1 Problem 1 (20 Pts)

1.1 Assume you are given a sample of n discrete observations \(x_1, \ldots, x_n\) which are drawn i.i.d from poisson distribution:

\[
P(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 1, 2, \ldots
\]

(a) Determine the maximum likelihood estimator of the Poisson parameter \(\lambda\).

(b) What is the conjugate prior distribution for the poisson distribution with unknown parameter \(\lambda\)? Calculate the posterior distribution of \(\lambda\) based on that prior distribution and using the observations.

1.2 Now, Consider the Gaussian distribution \(\mathcal{N}(x|\mu, \sigma^2)\) and suppose that both the mean \(\mu\) and the variance \(\sigma^2\) are unknown.

(a) Show that the conjugate prior for this distribution is a normal inverse gamma distribution:

\[
P(\mu, \sigma^2|\mu_0, V, a_0, b_0) = \mathcal{N}(\mu|\mu_0, V \sigma^2) InvGam(\sigma^2|a_0, b_0)
\]

Where the inverse gamma distribution is defined as:

\[
InvGam(x|a, b) = \frac{1}{\Gamma(a)} b^a x^{(-a-1)} \exp(-b/x)
\]
2

b) Let $D = \{x_1, \ldots, x_N\}$ be a set of i.i.d. samples. Plot the graphical model corresponding to the above probabilistic model for the data.

c) When the mean $\mu$ of the Gaussian is assumed known, the conjugate prior on $\sigma^2$ will be $\text{InvGam}(\sigma^2 | a_0, b_0)$. In this case, show that the predictive distribution $P(x|D)$ will be a Students t-distribution:

$$St(x|\mu, \lambda, v) = \frac{\Gamma(v/2 + 1/2)}{\Gamma(v/2)} \left( \frac{\lambda}{\pi v} \right)^{1/2} [1 + \frac{\lambda(x - \mu)^2}{v}]^{-(v/2+1/2)}$$ (4)

2. Problem 2 (15 Pts)

Exponential family includes the probability distributions that can be formulated as:

$$P(x|\eta) = h(x) \exp \{\eta^T T(x) - \ln Z(\eta)\}$$ (5)

$$Z(\eta) = \int h(x) \exp \{\eta^T T(x)\} dx$$ (6)

where $\eta$ denotes the canonical parameters, $T(x)$ shows the sufficient statistic function, and $Z(\eta)$ is the partition function.

2.1 Determine which of the following distributions are in the exponential family and also find $h(x)$, $T(x)$, and $\ln Z(\eta)$ functions for them.

a) Multivariate gaussian $\mathcal{N}(\mu, \Sigma)$

b) Dirichlet $\text{Dir}(\alpha_1, \ldots, \alpha_K)$

c) Boltzmann distribution: an undirected graphical model $G = (V, E)$ where each node is a binary random variable taking values $\{0, 1\}$ and $P(x) \propto \exp \{\sum_{i \in V} u_i x_i + \sum_{(i,j) \in E} w_{ij} x_i x_j\}$

2.2 Maximum Likelihood is one way to set parameters of a distribution. Another way of estimating the parameters of a distribution is to match the moments of the distribution to the empirical moments (moment matching algorithm). For some distributions such as Gaussian distribution, moment matching algorithm corresponds to maximum likelihood, but, this is not generally true for most of distributions.

Assume we have some i.i.d data with empirical mean $M$ and variance $S$. Show that to fit a Beta distribution $\text{Beta}(x|a, b)$ by moment matching algorithm, we have

$$a = \frac{M(M^2 - S)}{S}, \quad b = a - \frac{1 - M}{M}$$ (7)

Does this correspond to maximum likelihood?
3 Problem 3 (20 Pts)

In this problem, you are going to learn an undirected graphical model of $N$ binary random variables $[x_i \in \{0, 1\}]_{i=1}^N$, we define the joint distribution of these variables as

$$P(X|\theta) \propto \exp\left(\sum_{i \in V} \theta_i x_i + \sum_{(i,j) \in E} \theta_{ij} x_i x_j\right)$$

where $\theta = [\theta_1, ..., \theta_N, \theta_{ij} ((i,j) \in V)]$ are the set of free parameters of the model and $V$ and $E$ are the set of graph nodes and graph edges respectively. The dataset for this problem consists of voting records from senators during the 112th United States Congress. We provided data in the binary matrix $\text{senatorVotes}$ which consists of $N = 13$ rows (senators) and $L = 486$ columns (bills). The $ij$-th element of this matrix is the vote of senator $i$ on bill $j$. For learning the joint distribution on Senate votes, we interpret the bills as $L$ independent samples.

3.1 Assume a fully connected pairwise MRF, for which $E$ contains an edge linking every pair of nodes. Write a function $\text{MLFullMRF.m}$ which gets the matrix $\text{senatorVotes}$ and learns the parameters of the MRF based on the ML algorithm. Plot the log-likelihood of the estimated model after each optimization iteration (Initialize the node parameters $\theta_i$ to $\frac{1}{|V|}$ and the edge parameters $\theta_{ij} = 0$).

3.2 Now, assume we place a laplacian prior distribution on each of the parameters as in Eq. 9. Write a function $\text{MAPFullMRF.m}$ which gets the matrix $\text{senatorVotes}$ and a scalar parameter $\lambda$ and returns maximum a posteriori (MAP) estimate of $\theta$.

$$P(\theta|\lambda) = \prod_{i \in V} \text{Lap}(\theta_i|\lambda) \prod_{(i,j) \in E} \text{Lap}(\theta_{ij}|\lambda)$$

$$\text{Lap}(\theta|\lambda) = \frac{\lambda}{2} \exp(-\lambda|\theta|)$$

4 Problem 4 (25 Pts)

Suppose we want to perform dimensionality reduction for a set of data points $X = [x_1, ..., x_N]_{d \times N}$. There are lots of algorithms for such reduction among which, the PCA is the most well known algorithm. The problem with PCA algorithm is that it is limited by its global linearity. A general solution is to use a combination of local linear PCA projections. In this problem, you are going to implement the Mixture of Probabilistic PCA (MPPCA) for image compression applications.

The generative model of the MPPCA with $K$ component goes as follows:

- Draw $c_i \sim \text{MultiNomial}(\Pi)$, for $i = 1, ..., N$.
- Draw $z_i \sim \mathcal{N}(0, I_d)$, for $i = 1, ..., N$. 

Homework 3
• Draw $x_i \sim \mathcal{N}(W_{c_i} z_i + \mu_{c_i}, \sigma_{c_i}^2)$, for $i = 1, \ldots, N$.

where $\Pi = \{\pi_1, \ldots, \pi_K\}$ are the mixing proportions such that $\sum_{k=1}^K \pi_k = 1$, $\pi_k > 0$, $c_i \in \{1, \ldots, K\}$ is the component assignment variable for data point $x_i$, the parameter $W_k (k = 1, \ldots, K)$ is called the factor loading parameter and $\mu_k (k = 1, \ldots, K)$ permits the data model to have non-zero mean.

4.1 Draw the graphical representation of this model using plate notation.

4.2 Derive the $E$ step and the $M$ step for learning the parameters $\theta = (\{W_k, \mu_k, \pi_k, \sigma_{k_i}^2\}_{k=1}^K)$ using EM algorithm:

$E$ step: $q(c_1, \ldots, c_N, z_1, \ldots, z_N) = P(c_1, \ldots, c_N, z_1, \ldots, z_N | x_1, \ldots, x_N, \theta^t)$

$M$ step: $\theta^{t+1} = \arg \max_{\theta} \sum q(c_1, \ldots, c_N, z_1, \ldots, z_N) \log P(x_1, \ldots, x_N, c_1, \ldots, c_N, z_1, \ldots, z_N | \theta)$

we consider an application of the MPPCA to image coding. Figure 1 demonstrates a $400 \times 800$ grayscale image. The image is segmented into $8 \times 8$ non-overlapping blocks, giving a total dataset of 5000 64-dimensional vectors. You must use half of this data, corresponding to the left half of the picture as training data. You will use the right half for testing.

4.3 Implement MPPCA using the above setup and based on the derived $E$ step and the $M$ step of section 4.2.

4.4 Apply your code on the provided image data. More precisely, consider the block data of the left half of the picture for learning the parameters (consider $K = 12$ number of local PCA with dimensionality 4), after the model likelihood had been maximized, perform image coding by allocating data of the right half of the picture to the component with lowest reconstruction error. Finally, reconstruct the image using the coded components and compare your results with the reconstructed image using single PCA algorithm.

5 Problem 5 (20 pts)

Tree Augmented Naive Bayes (TAN):

TAN models are formed by adding directional edges between features $X_1, \ldots, X_D$ in the naive Bayes classifier where these directional edges impose a tree structure (with no v-structures) on the features.

5.1 Use the Chow-Liu algorithm to learn the tree structure for the features and also find all the parameters of the TAN model (i.e., $\theta_Y, \theta_{X_i|Y} (i = 1, \ldots, D), \theta_{X_i|X_j} (X_i, X_j) \in T$) using MLE on the training data.

Hint: You can use the MWST matlab function to find a maximum weight spanning tree of a weighted graph.
5.2 Based on the above structure and parameters, classify the test data and report the classification accuracy on the test data. For the classification purpose, you must perform the MAP inference to find \( \text{argmax}_y P(y|x, \theta) \).

5.3 Find also the accuracy of the naïve Bayes classifier on the test data and compare the obtained results.

In the training and test data files the first column shows the label of the data.