Midterm Exam - Part 2

Question Responses

1. **What is your experimental design to determine the best model and hyperparameter settings for the classifier? State all controlled and uncontrolled variables, define how parameters are being determined, and describe training/validation/testing strategies. Justify your experimental design and approach.**

A1: The experimental design involves the following steps to determine the best model for the classifier:

(1) **Split a data set into three groups: training (80 %), validation (10 %) and testing (10 %)**. Concerning that the data set covers data samples from three classes ("Iris-setosa", "Iris-versicolor", "Iris-virginica"), we attempted two methods to select data samples as training, validation and testing sets: a) to select 80 %, 10 % and 10 % of all data samples in the data set; b) to separately select 80 %, 10 % and 10 % of data samples in each class and then integrate together as training, validation and testing sets.

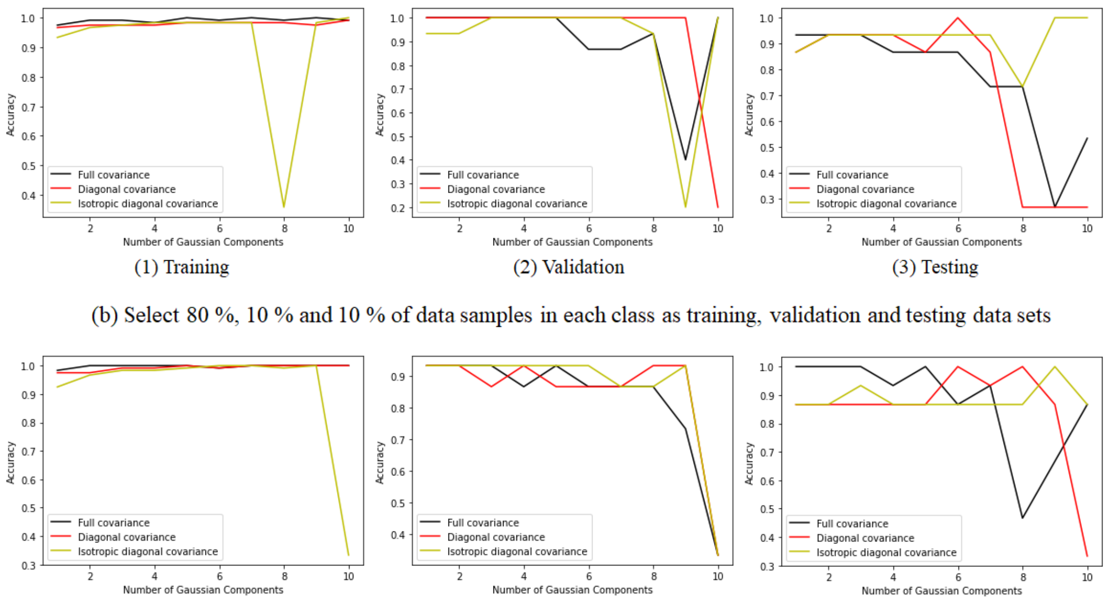
(2) **Clarify a goal of the classifier model: to correctly classify data samples into classes or groups as possible as it can**. To evaluate the model performance, we introduced a concept of confusion matrix to count how many data samples are correctly or wrongly classified for each class.

(3) **Select one suitable classifier model: to implement a Probabilistic Generative Classifier on the data set**. For each class, we used and trained a Gaussian Mixture Model to estimate the generating distribution, respectively. Then we used outputs of mean vector, covariance matrix, probability vector from three Gaussian Mixture Models to calculate the probability of every data sample being classified into each class. The data sample was classified into the class with the highest probability.

(4) **Vary hyperparameters and investigate how they impact the performance of the classifier model**. We varied the number of Gaussian components from 1 to 10, selected two splitting methods of training, validation and testing data sets, used three types of covariance (a full covariance, diagonal covariance and isotropic diagonal covariance) to determine how they impact the performance of the classifier model. The classification accuracy can be calculated as the ratio of correctly classified data samples to the total data samples. We set maximum number of iterations and difference threshold between two iterations as 100, 1e-4, respectively.

(5) **Train the classifier model and run the experiment**. According to the above experimental design, the parameters in (4) were determined depending on the optimal classification results on training and validation data sets. As illustrated in Figure 1, we found that a) the optimal number of Gaussian components were 2, b) selecting 80 %, 10 % and 10 % of data

samples in each class, rather than all data samples, was a better way to split training, validation and testing sets, c) a full covariance was better than the other two types of covariance in the classifier model.



(a) Select 80 %, 10 % and 10 % of all data samples as training, validation and testing data sets

(1) Training (2) Validation (3) Testing

(b) Select 80 %, 10 % and 10 % of data samples in each class as training, validation and testing data sets

(1) Training (2) Validation (3) Testing

Figure 1 Classification accuracy when number of Gaussian components varies

(6) **Test the classifier model and evaluate the model**. Based on the optimal parameters in (5), we tested the classifier model using the testing set and calculated the confusion matrix and classification accuracy on the testing set to evaluate the performance of the classifier model.

(7) **Describe all controlled and uncontrolled variables**. *All controlled variables* include: a) number of Gaussian components, b) the selection of covariance type in a Gaussian Mixture Model, c) parameters including maximum number of iterations and difference threshold between two iterations in a Gaussian Mixture Model, d) parameter initialization of mean vector in a Gaussian Mixture Model, e) the splitting methods of training, validation and testing data sets. *All uncontrolled variables* refer to the results of random selection of data samples in training, validation, testing sets and mean vector initialization in a Gaussian Mixture Model.

2. **One choice that needs to be made is whether you will use a full covariance, diagonal covariance, or isotropic diagonal covariances in your Gaussians in the mixture model. What are the advantages and disadvantages of these various options? Be sure to address issues related to small data set size and flexibility of the learned model.**

A2: Full covariance, diagonal covariance and isotropic diagonal covariance were used in each Gaussian component of the three Gaussian Mixture Models to investigate the impact of covariance types on the classifier model performance. Results indicate that full covariance was the best one in the Gaussian Mixture Model, followed by diagonal covariance and isotropic diagonal covariance. This is mainly due to the fact that a full covariance can capture more patterns from data samples on a small data set size, including relationships among different features and covariances of single feature.

Specifically, this phenomenon could be further explained by individual characteristics of the three covariance types, namely advantages and disadvantages. Full covariance fully considers the correlations and relationships among different features in data samples while partly weakening the impact of covariances of single feature on the model performance. By contrast, diagonal covariance takes the relationships among different features into no account but fully emphasizes on the value of covariances of single feature in data samples. This is favorable for selecting the most principal feature and strengthening its impacts on the model performance. In terms of isotropic diagonal covariance, it neither considers the relationships among different features nor the discrepancy in covariances of different features but it is very easy to implement in the Gaussian Mixture Model.

3. **What is your final confusion matrix on your training, validation and test data?**

A3: Based on the optimal parameter settings in Gaussian Mixture Models, we used data samples in training, validation and testing sets to evaluate the performance of the classifier model, in terms of confusion matrix and classification accuracy. Then we can easily calculate the final confusion matrix on the training, validation and testing sets as follows:

(1) Confusion matrix on the training set.

Table 1 Confusion matrix on the training set (80 %)

Class	Iris setosa	Iris versicolor	Iris virginica
Iris setosa	40	0	0
Iris versicolor	0	40	0
Iris virginica	0	0	40

Note: Every row and column refer to the actual and estimated classes, respectively.

(2) Confusion matrix on the validation set.

Table 2 Confusion matrix on the validation set (10 %)

Class	Iris setosa	Iris versicolor	Iris virginica
Iris setosa	5	0	0
Iris versicolor	0	4	1
Iris virginica	0	0	5

(3) Confusion matrix on the testing set.

Table 3 Confusion matrix on the testing set (10 %)

Class	Iris setosa	Iris versicolor	Iris virginica
Iris setosa	5	0	0
Iris versicolor	0	5	0
Iris virginica	0	1	4