MACHINE LEARNING

UNIT-1

Ingredients

of

Machine Learning

Topics

- **Definition of Machine Learning**

- **Ingredients of Machine Learning**

Machine Learning Definition

- The field of study that gives computers, the ability to learn without being explicitly programmed.

- Machine learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience.

Machine Learning Definition

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Machine Learning Definition

- **A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.**

- Classify e-mails as spam or not spam

- What type of e-mails are spam or not spam

- Number of e-mails correctly classified as spam or not spam

Ingredients of Machine Learning

- The three ingredients of Machine Learning are:
  - **Task**
  - **Model**
  - **Features**

- Machine Learning is concerned with using the right features to build the right models that achieve the right tasks.

- Models lend the machine learning field diversity, but tasks and features give it unity.

Ingredient-1: Tasks

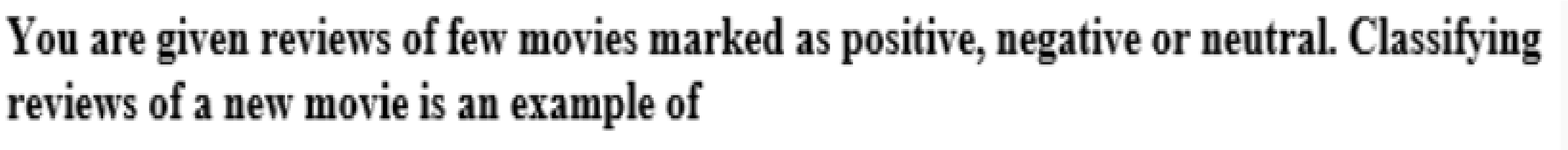
Tasks

- **Tasks are the problems that can be solved with machine learning**

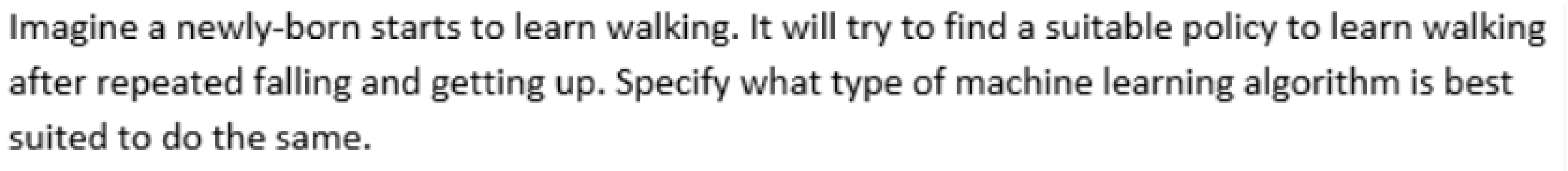
- The most common tasks in ML are
  - **Supervised Learning Tasks**
    - Classification and
    - Regression
  - **Unsupervised Learning Tasks**
    - Clustering
    - Association Rules

- Regression and Classification algorithms are Supervised Learning algorithms. But the difference between both is how they are used for different machine learning problems.

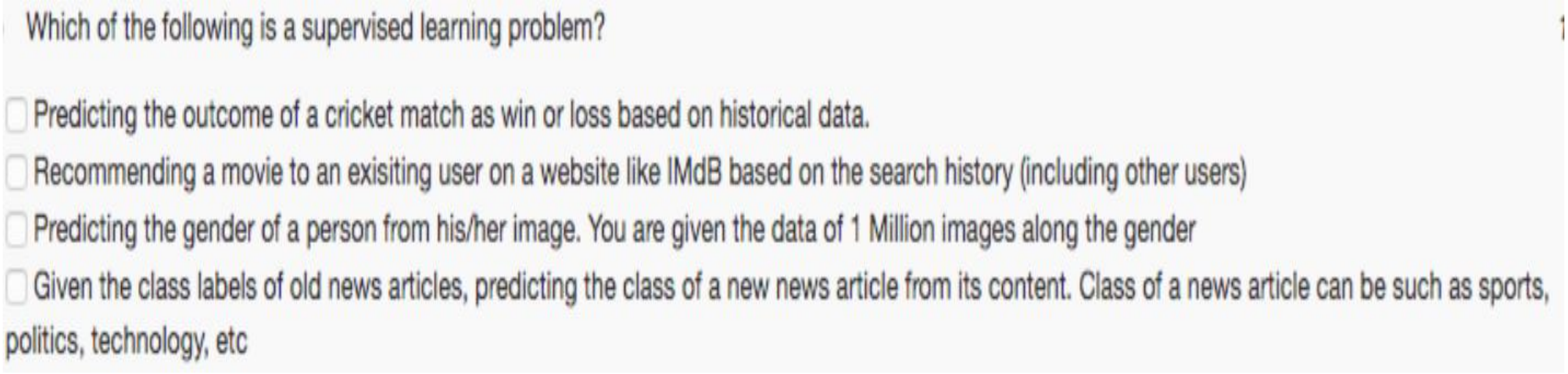
Quiz

- **Q1:** Given e-mails labeled as spam/not spam, learn a spam filter to decide whether a mail is spam or not spam.
- **Q2:** Given a set of news articles on web, group them in to set of articles about same story.
- **Q3:** Given customer data automatically discover market segments and group customers in to different market segments.
- **Q4:** Given database of patients having diabetics or not. Predict whether a new patient having diabetics or not.
- 

  You are given reviews of few movies marked as positive, negative or neutral. Classifying reviews of a new movie is an example of

- 

  Imagine a newly-born starts to learn walking. It will try to find a suitable policy to learn walking after repeated falling and getting up. Specify what type of machine learning algorithm is best suited to do the same.



Which of the following is a supervised learning problem?

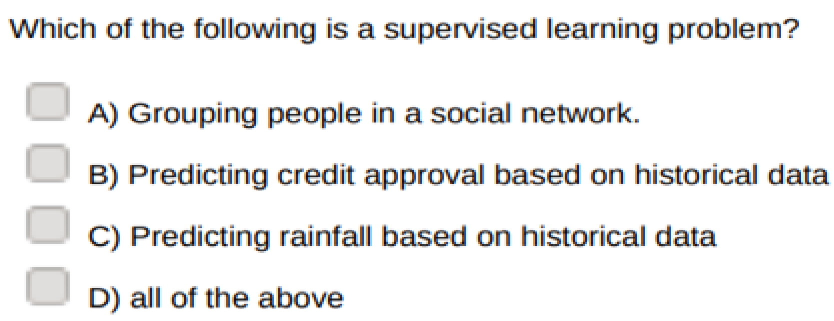
□ Predicting the outcome of a cricket match as win or loss based on historical data.

□ Recommending a movie to an exisiting user on a website like IMdB based on the search history (including other users)

□ Predicting the gender of a person from his/her image. You are given the data of 1 Million images along the gender

□ Given the class labels of old news articles, predicting the class of a new news article from its content. Class of a news article can be such as sports, politics, technology, etc

Quiz



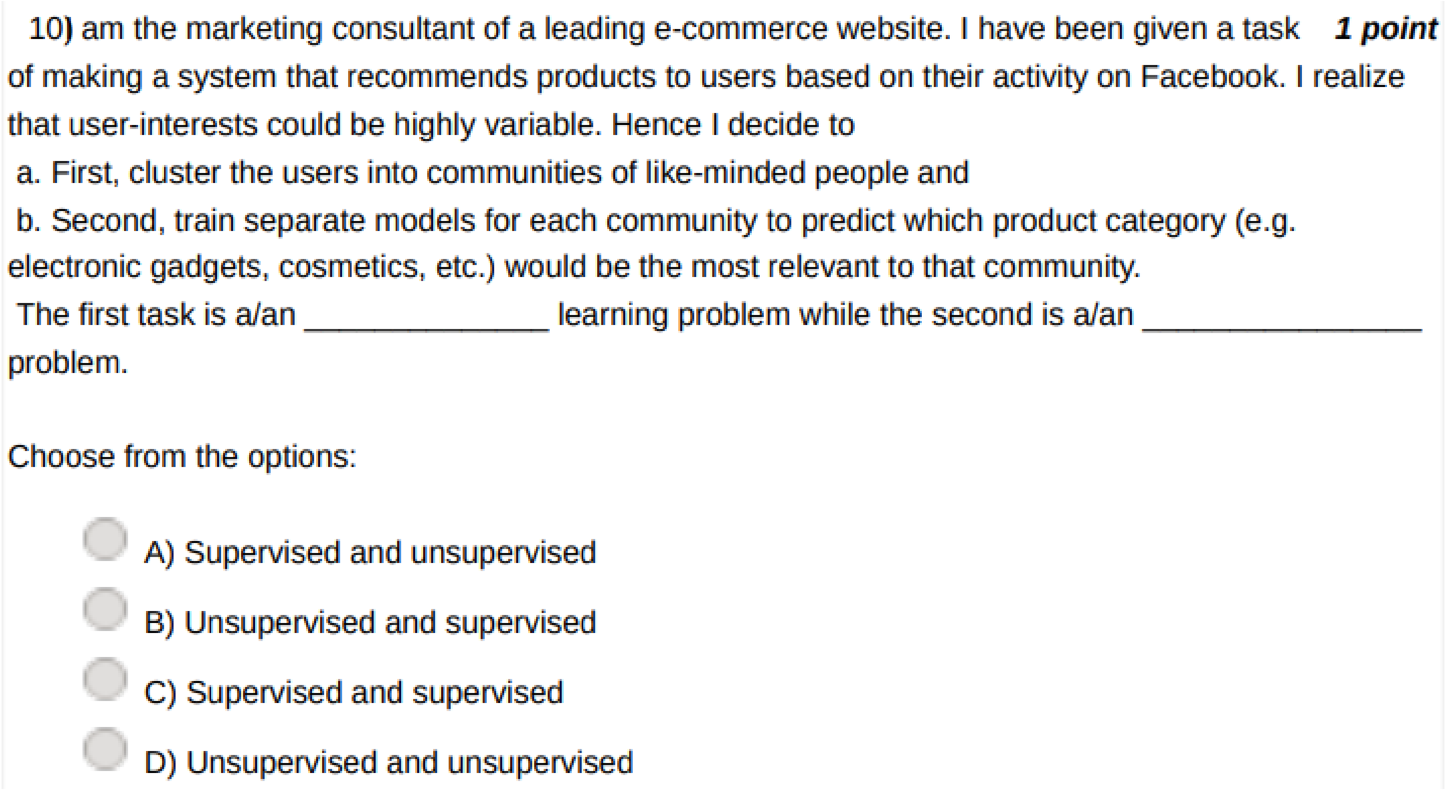
Which of the following is a supervised learning problem?

□ A) Grouping people in a social network.

□ B) Predicting credit approval based on historical data

□ C) Predicting rainfall based on historical data

□ D) all of the above



10**)** am the marketing consultant of a leading e-commerce website. I have been given a task ***1 point***

of making a system that recommends products to users based on their activity on Facebook. I realize that user-interests could be highly variable. Hence I decide to

a. First, cluster the users into communities of like-minded people and

b. Second, train separate models for each community to predict which product category (e.g. electronic gadgets, cosmetics, etc.) would be the most relevant to that community.

The first task is a/an ______________ learning problem while the second is a/an ________________ problem.

Choose from the options:

○ A) Supervised and unsupervised

○ B) Unsupervised and supervised

○ C) Supervised and supervised

○ D) Unsupervised and unsupervised

Tasks-classification

- Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters, i.e., The task of the classification is to find the **mapping function to map the input(x) to the discrete output(y).**

- In Classification, a computer program is trained on the training dataset and based on that training, it categorizes the data into different classes.

- **Example:** The best example to understand the Classification problem is Email Spam Detection.

- **The most common types of classification are**
  - Binary classification
  - Non-binary in terms of two binary classifications
  - Multi-class Classification

Tasks-Regression

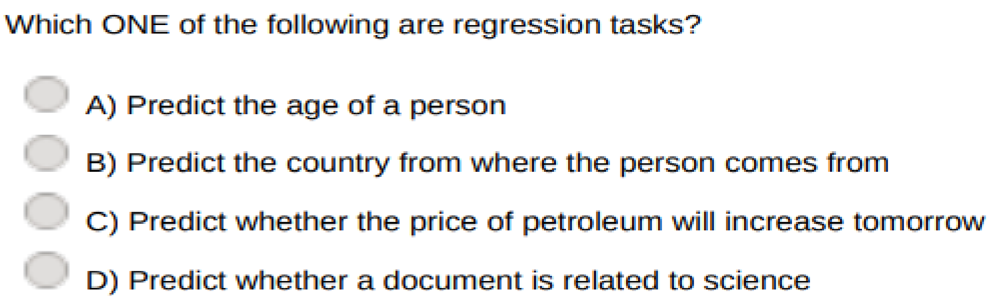
- Regression is a process of finding the correlations between dependent and independent variables.

- The task of the Regression algorithm is to **find the mapping function to map the input variable(x) to the continuous output variable(y).**

- It helps in predicting the continuous variables such as prediction of **Market Trends**, **prediction of tomorrow temparature** etc.

- **Example:** Suppose we want to do weather forecasting. In weather prediction, the model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.

Quiz

- 

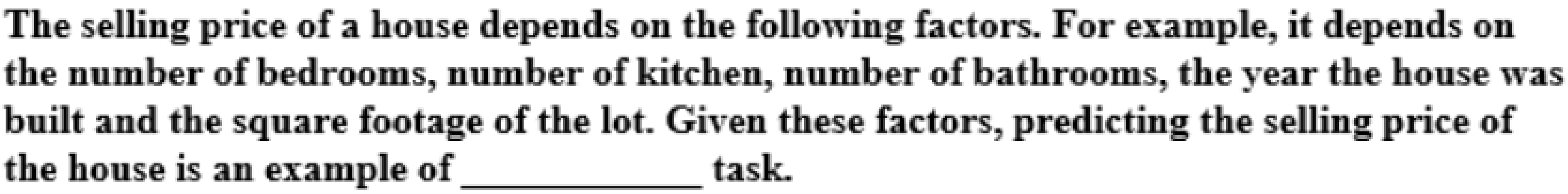
Which ONE of the following are regression tasks?

○ A) Predict the age of a person

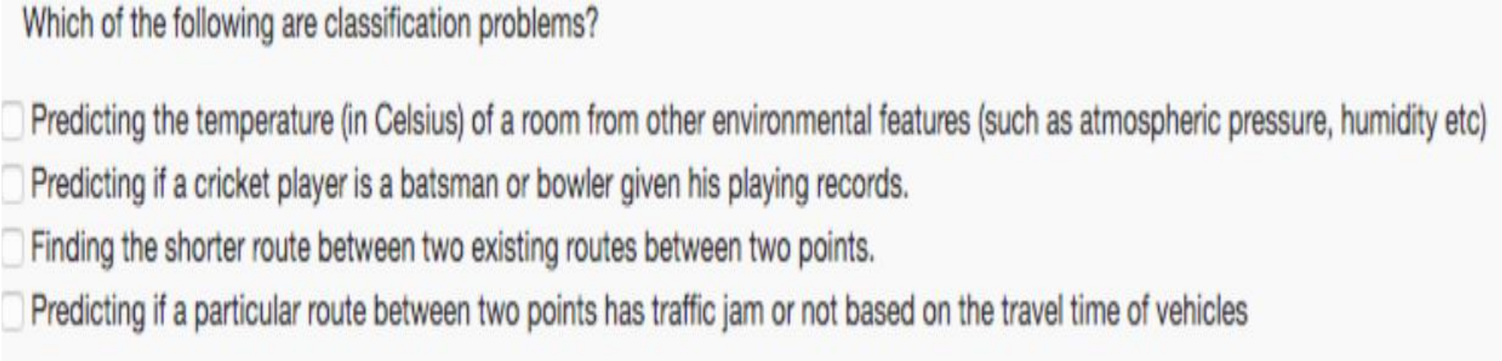
○ B) Predict the country from where the person comes from

○ C) Predict whether the price of petroleum will increase tomorrow

○ D) Predict whether a document is related to science



The selling price of a house depends on the following factors. For example, it depends on the number of bedrooms, number of kitchen, number of bathrooms, the year the house was built and the square footage of the lot. Given these factors, predicting the selling price of the house is an example of ____________ task.



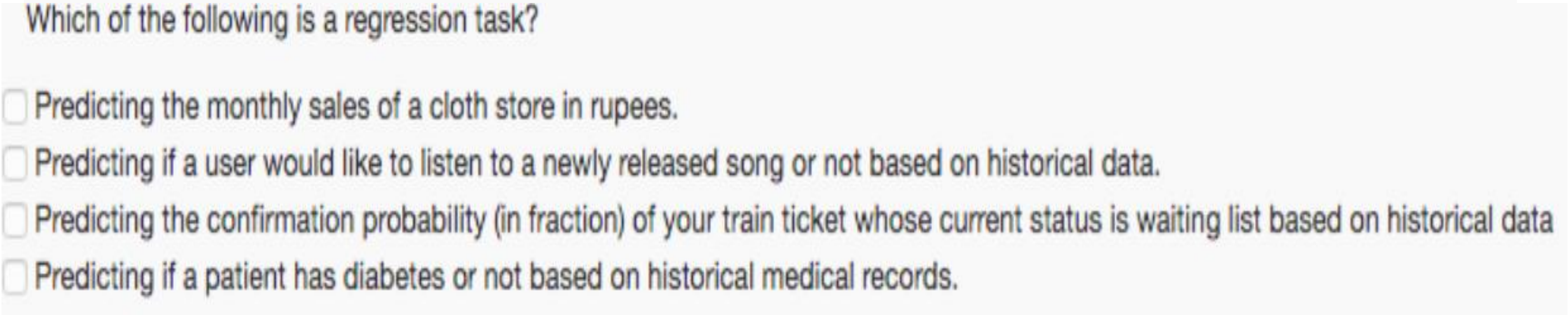
Which of the following are classification problems?

□ Predicting the temperature (in Celsius) of a room from other environmental features (such as atmospheric pressure, humidity etc)

□ Predicting if a cricket player is a batsman or bowler given his playing records.

□ Finding the shorter route between two existing routes between two points.

□ Predicting if a particular route between two points has traffic jam or not based on the travel time of vehicles



Which of the following is a regression task?

□ Predicting the monthly sales of a cloth store in rupees.

□ Predicting if a user would like to listen to a newly released song or not based on historical data.

□ Predicting the confirmation probability (in fraction) of your train ticket whose current status is waiting list based on historical data

□ Predicting if a patient has diabetes or not based on historical medical records.

Tasks-Clustering

- **Can we learn to separate the data without a labelled training set?**

- **The task of grouping data without prior information (labels) of data is called *clustering*.**

- clustering algorithm works by assessing the similarity between instances (the things we’re trying to cluster) and putting similar instances in the same cluster and ‘dissimilar’ instances in different clusters.

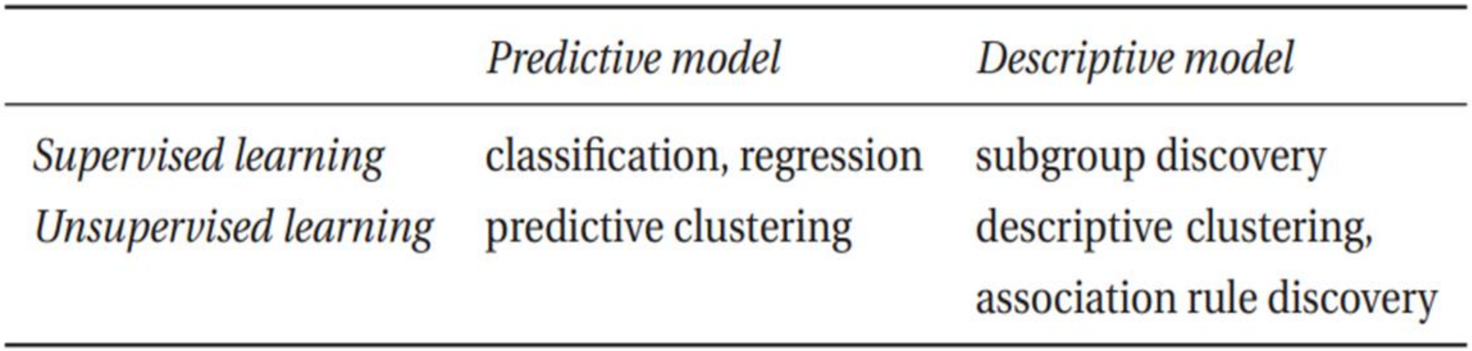
Tasks-Evaluating the Performance

- Test Data
- Cross-validation
- Performance measures

Ingredient-2: Models

Models

- **Models are the output of machine learning.** Models form the central concept in machine learning as they are **what is being learned from the data, in order to solve a given task**.
- **Categorization of Models:** Predictive Vs Descriptive
- whether the model output involves the target variable or not?
- Predictive model if it does, and a descriptive model if it does not.



	Predictive model	*Descriptive model*
Supervised learning	classification, regression	subgroup discovery
Unsupervised learning	predictive clustering	descriptive clustering, association rule discovery

Models

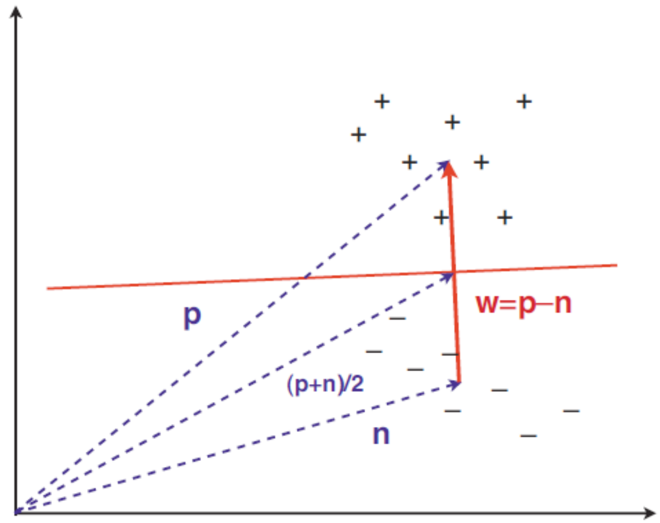
- Another Categorization of Models: based on the process used

- **Three groups of models:**
  - Geometric models
  - probabilistic models and
  - logical models

Geometric Model

- The ***instance space*** is the set of all possible or describable instances, whether they are present in our data set or not.
- A geometric model is constructed directly in instance space, **using geometric concepts such as lines, planes and distances.**
- Geometric Models are Easy to Visualize.
- Geometric concepts that potentially apply to high-dimensional spaces are usually prefixed with ‘hyper-’: for instance,
- A decision boundary in an unspecified number of dimensions is called a *hyperplane*.

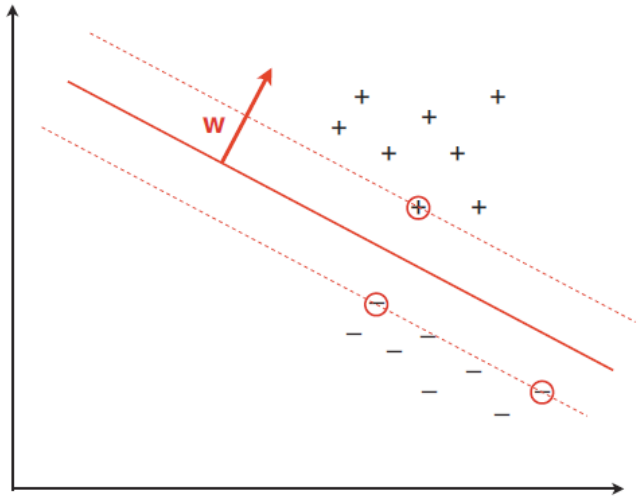
Geometric Models



Basic Linear Classifier

w·**x** =*t*, with **w**= **p**–**n**; the decision threshold can be found by (**p**+**n**)/2

P= Center of mass of +
N= Center of mass of -



Support Vector Machine

$$\mathbf{w}^T\mathbf{x} + b = 0$$

- **w is a weight vector**
- **x is input vector**
- **b is bias**

Allows us to write

$$\mathbf{w}^T\mathbf{x} + b \geq 0 \text{ for } d_i = +1$$
$$\mathbf{w}^T\mathbf{x} + b < 0 \text{ for } d_i = -1$$

Geometric Model

- A very useful geometric concept in machine learning is the notion of *distance*.

- Euclidian distance ED=$\sqrt{\sum(x_i - y_i)^2}$

- Manhattan distance MD=$\sum|x_i - y_i|$

- Examples
  - Classification :*nearest-neighbour classifier*.
  - Clustering: K-means

Probabilistic Models

- Let X denote the variables we know about, (feature values) and let Y denote the target variables we’re interested in, (the instance’s class).

- The key question in machine learning is how to model the relationship between X and Y i.e., conditional probability $P(Y|X)$.

- This is called a *posterior probability* because it is used *after* the features X

- There exists another conditional probability, called as *likelihood function* $P(X|Y)$ are observed.

Probabilistic Models: Bayes Theorem

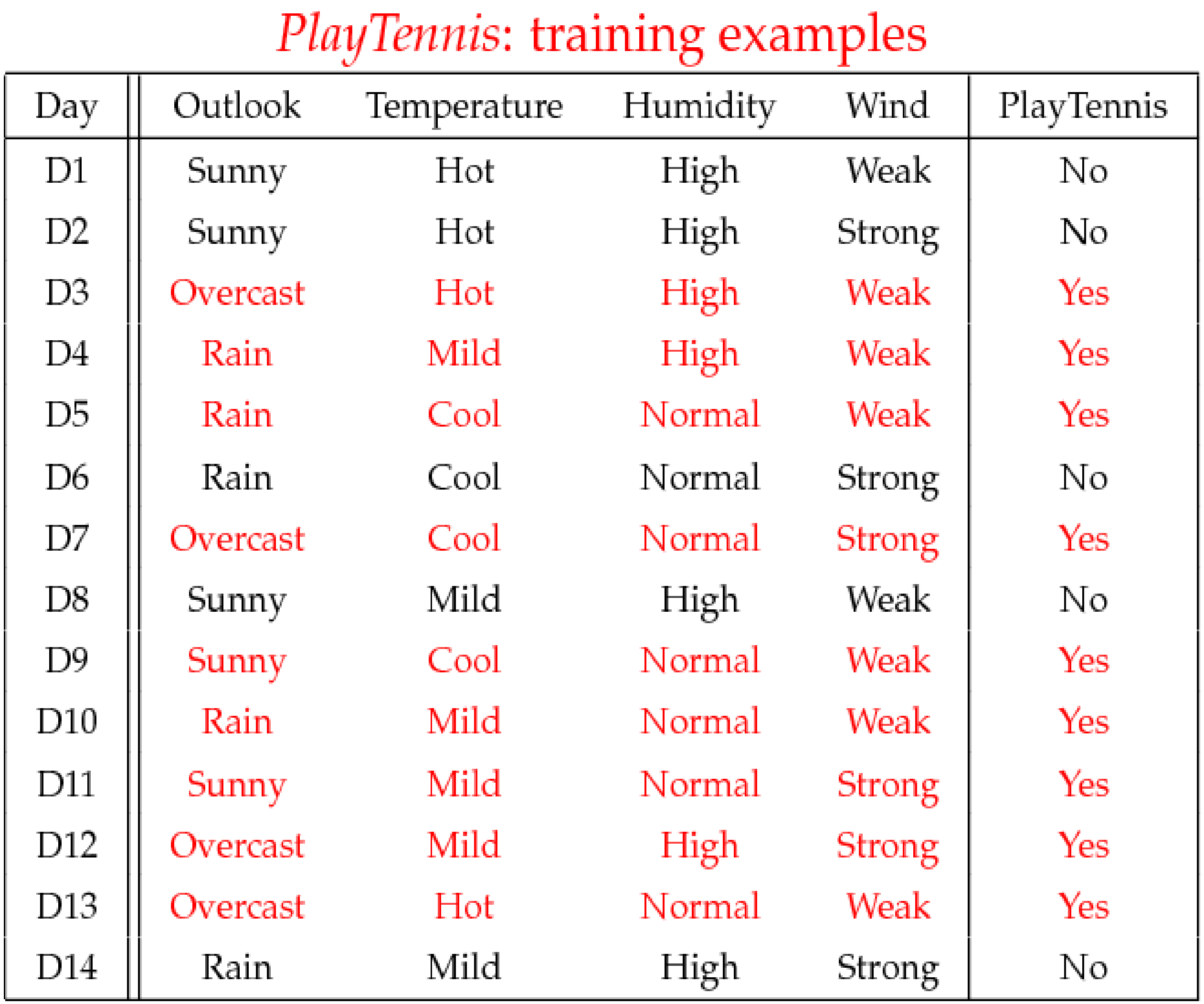
- $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
  - P(A | B) Posterior probability
  - P(A) Prior Probability
  - P(B | A) Likellyhood

- Maximum A Posteriori (MAP)

$$y_{MAP} = \underset{Y}{\arg\max} P(Y|X) = \underset{Y}{\arg\max} \frac{P(X|Y)P(Y)}{P(X)} = \underset{Y}{\arg\max} P(X|Y)P(Y)$$

- Example:
  - Naïve Bayes Classifier

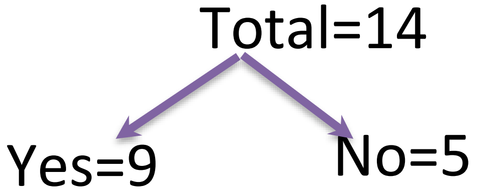
Example for Probabilistic Model



PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

D15 Sunny Cool High Strong ?

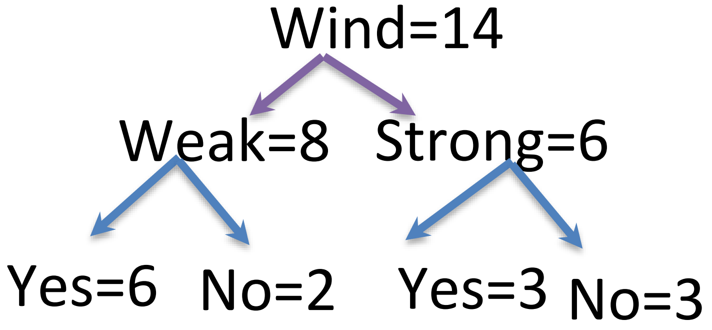


Total=14

Yes=9

No=5

P(Yes)=9/14 P(No)=5/14



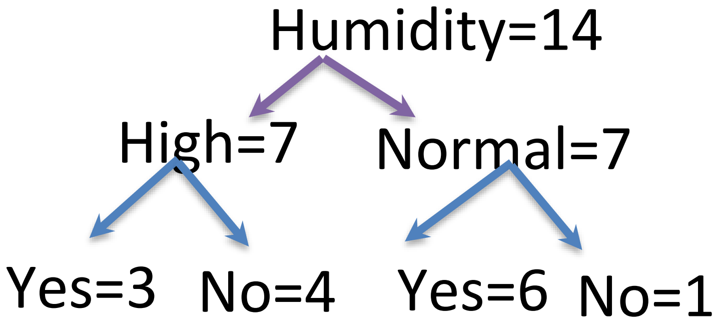
Wind=14

Weak=8 Strong=6

Yes=6 No=2 Yes=3 No=3

P(Weak/Yes)=6/9 P(Strong/Yes)=3/9

P(Weak/No)=2/5 P(Strong/No)=3/5



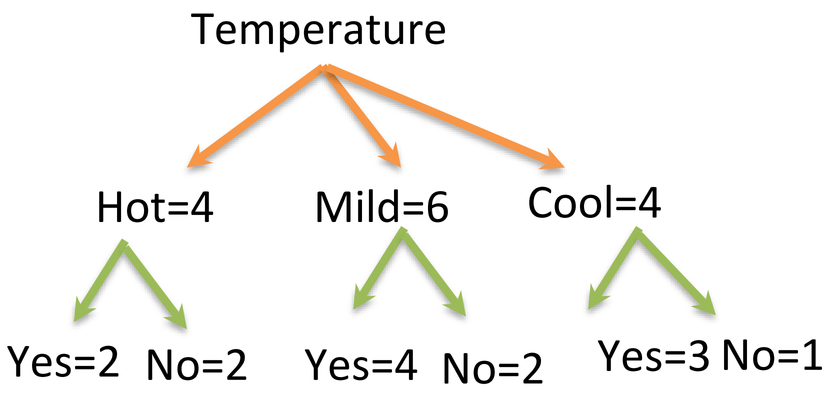
Humidity=14

High=7 Normal=7

Yes=3 No=4 Yes=6 No=1

P(High/Yes)=3/9 P(Normal/Yes)=6/9

P(High/No)=4/5 P(Normal/No)=1/5



P(Hot/Yes)=2/9

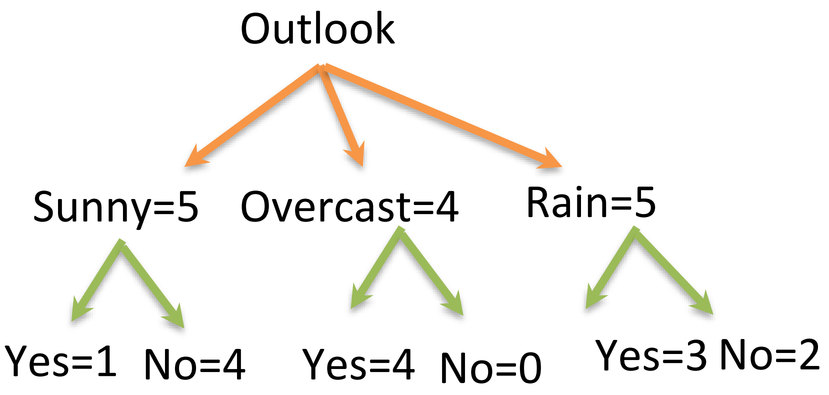
P(Hot/No)=2/5

P(Mild/Yes)=4/5

P(Mild/No)=2/5

P(Cool/Yes)=3/9

P(Cool/No)=1/5



P(Sunny/Yes)=1/9

P(Sunny/No)=4/5

P(Overcast/Yes)=4/9

P(Overcast/No)=0/5

P(Rain/Yes)=3/9

P(Rain/No)=2/5

Example for Probabilistic Model

- X={sunny, cool, high, strong}

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}$$

- P(X/Yes)=

  P(Yes)*P(Sunny/Yes))*P(Cool/Yes) *P(high/Yes) *P(strong/Yes)

  =0.0053

- P(X/No)=

  P(No)*P(Sunny/No)*P(Cool/No) *P(high/No) *P(strong/No)

  =0.206

- Maximum A Posteriori (MAP): $y_{MAP}=\underset{Y}{\arg\max}\,P(Y|X)=\underset{Y}{\arg\max}\frac{P(X|Y)P(Y)}{P(X)}=\underset{Y}{\arg\max}\,P(X|Y)P(Y)$

According to Majority class Rule: Max(0.0053,0.206)=0.206

The Answer is NO

Logical Models

- The third type of model are more algorithmic in nature.

- They are called **‘logical’ because models of this type can be easily translated into rules that are understandable by humans**, such as “If *condition* then *class = C1*.”

- Such rules are easily organized from a tree structure, which is called as feature tree.

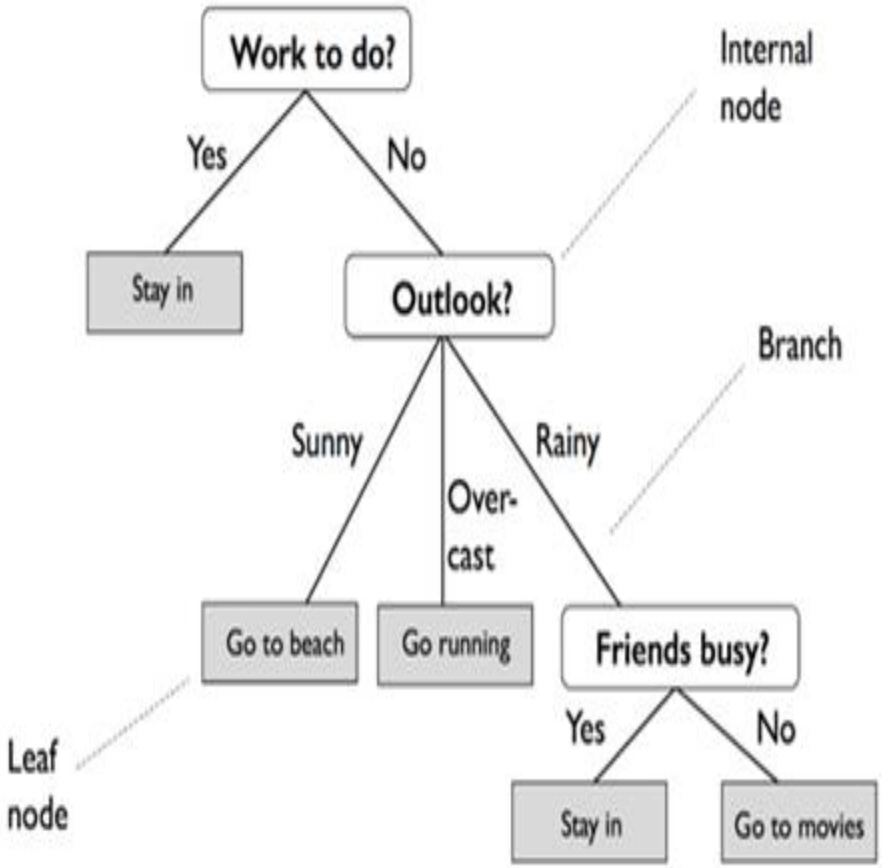
Logical Models

- Feature Tree
  - Features are used to iteratively partition the instance space.
  - Leaves denote instance class.

- Feature Trees whose leaves are labeled with classes area commonly called as decision trees.

- Example:
  - Decision Tree Induction Algorithm

Example for Logical Models



- Disjunction(or,V) or Conjunction(and,^) are used to form the rules from the feature tree.

Logical Models

- An interesting aspect of logical models, which sets them aside from most geometric and probabilistic models, is that they can, to some extent, provide *explanations* for their predictions.

- For example, a prediction assigned by a decision tree could be explained by reading off the conditions that led to the prediction from root to leaf.

- The model itself can also easily be inspected by humans, which is why they are sometimes called *declarative*.

Another Categorization of Models

- Based om the way of handling the instance space.
- Groping and Grading Models.
- Grouping models do this by breaking up the instance space into groups or *segments*, the number of which is determined at training time.
- Grading models, on the other hand, do not employ such a notion of segment. Rather than applying very simple, local models, they form one global model over the instance space.

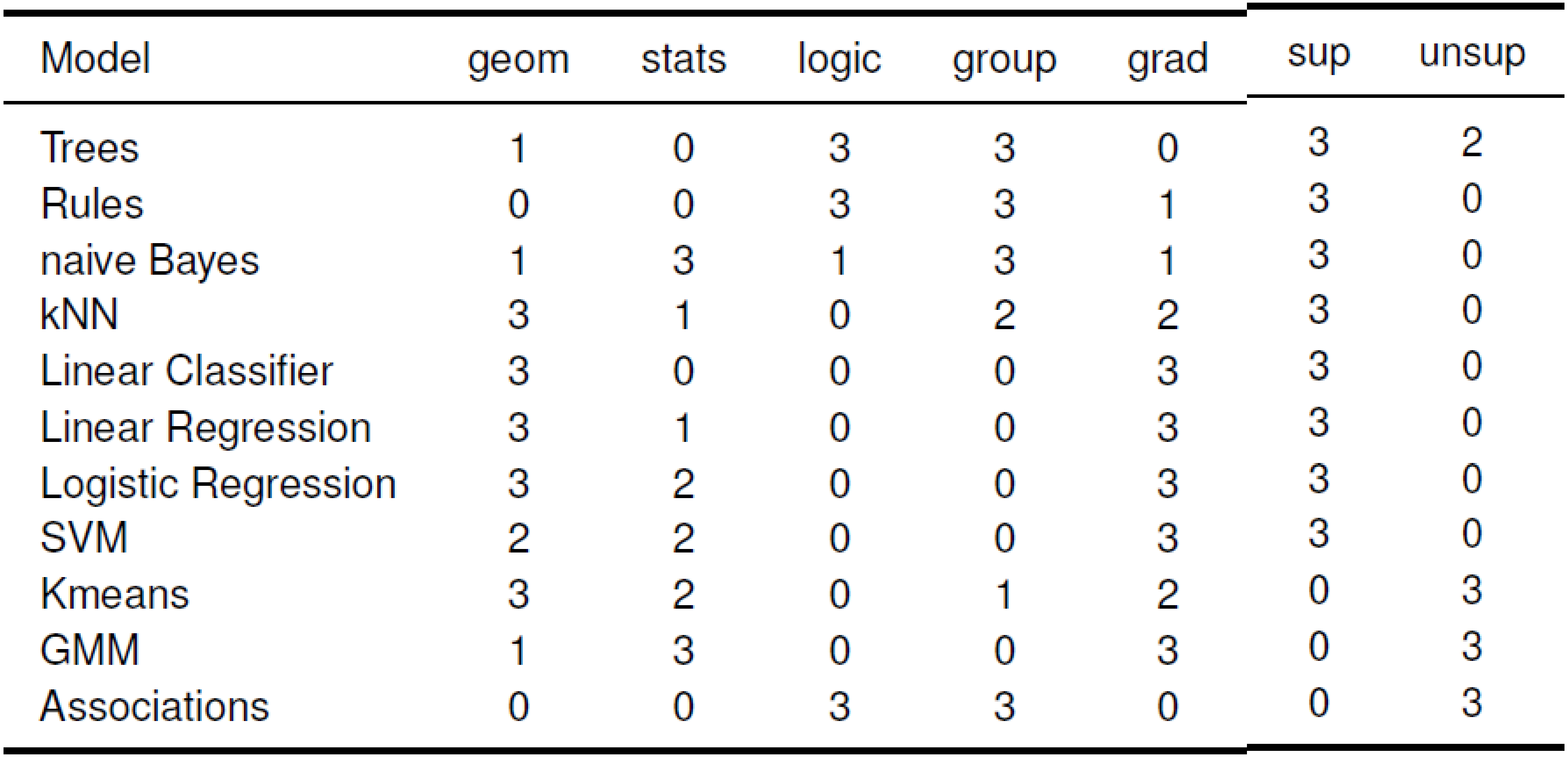
Algorithms Vs Models

	Geometric	Probabilistic	Logical	Grouping	Grading
Decision Trees					
Naïve Bayes					
SVM					
K-means					
Associations					
ANN					

Algorithms Vs Models

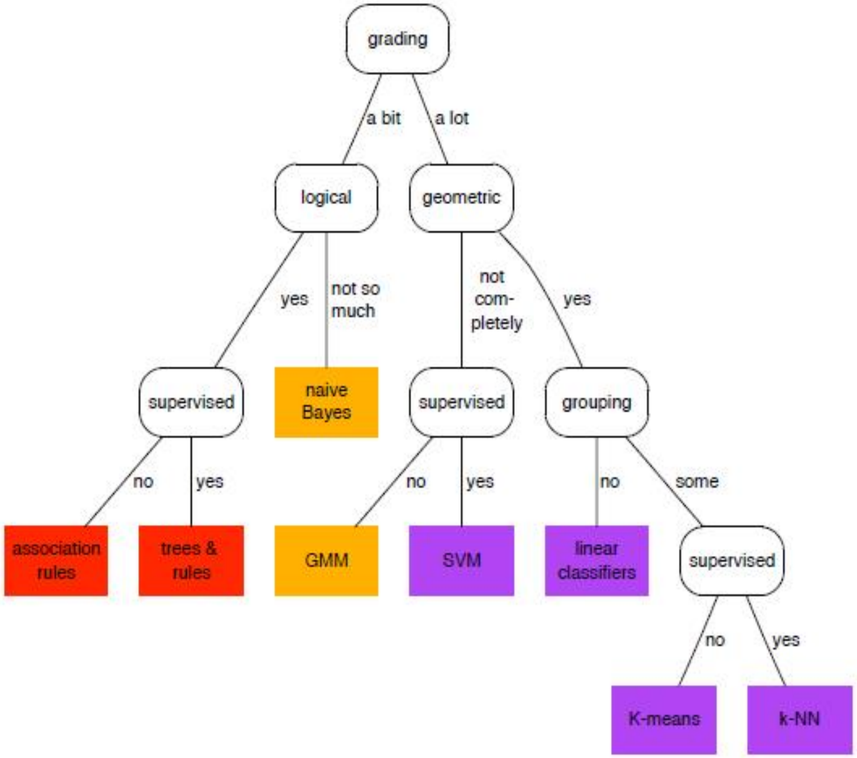
	Geometric	Probabilistic	Logical	Grouping	Grading
Decision Trees			✓	✓	
Naïve Bayes		✓			✓
SVM	✓				✓
K-means	✓				✓
Associations			✓	✓	
ANN			✓		✓

Algorithms Vs Models



Model	geom	stats	logic	group	grad	sup	unsup
Trees	1	0	3	3	0	3	2
Rules	0	0	3	3	1	3	0
naive Bayes	1	3	1	3	1	3	0
kNN	3	1	0	2	2	3	0
Linear Classifier	3	0	0	0	3	3	0
Linear Regression	3	1	0	0	3	3	0
Logistic Regression	3	2	0	0	3	3	0
SVM	2	2	0	0	3	3	0
Kmeans	3	2	0	1	2	0	3
GMM	1	3	0	0	3	0	3
Associations	0	0	3	3	0	0	3

Algorithms Vs Models



A taxonomy describing machine learning methods in terms of the extent to which they are grading or grouping models, logical, geometric or a combination, and supervised or unsupervised. The colours indicate the type of model, from left to right: logical (red), probabilistic (orange) and geometric (purple).

Ingredient 3: Features

Features

- **Features: the workhorses of machine learning**
- Features determine much of the success of a machine learning application, because a **model is only as good as its features.**
- A feature can be thought of as a kind of measurement that can be easily performed on any instance.
- *domain* of the feature
- Univariate Model

Features-Example

Category	Features
Housing Prices	No. of Rooms, House Area, Air Pollution, Distance from facilities, Economic Index city, Security Ranking etc.
Spam Detection	presence or absence of certain email headers, the email structure, the language, the frequency of specific terms, the grammatical correctness of the text etc.
Speech Recognition	noise ratios, length of sounds, relative power of sounds, filter matches
Cancer Detection	Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion, Single epithelial cell size, Number of bare nuclei, Bland chromatin, Number of normal nuclei, Mitosis etc.
Cyber Attacks	IP address, Timings, Location, Type of communication, traffic details
Video Recommendations	Text matches, Ranking of the video, Interest overlap, history of seen videos, browsing patterns etc.
Image Classification	Pixel values, Curves, Edges etc.

Two uses of Features

- **Use-1: Feature as a Split.**
  - features are used to zoom in on a particular area of the instance space.
- Let *f* be a feature counting the number of occurrences of the word ‘xxxx’ in an e-mail, and let *x* stand for an arbitrary e-mail
  - Condition $f(x) = 0$ selects e-mails that don’t contain the word ‘xxxx’
  - $f(x)$!=0 selects e-mails that do
  - $f(x) \geq 2$ selects e-mails that contain the word at least twice.
- Such conditions are called *binary splits*, because they divide the instance space into two groups: those that satisfy the condition, and those that don’t.
- Non-binary splits are also possible: for instance, if *g* is a feature that has the word ‘tweet’
  - ‘Tiny’ for e-mails with up to 20 words
  - ‘short’ for e-mails with 21 to 50 words
  - ‘medium’ for e-mails with 51 to 200 words
  - ‘long’ for e-mails with more than 200 words
  - Then the expression $g(x)$ represents a four-way split of the instance space.

Two uses of Features

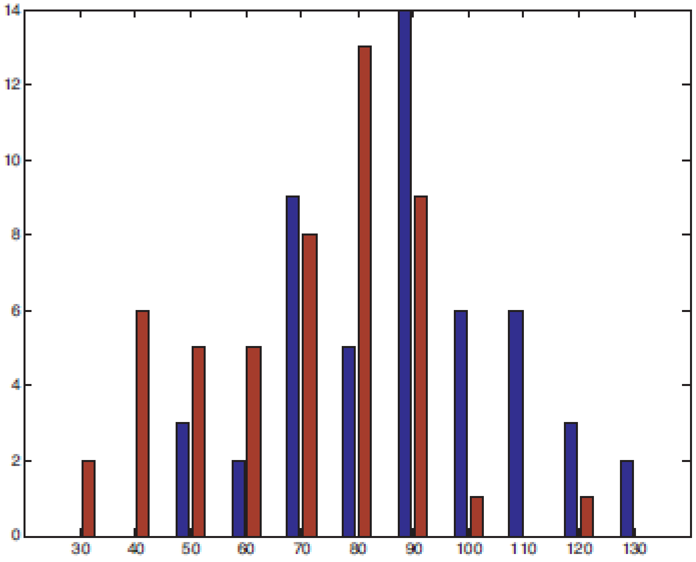
- **Use-2: Features as a Predictors**
- linear classifier employs a decision rule of the form $\sum w_i x_i > t$, where x_i is a numerical feature.

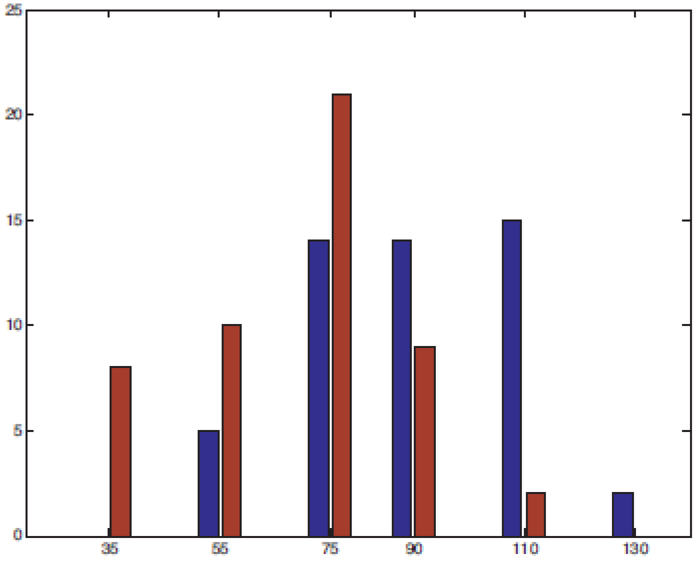
- The linearity of this decision rule means that each feature makes an independent contribution to the score of an instance.
  - This contribution depends on the weight w_i :
  - if this is large and positive, a positive x_i increases the score;
  - if *wi* <0, a positive x_i decreases the score;
  - if *wi* ≈ 0, *xi* ’s influence is negligible.

- Thus, the feature makes a precise and measurable contribution to the final prediction.

Feature Construction and Transformation

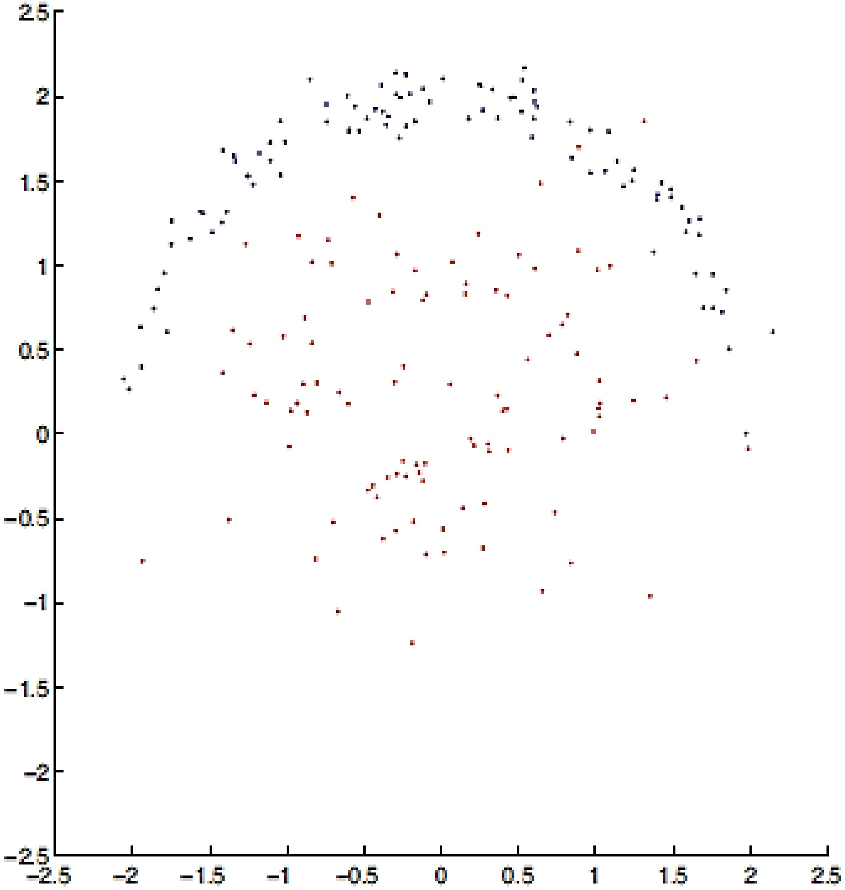
- It is often natural to build a model in terms of the given features. However, we are free to change the features, or even to introduce new features.
- For instance, real-valued features often contain unnecessary detail that can be removed by *discretization*.

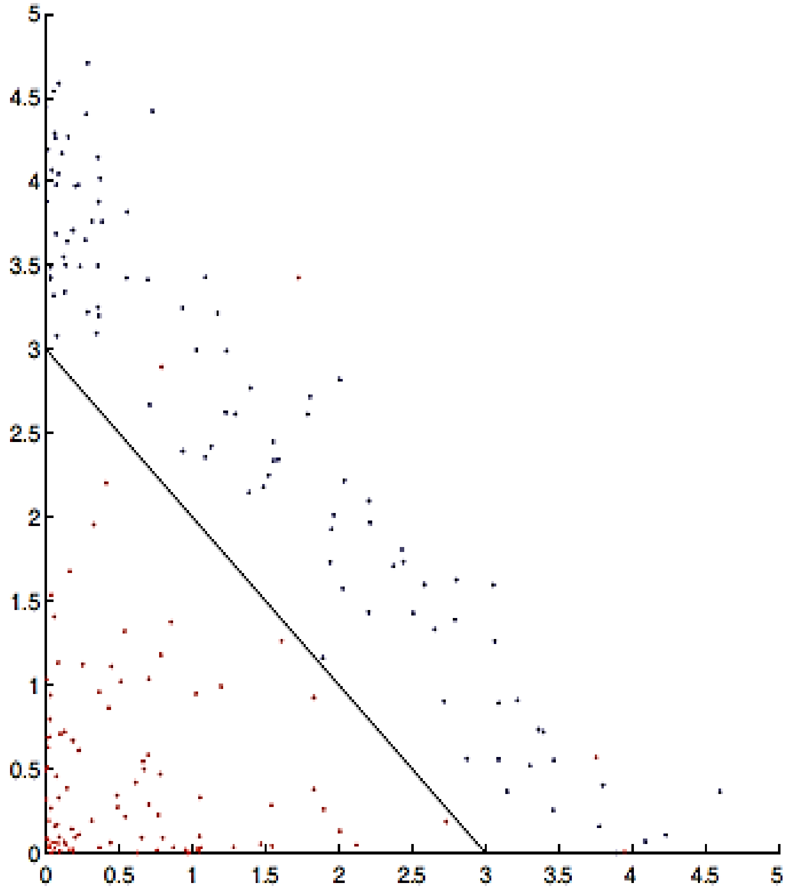




Feature Construction and Transformation

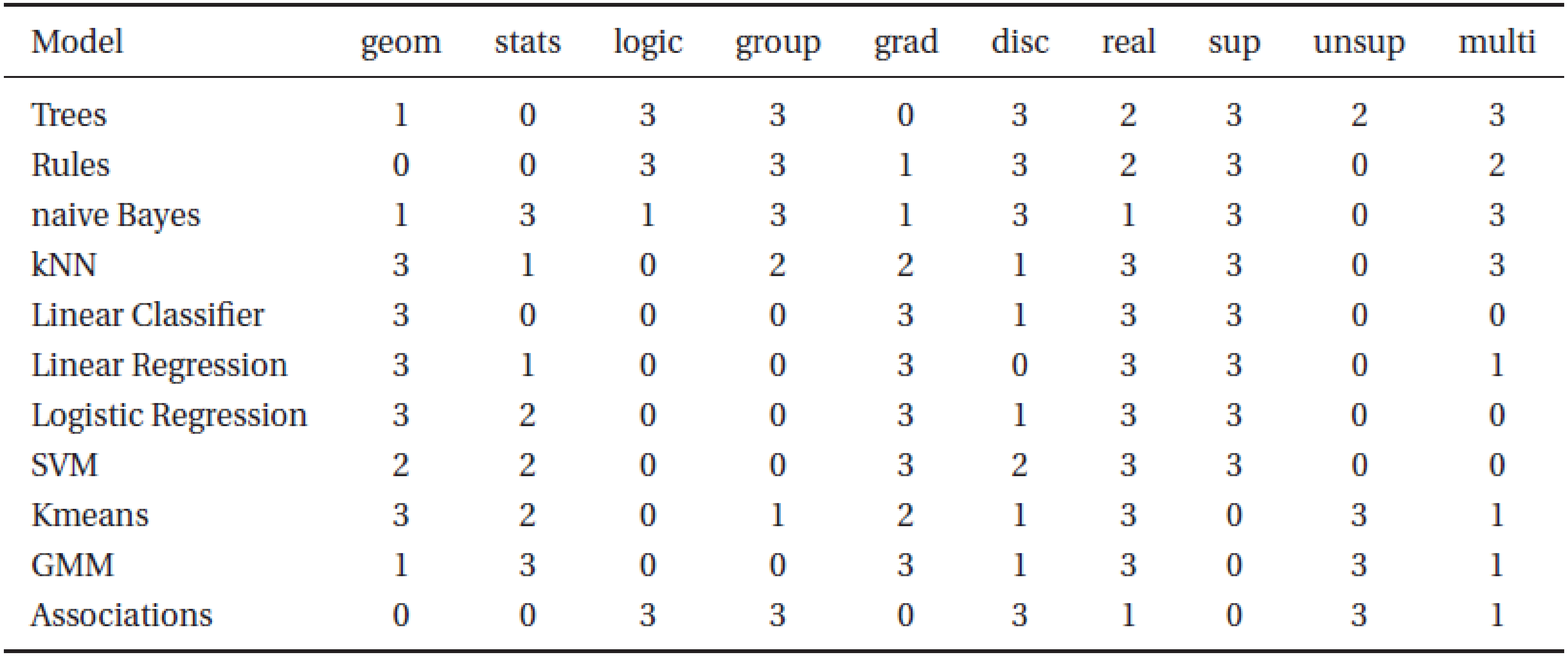
- **Feature Transformation**





Interaction between Features

- Example:



Model	geom	stats	logic	group	grad	disc	real	sup	unsup	multi
Trees	1	0	3	3	0	3	2	3	2	3
Rules	0	0	3	3	1	3	2	3	0	2
naive Bayes	1	3	1	3	1	3	1	3	0	3
kNN	3	1	0	2	2	1	3	3	0	3
Linear Classifier	3	0	0	0	3	1	3	3	0	0
Linear Regression	3	1	0	0	3	0	3	3	0	1
Logistic Regression	3	2	0	0	3	1	3	3	0	0
SVM	2	2	0	0	3	2	3	3	0	0
Kmeans	3	2	0	1	2	1	3	0	3	1
GMM	1	3	0	0	3	1	3	0	3	1
Associations	0	0	3	3	0	3	1	0	3	1

- **Positively correlated:** ‘logic’ and ‘disc’

- **Negatively correlated:** ‘logic’ and ‘grad’.

Summary

- A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E.

- The three ingredients of Machine Learning are:
  - Task
    - Supervised
    - unsupervised
  - Model
    - Predictive Vs Descriptive
    - Geometric Vs Probabilistic Vs Logical
    - Grouping Vs Grading
  - Features
    - Feature as a split
    - Feature as a predictor
    - Feature construction and transformation
    - Interaction between the features