Assignment 06 Solutions

1. In the sense of machine learning, what is a model? What is the best way to train a model?

Ans: A Machine Learning Model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

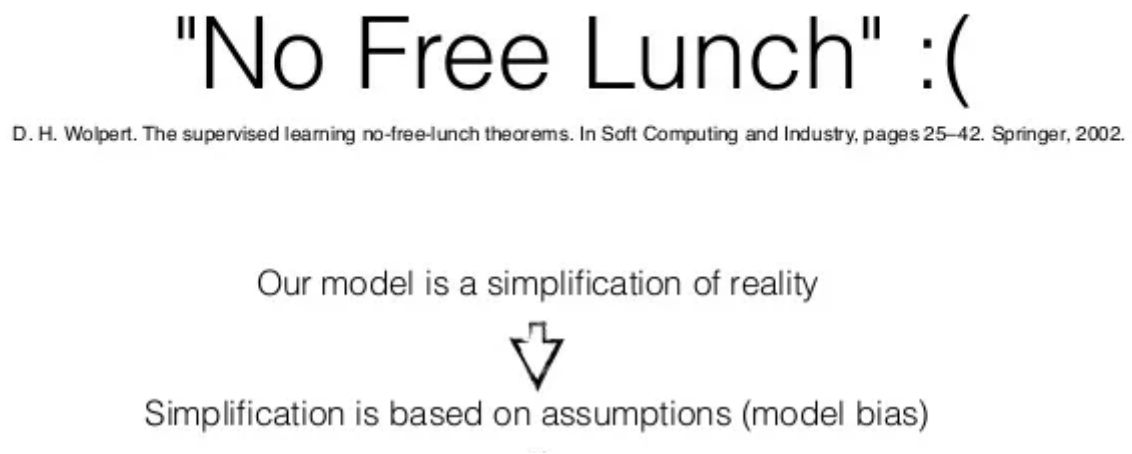
1. **Model Naming** — Give Your Model a Name, Let's start with giving your model a name, describe your model and attach tags to your model. Tags are to make your model searchable.
2. **Data Type Selection** — Choose data type(Images/Text/CSV), It's time to tell us about the type of data you want to train your model. ML Models support Images, Text and *.CSV (categorical data) data types.
3. **Data Upload** — Upload your data or choose from Public Data Sets: Choose from public datasets like Jewellery Data set (Images), Gender Data Set (Images), Question or Sentence Data Set (Text), Numerai Data Set (CSV) or upload your data.
4. **Type category(label)** for the files (images/text file) that you have uploaded and click on submit to begin upload. Wait for some time till our web app uploads all the files. You can upload images for as many categories as possible.
5. **Start Training -** Push the button, to start the training. Now Mateverse's intelligent backend will start with processing the data that you have uploaded and preparing it for the training.

2. In the sense of machine learning, explain the "No Free Lunch" theorem.

Ans: The No Free Lunch Theorem, often abbreviated as NFL or NFLT, is a theoretical finding that suggests all optimization algorithms perform equally well when their performance is averaged over all possible objective functions. In computational complexity and optimization the no free lunch theorem is a result that states that for certain types of mathematical problems, the computational cost of finding a solution, averaged over all problems in the class, is the same for any solution method.

In Simple Words "No Free Lunch" theorem means we can't rely on one model to be best of all models. We have to understand data properly and make use of ML understanding and make use of models to find best out of it.

"

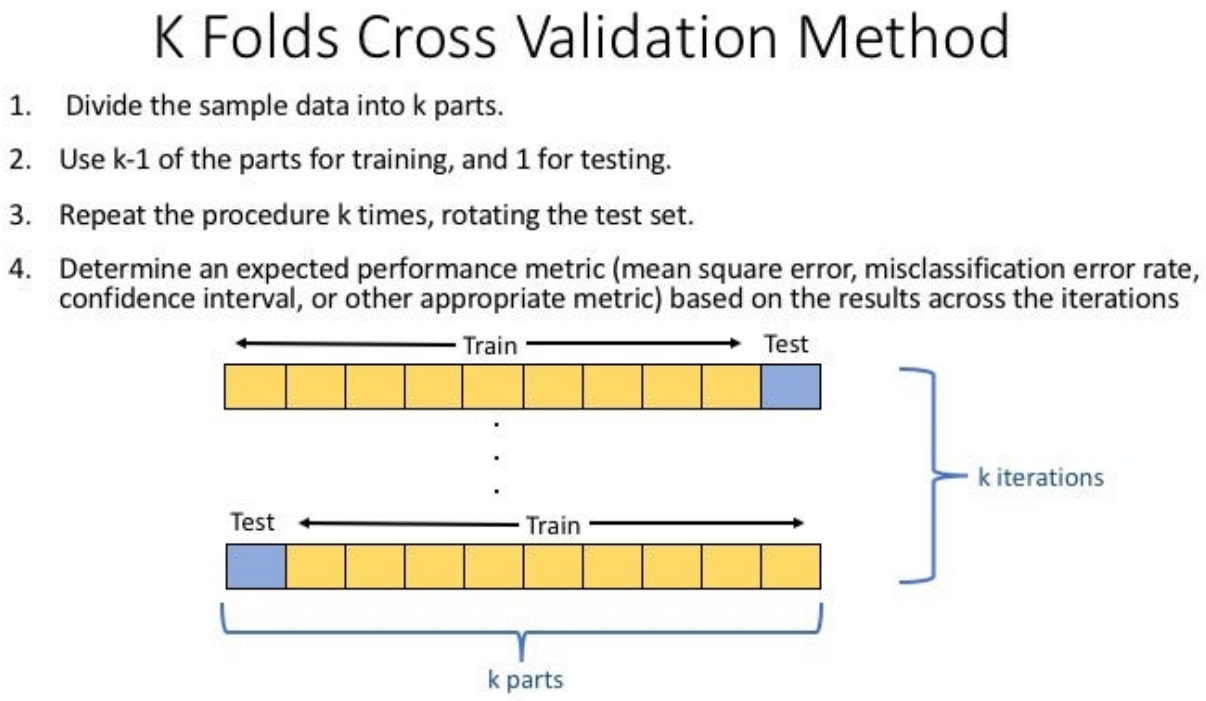


3. Describe the K-fold cross-validation mechanism in detail.

Ans: In K-fold cross validation, data D is subset into k subsets randomly. Let us assume S1...Sk are the subsets where Sk is the kth randomly split subset of data D. In the first iteration, D-S1 is used for training and S1 for testing the model. When the model has been trained and tested, evaluation can be done, score is noted elsewhere and the trained model is discarded.

These k-iterations go on where 1/k subset of D is always set aside for testing the data and D-1/k subsets are used for training, evaluating and discarding the model. At the end of all the iterations, average of all the evaluation scores is taken and used as output.

"



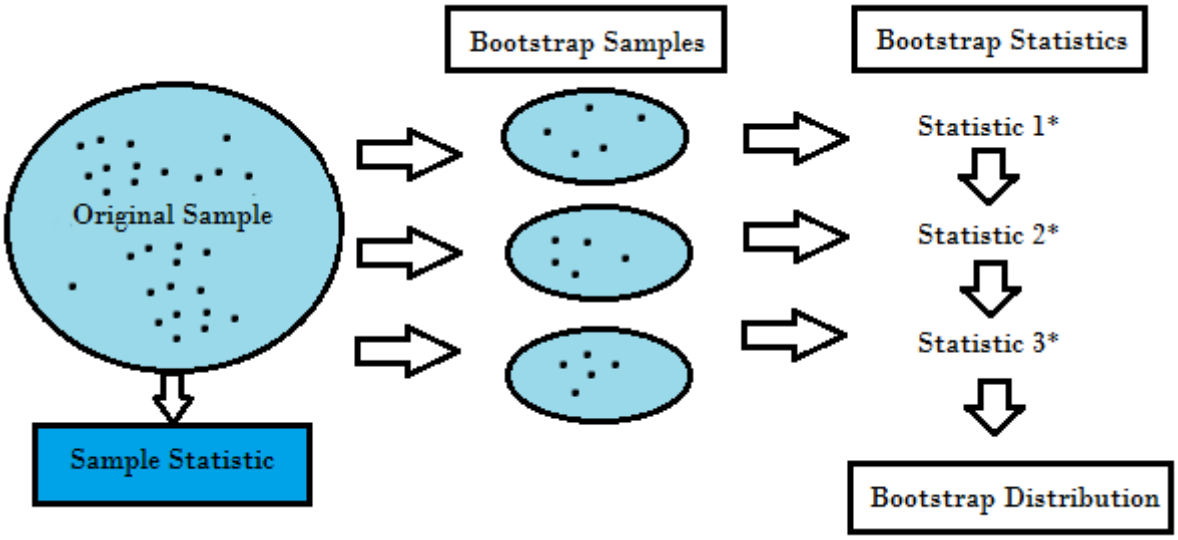
"

4. Describe the bootstrap sampling method. What is the aim of it?

Ans: The bootstrap method is a statistical technique for estimating quantities about a population by averaging estimates from multiple small data samples.

Importantly, samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen. This allows a given observation to be included in a given small sample more than once. This approach to sampling is called sampling with replacement.

"



"

5. What is the significance of calculating the Kappa value for a classification model? Demonstrate how to measure the Kappa value of a classification model using a sample collection of results.

Ans: Kappa value or Cohen's Kappa coefficient is an evaluation metric for classification models. Its significance as an evaluation metric is that it can be used to evaluate multi class classification models and also works on models trained on imbalanced datasets(scores like accuracy scores fail for imbalanced datasets).

In simpler words It basically tells you how much better your classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. Cohen's kappa is always less than or equal to 1. Values of 0 or less, indicate that the classifier is useless Cohen suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

"

Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement

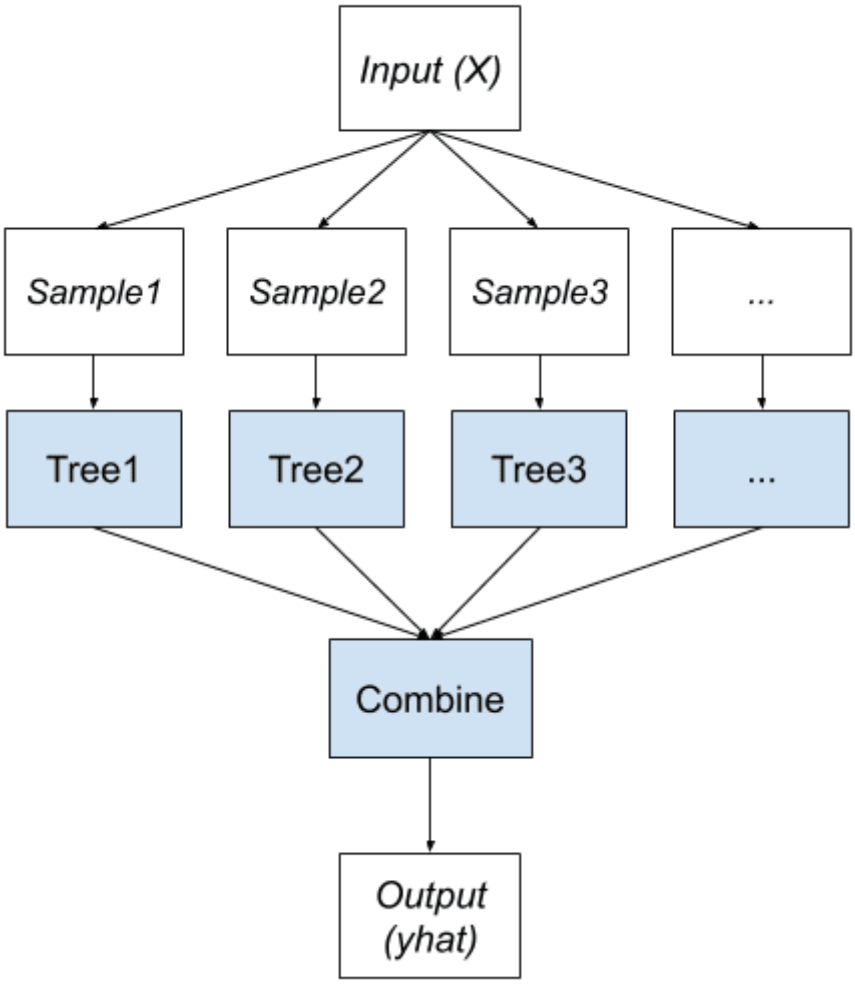
6. Describe the model ensemble method. In machine learning, what part does it play?

Ans: Ensemble methods or ensemble machine learning models are models where more than one models are being used spontaneously to produce better results than individually trained models. Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data.

The three main classes of ensemble learning methods are bagging, stacking, and boosting

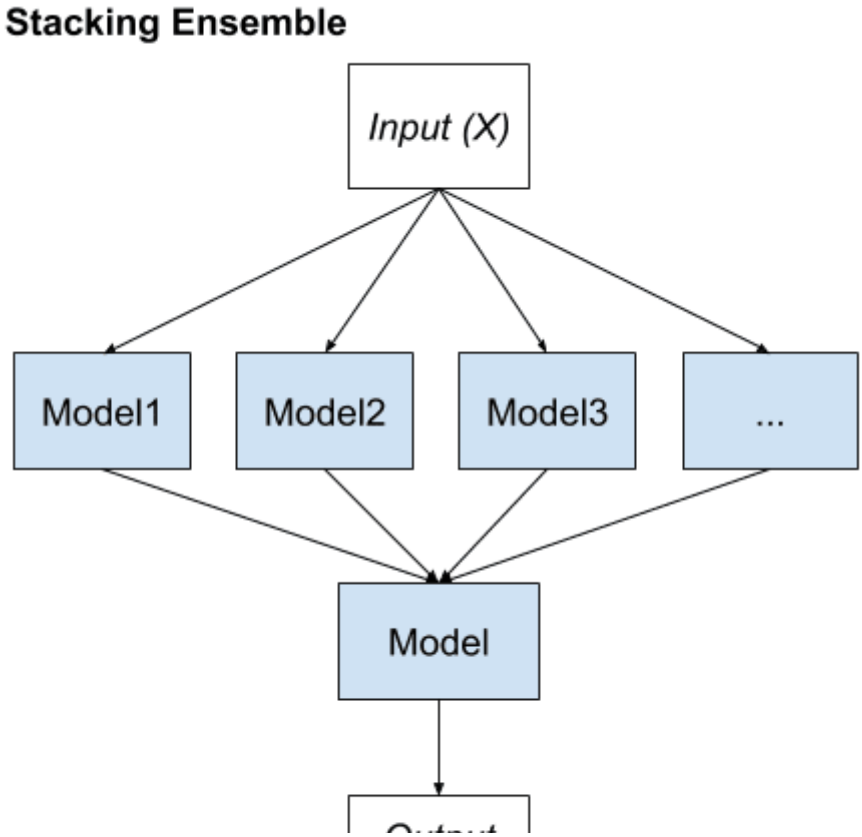
"

Bagging Ensemble



"

"



7. What is a descriptive model's main purpose? Give examples of real-world problems that descriptive models were used to solve.

Ans: A descriptive model is used for tasks that would benefit from the insight gained from summarizing data in new and interesting ways. As opposed to predictive models that predict a target of interest, in a descriptive model, no single feature is more important than any other. In fact, because there is no target to learn, the process of training a descriptive model is called unsupervised learning.

It is used in customer classification as real life problem .

8. Describe how to evaluate a linear regression model.

Ans: Evaluation of a linear regression model can be done using R-square. R square is calculated as the sum of squared errors in predictions made, divided by summation of all sum of squares. R square measures how much of the change in target variable can be explained by the linear regressor. Its value ranges from 0 to 1 where 0 means poor performance and 1 means good. Some other techniques which can be used to evaluate a linear regression model are:

1. Mean Square Error(MSE)/Root Mean Square Error(RMSE)
2. Mean Absolute Error(MAE) Others are mentioned in the picture below:

"

Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$

9. Distinguish :

1. Descriptive vs. predictive models
2. Underfitting vs. overfitting the model
3. Bootstrapping vs. cross-validation

Ans: The differences between:

- **Descriptive vs. predictive models**
  - Descriptive models are built to identify trends and underlying patterns.
  - Predictive models are built to predict a dependent variable value.
  - Most of descriptive models are built using unsupervised machine learning.
  - Most of predictive models are built using classification and regression models.
  - Example for descriptive model: Finding why consumers are engaging more with a social media post.
  - Example for predictive model: Predicting the chances of cancer in a patient.

- **Underfitting vs. overfitting the model**
  - Underfitting is a situation arising when the hypothesis is way too simple, or when the machine learning model is way too simple to produce good results.
  - Overfitting is a situation arising when the hypothesis is way too complex, or when the machine learning model is way too complex to produce good results.
  - Underfitting causes a model to produce poor results due to heavily simplified algorithm reacting lightly to changes in the unseen data for independent variables from the training data.
  - Overfitting makes a model produce poor results due to slightest variations in the unseen data for independent variables from the training data
  - Underfitting is also called High Bias.
  - Overfitting is also called High variance

- **Bootstrapping vs cross-validation**
  - Boostrap sampling is a method of sampling in which the repeated sampling is done with replacement using a data D in random draws over which machine learning models are trained for better performance.

- Cross validation is a method used to check the efficacy of the machine learning model on test data.
- End goal of bootstrapping is to reduce overfitting and increase performance.
- End goal of cross validation is only to produce test scores to check efficacy of model
- Bootstrapping is best employed in Random Forest Classifier.

10. Make quick notes on:

1. LOOCV.
2. F-measurement
3. The width of the silhouette
4. Receiver operating characteristic curve

Ans: The Quick notes on: LOOCV or Leave One Out Cross Validation is a form of K-fold cross validation where only one observation is left out for validation purpose while the rest of the data is used for model training each iteration. It is computationally taxing and should only be used for data with low dimensionality.

Harmonic mean of Precision score and recall score is called F-measurement or F-score. It is formulated as 2 (pr re)/pr +re where pr is precision score and re is recall score.

Estimate of average inter cluster distance to give efficacy/performance of cluster algorithms is called width of the silhouette. It can also be defined as how identical/similar a data point 'x' is to the data points inside the cluster to which x is assigned. Its value ranges from -1 to 1 where 1 means good and -1 means bad.

Curve plotted between True Positive Rate and False Positive Rate is Receiver Operating Characteristics curve and is used to find the area under the curve for ROC-AUC score for binary classification evaluation. True Positive Rate and False Positive Rate are calculated for different thresholds values where thresholds take values starting from the highest probability scores assigned to data points and goes up to the lowest probability score. The curve is impacted by presence of outliers, and simple models. Extensions can be made to this curve to suit multiclass classification evaluation requirements.