Homework 1

Probabilistic and Bayesian Machine Learning

Yee Whye Teh, Gatsby Computational Neuroscience Unit, University College London

1. **[30 points] Successful Applications**. Find a successful, interesting, and/or cool application of probabilistic modelling techniques that has been reported in the literature. You may find these in good venues for machine learning or statistics. Journal venues include: Journal of Machine Learning Research, Machine Learning Journal, Journal of the American Statistical Association, Journal of the Royal Statistical Society B, Statistical Science, Bayesian Analysis. Conference venues include: NIPS, ICML, UAI, ECML, AISTATS, ISBA.

   (a) What problem did authors address with probabilistic models? Why did you find this application successful/interesting/cool?

   (b) What are the observed variables, latent variables, and the joint distribution?

   (c) Can you draw the joint distribution as a graphical model? What is it?

   (d) Did the authors attempt to estimate or learn the parameters? How is learning or estimation achieved?

   (e) Were there approximations involved in inference or learning, and if so, what?

2. **[20 points] Gaussian 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 $2X^2$?

   (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 $x_1/3 + 2/3x_2$?

   (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?

   (d) Show that a Gaussian distribution is a conjugate prior for μ (if σ^2 is kept fixed).

   (e) Show that a gamma distribution is a conjugate prior for σ^{-2} (if μ is kept fixed). The induced distribution on σ^2 is called an inverse gamma distribution.

   (f) Show that a prior which assumes that μ and σ^2 are independent, with μ being Gaussian and σ^2 being inverse gamma, is **not** a conjugate prior.

3. **[20 points] Maximum entropy and exponential families**.

   Suppose we have a data set $x_1, \dots, x_N$ of iid samples from some distribution. Assume that we have a number of functions $f_1(x), \dots, f_K(x)$ which summarize what we believe are important about the distribution, and we collect the statistics

$$\hat{s}_k = \frac{1}{N} \sum_{i=1}^{N} f_k(x_i) \quad \text{for each } k = 1, \dots, K.$$

For example, if $f_1(x) = x$ then the statistics is the empirical mean, and if $f_2(x) = x^2$ then the corresponding statistics is the empirical spread of the data set around 0. The maximum entropy (MaxEnt) approach says that if these statistics are what is important about the data set, then we should not make any further assumption about the distribution, so should model the distribution as the maximum entropy distribution subject to constraints on the statistics:

$$p^{\text{MaxEnt}} = \underset{p(x)}{\operatorname{argmax}}\, H[p(x)] \text{ subject to } E_p[f_k(x)] = \hat{s}_k \text{ for each } k$$

Show that the maximum entropy distribution is a member of the exponential family of distributions with sufficient statistics functions $\mathsf{T}(x) = [f_1(x), \ldots, f_K(x)]^\top$.

4. **[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 $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ and each image has D pixels, where D is (number of rows × 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 $d \in \{1, \ldots, D\}$.

   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.

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

   (b) 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 $\mathbf{p}$.

   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).

   (c) Write code to learn the ML parameters of a multivariate Bernoulli from this data set and display these parameters as an 8×8 image.

   (d) 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. **[40 points] Bonus: EM for mixture of multivariate Bernoullis.**

   (a) Write down the likelihood for a model consisting of a mixture of K multivariate Bernoulli distributions. Use the parameters $\pi_1, \ldots, \pi_K$ to denote the mixing proportions ($0 \leq \pi_k \leq 1; \sum_k \pi_k = 1$) and arrange the K Bernoulli parameter vectors into a matrix P with elements p_{kd} denoting the probability that pixel d takes value 1 under mixture component k.

(b) Just like in a mixture of Gaussians we can think of this model as a latent variable model, with a discrete hidden variable $s^{(n)} \in \{1, \ldots, K\}$ where $P(s^{(n)} = k|\mathbf{pi}) = \pi_k$. Derive the E and M steps of the EM algorithm for a mixture of K multivariate Bernoullis.

(c) Implement the EM algorithm for a mixture of K multivariate Bernoullis. The algorithm should take as input K, a matrix X containing the data set, and a number of iterations. The algorithm should run for that number of iterations or until the log likelihood converges (does not increase by more than a very small amount). Beware of numerical problems as likelihoods can get very small, it is better to deal with log likelihoods. Also be careful with numerical problems when computing responsibilities — it might be necessary to multiply the top and bottom of the equation for responsibilities by some constant to avoid problems. Hand in code and a high level explanation of what you algorithm does.

(d) Run yor algorithm on the data set for varying $K = 2, 3, 4$. Verify that the log likelihood increases at each step of EM. Report the log likelihoods obtained (measured in *bits*) and display the parameters found as 8×8 images.

(e) Comment on how well the algorithm works, whether it finds good clusters (look at the responsibilities and try to interpret them), and how you might improve the model.