**Part B**

Q1.

1.1 At the above BPTT function, it is observably that ‘self,ForwardPass(sample, slide)’ includes two parameter ‘sample’ and ‘slide’, the reasons will list below:

1. Processing sequence data: BPTT is designed for processing sequence data, especially in recurrent neural networks (RNN). In this case, 'sample' stands for a sample in a sequence, such as a time series of data or a sentence.
2. Processing time step: The slide parameter represents the time step or a specific moment in the sequence. In sequence processing the network needs to process a portion of the input data at each time step. For example, when working with a piece of text, each "slide" might represent a word in the sentence.

In conclusion, the two parameters ‘sample’ and ‘slide’ in self.ForwardPass(sample, slide) enable the algorithm to process the input correctly at each point in time in the d sequence and maintain continuity and dependency in time, which is critical for processing time series data or any form of sequence data.

**Reference:**

1. <https://www.wrike.com/project-management-guide/faq/what-is-forward-pass-in-project-management/>
2. <https://datascience.stackexchange.com/questions/116973/whats-a-forward-pass-in-a-neural-net>

1.2 In the provided K-Means algorithm Python code, the formation and refinement of clusters are explained through the following steps:

1. \*\*Initialization of Centroids (Lines 11-13):\*\*

- The `fit` function starts by selecting the first `k` points from the dataset as the initial centroids.

2. \*\*Iteration Process (Lines 15-50):\*\*

- The algorithm performs a loop for a maximum number of iterations, controlled by the `max\_iter` parameter(in line8, 11, 21). Each iteration updates the cluster assignments and the positions of the centroids.

3. \*\*Assigning Points to the Nearest Centroid (Lines 22-29):\*\*

- For each point (`featureset`) in the dataset, the Euclidean distance to each centroid is calculated. The point is then assigned to the cluster of the closest centroid.

4. \*\*Updating Centroid Positions (Lines 33-35):\*\*

- For each cluster, the average position of all its points is calculated, and this average is set as the new position of the centroid.

5. \*\*Checking for Convergence (Lines 37-46):\*\*

- The algorithm checks if the movement of all centroids is less than a set tolerance (`tol`). If the movement of all centroids is below this threshold, it indicates that the algorithm has converged, and the iteration ends.

6. \*\*Termination and Returning Centroids (Lines 48-49):\*\*

- Once the algorithm converges or reaches the maximum number of iterations, the iteration stops, and the current positions of the centroids are returned.

This process effectively involves iteratively reassigning data points to their nearest centroids and updating the positions of centroids based on these assignments. This repeated process ensures that over time, clusters gradually stabilize and ultimately converge to a relatively stable state, forming the final clusters.

**Reference:**

1. [**https://en.wikipedia.org/wiki/K-means\_clustering**](https://en.wikipedia.org/wiki/K-means_clustering)
2. [**https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/**](https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/)
3. [**https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1**](https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1)
4. [**https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html**](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
5. [**https://www.stata.com/manuals13/rmaximize.pdf**](https://www.stata.com/manuals13/rmaximize.pdf)

**Q2.**

2.1 Choosing the Appropriate Prior Distribution:

The appropriate prior distribution for a model depends on the prior knowledge about the parameters and the problem domain. For instance, if you know a parameter is positive and likely to be in a specific range, a uniform or log-normal prior might be appropriate. If there's no specific prior knowledge, a non-informative or weakly informative prior, such as a normal distribution with a large variance, could be used to reflect this uncertainty. In a Bayesian neural network, priors over weights could be chosen based on the expected size of the weights or based on regularization considerations—smaller weights might suggest a prior that favors smaller absolute values, like a Gaussian centered at zero.

2.2 Types of Priors in the Given Equation:

The equation shows two priors that are derived from the inverse gamma distribution, which is a type of conjugate prior for the variance in a Gaussian distribution. Conjugate priors are often used in Bayesian statistics because they simplify the posterior updating process. The reason for using two different priors (parameters v1 and v2 for the inverse gamma distributions) in the Gaussian likelihood for regression problems is to separately model the uncertainty in the scale of the data (the variance of the Gaussian likelihood) and the scale of the weights (from the regularization term). These separate priors allow for more flexibility in modeling; the data variance and weight variance can be independently controlled and updated based on observed data.

2.3 Suitability of Priors for Multinomial Likelihood:

The given priors are not suitable for a multinomial likelihood because a multinomial distribution models discrete outcomes and typically requires different types of priors, like the Dirichlet distribution, which is the conjugate prior for the parameters of a multinomial distribution. If one were to use these priors for a multinomial likelihood, they would need to change to reflect the discrete nature of the data and the different parameter space—for example, using a Dirichlet prior instead of an inverse gamma, to model the probabilities associated with the different possible outcomes of the multinomial distribution. This change is necessary because the support and nature of the parameters (probabilities in multinomial vs. variances in Gaussian) are fundamentally different.

**PART C OPTION 1**





















