EECS 545 – Machine Learning - Homework #5

Daniel LeJeune & Benjamin Bray Due: 11:00pm 04/04/2016

Homework Policy: Working in groups is fine, but each member must submit their own writeup. Please write the members of your group on your solutions. There is no strict limit to the size of the group but we may find it a bit suspicious if there are more than 4 to a team. **For coding problems, please include your code and report your results (values, plots, etc.)** in your PDF submission. You will lose points if your experimental results are only accessible through rerunning your code. Homework will be submitted via Gradescope (https://gradescope.com/).

1) **Forwards vs. Reverse KL Divergence, (20 pts).**

Consider a factored approximation $q(x, y) = q_1(x)q_2(y)$ to a joint distribution $p(x, y)$.

(a) Show that to minimize the forwards divergence $D_{KL}(p||q)$, we should set $q_1(x) = p(x)$ and $q_2(y) = p(y)$, that is, the optimal approximation is a product of marginals.

(b) Now consider the joint distribution shown in the table. Show that the reverse divergence $D_{KL}(q||p)$ has three distinct minima. Identify those minima and evaluate $D_{KL}(q||p)$ at each.

	x_1	x_2	x_3	x_4
y_1	1/8	1/8		
y_2	1/8	1/8		
y_3			1/4	
y_4				1/4

Table 1: Joint probability table for $p(x, y)$.

(c) What is the value of $D_{KL}(q||p)$ if we set $q(x, y) = p(x)p(y)$ using the joint distribution in Table 1?

2) **Gibbs Sampling from a 2D Gaussian, (20 pts).**

Suppose $X \sim \mathcal{N}(\mu, \Sigma)$, where

$$\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$$

(a) Derive the full conditionals $p(x_1|x_2)$ and $p(x_2|x_1)$. You should not use already existing results on conditional probabilities for jointly Gaussian random variables.

(b) Implement a Gibbs sampling algorithm for estimating $p(x_1, x_2)$ using the conditionals determined in part **(a)**.

Implementation details:

- Start with $x_1 = 0$, then sample x_2 conditioned on the current value of x_1. Then sample x_1 conditioned on the current value of x_2, and so on.
- Store the values of x_1 and x_2 as you go, because you will plot histograms later.
- Sample points in this manner until you have 5000 points for each x_1 and x_2.

Deliverables:

- Plots of the one-dimensional marginals $p(x_1)$ and $p(x_2)$ as histograms. Superimpose a plot of the exact (true) marginals on each.
- Please submit your code, as usual.

3) **Hidden Markov Models, (20 pts).**

Consider the following HMM with 3 states (0,1,2) and binary observations (0,1):

$$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.2 & 0.4 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \qquad \phi = \begin{bmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \\ 0.5 & 0.5 \end{bmatrix} \qquad \pi_0 = \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix}$$

A is the transition matrix, where A_{ij} (using 0-based indexing) is the probability of transition from state i to state j. ϕ is the observation matrix consisting of emission probabilities, i.e. ϕ_{ij} is the probability of seeing observation j at state i. π_0 specifies the initial distribution over states.

(a) Given the sequence of observations 0101, compute the posterior distribution over the sequence of states and report the 3 most probable sequences. Fill out the table below. In the table, the prior probability means the unconditional probability of the state sequence in the row. The likelihood probability means the conditional probability of the observation sequence given the state sequence. The posterior probability means the conditional probability of the state sequence in the row given the sequence of observations. We suggest writing code for this, rather than evaluating all of these things by hand, and please turn in your code.

Most Probable State Sequences	Prior Probability	Likelihood	Posterior Probability

(b) Sample 5000 observation sequences of length 4 from the HMM (or see `hw5p3.py`). Then, treat the first N sequences as training data and learn the HMM parameters by the Baum–Welch algorithm (Please refer to the exercise 13.12 in Bishop's book for the E step and M step details).

Implementation details:

- Run the experiment for $N = 500, 1000, 2000, 5000$.
- Initialize your parameter estimates by sampling uniformly from $[0, 1]$ and then scaling them so that your distributions meet the summation constraints.
- Run EM for 50 iterations and after each iteration, compute the unconditional distributions over all possible observation sequences of length 4 given by the current parameters and compare to the distribution given by the true parameters.
- Use the same initial values of parameters for different N values in the EM algorithm.

Deliverables:

- A plot of the distance defined below between the distributions as a function of the number of iterations [1]. Draw the curves for $N = 500, 1000, 2000, 5000$ on the same figure.
- Please submit your code, as usual.

4) **Kaggle Challenge, (20 pts).**

You can use any algorithm and design any features to remove the noise from handwritten digits from the MNIST dataset. Please refer to `https://inclass.kaggle.com/c/handwritten-digit-denoising` for details. This problem will be graded separately based on your performance on the public leaderboard.

5) **Independent Component Analysis, (20 pts).**

Consider the scenario where we observe a random vector $\mathbf{x} \in \mathbb{R}^d$ generated according to

$$\mathbf{x} = A\boldsymbol{s}$$

where $A \in \mathbb{R}^{d\times d}$ is unknown, and $\boldsymbol{s} \in \mathbb{R}^d$ is a vector of *independent* random variables, each of unknown distribution. Given *iid* observances $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$, we would like to recover $(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)})$. Collect the observations $\mathbf{x}^{(i)}$ into a matrix $X \in \mathbb{R}^{d\times N}$, and collect $\boldsymbol{s}^{(i)}$ into a matrix $S \in \mathbb{R}^{d\times N}$. Now our model is

$$X = AS$$

Our goal in this problem is to obtain a matrix $Y \in \mathbb{R}^{d\times N}$ satisfying

$$Y = WX$$

for some $W \in \mathbb{R}^{d\times d}$, such that, if viewed as a collection of *iid* observations $(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(N)})$ of a random variable $\mathbf{y}$, maximizes the pairwise independence of the elements of $\mathbf{y}$. Our hope is that the recovered Y is similar to S; however, because A is unknown, it is impossible to recover the scale or permutations of the rows of S. We will consider all random variables to have zero mean in this problem. This problem is known as independent component analysis (ICA).

(a) If the elements of $\mathbf{y}$ are independent, then they are uncorrelated, so we can restrict our solution to those with diagonal covariance. Further, since we can't recover the scale of the rows of S, we can restrict each row of Y to have unit variance. Combining these two, we restrict our set of possible solutions to the set of matrices with identity covariance; i.e.,

$$\frac{1}{N}YY^T = I$$

Data with this property is said to be *white*. If X is also white, then any solution to $Y = WX$ must have W as an orthogonal matrix, which reduces the search space of our problem significanly. To that end, we define $\tilde{X} = DX$ for some $D \in \mathbb{R}^{d\times d}$, such that $\tilde{X}$ is white, and then solve the new version of the problem, where $Y = W\tilde{X}$.

[1] Given two distributions μ, μ' over a finite set X, the distance is defined as

$$\delta(\mu, \mu') = \frac{1}{2}\sum_{x\in X} |\mu(x) - \mu'(x)|$$

Provide such a whitening transformation matrix D that will whiten X. You must be able to construct D using X.

(b) There are several ways to quantify how independent the elements of $\mathbf{y}$ are (negentropy and mutual information, for example), but it turns out that they are all equivalent to measuring the non-Gaussianity of $\mathbf{y}$.[2] So, to maximize independence between the elements of $\mathbf{y}$, we can maximize the non-Gaussianity of the elements of $\mathbf{y}$. We will use as our measure of non-Gaussianity the following function:

$$J(y) = (\mathbb{E}[G(y)] - \gamma)^2$$

where $\gamma \triangleq \mathbb{E}[G(\nu)]$ for a Gaussian random variable ν with zero mean and unit variance, and $G(y)$ is a well-chosen [non-quadratic] function. Letting $\mathbf{w}_k$ denote the k^{th} row of W, our total non-Gaussianity is

$$J(\mathbf{y}) = \sum_{k=1}^{d}(\mathbb{E}[G(y_k)] - \gamma)^2 = \sum_{k=1}^{d}(\mathbb{E}[G(\mathbf{w}_k^T\mathbf{x})] - \gamma)^2$$

Let's try this out. Generate data as follows (or see `hw5p5.py`):

- Use $d = 2$, $N = 10,000$.
- Let $s_1^{(i)} = \sin(i/200)$ (a sinusoidal wave).
- Let $s_2^{(i)} = \text{remainder}(i/200, 2) - 1$ (a sawtooth wave).
- Let $A = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$
- Compute $X = AS$.

Obtain the matrix Y with maximal independence using the total measure of non-Gaussianity above. Do this by finding the orthogonal matrix W that achieves the maximum. Since W is a 2×2 orthogonal matrix, let

$$W(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Implementation details:

- Whiten X using your whitening matrix D from part **(a)** first.
- Use $G(y) = \log\cosh y$.
- Estimate the values of all expectations using empirical means.
- Estimate γ with the empirical mean of $G(\cdot)$ applied to 10^6 random standard normal values.
- Use a grid search on $\theta \in [0, \pi/2]$ to select the optimal θ.

Deliverables:

- A plot of your estimate of $J(\mathbf{y})$ versus θ for $\theta \in [0, \pi/2]$.
- A plot of each row of the recovered Y, preferably in the same plot but not overlapping.
- Please submit your code, as usual.

[2]For more details, see Hyvärinen, a., & Oja, E. (2000). Independent component analysis: Algorithms and applications. Neural Networks, 13(4-5), 411-430. http://doi.org/10.1016/S0893-6080(00)00026-5