CSE 4553
Machine Learning
Lecture 4: Basic Practices in ML

Winter 2022

Hasan Mahmud | hasan@iut-dhaka.edu

Contents

- Feature Engineering
- Selection of ML algorithms
- Training, Validation, and Test set
- Underfitting and Overfitting
- Regularization
- ML Model Performance Analysis
  - Confusion Matrix
  - Accuracy
  - Precision, Recall
  - ROC, AUC curves
- Hyperparameter Tuning, Cross-Validation

Feature Engineering

- The problem of transforming raw data into a dataset is called feature engineering.
- Logs of user interaction with a computer system may contain the following features:
  - Price of the subscription
  - Frequency of connection per day, per week, and per year
  - Average session duration in seconds
  - Average response time and so on.
- Informative features: Help learning algorithm to build a model that predicts well labels of the data used for training.
- A model has a low bias when it predicts well the training data.

Feature Engineering…

- One-Hot Encoding
  - Used for categorical features when order of the feature is not important
  - E.g.

$$red = [1, 0, 0],$$
$$yellow = [0, 1, 0],$$
$$green = [0, 0, 1]$$

Feature Engineering…

- Binning
  - Used to convert numerical feature in to categorical feature.
  - Binning (also called bucketing) is the process of converting a continuous feature into multiple binary features called bins or buckets, typically based on value range
  - Three common approach of binning: Equal width binning, Equal frequency binning, and a K-means approach
  - For example, instead of representing age as a single real-valued feature, the analyst could chop ranges of age into discrete bins: all ages between 0 and 5 years-old could be put into one bin, 6 to 10 years-old could be in the second bin, 11 to 15 years-old could be in the third bin, and so on.

Feature Engineering…

- Normalization
  - Normalization is the process of converting an actual range of values which a numerical feature can take, into a standard range of values, typically in the interval [−1, 1] or [0, 1].
  - More generally, the normalization formula looks like this:

$$\bar{x}^{(j)} = \frac{x^{(j)} - min^{(j)}}{max^{(j)} - min^{(j)}},$$

where $min^{(j)}$ and $max^{(j)}$ are, respectively, the minimum and the maximum value of the feature j in the dataset.

Feature Engineering…

- Standardization
  - Standardization (or z-score normalization) is the procedure during which the feature values are rescaled so that they have the properties of a standard normal distribution with $\mu = 0$ and $\sigma = 1$, where μ is the mean (the average value of the feature, averaged over all examples in the dataset) and σ is the standard deviation from the mean.
  - Standard scores (or z-scores) of features are calculated as follows:

$$\hat{x}^{(j)} = \frac{x^{(j)} - \mu^{(j)}}{\sigma^{(j)}}.$$

Feature Engineering…

- Dealing with missing features:

  The typical approaches of dealing with missing values for a feature include:

  - Removing the examples with missing features from the dataset. This can be done if your dataset is big enough so you can sacrifice some training examples.
  - Using a learning algorithm that can deal with missing feature values (depends on the library and a specific implementation of the algorithm).
  - Using a data imputation technique.

Feature Engineering…

- Data Imputation Techniques
- One technique consists in replacing the missing value of a feature by an average value of this feature in the dataset:

$$\hat{x}^{(j)} = \frac{1}{N} x^{(j)}.$$

- Another technique is to replace the missing value by the same value outside the normal range of values. For example, if the normal range is [0, 1], then you can set the missing value equal to 2 or −1.

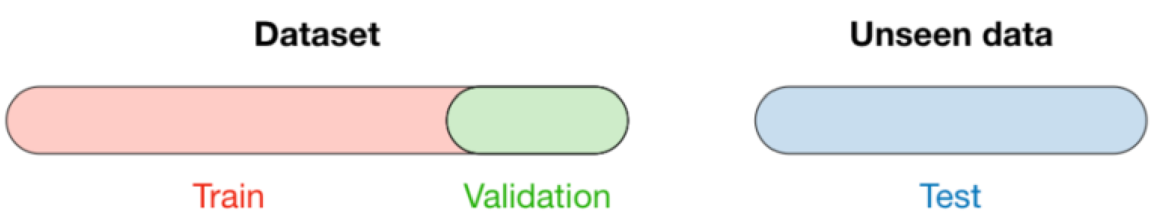
Learning Algorithm Selection

- Explainability
  - ML Models need to be explainable to the non-technical audience.
  - Neural Network/DNN VS KNN, linear regression/logistic regression
- In-memory vs out-of-memory
  - Can the dataset be loaded fully into the RAM or incremental learning procedure should be applied.
- Number of features and examples
- Categorical vs numerical features
- Nonlinearity of the data
- Training speed
- Prediction speed

Training set, validation set, test set

Training set	Validation set	Testing set
- Model is trained - Usually 80% of the dataset	- Model is assessed - Usually 20% of the dataset - Also called hold-out or development set	- Model gives predictions - Unseen data

Once the model has been chosen, it is trained on the entire dataset and tested on the unseen test set. These are represented in the figure below:

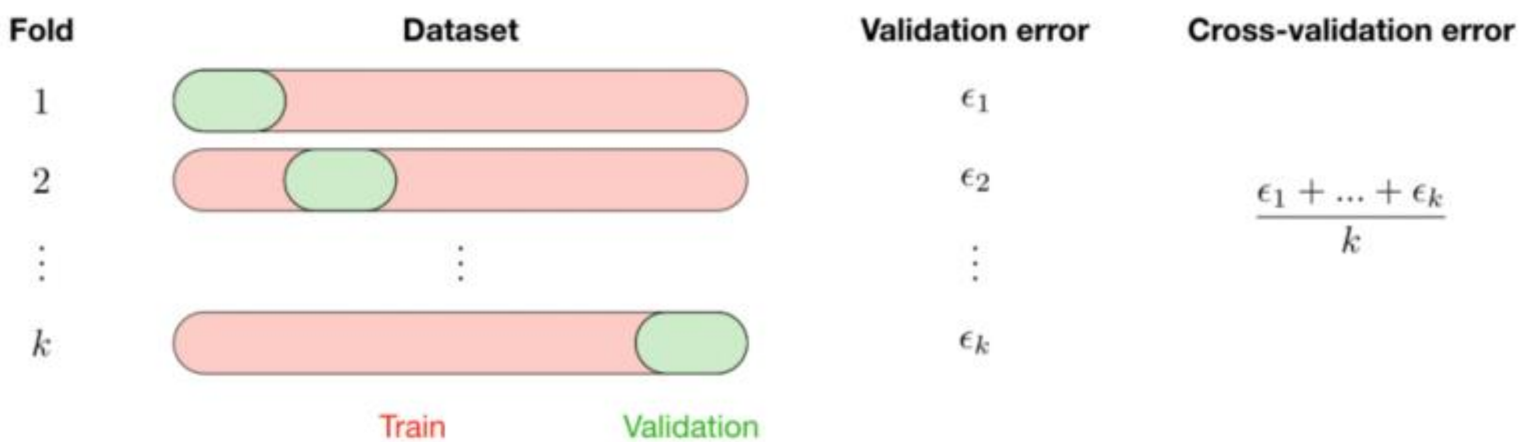


Cross-validation

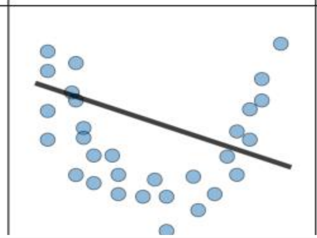
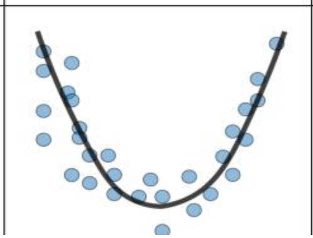
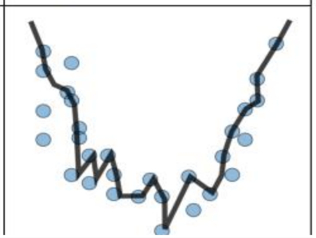
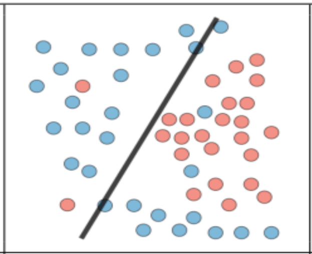
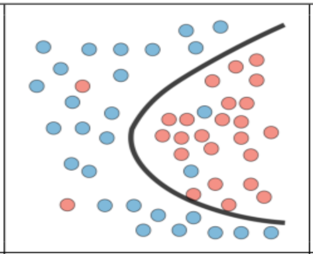
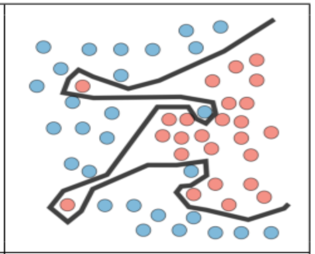
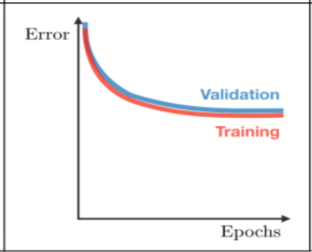
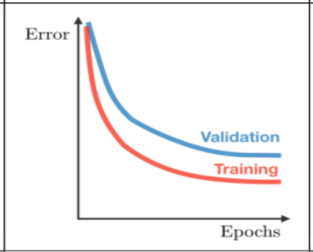
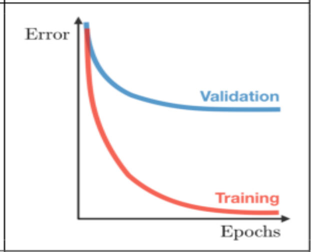
❒ **Cross-validation** – Cross-validation, also noted CV, is a method that is used to select a model that does not rely too much on the initial training set. The different types are summed up in the table below:

k-fold	Leave-p-out
- Training on $k-1$ folds and assessment on the remaining one - Generally $k = 5$ or 10	- Training on $n-p$ observations and assessment on the p remaining ones - Case $p = 1$ is called leave-one-out

The most commonly used method is called k-fold cross-validation and splits the training data into k folds to validate the model on one fold while training the model on the $k-1$ other folds, all of this k times. The error is then averaged over the k folds and is named cross-validation error.



Overfitting vs underfitting

	Underfitting	Just right	Overfitting
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			
Classification			
Deep learning			
Remedies	- Complexify model - Add more features - Train longer		- Regularize - Get more data

Model Performance Assessment

- Confusion Matrix

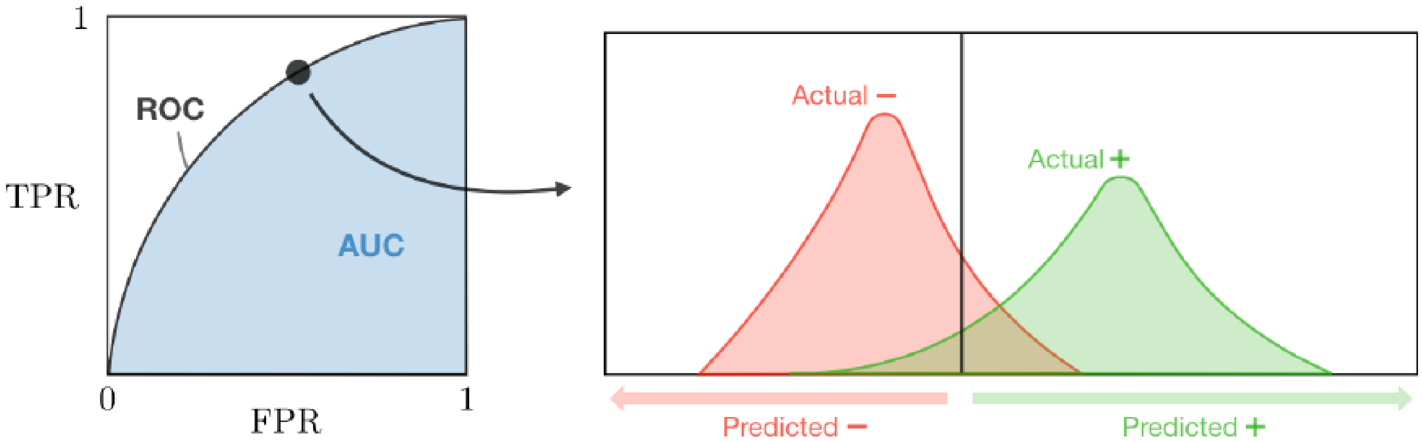
Metric	Formula	Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall performance of model
Precision	$\frac{TP}{TP+FP}$	How accurate the positive predictions are
Recall Sensitivity	$\frac{TP}{TP+FN}$	Coverage of actual positive sample
Specificity	$\frac{TN}{TN+FP}$	Coverage of actual negative sample
F1 score	$\frac{2TP}{2TP+FP+FN}$	Hybrid metric useful for unbalanced classes

		Predicted class	
		+	–
Actual class	+	**TP** True Positives	**FN** False Negatives Type II error
	–	**FP** False Positives Type I error	**TN** True Negatives

ROC, AUC

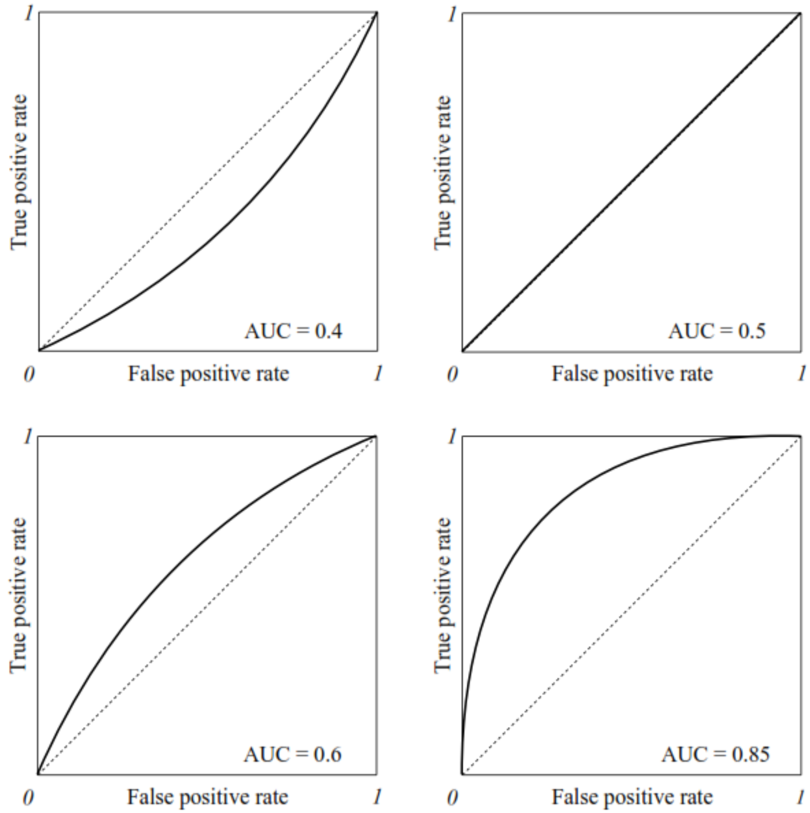
- ROC: The receiver operating curve, also noted ROC, is the plot of TPR versus FPR by varying the threshold.
- AUC: The area under the receiving operating curve, also noted AUC or AUROC

Metric	Formula	Equivalent
True Positive Rate TPR	$\frac{TP}{TP + FN}$	Recall, sensitivity
False Positive Rate FPR	$\frac{FP}{TN + FP}$	1-specificity



ROC, AUC

- ROC: The receiver operating curve, also noted ROC, is the plot of TPR versus FPR by varying the threshold.
- AUC: The area under the receiving operating curve, also noted AUC or AUROC



Hyperparameter Tuning

- Hyperparameters are defined as the parameters that are explicitly defined by the user to control the learning process.
- These are external to the model, and their values cannot be changed during the training process.
- Few examples:
  - The k in kNN or K-Nearest Neighbour algorithm
  - Learning rate for training a neural network
  - Train-test split ratio
  - Batch Size
  - Number of Epochs
  - Branches in Decision Tree
  - Number of clusters in Clustering Algorithm