### Latent Variables, Linear Models and Model Assessment

Train-test split for ML models: The train-test split is a method of evaluating the performance of a machine learning model by splitting the data set into two mutually exclusive sets: the training set and the testing set.

Advantages of train-test split:

1. Helps in evaluating model performance: By dividing the data into a training and testing set, we can validate how well the model has been trained.
2. Prevents overfitting: it helps in identifying overfitting by testing the model on unseen data.
3. Enables hyperparameter tuning: we can optimize hyperparameters and fine-tune the model's performance.

Disadvantages of train-test split:

1. Data Variability: The random selection of data could lead to variability in model performance.
2. Information loss: By splitting the data into training and testing sets, we lose some data for training. If the dataset is small, this could impact the model's performance.
3. Limited evaluation: It does not give insight into how well the model performs on different types of data.

### When to Use the Train-Test Split: The train test split method is a good evaluation procedure for machine learning algorithms when dealing with large data sets. It is also reliable when training costly models or estimating model performance quickly.

Configuring Train-test split:

### train, test = train\_test\_split(dataset, …)

### X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, …)

Latent Variable Models:

In a latent variable model, the observed data is assumed to be generated by a set of latent variables that are not directly observable but can be inferred from the observed data. The relationship between the observed data and the latent variables is modeled through a set of parameters, which are learned from the data using various estimation techniques.

Advantages of latent variable models:

1. Uncovering hidden structures: Latent variable models can help uncover hidden structures, relationships, or patterns in data that are not directly observable.
2. Dimensionality reduction: it can be used to reduce the dimensionality of the data by representing it in terms of a smaller number of latent variables.
3. Improved modeling: it can lead to improved modeling accuracy and interpretability by accounting for the relationships between the observed data and the latent variables.

Disadvantages of latent variable models:

1. Complex modeling:it can be complex to model and estimate, and may require significant computational resources.
2. Model selection: Selecting an appropriate model structure and determining the number of latent variables can be challenging and may require iterative experimentation.
3. Interpretability: The interpretability of the latent variables may be limited, particularly in complex models with many latent variables.

Applications:

1. Dimensionality reduction: it can be used to reduce the dimensionality of high-dimensional data, such as images, text, or genomic data.
2. Clustering: it can be used to cluster data based on shared underlying structures or relationships.
3. Prediction: it can be used to predict future outcomes based on past observations and underlying structures.
4. Causal inference: it can be used to infer causal relationships between variables by identifying the underlying structures that generate the observed data.

AUC ROC Curve:

the AUC-ROC curve is a powerful tool for evaluating the performance of binary classification models. The ROC (receiver operating characteristic) curve is a graphical representation of the performance of a binary classification model as the discrimination threshold is varied. The AUC (area under the curve) is a single scalar that summarizes the overall performance of the model, irrespective of the choice of threshold.

1. True Positive (TP): The number of positive cases that are correctly classified as positive by the model.
2. False Positive (FP): The number of negative cases that are incorrectly classified as positive by the model.
3. True Negative (TN): The number of negative cases that are correctly classified as negative by the model.
4. False Negative (FN): The number of positive cases that are incorrectly classified as negative by the model.
5. Sensitivity or True Positive Rate (TPR): The proportion of actual positive cases that are correctly identified as positive by the model. TPR = TP / (TP + FN)
6. Specificity or True Negative Rate (TNR): The proportion of actual negative cases that are correctly identified as negative by the model. TNR = TN / (TN + FP)
7. False Positive Rate (FPR): The proportion of actual negative cases that are incorrectly identified as positive by the model. FPR = FP / (FP + TN)
8. Precision or Positive Predictive Value (PPV): The proportion of cases that the model classifies as positive that are actually positive. PPV = TP / (TP + FP)
9. Negative Predictive Value (NPV): The proportion of cases that the model classifies as negative that are actually negative. NPV = TN / (TN + FN)

The AUC ranges from 0 to 1, with a value of 0.5 indicating that the model is no better than random, and a value of 1 indicating that the model is perfect.

Confusion Matrix:

It is a table with 4 different combinations of predicted and actual values.

It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.