CPSC 540 Assignment 3 (due February 21 at midnight)

The assignment instructions are the same as for the previous assignment.

1. Name(s):
2. Student ID(s):

1 Discrete and Gaussian Variables

1.1 MLE for General Discrete Distribution

Consider a density estimation task, where we have two variables ($d = 2$) that can each take one of k discrete values. For example, we could have

$$X = \begin{bmatrix} 1 & 3 \\ 4 & 2 \\ \text{k} & 3 \\ 1 & \text{k-1} \end{bmatrix}.$$

The likelihood for example x^i under a general discrete distribution would be

$$p(x_1^i, x_2^i|\Theta) = \theta_{x_1^i, x_2^i},$$

where θ_{c_1,c_2} gives the probability of x_1 being in state c_1 and x_2 being in state c_2, for all the k^2 combinations of the two variables. In order for this to define a valid probability, we need all elements θ_{c_1,c_2} to be non-negative and they must sum to one, $\sum_{c_1=1}^{k}\sum_{c_2=1}^{k}\theta_{c_1,c_2} = 1$.

1. Given n training examples, derive the MLE for the k^2 elements of Θ.

   Answer: The log-likelihood is given by

   $$-\log p(X|\Theta) = -\sum_{i=1}^{n}\sum_{c\in[k]^2}\mathcal{I}[x^i = c]\log\theta_c + \text{const.}$$

   By re-arranging the sums and ignoring the constant term the Lagrangian is given by

   $$L(\Theta, z) = -\sum_{c\in[k]^2}\log\theta_c\sum_{i=1}^{n}\mathcal{I}[x^i = c] + z(\sum_{c\in[k]^2}\theta_c - 1).$$

   We have

   $$\nabla_{\theta_c}L(\Theta, z) = -\frac{1}{\theta_c}\sum_{i=1}^{n}\mathcal{I}[x^i = c] + z,$$

   and setting this equal to 0 we get

   $$\theta_c = \frac{\sum_{i=1}^{n}\mathcal{I}[x^i = c]}{z}.$$

From the sum-to-one constraint we have

$$1 = \sum_{c \in [k]^2} \theta_c = \sum_{c \in [k]^2} \sum_{i=1}^{n} \frac{\mathcal{I}[x^i = c]}{z},$$

or that

$$z = \sum_{i=1}^{n} \sum_{c \in [k]^2} \mathcal{I}[x^i = c] = \sum_{i=1}^{n} 1 = n.$$

This gives

$$\theta_c = \frac{N_c}{n},$$

where N_c is the number of times we have $x^i = (c_1, c_2)$.

2. Because of the sum-to-1 constraint, there are only (k^2-1) degrees of freedom in the discrete distribution, and not k^2. Derive the MLE for this distribution assuming that

$$\theta_{k,k} = 1 - \sum_{c_1=1}^{k} \sum_{c_2=1}^{k} \mathcal{I}[c_1 \neq k, c_2 \neq k] \theta_{c_1,c_2},$$

so that the distribution only has $(k^2 - 1)$ parameters.

Answer: The Lagrangian under this constraint is the same, and all the other steps are the same so we get the same MLE.

3. If we had separate parameter θ_{c_1} and θ_{c_2} for each variables, a reasonable choice of a prior would be a product of Dirichlet distributions,

$$p(\theta_{c_1}, \theta_{c_2}) \propto \theta_{c_1}^{\alpha_{c_1}-1} \theta_{c_2}^{\alpha_{c_2}-1}.$$

For the general discrete distribution, a prior encoding the same assumptions would be

$$p(\theta_{c_1,c_2}) \propto \theta_{c_1,c_2}^{\alpha_{c_1}+\alpha_{c_2}-2}.$$

Derive the MAP estimate under this prior (assuming we use k^2 variables to parameterize Θ)

Answer: For the MAP estimate we'll replace $\sum_{i=1}^{n} \mathcal{I}[x^i = c]$ in the Lagrangian by

$$\left[\sum_{i=1}^{n} \mathcal{I}[x^i = c]\right] + \alpha_{c_1} + \alpha_{c_2} - 2.$$

Following the steps above we obtain

$$\theta_c = \frac{N_c + \alpha_{c_1} + \alpha_{c_2} - 2}{z} = \frac{N_c + \alpha_{c_1} + \alpha_{c_2} - 2}{n + \sum_{c \in [k]^2} [\alpha_{c_1} + \alpha_{c_2} - 2]}$$

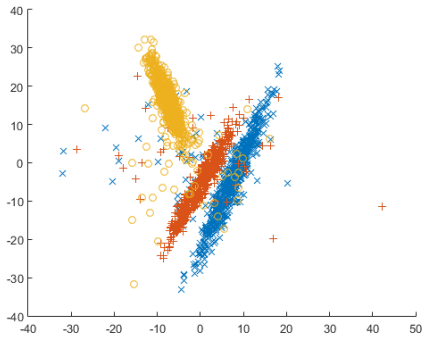
Hint: it is convenient to write the likelihood for an example i in the form

$$p(x^i | \Theta) = \prod_{c \in [k]^2} \theta_c^{\mathcal{I}[x^i = c]},$$

where c is a vector containing (c_1, c_2), $[x^i = c]$ evaluates to 1 if all elements are equal, and $[k]^2$ is all ordered pairs (c_1, c_2). You can use the Lagrangian to enforce the sum-to-1 constraint on the log-likelihood, and you may find it convenient to define $N_c = \sum_{i=1}^{n} \mathcal{I}[x^i = c]$.

1.2 Generative Classifiers with Gaussian Assumption

Consider the 3-class classification dataset in this image:



In this dataset, we have 2 features and each colour represents one of the classes. Note that the classes are highly-structured: the colours each roughly follow a Gausian distribution plus some noisy samples.

Since we have an idea of what the features look like for each class, we might consider classifying inputs x using a *generative classifier*. In particular, we are going to use Bayes rule to write

$$p(y^i = c|x^i, \Theta) = \frac{p(x^i|y^i = c, \Theta) \cdot p(y^i = c|\Theta)}{p(x^i|\Theta)},$$

where Θ represents the parameters of our model. To classify a new example $\tilde{x}^i$, generative classifiers would use

$$\hat{y}^i = \underset{y \in \{1,2,\dots,k\}}{\arg\max}\ p(\tilde{x}^i|y^i = c, \Theta)p(y^i = c|\Theta),$$

where in our case the total number of classes k is 3.[1] Modeling $p(y^i = c|\Theta)$ is easy: we can just use a k-state categorical distribution,

$$p(y^i = c|\Theta) = \theta_c,$$

where θ_c is a single parameter for class c. The maximum likelihood estimate of θ_c is given by n_c/n, the number of times we have $y^i = c$ (which we've called n_c) divided by the total number of data points n.

Modeling $p(x^i|y^i = c, \Theta)$ is the hard part: we need to know the *probability of seeing the feature vector x^i given that we are in class c*. This corresponds to solving a density estimation problem for each of the k possible classes. To make the density estimation problem tractable, we'll assume that the distribution of x^i given that $y^i = c$ is given by a $\mathcal{N}(\mu_c, \Sigma_c)$ Gaussian distribution for a class-specific μ_c and Σ_c,

$$p(x^i|y^i = c, \Theta) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_c|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^i - \mu_c)^T\Sigma_c^{-1}(x^i - \mu_c)\right).$$

Since we are distinguishing between the probability under k different Gaussians to make our classification, this is called *Gaussian discriminant analysis* (GDA). In the special case where we have a constant $\Sigma_c = \Sigma$ across all classes it is known as *linear discriminant analysis* (LDA) since it leads to a linear classifier between any two classes (while the region of space assigned to each class forms a convex polyhedron as in k-means clustering and softmax classification). Another common restriction on the Σ_c is that they are diagonal matrices, since this only requires $O(d)$ parameters instead of $O(d^2)$ (corresponding to assuming that the features are independent univariate Gaussians given the class label). Given a dataset $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^n$,

[1] The denominator $p(\tilde{x}^i|\Theta)$ is irrelevant to the classification since it is the same for all y.

where $x^i \in \mathbb{R}^d$ and $y^i \in \{1, \ldots, k\}$, the maximum likelihood estimate (MLE) for the μ_c and Σ_c in the GDA model is the solution to

$$\arg\max_{\mu_1,\mu_2,\ldots,\mu_k,\Sigma_1,\Sigma_2,\ldots,\Sigma_k} \prod_{i=1}^{n} p(x^i|y^i, \mu_{y^i}, \Sigma_{y^i}).$$

This means that the negative log-likelihood will be equal to

$$\begin{aligned} -\log p(X|y,\Theta) &= -\sum_{i=1}^{n} \log p(x^i|y^i|\mu_{y^i}, \Sigma_{y^i}) \\ &= \sum_{i=1}^{n} \frac{1}{2}(x^i - \mu_{y^i})^T \Sigma_{y^i}^{-1}(x^i - \mu_{y^i}) + \frac{1}{2}\sum_{i=1}^{n} \log|\Sigma_{y^i}| + \text{const.} \end{aligned}$$

1. Derive the MLE for the GDA model under the assumption of *common diagonal covariance* matrices, $\Sigma_c = D$ (d parameters). (Each class will have its own mean μ_c.)

   Answer: First let's derive the mean μ_c for class c. If we take the gradient with respect to a particular μ_c and set it to 0 then all terms outside of y_c are zero and we get

   $$\begin{aligned} 0 &= -\sum_{i \in y_c} \Sigma_c(x^i - \mu_c) \\ \Sigma_c^{-1}\mu_c \sum_{i \in y_c} 1 &= \Sigma_c^{-1} \sum_{i \in y_c} x^i. \end{aligned}$$

   Pre-multiplying by Σ_c we get

   $$\mu_c = \frac{\sum_{i \in y_c} x^i}{n_c},$$

   which (as expected) is the mean of the examples in class c. With Σ_c set to a diagonal D for all classes c, the log-likelihood simplifies to

   $$\begin{aligned} -\log p(X|y,\Theta) + \text{const.} &= \sum_{i=1}^{n} \frac{1}{2}(x^i - \mu_{y^i})^T D^{-1}(x^i - \mu_{y^i}) + \frac{1}{2}\sum_{i=1}^{n} \log|D| \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}(x_j^i - (\mu_{y_i})_j)^2/D_{jj} + \frac{n}{2}\log(\prod_{j=1}^{d} D_{jj}) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}(x_j^i - (\mu_{y_i})_j)^2/D_{jj} + \frac{n}{2}\sum_{j=1}^{d}\log(D_{jj}) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}(x_j^i - (\mu_{y_i})_j)^2\Lambda_{jj} - \frac{n}{2}\sum_{j=1}^{d}\log(\Lambda_{jj}) \end{aligned}$$

   where we've that the determinant of a diagonal matrix is the product of the diagonal elements, and we've re-parameterized in terms of the inverse Λ. This function is separable in the Λ_{jj} so we can solve it independently for each of the Λ_{jj}. Taking the derivative with respect to Λ_{jj} and equating it with zero we get

   $$0 = \sum_{i=1}^{n} \frac{1}{2}(x_j^i - (\mu_{y_i})_j)^2 - \frac{n}{2\Lambda_{jj}}$$

or

$$\frac{1}{\Lambda_{jj}} = \frac{1}{n}\sum_{i=1}^{n}(x_j^i - (\mu_{y_i})_j)^2,$$

which is solved by setting $D_{jj} = \frac{1}{n}\sum_{i=1}^{n}(x_j^i - (\mu_{y_i})_j)^2$, the MLE for the variance along coordinate j around its class-specific mean over all i.

2. Derive the MLE for the GDA model under the assumption of *individual scale-identity* matrices, $\Sigma_c = \sigma_c^2 I$ (k parameters).

   Answer: If this case we have

$$-\log p(X|y,\Theta) + \text{const.} = \sum_{i=1}^{n}\frac{1}{2\sigma_{y^i}^2}(x^i - \mu_{y^i})^T(x^i - \mu_{y^i}) + \frac{1}{2}\sum_{i=1}^{n} d\log(\sigma_{y^i}^2)$$

   Parameterizing in terms of the squared inverse, we an solve for a particular σ_c using

$$0 = \frac{1}{2}\sum_{i\in y_c}\|x^i - \mu_c\|^2 - \frac{n_c}{\sigma_c^2},$$

   implying that $\sigma_c^2 = \frac{1}{n_c d}\sum_{i\in y_c}\|x^i - \mu_c\|^2$, the average variance across variables for examples in the class c.

3. It's painful to derive these from scratch, but you should be able to see a pattern that would allow other common restrictions. Without deriving the result from scratch (hopefully), give the MLE for the case of *individual full* covariance matrices, Σ_c ($O(kd^2)$ parameters).

   Answer: From class we know that the MLE for a Gaussian is $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x^i - \mu)(x^i - \mu)^T$. Thus if we worked out the general case, we would see that we need to do this for each class:

$$\Sigma_c = \frac{1}{n_c}\sum_{i\in y_c}(x^i - \mu_c)(x^i - \mu_c)^T.$$

4. When you run *example_generative* it loads a variant of the dataset in the figure that has 12 features and 10 classes. This data has been split up into a training and test set, and the code fits a k-nearest neighbour classifier to the training set then reports the accuracy on the test data (around $\sim 63\%$ test error). The k-nearest neighbour model does poorly here since it doesn't take into account the Gaussian-like structure in feature space for each class label. Write a function *gda* that fits a GDA model to this dataset (using individual full covariance matrices). Hand in the function and report the test set accuracy.

   Answer: The Gaussian generative classifiers improves the performance from 0.633 (with KNN) to 0.369. The training code should look roughly like this:

```
function gda(X,y)

   (n,d) = size(X)
   k = maximum(y)

  nc = zeros(k)
  theta = zeros(k)
  mu = zeros(d,k)
  Sigma = zeros(d,d,k)

  for i in 1:n
    c = y[i]
    nc[c] += 1
    mu[:,c] += X[i,:]
  end
  for c in 1:k
    theta[c] = nc[c]/n
    mu[:,c] ./= nc[c]
  end

  for i in 1:n
    c = y[i]
    xCentered = X[i,:] - mu[:,c]
    Sigma[:,:,c] += xCentered*xCentered'
  end
  for c in 1:k
    Sigma[:,:,c] ./= nc[c]
  end
```

The predict function should look roughly like this:

```
function predict(Xhat)
  (t,d) = size(Xhat)
  yhat = zeros(t,1)

  for i in 1:t
      logp = zeros(k)
      for c in 1:k
          xCentered = Xhat[i,:] - mu[:,c]
          logp[c] = log(theta[c]) - (1/2)*dot(xCentered,Sigma[:,:,c]\xCentered) - (1/2)*logdet(Sigma[:,:,c])
      end
      (~,yhat[i]) = findmax(logp)
  end

  return yhat
end
```

5. In this question we would like to replace the Gaussian distribution of the previous problem with the more robust multivariate-t distribution so that it isn't influenced as much by the noisy data.

Unlike the previous case, we don't have a closed-form solution for the parameters. However, if you run *example_student* it generates random noisy data and fits a multivariate-t model. By using the *studentT* model, write a new function *tda* that implements a generative model that is based on the multivariate-t distribution instead of the Gaussian distribution. Report the test accuracy with this model.

Answer: This robust generative classifiers reduces the error further down to 0.194. The code wasn't asked for, but here is my implementation:

```
theta = zeros(k)
subModel = Array{DensityModel}(k)
for c in 1:k
    @show c
   theta[c] = sum(y.==c)/n
   subModel[c] = studentT(X[y.==c,:])
end
@show(theta)

function predict(Xhat)
      (t,d) = size(Xhat)
      yhat = zeros(t,1)

      PDFs = zeros(t,k)
      for c in 1:k
          PDFs[:,c] = subModel[c].pdf(Xhat)
      end

      for i in 1:t
          logp = zeros(k)
          for c in 1:k
              logp[c] = log(theta[c]) + log(PDFs[i,c])
          end
          (~,yhat[i]) = findmax(logp)
      end

      return yhat
  end

return GenericModel(predict)
```

Hints: you will be able to substantially simplify the notation in parts 1-3 if you use the notation $\sum_{i \in y_c}$ to mean the sum over all values i where $y^i = c$. Similarly, you can use n_c to denote the number of cases where $y_i = c$, so that we have $\sum_{i \in y_c} 1 = n_c$. Note that the determinant of a diagonal matrix is the product of the diagonal entries, and the inverse of a diagonal matrix is a diagonal matrix with the reciprocals of the original matrix along the diagonal.

For part three you can use the result from class regarding the MLE of a general multivariate Gaussian. At test time for GDA, you may find it more numerically reasonable to compare log probabilities rather than probabilities of different classes, and you may find it helpful to use the *logdet* function to compute the log-determinant in a more numerically-stable way than taking the log of the determinant. (Also, don't forget to center at training and test time.)

For the last question, you can define an empty array that can be filled with k *DensityModel* objects using:

```
subModel = Array{DensityModel}(k)
```

1.3 Self-Conjugacy for the Mean Parameter

If x^i is distributed according to a Gaussian with mean μ,

$$x^i \sim \mathcal{N}(\mu, \sigma^2),$$

and we assume that μ itself is distributed according to a Gaussian

$$\mu \sim \mathcal{N}(\alpha, \gamma^2),$$

then the posterior $\mu|x^i$ also follows a Gaussian distribution.[2] Derive the form of the (Gaussian) distribution for $p(\mu|x^i, \alpha, \sigma^2, \gamma^2)$.

Hints: Use Bayes rule and use the $\propto$ sign to get rid of factors that don't depend on μ. You can then "complete the square" to make the product look like a Gaussian distribution. In particular, when you have $\exp(ax^2 - bx + \text{const})$ you can factor out an a and add/subtract $(b/2a)^2$ to re-write it as

$$\begin{aligned}\exp\left(ax^2 - bx + const\right) &\propto \exp\left(ax^2 - bx\right) = \exp\left(a(x^2 - (b/a)x)\right) \\ &\propto \exp\left(a(x^2 - (b/a)x + (b/2a)^2)\right) = \exp\left(a(x - (b/2a))^2\right).\end{aligned}$$

Note that multiplying by factors that do not depend on μ within the exponent does not change the distribution. In this question you will want to complete the square to get the distribution on μ, rather than x^i. You may find it easier to solve thie problem if you parameterize the Gaussians in terms of their "precision" parameters (e.g., $\lambda = 1/\sigma^2$, $\lambda_0 = 1/\gamma^2$) rather than their variances σ^2 and γ^2.

Answer: From the question description (and keeping in mind that factors in the product that don't depend on μ don't change the distribution),

$$\begin{aligned}p(\mu|x, \alpha, \sigma^2, \gamma^2) &\propto p(x|\mu, \sigma^2)p(\mu|\alpha, \gamma^2) \\ &\propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\exp\left(-\frac{(\mu-\alpha)^2}{2\gamma^2}\right) \\ &= \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\alpha)^2}{2\gamma^2}\right)\end{aligned}$$

[2] We say that the Gaussian distribution is the 'conjugate prior' for the Gaussian mean parameter (we'll formally discuss conjugate priors later in the course). Another reason the Gaussian distribution is important is that is the only (non-trivial) continuous distribution that has this "self-conjugacy" property.

We now re-parameterize in terms of the precisions $\lambda = 1/\sigma^2$ and $\lambda_0 = 1/\gamma^2$,

$$\begin{aligned}
p(\mu|x,\alpha,\lambda,\lambda_0) &\propto \exp\left(-\frac{\lambda}{2}(x-\mu)^2 - \frac{\lambda_0}{2}(\mu-\alpha)^2\right) \\
&= \exp\left(-\frac{1}{2}\left[\lambda(x^2 - 2x\mu + \mu^2) + \lambda_0(\mu^2 - 2\mu\alpha + \alpha^2)\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\lambda(-2x\mu + \mu^2) + \lambda_0(\mu^2 - 2\mu\alpha)\right]\right) \\
&= \exp\left(-\frac{1}{2}\left[\mu^2(\lambda+\lambda_0) - 2\mu(x\lambda + \alpha\lambda_0)\right]\right) \\
&= \exp\left(-\frac{(\lambda+\lambda_0)}{2}\left[\mu^2 - 2\mu\frac{(x\lambda+\alpha\lambda_0)}{(\lambda+\lambda_0)}\right]\right) \\
&\propto \exp\left(-\frac{(\lambda+\lambda_0)}{2}\left[\mu^2 - 2\mu\frac{(x+\alpha)}{(\lambda+\lambda_0)} + \frac{(x\lambda+\alpha\lambda_0)^2}{(\lambda+\lambda_0)^2}\right]\right) \\
&= \exp\left(-\frac{(\lambda+\lambda_0)}{2}\left(\mu - \frac{x\lambda+\alpha\lambda_0}{\lambda+\lambda_0}\right)^2\right)
\end{aligned}$$

This is proportional to a Gaussian distribution with a mean of $\frac{x\lambda-\alpha\lambda_0}{\lambda+\lambda_0}$ and a precision of $\lambda+\lambda_0$, so we have

$$\mu|x,\alpha,\lambda,\lambda_0 \sim \mathcal{N}\left(\frac{x\lambda+\alpha\lambda_0}{\lambda+\lambda_0}, (\lambda+\lambda_0)^{-1}\right).$$

Thus, we add the precision values together to get the final precision, and the final mean is the sum of the data point and the prior mean (weighted by their precisions) divided by the new precision. Equivalently, we could write

$$\mu|x,\alpha,\sigma^2,\gamma^2 \sim \mathcal{N}\left(\frac{\frac{x}{\sigma^2}+\frac{\alpha}{\gamma^2}}{\frac{1}{\sigma^2}+\frac{1}{\gamma^2}}, \left(\frac{1}{\sigma^2}+\frac{1}{\gamma^2}\right)^{-1}\right),$$

and $p(\mu|x,\alpha,\sigma^2,\gamma^2)$ is the Gaussian PDF with the above mean and variance. If you assume that $\sigma^2 = \gamma^2 = 1$, this would simplify to

$$\mu|x,\alpha,\sigma^2,\gamma^2 \sim \mathcal{N}\left(\frac{x+\alpha}{2}, \frac{1}{2}\right),$$

2 Mixture Models and Expectation Maximization

2.1 Semi-Supervised Gaussian Discriminant Analysis

Consider fitting a GDA model where some of the y^i values are missing at random. In particular, let's assume we have a set of n labeled examples (x^i, y^i) and another set of t unlabeled examples (x^i). This is a special case of *semi-supervised learning*, and fitting generative models with EM is one of the oldest semi-supervised learning techqniues. When the classes exhibit clear structure in the feature space, it can be very effective even if the number of labeled examples is very small.

1. Derive the EM update for fitting the parameters of a GDA model (with individual full covariance matrices) in the semi-superivsed setting where we have n labeled examples and t unlabeled examples.

Answer: The update for the categorical distribution of the y^i is given by

$$\theta_c^{t+1} = \frac{n_c + \sum_{i=1}^t r_c^i}{n+t},$$

but they don't need to explicitly write that because it's in the EM notes. The update for the Gaussian mean parameters is given by

$$\mu_c = \frac{\sum_{i \in y_c} x^i + \sum_{i=1}^t r_c^i \tilde{x}^i}{n_c + \sum_{i=1}^t r_c^i},$$

and for the covariance matrices is given by

$$\Sigma_c = \frac{\sum_{i \in y_c} (x^i - \mu_c)(x^i - \mu_c)^T + \sum_{i=1}^t r_c^i (\tilde{x}^i - \mu_c)(\tilde{x}^i - \mu_c)^T}{n_c + \sum_{i=1}^t r_c^i}.$$

2. If you run the demo *example_SSL.jl*, it will load a variant of the dataset from the previous question, but where the number of labeled examples is small and a large number of unlabeled examples are available. The demo first fits a KNN model and then a generative Gaussian model (once you are finished Question 1 and assuming that you put your *gda* function in a file named *gda.jl*). Because the number of labeled examples it quite small, the performance is worse than in Question 2. Write a function *generativeGaussianSSL* that fits the generative Gaussian model of the previous question using EM to incorporate the unlabeled data. Hand in the function and report the test error when training on the full dataset.

Answer: Here is one possible implementation of the EM inner loop (the update for Σ_c is similar)

```
# Comptue responsibilities
for i in 1:t
  for c in 1:k
    r[i,c] = theta[c]*exp(logMVNpdf(Xbar[i,:],mu[:,c],Sigma[:,:,c]))
  end
  r[i,:] ./= sum(r[i,:])
end

# Update theta
for c in 1:k
  z[c] = nc[c] + sum(r[:,c])
  theta[c] = z[c]/(n+t)
end

# Update mu
mu = zeros(d,k)
for i in 1:n
  c = y[i]
  mu[:,c] += X[i,:]
end
for i in 1:t
  for c in 1:k
    mu[:,c] += r[i,c]*Xbar[i,:]
  end
end
for c in 1:k
  mu[:,c] ./= z[c]
end
```

The test error when training with EM decreases from 0.469 (which was worse than if had all the labels) to 0.249 which non-intuitively is actually *better than if we had all the true labels*. The reason that EM does better is probably that the true labels are very noisy so by integrating over their values the Gaussians aren't chasing as much of this noise.

3. Repeat the previous part, but using the imputation approach ("hard"-EM) where we explicitly classify all the unlabeled examples before each model update. How does this change the performance and the number of iterations?

   Answer: You can modify the EM code to do "hard"-EM by just changing the responsibility variables r so that they are one for the most likely class and zero for the others. I found that this gave the same accuracy (0.248) but converged in fewer iterations.

Hint: for the first question most of the work has been done for you in the EM notes on the course webpage. You can use the result (**) and the update of θ_c from those notes, but you will need to work out the update of the parameters of the Gaussian distribution $p(x^i|y^i, \Theta)$.

For the second question, although EM often leads to simple updates, implementing them correctly can often be a pain. One way to help debug your code is to compute the observed-data log-likelihood after every iteration. If this number goes down, then you know your implementation has a problem. You can also test your updates of sets of variables in this way too. For example, if you hold the μ_c and Σ_c fixed and only update the θ_c, then the log-liklihood should not go down. In this way, you can test each of combinations of updates on their own to make sure they are correct.

2.2 Poisson Mixture Model

Consider a density estimation with examples $x^i \in \{\mathbb{Z}_{\geq 0}\}^d$ representing *counts* of d variables. In this setting, a natural way to model an individual variable x_j^i would be with a Poisson distribution,[3]

$$p(x_j^i = \alpha \mid \lambda_j) = \frac{\lambda_j^\alpha e^{-\lambda_j}}{\alpha!}.$$

However, if we assume this structure for each variable then the variables would be independent. One way to model dependent count variables would be with a mixture of independent Poisson distributions,

$$p(x^i|\Theta) = \sum_{c=1}^{k} \theta_c \prod_{j=1}^{d} p(x_j^i \mid \lambda_{jc}),$$

where Θ contains all the θ_c and λ_{jc} values. Derive the EM update for this.

Hint: most of the work has been done for you in the EM notes on the course webpage.

Answer: From the notes we have that the "responsibilities" are

$$r_c^i = p(z^i = c|x^i, \Theta^t) = \frac{p(z^i = c, x^i|\Theta^t)}{\sum_{c'=1}^{k} p(z^i = c', x^i|\Theta^t)},$$

and the update for θ_c is

$$\theta_c = \frac{\sum_{i=1}^{n} r_c^i}{n},$$

since these are true for any mixture model. So what remains is maximizing the EM objective,

$$Q(\Theta|\Theta^t) = \sum_{i=1}^{n} \sum_{z^i=1}^{k} r_{z^i}^i \log p(z^i|\Theta) + \sum_{i=1}^{n} \sum_{z^i=1}^{k} r_{z^i}^i \log p(x^i|z^i, \Theta),$$

[3]Note that the Poisson distribution makes certain assumptions about count data, which may not always be true, so in practice you may want to think about other possible likelihoods.

in terms of the λ_{jc}. But note that this function is separable in λ_{jc} and the objective with respsect to a particular λ_{jc} is given by

$$\sum_{i=1}^{n} r_c^i \log\left(\frac{\lambda_{jc}^{x_j^i} e^{-\lambda_{jc}}}{x_j^i!}\right)$$
$$= \sum_{i=1}^{n} r_c^i (x_j^i \log(\lambda_{jc}) - \lambda_{jc} - \log(x_j^i!)).$$

Setting the derivative of this expression with respect to λ_{jc} to zero gives

$$0 = \sum_{i=1}^{n} r_c^i \left(\frac{x_j^i}{\lambda_{jc}} - 1\right)$$
$$0 = \sum_{i=1}^{n} r_c^i \frac{x_j^i}{\lambda_{jc}} - n\theta_c.$$

or that

$$\lambda_{jc} = \frac{\sum_{i=1}^{n} r_c^i x_j^i}{n\theta_c},$$

which is the average of the counts of x_j^i over i that have been "soft-assigned" to cluster c.

3 Very-Short Answer Questions

Give a short and concise 1-sentence answer to the below questions.

1. Which of the following models cannot be kernelized?
   (a) K-nearest neighbours.
   (b) L2-regularized least squares.
   (c) L1-regularized least squares.
   (d) L2-regularized logistic regression.
   (e) PCA.

   Answer: L1-regularized least squares. (All the others are distanced-based, L2-regularized linear models, or eigenvalue-based.)

2. How can we use density estimation for supervised learning?

   Answer: Fit a density estimation model for $p(x^i, y^i)$, then use conditional probabilities $p(y^i|x^i)$ to do supervised learning. (A varation is to write $p(y^i|x^i) \propto p(x^i|y^i)p(y^i)$ and fit a density estimator for $p(y^i)$ and $p(x^i|y^i = c)$ separately for each c.)

3. For categorical variables, which prior leads to Laplace smoothing as the MAP estimate?

   Answer: Dirichlet.

4. What is an advantage and a disadvantage of density estimation using a product of independent distributions?

   Answer: Advantages are that it's fast or that you can fit each column independently, disadvantage is that it's a very restrictred class of densities.

5. Describe a one-dimensional dataset where fitting a Gaussian distribution to the data would be innappropriate.

   Answer: A dataset that has outliers, or just heavy-tails, or is multi-modal (like having clusters).

6. In the graphical LASSO, what do edges in the graph correspond to?

   Answer: Non-zeroes in the precision/inverse-covariance matrix $\Theta = \Sigma^{-1}$.

7. List the 4 operations we discussed that give Gaussian distributions by transforming existing Gaussian variables/distributions.

   Answer: Affine transformation of Gaussian variable, marginalization in joint Gaussian, conditioning in joint Gaussian, product of Gaussian PDFs.

8. Why do the π_c need to be a convex combination in mixture models?

   Answer: Makes sure that distribution sums/integrates to 1 (so it's a probability)

9. What is the difference between "missing at random" and "missing completely at random"?

   Answer: In MAR the locations of the missing values are not random (there could be a pattern in the missing values, although this pattern can't depend on the values that are missing).

10. In what setting does it make sense to use the EM algorithm?

    Answer: When knowing the values of the hidden variables makes the problem easy.

11. What is an advantage and a disadvantage of using a homogeneous Markov chain instead of an inhomogeneous Markov chain?

    Answer: Advantages of using a homogeneous chain could be that you need fewer parameters or that you can handle chains or arbitrary length, but a disadvantage is that you can't model process that change over time.

4 Project Proposal

For the final part of this assignment, you must a submit a project proposal for your course project. The proposal should be a maximum of 2 pages (and 1 page or half of a page is ok if you can describe your plan concisely). The proposal should be written for me and the TAs, so you don't need to introduce any ML background but you will need to introduce non-ML topics. The projects must be done in groups of 2-3. If you are doing your assignment in a group that is different from your project group, only 1 group member should include the proposal as part of their submission (we'll do the merge across assignments, and this means that assignments could have multiple proposals). Please state clearly who is involved with each project proposal.

There is quite a bit of flexibility in terms of the type of project you do, as I believe there are many ways that people can make valuable contributions to research. However, note that ultimately the project will have three parts:

1. A very shot paper review summarizing the pros and cons of a particular paper on the topic (due with Assignment 4).
2. A short literature review summarizing at least 10 papers on a particular topic (due with Assignment 5).
3. A final report containing at most 6 pages of text (the actual document can be longer due to figures, tables, references, and proofs) that emphasizes a particular "contribution" (i.e., what doing the project has added to the world).

The reason for this, even though it's strange for some possible projects, is that this is the standard way that results are communicated to the research community.

The three mains ingredients of the project proposal are:

1. What problem you are focusing on.
2. What you plan to do.
3. What will be the "contribution".

Also, note that for the course project that negative results (i.e., we tried something that we thought we would work in a particular setting but it didn't work) are acceptable (and often unavoidable).

Here are some standard project "templates" that you might want to follow:

- **Application bake-off**: you pick a specific application (from your research, personal interests, or maybe from Kaggle) or a small number of related applications, and try out a bunch of techniques (e.g., random forests vs. logistic regression vs. generative models). In this case, the contribution would be showing that some methods work better than others for this specific application (or your contribution could be that everything works equally well/badly).
- **New application**: you pick an application where ML methods where people aren't using ML, and you test out whether ML methods are effective for the task. In this case, the contribution would be knowing whether ML is suitable for the task.
- **Scaling up**: you pick a specific machine learning technique, and you try to figure out how to make it run faster or on larger datasets (for example, how do we apply kernel methods when n is very large). In this case, the contribution would be the new technique and an evaluation of its performance, or could be a comparison of different ways to address the problem.
- **Improving performance**: you pick a specific machine learning technique, and try to extend it in some way to improve its performance (for example, how can we efficiently use non-linearity within graphical models). In this case, the contribution would be the new technique and an evaluation of its performance.
- **Generalization to new setting**: you pick a specific machine learning technique, and try to extend it to a new setting (for example, making a graphical-model version of random forests). In this case, the contribution would be the new technique and an evaluation of its performance, or could be a comparison of different ways to address the problem.
- **Perspective paper**: you pick a specific topic in ML, read a larger number of papers on the topic, then write a report summarizing what has been done on the topic and what are the most promising directions of future work. In this case, the contribution would be your summary of the relationships between the existing works, and your insights about where the field is going.
- **Coding project**: you pick a specific method or set of methods (like independent component analysis), and build an implementation of them. In this case, the contribution could be the implementation itself or a comparison of different ways to solve the problem.
- **Theory**: you pick a theoretical topic (like the variance of cross-validation or the convergence of proximal stochastic gradient in the non-convex setting), read what has been done about it, and try to prove a new result (usually by relaxing existing assumptions or adding new assumptions). The contribution could be a new analysis of an existing method, or why some approaches to analyzing the method will not work.

The above are just suggestions, and many projects will mix several of these templates together, but if you are having trouble getting going then it's best to stick with one of the above templates. Also note that the

above includes topics not covered in the course (like random forests), so there is flexibility in the topic, but the topic should be closely-related to ML.

This question is mandatory but will not be formally marked: it's just a sanity check that you have at least one project idea that fits within the scope of 540 course project, and it's an excuse for you to allocate some time to thinking about the project. Also, there is flexibility in the choice of project topics even after the proposal: if you want to explore different topics you can ultimately choose to do a project that is unrelated to the one in your proposal/paper-review/literature-review, although it will likely be easier to do all 4 parts on the same topic.