral properties**. Let A be a symmetric $n \times n$-matrix, with eigenvalues $\lambda_1, \ldots, \lambda_n$.

(a) Show that the matrix $B = A + cI$, where I is the identity matrix and $c \in \mathbb{R}$, has eigenvalues $\lambda_1 + c, \ldots, \lambda_n + c$. [3 marks]

(b) Suppose v and w are eigenvectors of A, both with the same eigenvalue λ. Show that any linear combination of v and w is again an eigenvector of A. What is its eigenvalue? [2 marks]

6. **[15 marks] Optimization**.

(a) Find the local (!) extrema of the function $f(x, y) := x + 2y$ subject to the constraint $y^2 + xy = 1$. For illustration, here are plots of the function f (left) and the set of points satisfying the constraints (right) on the square $[-3, 3]^2$:

Please derive your solution using a Lagrange multiplier, and denote this multiplier by λ. We are asking for the points at which the local extrema occur, not for the function values at these points. [9 marks]

(b) Suppose we have a numerical routine to evaluate the exponential function $\exp(x)$. How can we compute the function $\ln(a)$, for a given $a \in \mathbb{R}_+$, using Newton's method?

i. Derive a function $f(x, a)$ to which Newton's method can be applied to find x such that $x = \ln(a)$.

ii. Specify the update equation $x_{n+1} = \ldots$ in Newton's algorithm for this problem.

[6 marks]

BONUS QUESTIONS: you must attempt the questions above before answering those below.

7. **[Bonus: 20 marks] Eigenvalues as solutions of an optimization problem**. Let A be a symmetric $n \times n$-matrix, and define

$$q_A(x) := x^\mathsf{T} A x \quad \text{and} \quad R_A(x) := \frac{x^\mathsf{T} A x}{x^\mathsf{T} x} = \frac{q_A(x)}{\|x\|^2} \quad \text{for } x \in \mathbb{R}^n \, .$$

We have already encountered the quadratic form q_A in class. The purpose of this problem is to verify the following fact:
If A is a symmetric $n \times n$-matrix, the optimization problem

$$x^* := \operatorname*{argmax}_{x \in \mathbb{R}^n} R_A(x)$$

has a solution, $R_A(x^*)$ is the largest eigenvalue of A, and x^* is a corresponding eigenvector.

This result is very useful in machine learning, where we are often interested in the largest eigenvalue specifically—it allows us to compute the largest eigenvalue without computing the entire spectrum, and it replaces an algebraic characterization (the eigenvalue equation) by an optimization problem. We will assume as known that the function q_A is continuous.

(a) Use the extreme value theorem of calculus (recall: a continous function on a compact domain attains its maximum and minimum) to show that $\sup_{x \in \mathbb{R}^n} R_A(x)$ is attained.
Hint: Since $\mathbb{R}^n$ is not compact, transform the supremum over $\mathbb{R}^n$ into an equivalent supremum over the unit sphere $S = \{x \in \mathbb{R}^n | \, \|x\| = 1\}$. The set S is compact (which you can assume as known). [6 marks]

(b) Let $\lambda_1 \geq \ldots \geq \lambda_n$ be the eigenvalues of A enumerated by decreasing size, and $\xi_1, \ldots, \xi_n$ corresponding eigenvectors that form an ONB. Recall from class that we can represent any vector $x \in \mathbb{R}^n$ as

$$x = \sum_{i=1}^{n} (\xi_i^\mathsf{T} x)\xi_i \ .$$

Show that $R_A(x) \leq \lambda_1$. [9 marks]

Since clearly $R_A(\xi_1) = \lambda_1$, we have in fact shown the existence of the maximum twice, using two different arguments! In summary, we now know the maximum exists, and that ξ_1 attains it. What we still have to show is that any vector in S that is *not* an eigenvector for λ_1 does not maximize R_A.

(c) Recall that there may be several linearly independent eigenvectors that all have eigenvalue λ_1. Let these be $\xi_1, \ldots, \xi_k$, for some $k \leq n$. Show that, if $x \in \mathbb{R}^n$ is not contained in $\mathrm{span}\{\xi_1, \ldots, \xi_k\}$, then $R_A(x) < \lambda_1$. [5 marks]