Chapter Two

Supervised Learning

Prepared by: Tsehay A. (B.Sc., and M.Sc., in Computer Science)

Department of Computer Science

2015 E.C

Contents

Introduction to supervised learning

- Supervised learning uses a set of labeled examples in learning to predict unseen examples.
- Input representation: we need to decide what attributes (features) to use to describe the input patterns (examples, instances).
- A training set of examples with the correct responses (targets) is provided.
- Based on the training set, the algorithm generalizes to respond correctly to all possible inputs.
- This is also called learning from examples.
- Supervised learning is a function that maps an input to an output.

Cont.

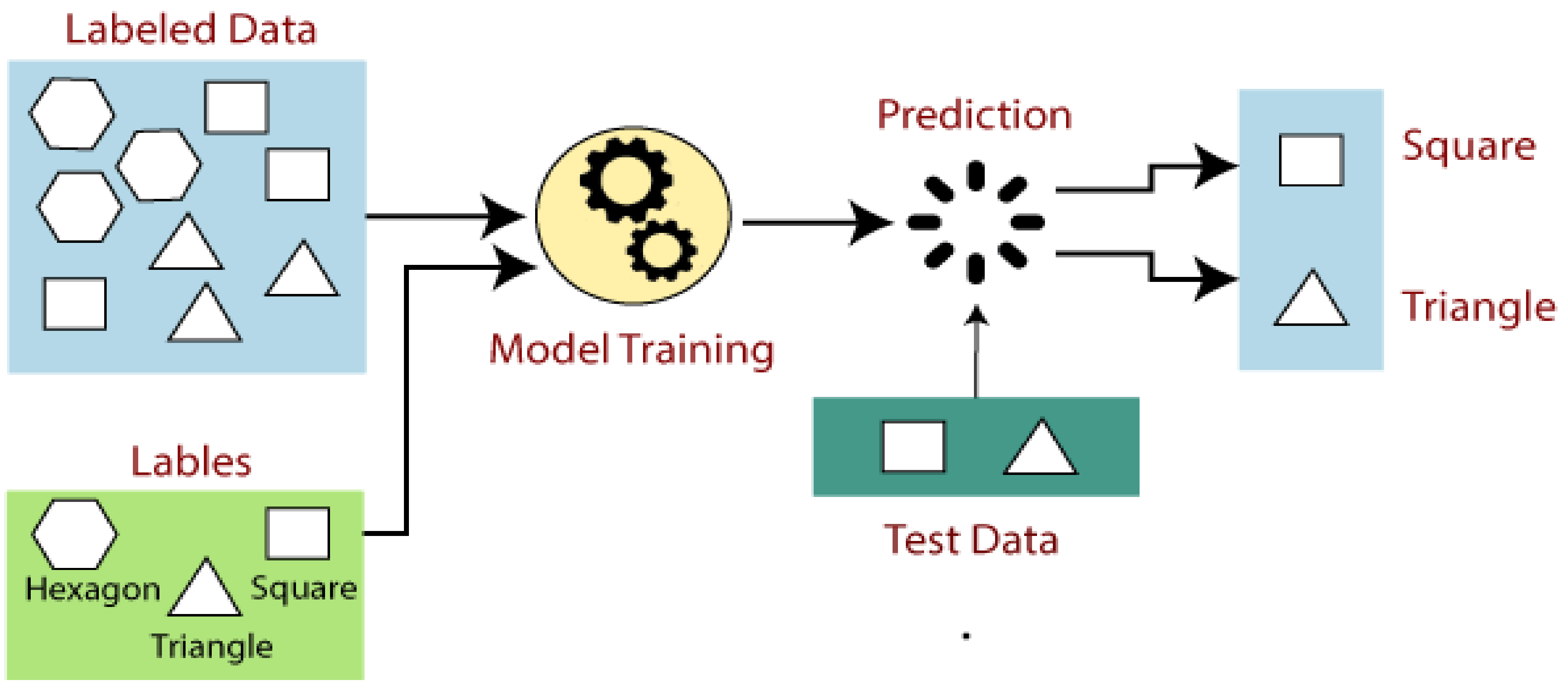


Fig. 2.1. Supervised learning

Supervised learning

- Consider the following data regarding patients entering a clinic.
- The data consists of the gender and age of the patients and each patient is labeled as "healthy" or "sick".

Table 2.1. Dataset for supervised learning

Gender	Age	Class label
Male	48	Sick
Male	67	Sick
Female	53	Healthy
Male	49	Healthy
Female	34	Sick

Linear model

- The basic type of learning could be that of fitting a straight line to data.
- Machine learning usually deals with much more flexible models than straight lines.
- Fitting a straight line to data can be used to conclude new data.
- If a model learns from a data set of 1000 puppy images, the model might tell whether another image (not among the 1000 used for learning) depicts a puppy or not.
- That is know as *generalization.*

Cont.

- Linear models are relatively simple.
- A linear model represents a function a linear combination of its inputs.
- The equation for linear function $f(x)$, is of form $f(x) = mx + c$ where c represents the intercept and m represents the slope.

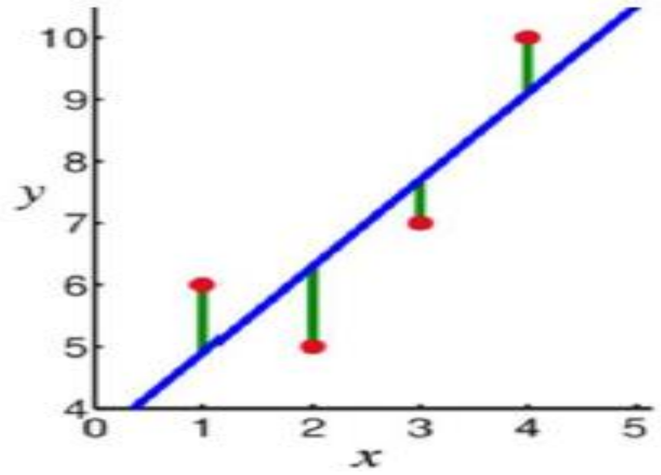


Fig. 2.2. The graph of linear function

Characteristics of liner model

- Linear models are stable.
- Linear models have low variance and high bias.
- Linear models are less likely to overfit the training data than some other models.
- However, they are more likely to underfit.
- For example, if we want to learn the boundaries between countries based on labeled data, then linear models are not likely to give a good approximation.

Regression

- Regression models the relationship between dependent and independent variables.
- Regression helps to understand how the value of the dependent variable changes corresponding to an independent variable.
- It predicts continuous or real values such as *temperature, age, salary, and price*.
- For example estimation of age from the weight of a person is a regression task.

Metrics used to evaluate regression

- Three metrics commonly used for evaluating and reporting the performance of a regression model;
- *Mean Squared Error (MSE)*: calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.
- *Root Mean Squared Error (RMSE)*: calculated as the square root of mean square error.
- *Mean Absolute Error (MAE)*: calculated as the average of the absolute MSE value.

A case study

- Design an algorithm that predicts the age of a person given the weight.
- The dataset consists of different age values and weights of people.
- Develop a regression model that predicts the age of a person based on his or her weight.

Table 2.2. Dataset of weight and height of people

Weight	**Age**
35	18
40	26
42	30
43	50
41	**?**

Y=-95.31+3*X, Thus, Y=-95.31+3*41=-95.31+123=27.69≈ 28

Classification

- The task of learning target function f that maps each attribute set x to one of the predicted class label y.
- The target function is known as *classification model*.

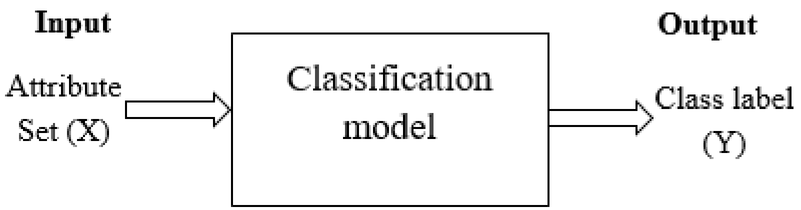


Fig. 2.3. Classification task

K-nearest neighbor (KNN)

- KNN algorithm is also called as instance based learning or lazy algorithm.
- Training phase: store the training examples.
- At prediction time: find the k training examples (x_1, y_2),…(x_k, y_k) that are closet to the test example x:
- Classification: predict the most frequent class among those y_is.
- Regression: predict the average among the y_is.

KNN-Example

- Consider that a company produced four paper tissues.
- The quality of paper tissue is classified as good or bad based on the acid durability and strength as tested in laboratory as indicated in Table 2.3.
- Determine the class label for new paper tissue produced by the company that pass laboratory test with X1=3, and X2=7.

Table 2.3. Dataset for KNN algorithm: case study

Acid durability=X1	Strength =X2	Class label=Y
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good
3	7	?

Steps in KNN prediction

- Step 1: Determine K=number of nearest neighbors.
- Step 2: Calculate distance between query instance and all training samples.
- Step 3: Rank the distance obtained in step 2.
- Step 4: Use simple majority voting of neighbors as predicted value.
- Suppose K=3, then the distance between query instance (3, 7) and all training sample is given in Table 2.4.

Table 2.4. Dataset for KNN algorithm: case study

Acid durability=X1	**Strength =X2**	**Distance**	**Rank the distance**	**Included instance**
7	7	$(7-3)^2+(7-7)^2=16$	3	Yes
7	4	$(7-3)^2+(4-7)^2=25$	4	No
3	4	$(3-3)^2+(4-7)^2=9$	1	Yes
1	4	$(1-3)^2+(4-7)^2=13$	2	Yes

Cont.

- Use simple majority of category to determine the prediction value of the query instance.
- As shown in Table 2.4, three sentence were included to determine the majority of the category, as K=3.
- Among the K nearest neighbors, 2 instances are Good, and 1 instance is bad, 2>1, thus, the new paper tissue is *Good*.

Table 2.5. The class labels of nearest neighbors

Acid durability=X1	**Strength =X2**	**Rank the distance**	**Rank the distance**	**Included instance**
7	7	3	Included instance	Bad
7	4	4	Yes	-
3	4	1	No	Good
1	4	2	Yes	Good

Decision trees

- Supervised Machine Learning where the data is continuously split according to a certain parameter.
- Decision tree is a hierarchical data structure implementing the divide-and-conquer strategy.
- It is nonparametric method, which can be used for classification and regression.
- The tree can be explained by two entities, namely *decision nodes and leaves*.
- The leaves are the decisions or the final outcomes, the decision nodes are where the data is split.

Cont.

- The basic idea behind any decision tree algorithm is as follows:

1. Select the best attribute to split the records.
2. Make that attribute a decision node and breaks the dataset into smaller subsets.
3. Start tree building by repeating this process recursively for each child until one of the conditions will match:

   ✓There are no more remaining attributes.

   ✓There are no more instances.

Decision tree example

Table 2.6. Dataset for decision tree classification

Day	Weather	Temperature	Humidity	Wind	Play football?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

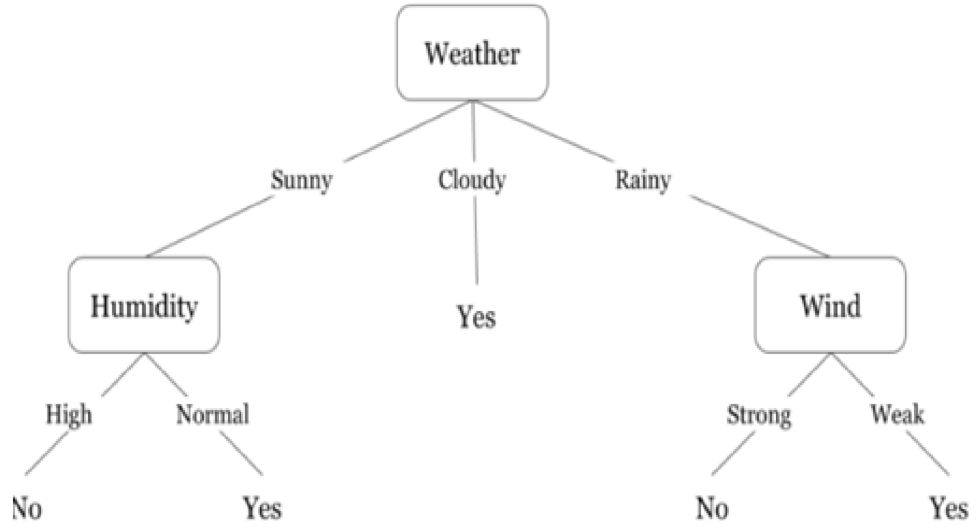


Fig. 2.4. Decision tree classifier

Ensemble learners

- Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a weighted vote of their predictions.
- Simplest approach:
    1. Generate multiple classifiers
    2. Each votes on test instance
    3. Take majority as classification

Ensemble learners

- An ensemble contains a number of learners which are usually called *base learners*.
- The generalization ability of an ensemble is stronger than that of base learners.
- Ensemble learning boosts weak learners to strong learners, and makes very accurate predictions.
- Consequently, “base learners” are also referred as “weak learners”.
- Example of ensemble learners include, Random forest, Adaboost, and Extreme boosting.

Support vector machine

- SVM is one of the most popular Supervised Learning algorithms,
- Can be used for classification as well as regression problems.
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.
- This best decision boundary is called a *hyperplane*.
- SVM chooses the extreme points/vectors that help in creating the hyperplane.
- These extreme cases are called as support vectors.

Cont.

- **Hyperplane:** The decision boundary to segregate the classes in n-dimensional space.
- **Support Vectors:** The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector

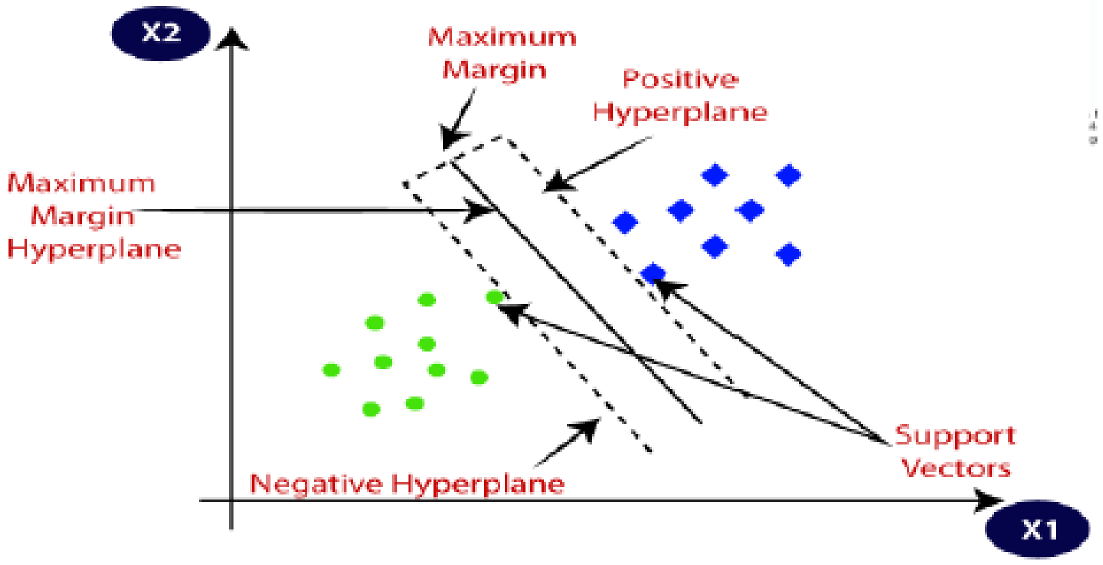


Fig. 2.5. Support vector machine

Case study classification

- Handwritten digit recognition using Classification Algorithms

Metrics used to evaluate classifiers

- There are many ways for measuring classification performance.
- Accuracy, confusion matrix, and AUC-ROC are some of the most popular metrics.
- Precision-recall is a widely used metrics for classification problems.
- Accuracy simply measures how often the classifier correctly predicts.
- Accuracy refers to the ratio of the number of correct predictions and the total number of predictions.

$$\text{Accuracy} = \frac{\mathrm{TP}+TN}{TP+TN+FP+FN} \ldots\ldots\ldots(1)$$

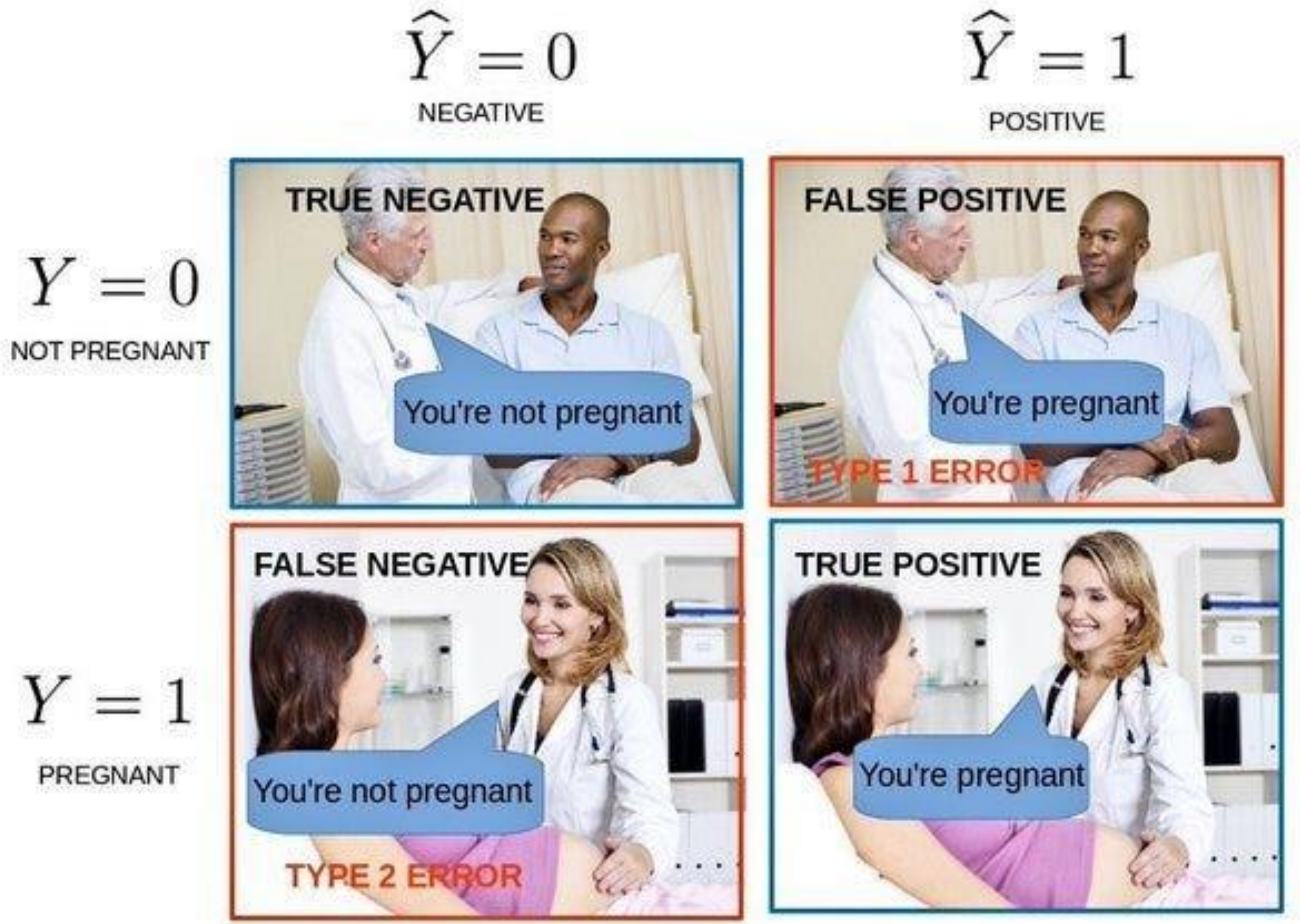
- **Recall(Sensitivity)**- explains how many of the actual positive cases we were able to predict correctly with our model

Cont.

- Confusion Matrix is a performance measurement for the machine learning classification problems where the output can be two or more classes.
- It is a table with combinations of predicted and actual values.

		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Cont.



$\hat{Y} = 0$
NEGATIVE
$\hat{Y} = 1$
POSITIVE
$Y = 0$
NOT PREGNANT
$Y = 1$
PREGNANT
TRUE NEGATIVE
You're not pregnant
FALSE POSITIVE
You're pregnant
TYPE 1 ERROR
FALSE NEGATIVE
You're not pregnant
TYPE 2 ERROR
TRUE POSITIVE
You're pregnant

Review questions

1) Define supervised learning.

2) Mention the methods for optimizing the KNN classifier.

3) What are the type of tasks that are suitable for KNN classifier.

4) Explain the concept of ensemble learning.

5) Differentiate between boosting and bagging.

6) Describe the real-world applications of linear regression model by giving an example.

7) Explain the metrics used in evaluation of the performance of classification and regression models.