HW8

Saturday, 23 March 2024 16:10

Q1)

Describe how the posterior predictive distribution is created for mixture models

We first need to define the mixture model for the data of interest, including specifying the likelihood and the prior and their parameters

To estimate the parameters we can use MCMC and these will include the mixing proportions, means and covariances of the components of the mixture

We know the posterior or prior. likelihood, so since we have estimated the parameters, we can now compute the posterior distribution of the parameters given the data.

For the predictive distribution, we can integrate over the parameter space, weighting the predictions from each component of the mixture model by its posterior possibility:

$$p(y \mid x) = \int p(y \mid \theta) \cdot p(\theta \mid x)\, d\theta$$

where y is the new data we want to predict
x is the observed data
and θ is the parameters

Now to generate new data points, samples are drawn from the predictive distribution and these will represent possible values of the new data points given observed data and the uncertainty in the model parameters.

Q2) Describe the posterior predictive distribution is created in general

In general, this represents the distribution of data that has not been observed before given previous data and the uncertainty in the model parameters

We must also specify the prior distributions and the likelihood

We must also specify the prior distributions and the likelihood and we can use these to calculate the posterior distribution using Bayes' thm:

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

Here $p(x)$ is the marginal likelihood of the observed data and it acts as a normalizing constant

Like before we can use the posterior to make predictions about new data by integrating over Θ:

$$p(y|x) = \int p(y|\theta) \cdot p(\theta|x) d\theta$$

Now we could sample from this posterior predictive distribution using something like MCMC

Q3) If we are doing regression of y on X but X has missing values, we could perform a Bayesian analysis without throwing away the rows with missing values in X.

By the hint, we need to use missing values as latent variables to be estimated along with the model parameters

We begin by assuming the data is missing at random, so there is no relationship between the missing data and the overall data patterns. We will validate the assumption later.

We would first specify the model, including the relationship between our dependent and independent variables and also introducing latent variables for the missing values, which will be treated as parameters to be inferred in the Bayesian analysis

Our likelihood function needs to incorporate both observed and missing data:

For observed data, use likelihood corresponding to the regression model

For missing data, introduce a likelihood that accounts for the uncertainty of the missing values

Next assign prior distributions for the model parameters and also priors for the latent variables

We use MCMC or other methods to sample from the joint posterior distribution of the model parameters and missing values We update this posterior iteratively, incorporating information from the observed data and the missing data.

Now for imputation, we impute missing values using this posterior predictive distribution and for each missing values, generate samples from the posterior distribution of the corresponding latent variables and use these to impute the missing values.

We must now assess the model using both observed and imputed data and validate the MCAR assumption