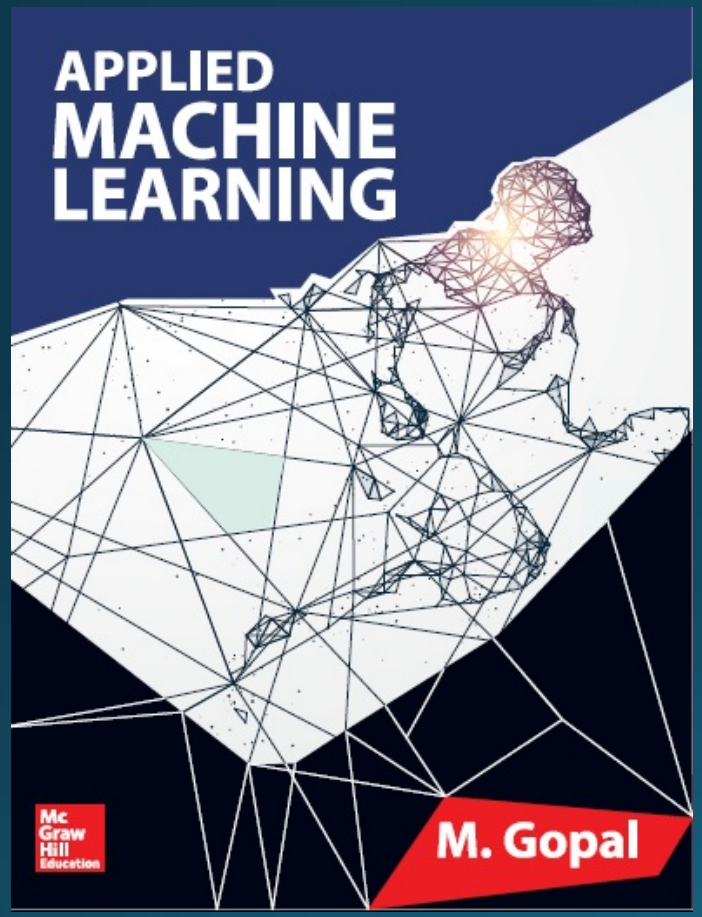


Module 2

Supervised Learning

Inductive learning

Task of inductive learning:

- Given a collection of examples $\left(\mathbf{x}^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right); i = 1,2, \ldots, N$, of a function $f(\mathbf{x})$, returns a function $h(\mathbf{x})$ that *approximates* $f(\mathbf{x})$.
- The approximating function $h(\mathbf{x})$ is called *hypothesis function*.
- The unknown *true function* $f(\mathbf{x})$ correctly maps the input space $\mathbf{X}$ (of the entire data) to the output space Y.
- Central aim of designing $h(\mathbf{x})$ is to suggest decisions for unseen patterns.

- Better approximation of $f(\cdot)$ leads to better generalization.
- *Generalization performance is the fundamental problem in inductive learning.*
- *Off-training set error*—the error on points not in the training set, is used as a measure of generalization performance.
- Inductive learning assumes that the best hypothesis regarding unseen patterns is the one *induced* by the observed training set

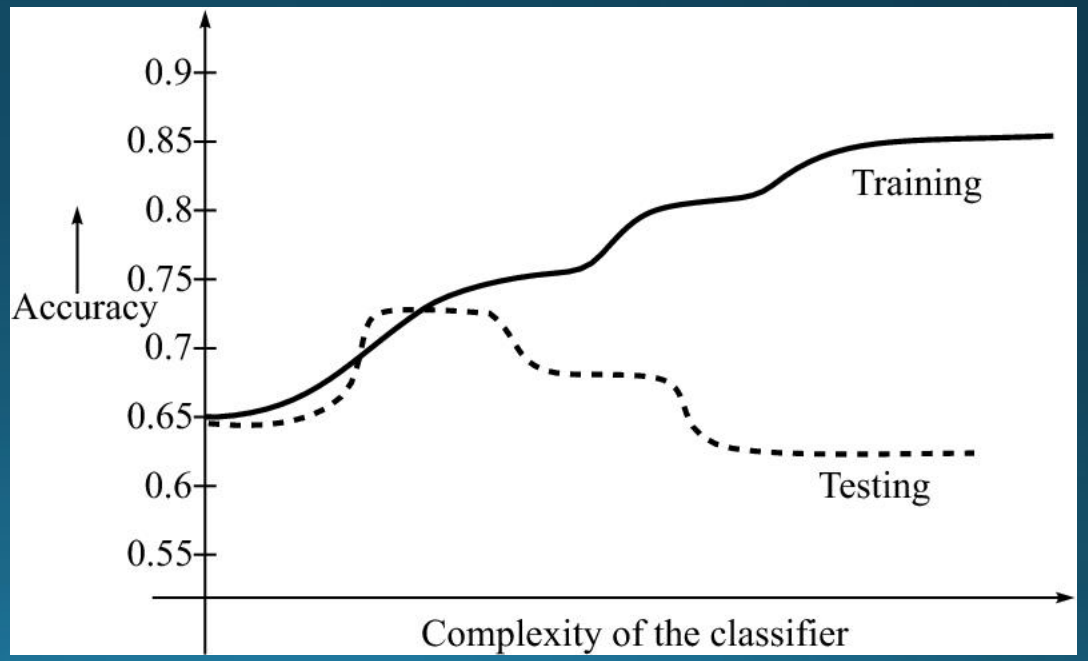
Occam's Razor Principle

- A simpler algorithm can be expected to perform better on a test set.
- "simpler" – may stand for fewer parameters, lesser training time, fewer features and so forth.
- Generally searching is stopped for a design when the solution is "good enough" and not the optimal one.
- Occam's razor principle recommends hypothesis functions that avoid *overfitting* of the training data.

Overfitting

- With increase in complexity, the training set's performance increases but performance of the test set decreases, then overfitting has happened.

The accuracy of the classifier over training examples increases monotonically as the classifier grows in complexity. However, the accuracy over the independent test examples first increases, then decreases.



Heuristic Search in Inductive Learning

Goal of machine learning:

- Not to learn exact representation of training data.
- To build a statistical model of process which generates the data.

Success of learning: Depends on hypothesis space complexity and sample complexity.

Search problem: *finding a hypothesis function of complexity consistent with the given training data*

Machine learning community depend on tools that appear to be *heuristic*, trail-and-error tools.

Estimating Generalization Errors

- Holdout method and random subsampling
  - Certain amount of data reserved for testing and rest is used for training.
  - To partition dataset $\mathcal{D}$, *randomly* sample a set of training examples from $\mathcal{D}$, and use the rest for testing.
  - For *time-series data*, use the earlier part for training and the later for testing.
  - Usually, one-third of the data is used for testing.

- This procedure of partitioning time-series data is suitable because the learning machine is used in the real world. Unseen data are from the future.
- Samples used for training and testing should have same distribution.
- It can not be identified whether a sample is representative or not since the distribution is unknown.
- Check: In classification problems, each class should be represented in about the right proportion in the training and test sets.

Mc Graw Hill Education

K-Fold Cross-Validation

- Data $\mathcal{D}$ randomly partitioned into K mutually exclusive subsets or "folds", $\mathcal{D}_k; k = 1, \ldots, K$, each of approximately equal size.
- In iteration k, partition $\mathcal{D}_k$ is test set and remaining partitions are collectively used to train the model.
- Error estimates obtained from K iterations are averaged to yield an overall error estimate.

K=10 folds is the standard number used for predicting the error rate of a learning technique.

Difference between Regression and Classification

Regression Algorithm	Classification Algorithm
In Regression, the output variable must be of continuous nature or real value.	In Classification, the output variable must be a discrete value.
The task of the regression algorithm is to map the input value (x) with the continuous output variable(y).	The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).
Regression Algorithms are used with continuous data.	Classification Algorithms are used with discrete data.
In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.

Assessing Regression Accuracy

Mean Square Error

- Most commonly used metric

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - h\left(\mathbf{w}, \mathbf{x}^{(i)}\right)\right)^2$$

Root Mean Square Error

- Same dimensions as the predicted value itself

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - h(\mathbf{w}, \mathbf{x}^{(i)})\right)^2}$$

Mc Graw Hill Education

Sum-of-Errors Squares

- Mathematical manipulation of MSE

$$Sum\text{-}of\text{-}Error\text{-}Squares = \sum_{i=1}^{N} \left(y^{(i)} - h\left(\boldsymbol{w}, \boldsymbol{x}^{(i)}\right) \right)^2$$

Assessing Classification Accuracy

- Misclassification Error
  - Metric for assessing the accuracy of classification algorithms is: *number of samples misclassified by the model* $h(\mathbf{w}, \mathbf{x})$.
  - For binary classification problems,

$$y^{(i)} \epsilon\ [0,1], \text{and } h(\mathbf{w}, \mathbf{x}) = \hat{y}^{(i)} \epsilon\ [0,1]; i = 1, \ldots, N$$

  - For 0% error, $\left(y^{(i)} - \hat{y}^{(i)}\right) = 0$ for all data points

$$Misclassification\ error = \frac{\text{Number of data points for which} \left(y^{(i)} - \hat{y}^{(i)}\right) \neq 0}{N}$$

Mc Graw Hill Education

Confusion Matrix

- Decisions made on classifications based on misclassification error rate lead to poor performance when data is *unbalanced*.
- For example, in case of financial fraud detection, the proportion of fraud cases is extremely small.
- In such classification problems, the interest is mainly in minority cases.
- The class that the user is interested in is commonly called *positive class* and the rest *negative class*.

- A single prediction on the *test set* has four possible outcomes.

1. The *true positive* (TP) and *true negative* (TN) are correct classifications.
2. A *false positive* (FP) occurs when the outcome is incorrectly predicted as positive when it is actually negative.
3. A *false negative* (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive.

		Hypothesized class (prediction)	
Actual Class (observation)		Classified +ve	Classified –ve
	Actual +ve	TP	FN
	Actual -ve	FP	TN

Confusion Matrix

- Misclassification Rate

$$Misclassification\ rate = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

True Positive Rate (*tp* rate)

$$tp\ rate \cong \frac{Positively\ correctly\ classified}{Total\ positives}$$

$$= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

- Determines sensitivity in detection of abnormal events
- Classification method with high sensitivity would rarely miss abnormal event.
- FP = FN = 0 is desired.

- True Negative Rate

$$tn\ rate \cong \frac{Negatively\ correctly\ classified}{Total\ negatives}$$

$$= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

- Determines the specificity in detection of the abnormal event
- High specificity results in low rate of false alarms caused by classification of a normal event as an abnormal one.

$$1 - specificity = 1 - \frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}$$

$$= \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$$

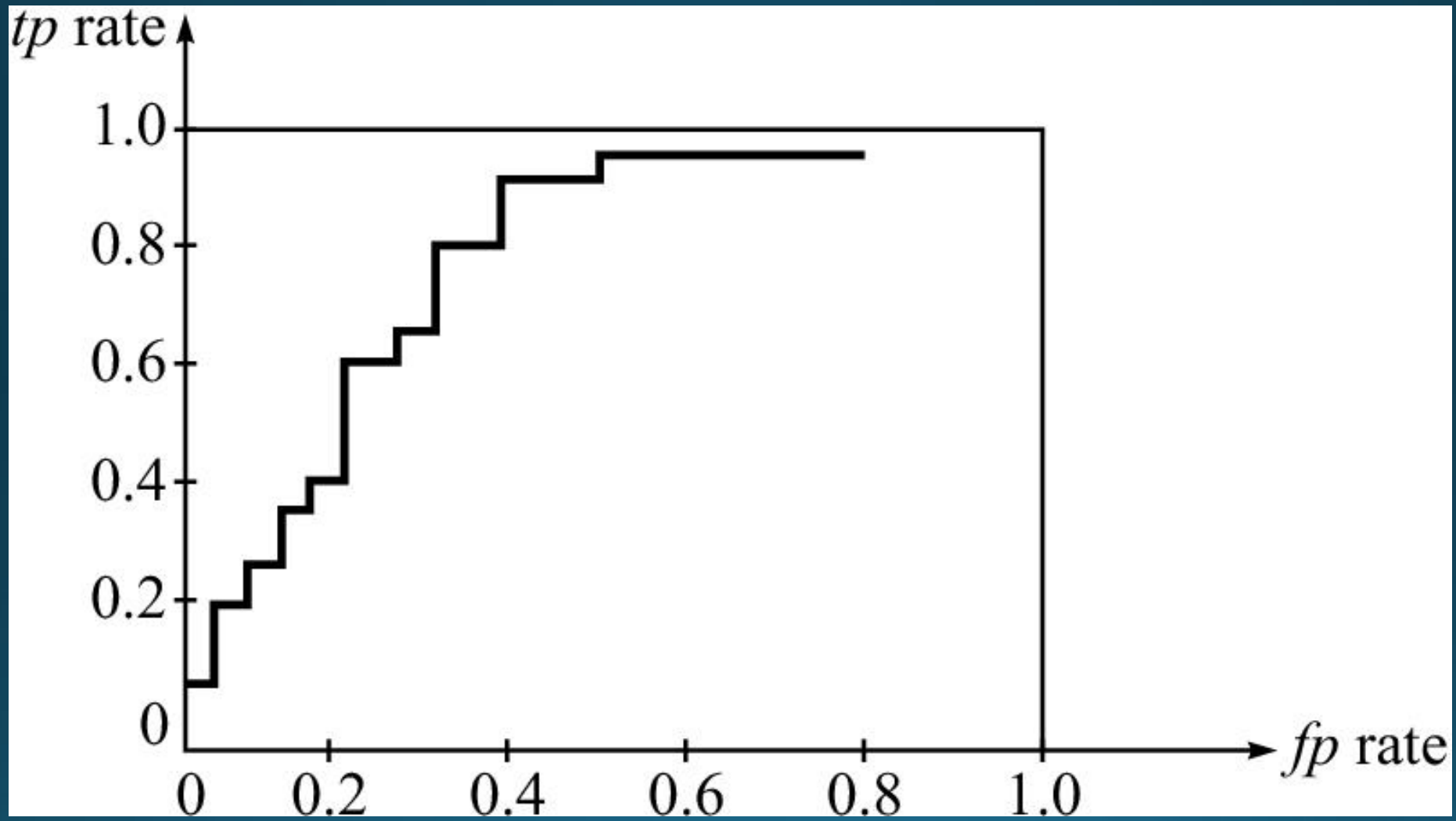
$$= \frac{Negatively\ incorrectly\ classified}{Total\ negstives}$$

$$= fp\ rate\ (\text{False positive rate})$$

- Simultaneously high sensitivity and high specificity is desired.

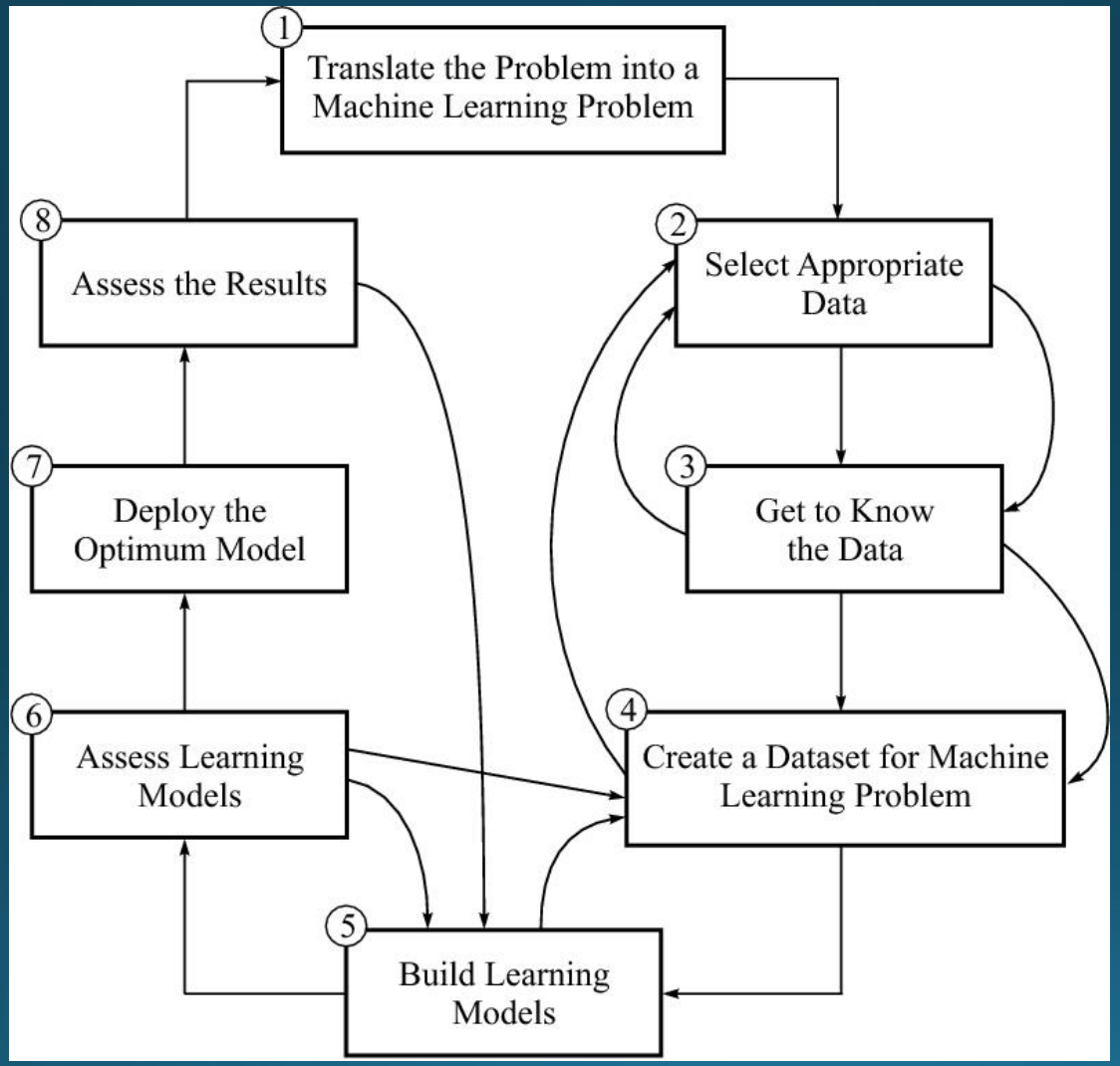
ROC Curves

- When a classifier algorithm is applied to test set, it yields a confusion matrix, which corresponds to one ROC point.
- An *ROC curve* is created by thresholding the classifier with respect to its complexity.
- Each level of complexity in the space of the hypothesis class produces a different point in the ROC space.
- Comparison of two learning schemes is done by analyzing ROC curves in the same ROC space for the learning schemes.



A sample ROC curve

An Overview of the Design Cycle



An overview of the design cycle