Module - 3

Module – 3:

Syllabus

- **Classification Using Decision Trees and Rules:**
  - **Divide and conquer strategy.**
  - **Decision tree algorithm.**

Regression Methods:

- **Simple linear regression:-**
  - **Ordinary least squares estimation**
  - **Correlations –**
  - **Multiple linear regression**

Decision Trees.

- **Machine learning method that make complex decisions from sets of simple choices.**
- **present their knowledge in the form of logical structures.**
- **Useful for business strategy and process improvement.**

Decision Trees.

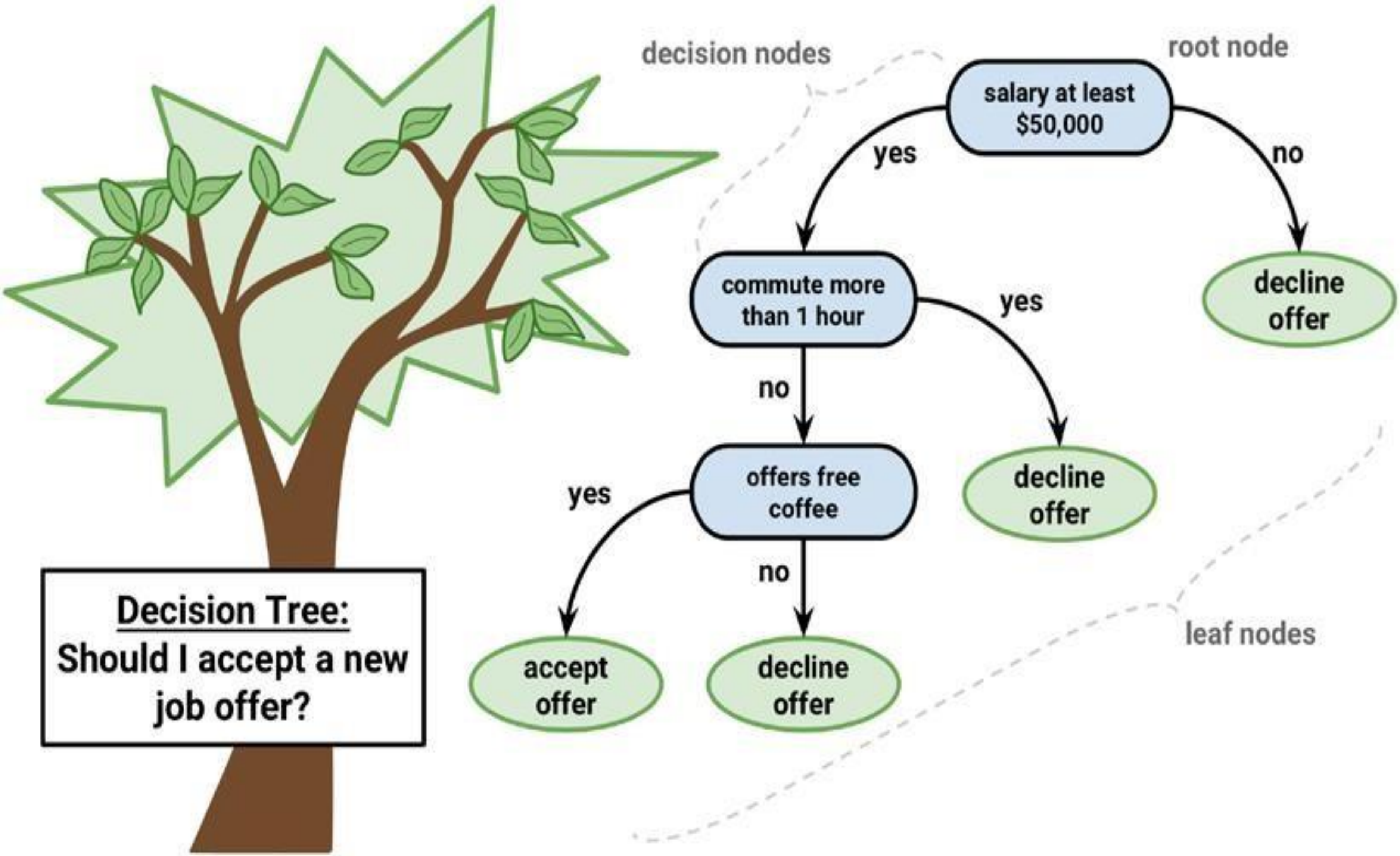
Decision tree learners :

- powerful classifiers
- utilize a tree structure to model the relationships among the features and the potential outcomes.
- a decision tree classifier uses a structure of branching decisions, which channel examples into a final predicted class value.

Decision Trees

- **Root node**
- **Decision nodes** (choices to be made based on the attributes of the job).
- **Branches** → potential outcomes of a decision (yes or no)
- **Leaf nodes**→ final decision (also known as terminal nodes).

Decision Trees



Decision tree : example.

Predicts whether a job offer should be accepted.

- Root node → job offer to be considered (begins at the root node).
- it is then passed through decision nodes → equire choices to be made based on the attributes of the job.
- These choices split the data across branches→ outcomes of a decision (depicted here as yes / no outcomes).
- more than two outcome is also possible.
- In the case a final decision can be made, the tree is terminated by leaf nodes (terminal nodes).
- leaf nodes → action to be taken as the result of the series of decisions.

benefit of decision tree algorithms:

- After the model is created, many decision tree algorithms output the resulting structure in a human-readable format.
- This provides tremendous insight into how and why the model works or doesn't work well for a particular task:
  - for future business practices.

Uses of decision tree algorithms:

1. Credit scoring models in which the criteria that causes an applicant to be rejected need to be clearly documented and free from bias.
2. Marketing studies of customer behavior such as satisfaction or not, which will be shared with management or advertising agencies.
3. Diagnosis of medical conditions based on laboratory measurements, symptoms, or the rate of disease progression.

Divide and Conquer

- Decision trees are built using a heuristic called ***recursive partitioning.***
- commonly known as ***divide and conquer***:
  - It **splits** the **data** into **subsets**, which are then **split repeatedly** into even **smaller subsets**, and so on &
  - Until the process **stops** when the algorithm determines the **data within the subsets are sufficiently homogenous**, or **another stopping criterion has been met.**

Divide and Conquer(cntd..)

- splitting a dataset can create a decision tree.
- At first, the root node → entire dataset, since no splitting has done.
- Next, the decision tree algorithm must choose a feature to split upon;
- ideally, it chooses the feature most predictive of the target class.
- The examples are then partitioned into groups according to the distinct values of this feature, and
- First set of tree branches are formed.

Divide and Conquer(cntd..)

- Working down each branch, the algorithm continues to divide and conquer the data, choosing the best candidate feature each time to create another decision node, until a stopping criterion is reached.

Divide and Conquer(cntd..)

- **Divide and conquer might stop at a node in a case that:**

➢ **All (or nearly all) of the examples at the node have the same class.**

➢ **There are no remaining features to distinguish among the examples.**

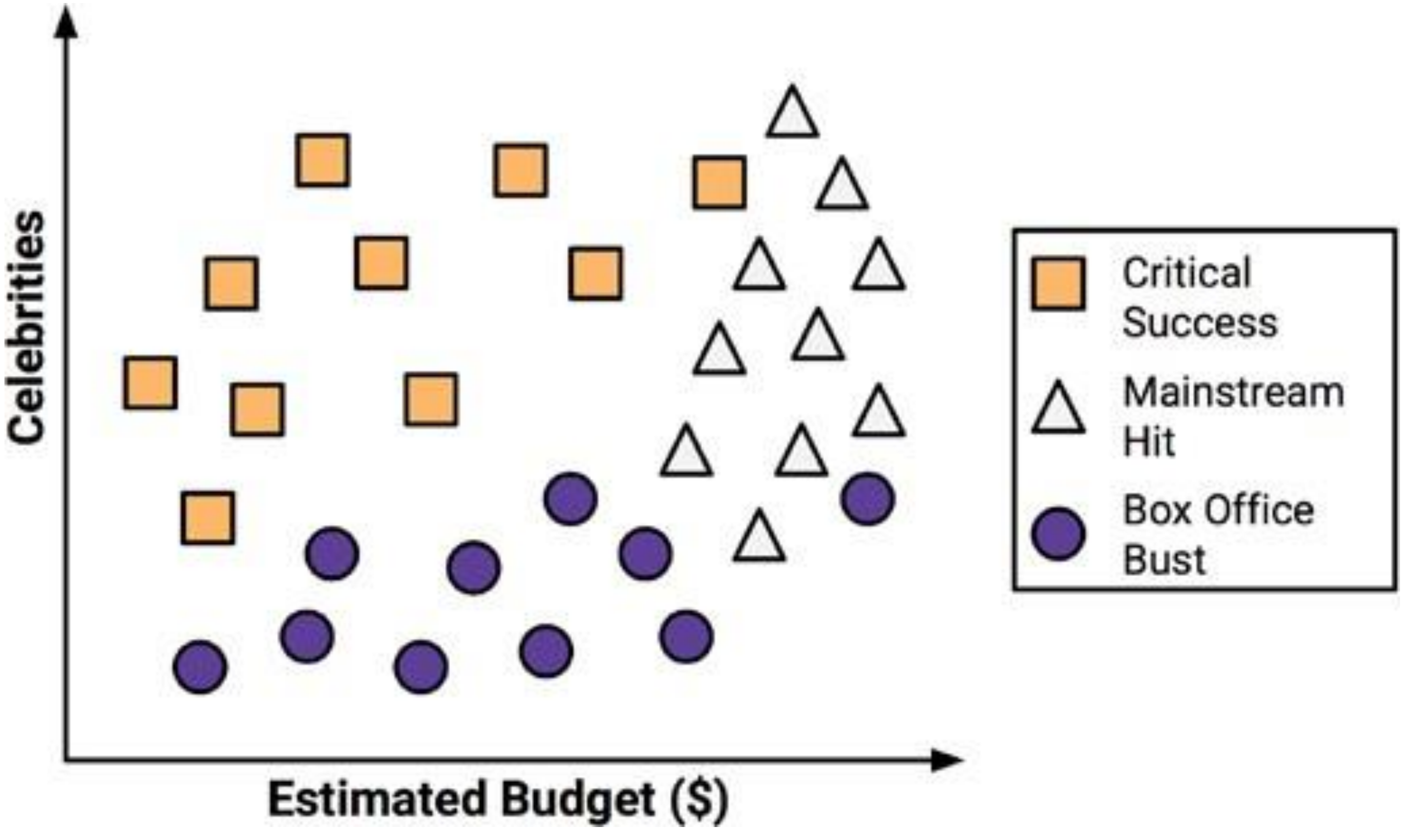
➢ **The tree has grown to a predefined size limit.**

?

- Decision tree algorithm to predict whether a movie would fall into one of three categories:
  - Critical Success
  - Mainstream Hit
  - Box Office Bust (flop/unsuccess).

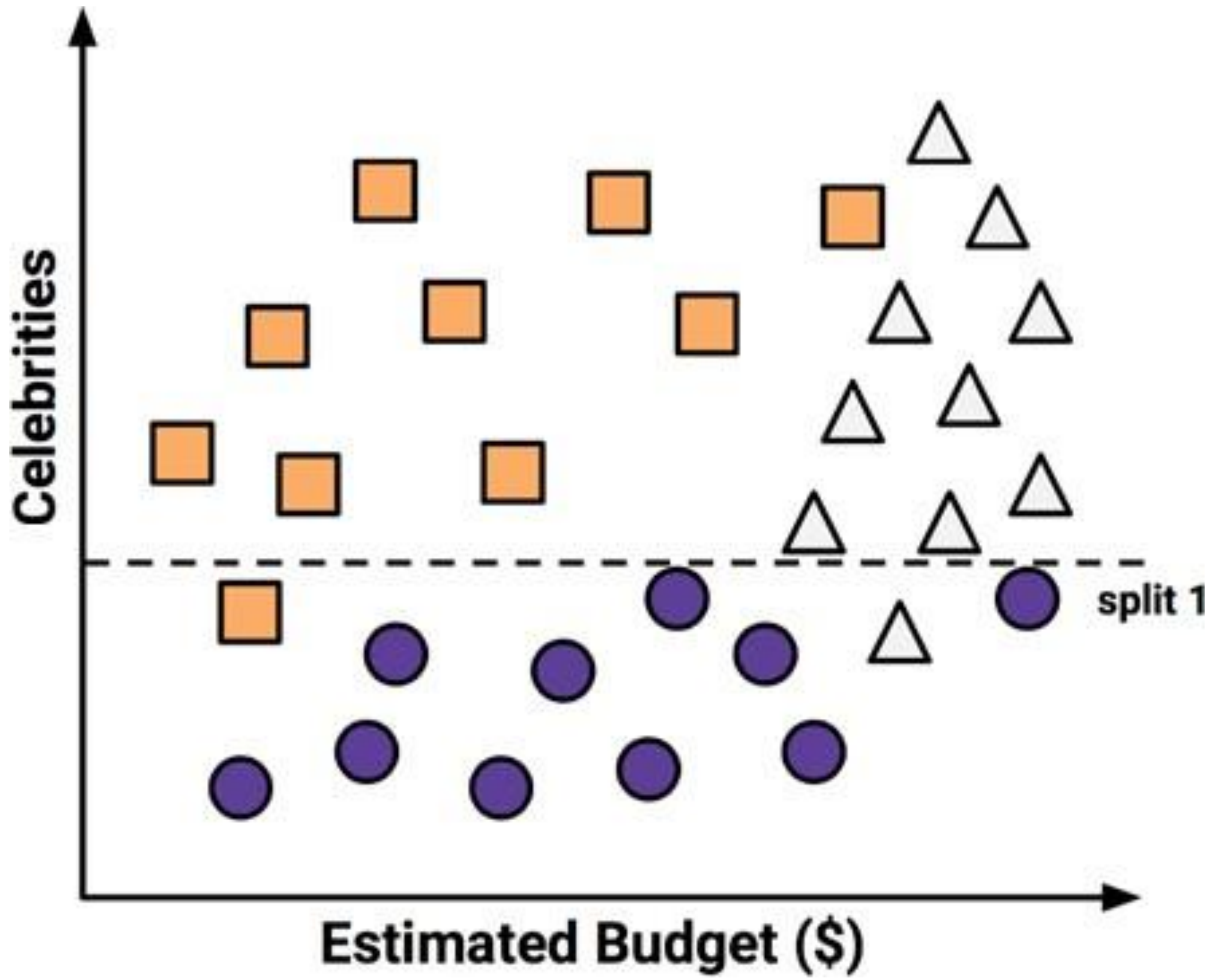
- To build the decision tree, we have to examine the factors leading to the success and failure of the company's 30 most recent releases.
- You quickly notice a relationship between the film's estimated shooting budget, the number of major celebrities lined up for starring roles, and the level of success.

- **Scatterplot** to illustrate the pattern:

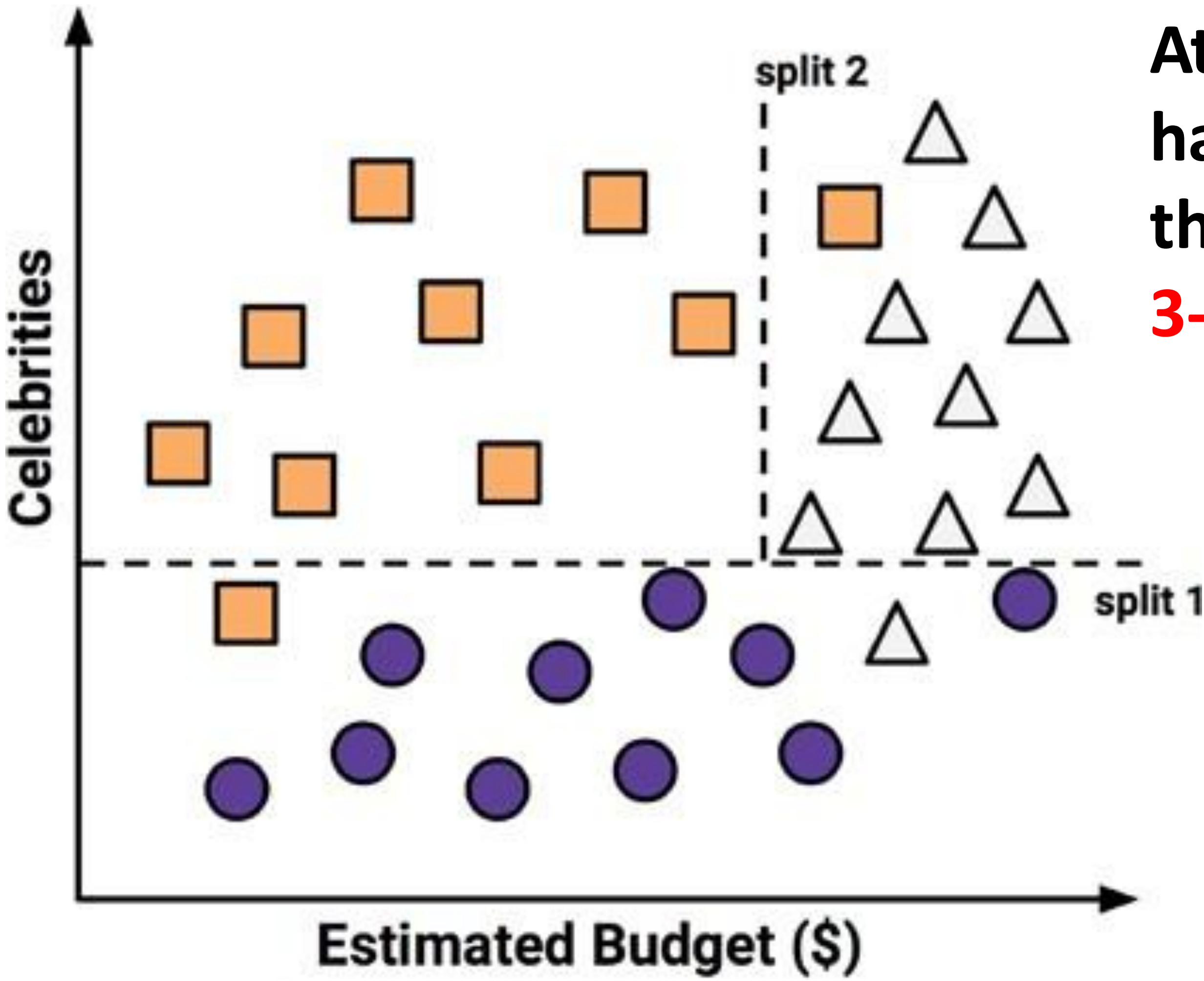


- Using the divide and conquer strategy, we can build a simple decision tree from this data.
- First, to create the tree's root node, we split the feature indicating the number of celebrities;
  - Partitioning the movies into groups with and without a significant number of major stars:

Partitioning the movies into groups with and without a significant number of major stars:

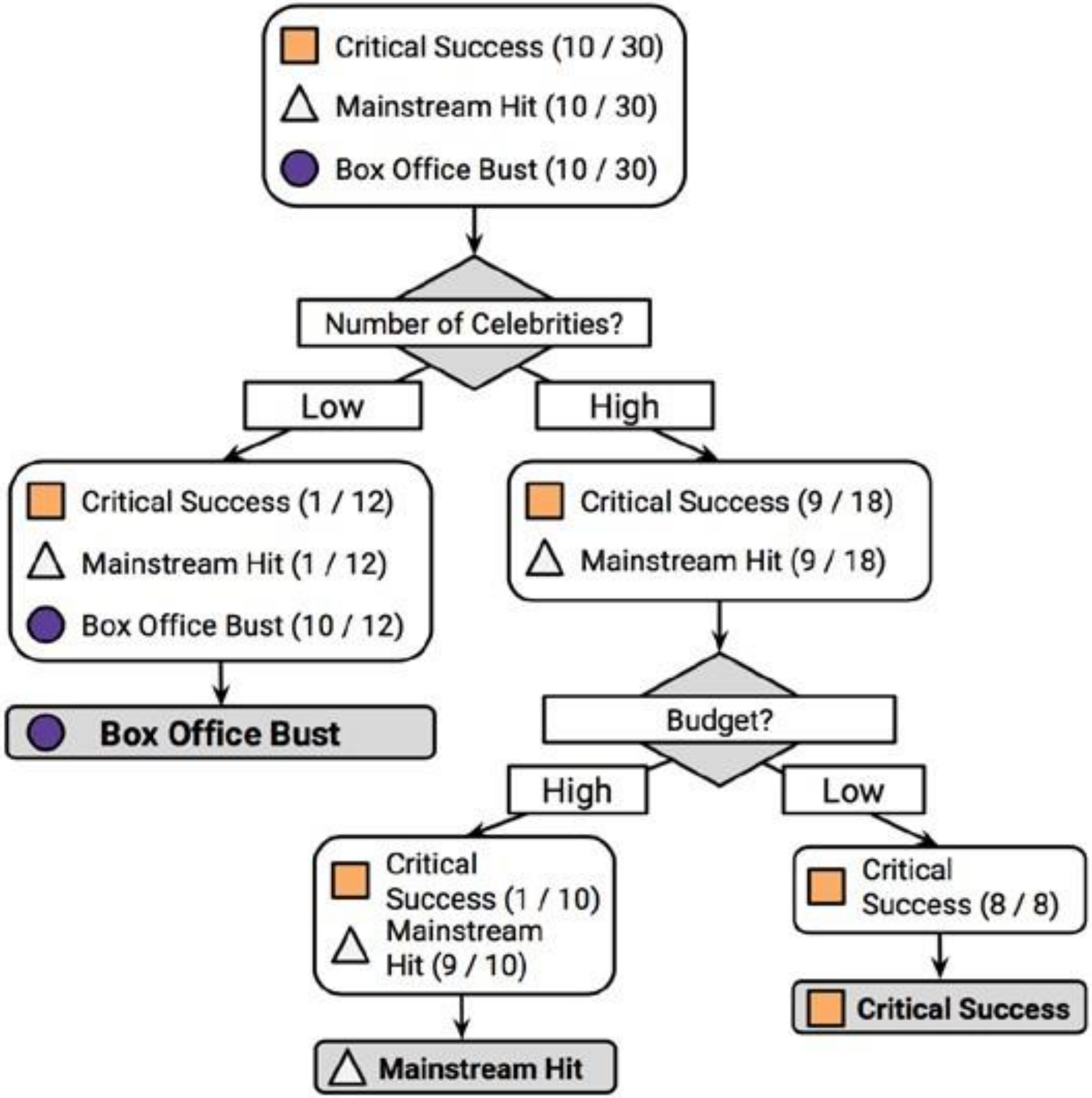


- **Among the group of movies with a larger number of celebrities, we can make another split between movies with and without a high budget:**

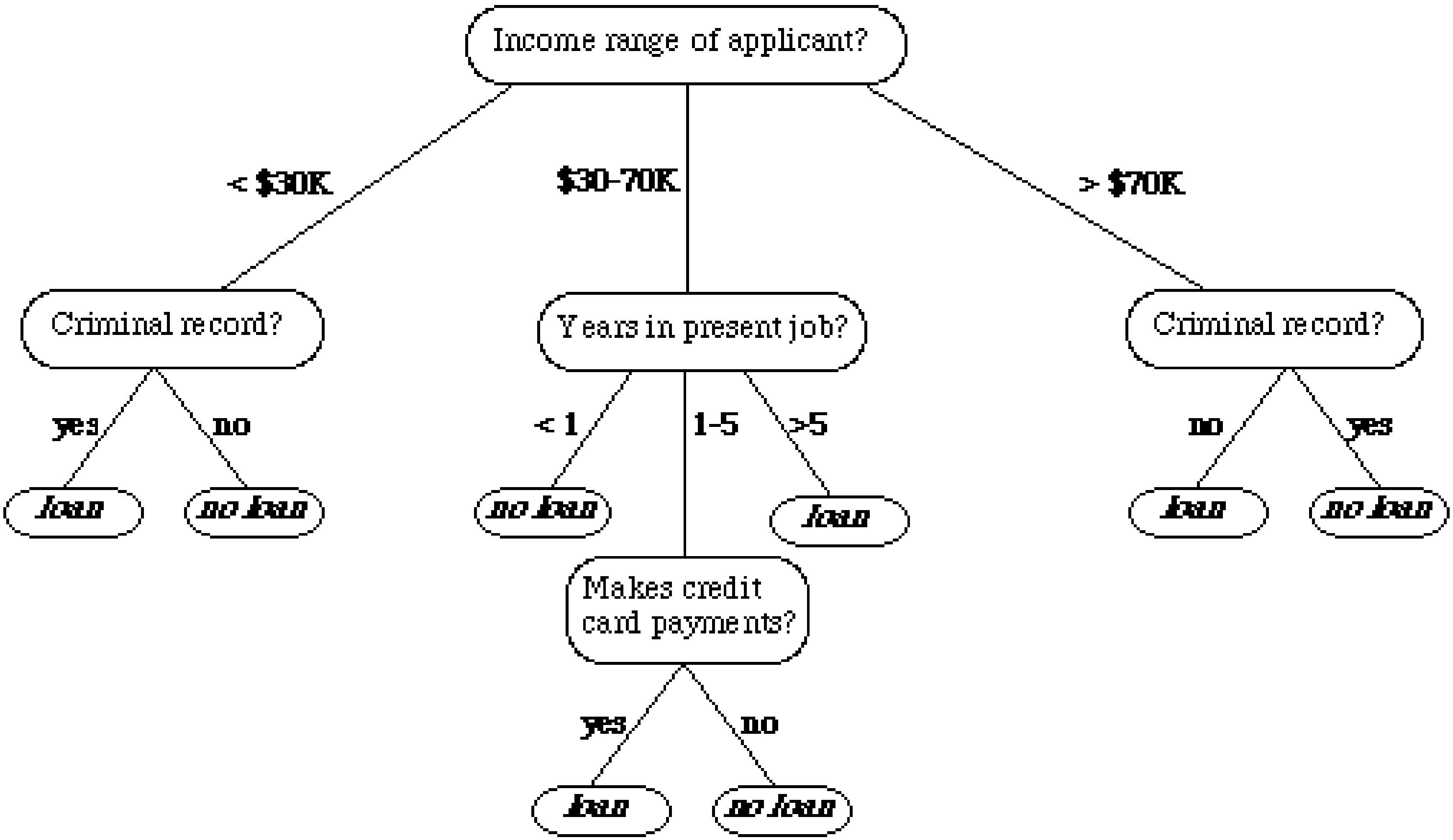


At this point, we have partitioned the data into 3- groups.

Decision Tree : Predicting the future success of movies



decision tree. eg:-



The C5.0 decision tree algorithm

- popular algorithm to build decision tree models automatically.
- one of the most well-known implementations is the C5.0 algorithm.
- developed by computer scientist ‘J. Ross Quinlan’.
  - Improved version of his prior algorithm, C4.5,
  - Which itself is an improvement over his Iterative Dichotomiser 3 (ID3) algorithm.

The C5.0 decision tree algorithm

- easier to understand and deploy.

Strengths & weaknesses of Decision Tree Algorithm.

Strengths	Weaknesses
• An all-purpose classifier that does well on most problems • Highly automatic learning process, which can handle numeric or nominal features, as well as missing data • Excludes unimportant features • Can be used on both small and large datasets • Results in a model that can be interpreted without a mathematical background (for relatively small trees) • More efficient than other complex models	• Decision tree models are often biased toward splits on features having a large number of levels • It is easy to overfit or underfit the model • Can have trouble modeling some relationships due to reliance on axis-parallel splits • Small changes in the training data can result in large changes to decision logic • Large trees can be difficult to interpret and the decisions they make may seem counterintuitive

Choosing the best split

- **The first challenge that a decision tree will face is to identify which feature to split upon.**
- **<u>Purity :</u>**
  - **The degree to which a subset of examples contains only a single class is known as purity. &**
- ***<u>Pure subset.</u>***
  - **Any subset composed of only a single class is called pure.**

Entropy

- **various measurements of purity:**
  - **To identify the best decision tree splitting candidate.**
- **C5.0 uses <u>Entropy</u>;**
  - **A concept borrowed from information theory that quantifies the randomness, or disorder, within a set of class values.**
- **Sets with <u>high entropy</u> are:**
  - **very diverse and**
  - **provide little information about other items that may also belong in the set;**
  - **there is no commonality.**
- **The decision tree hopes to find splits that reduce entropy, ultimately increasing homogeneity within the groups.**

Entropy(cntd..)

- Measured in bits.
- If there are only 2 - possible classes, entropy values can range from 0 to 1.
- For *n* - classes:
  - Entropy ranges from 0 to *log2(n)*.
- In each case, the minimum value → that the sample is completely homogenous;
- maximum value → that the data are as diverse as possible, and
  - no group has even a small plurality.

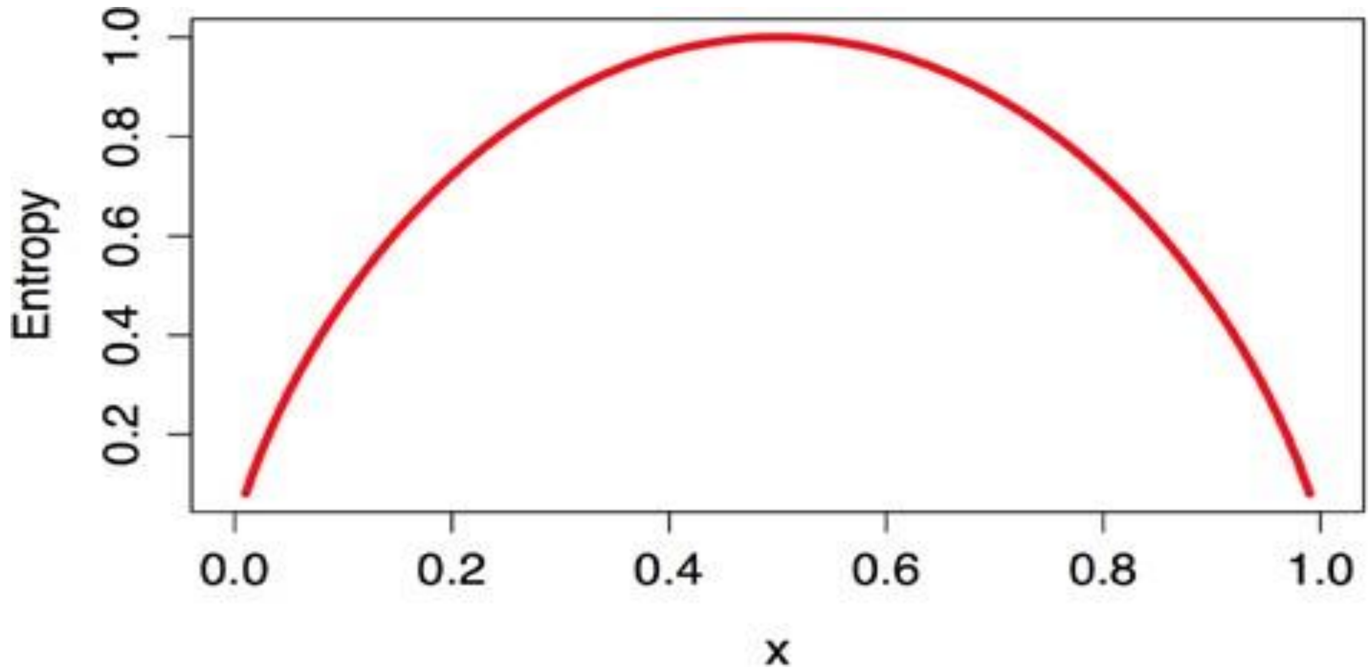
$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \, log_2(p_i)$$

- **S→ segment of data *(S)* ;**
- ***c* → number of class levels ; and**
- ***pi* → proportion of values falling into class level *i*.**

- **Eg: - a partition of data with two classes:**
  - **red (60 percent) and**
  - **white (40 percent).**
  - **entropy = -0.60 * log2(0.60) - 0.40 * log2(0.40) ;**
- **= 0.9709506**

- **We can examine the entropy for all the possible two-class arrangements.**
- **If we know that the proportion of examples in one class is *x*, then**
  - **Proportion in the other class is *(1 – x)*.**

- **Using the curve() function, we can then plot the entropy for all the possible values of *x*:**



- **As illustrated by the peak in entropy at *x = 0.50*, a 50-50 split results in maximum entropy.**
- **As one class increasingly dominates the other, the entropy reduces to zero.**

Information Gain.

- The information gain for a feature *F* is calculated as the difference between the entropy in the segment before the split *(S1)* and the partitions resulting from the split *(S2)*:

$$\mathrm{InfoGain}(F) = \mathrm{Entropy}(S_1) - \mathrm{Entropy}(S_2)$$

- **One complication is that:**
  - **after a split, the data is divided into more than one partition.**
  - **Therefore, the function to calculate *Entropy(S2)* needs to consider the total entropy across all of the partitions.**
  - **It does this by weighing each partition's entropy by the proportion of records falling into the partition.**

- **Total Entropy:**

$$\text{Entropy}(S) = \sum_{i=1}^{n} w_i \text{ Entropy}(P_i)$$

- **Ie, total entropy resulting from a split is the sum of the entropy of each of the *n* partitions weighted by the proportion of examples falling in the partition (*wi*).**

- **Higher information gain:**
  - the better a feature is at creating homogeneous groups after a split on this feature.
- **Information gain = 0:**
  - there is no reduction in entropy for splitting on this feature.
- On the other hand, the maximum information gain is equal to the entropy prior to the split.
- This would imply that the entropy after the split is zero, which means that the split results in completely homogeneous groups.

Pruning the decision tree

- process of pruning a decision tree involves:
  - reducing the size of decision tree such that it generalizes better to unseen data.
- 2 types:
  - pre-pruning (early stopping)
  - post-pruning

Pruning the decision tree

Pre-pruning (early stopping):

- stop growing the tree earlier, before it perfectly classifies the training set.
- To stop the tree from growing:
  - once it reaches a certain number of decisions; or
  - when the decision nodes contain only a small number of examples.

Pruning the decision tree

Post-pruning:

- Allows the tree to perfectly classify the training set, and then post prune the tree.
- Involves growing a tree that is intentionally too large &
- pruning leaf nodes to reduce the size of the tree to a more appropriate level.
- more effective approach than pre-pruning:
  - because it is quite difficult to determine the optimal depth of a decision tree without growing it first.

Regression Methods:

- ✓ Simple linear regression
- ✓ Ordinary least squares estimation
- ✓ Correlations
- ✓ Multiple linear regression

Regression Methods:

- Mathematical relationships help us to understand many aspects of everyday life.
- Eg:- body weight is a function of one's calorie intake.
  - Income is often related to education and job experience.
- When such relationships are expressed with exact numbers, we gain additional clarity.
- Eg:- each year of job experience may be worth an additional $1,000 in yearly salary;

Understanding Regression

- Regression is concerned with **specifying** the **relationship between** a **single numeric dependent variable** (the value to be predicted) and **one or more numeric independent variables** (the predictors).
- ***Dependent Variable:***
  - **depends upon** the **value of the independent variable** or variables.
- The simplest forms of regression assume that:
  - The **relationship between** the **independent** and **dependent variables** follows a **Straight Line**.

Understanding Regression(cntd.)

- Regression equations model data using a slope-intercept format.

Slope-intercept form:

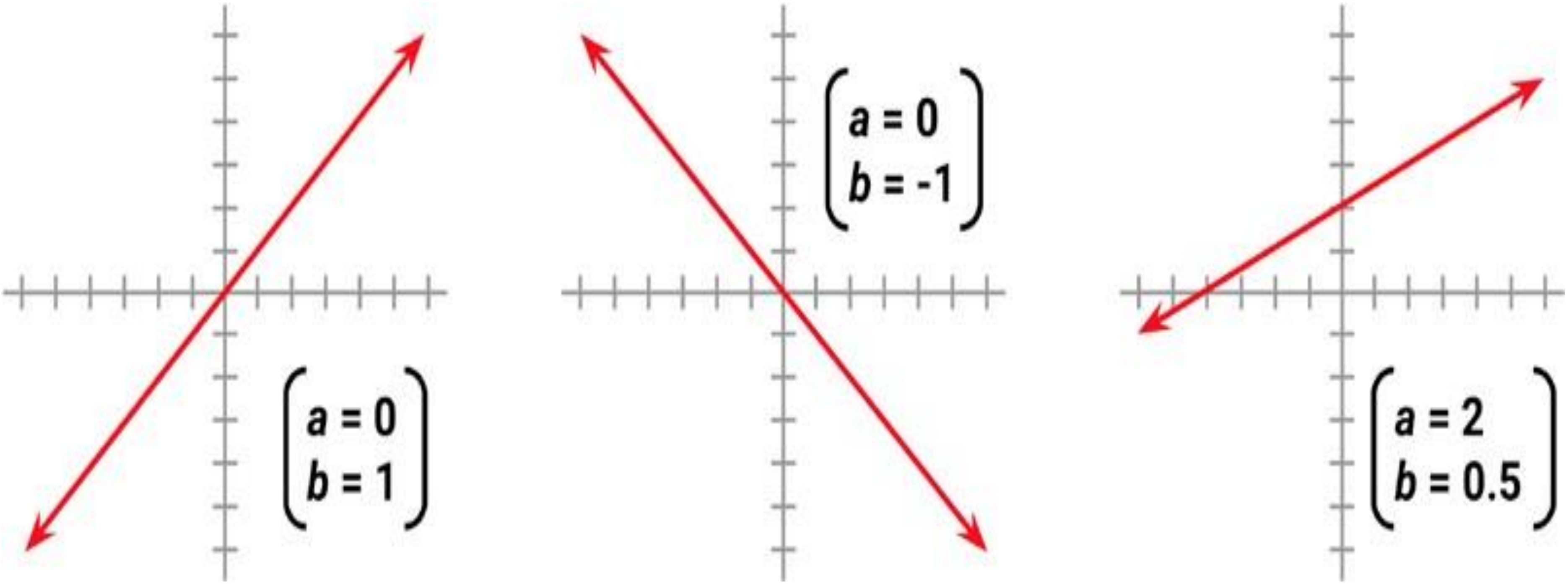
  - $y = a + bx$.

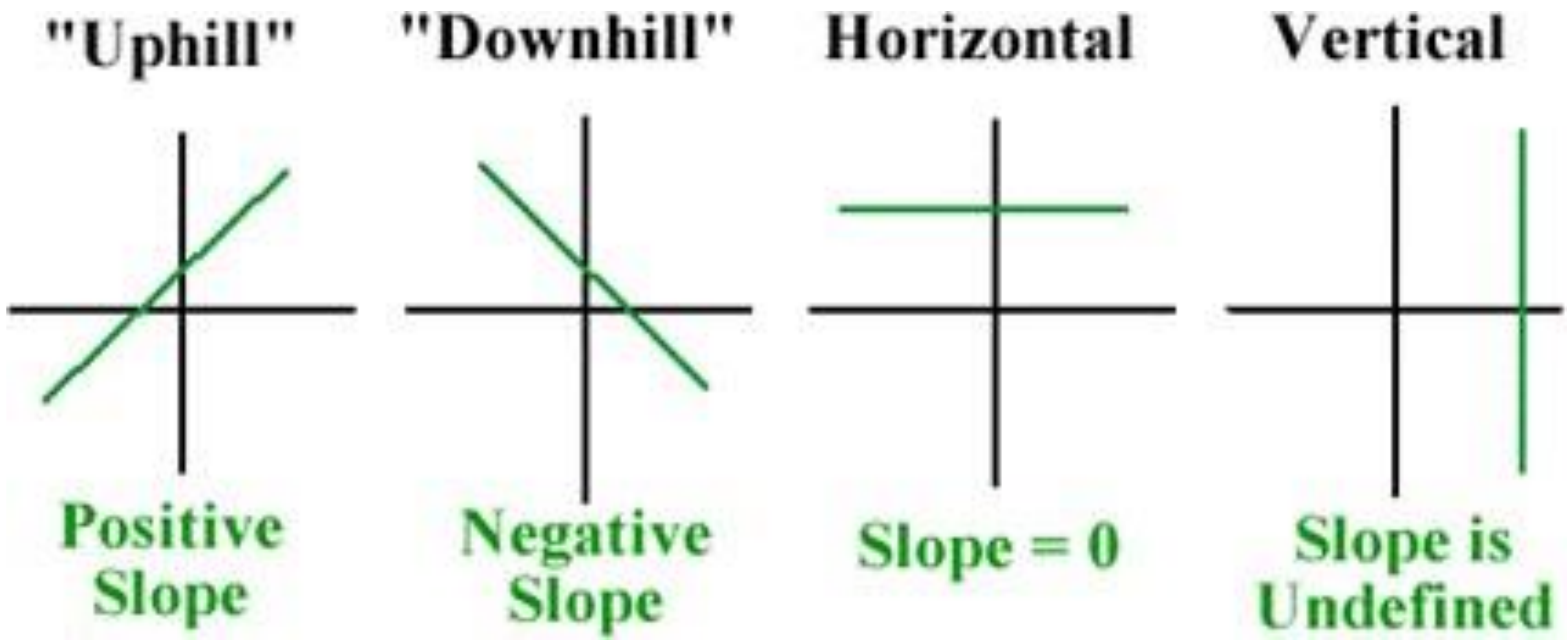
- y → dependent variable
- x → independent variable.
- b → slope (specifies how much the line rises for each increase in x).
- a → intercept (specifies the point where the line crosses, or intercepts, the vertical y axis).
  - It indicates the value of y when $x = 0$.

Understanding Regression(cntd.)

- Positive values → lines that slope upward.
- negative values → lines that slope downward.

Understanding Regression(cntd.)




"Uphill"
"Downhill"
Horizontal
Vertical
Positive Slope
Negative Slope
Slope = 0
Slope is Undefined

Understanding Regression(cntd.)

- Regression equations model data using a similar slope-intercept format.
- The machine's job is to identify values of *a* and *b* so that the specified line is best able to relate the supplied *x* values to the values of *y*.

Regression analysis - Uses

- For modeling complex relationships among data elements.
- estimating the impact of a treatment on an outcome, and extrapolating into the future.
- Quantifying the causal relationship between an event and the response(in clinical drug trials, engineering safety tests, or marketing research).
- Identifying patterns that can be used to forecast future behavior given known criteria(predicting insurance claims, natural disaster damage, election results, and crime rates).
- For statistical hypothesis testing (determines whether a premise is likely to be true or false in light of the observed data).

Basic Linear Regression Models

– those that use straight lines.

- ***Simple Linear Regression:***
  – there is only a single independent variable.
- ***Multiple Linear Regression*** (multiple Regression):
  – Two or more independent variables.

Other Regression Methods:

- **Logistic Regression:**
  - **is used to model a binary categorical outcome.**
- **Poisson Regression:**
  - **models integer count data.**
- **Multinomial Logistic Regression:**
  - **models a categorical outcome;**

Simple Linear Regression - Example.

- Eg: - Rocket Failure:
  - A regression model that demonstrated a link between temperature and O-ring failure, and
  - Could forecast (predict) the chance of failure given the expected temperature at launch.
- A component distress → one of the two types of problems:

➢ *Erosion*: occurs when excessive heat burns up the O-ring.

➢ *Blowby:* occurs when hot gases leak through or "blow by" a poorly sealed O-ring.

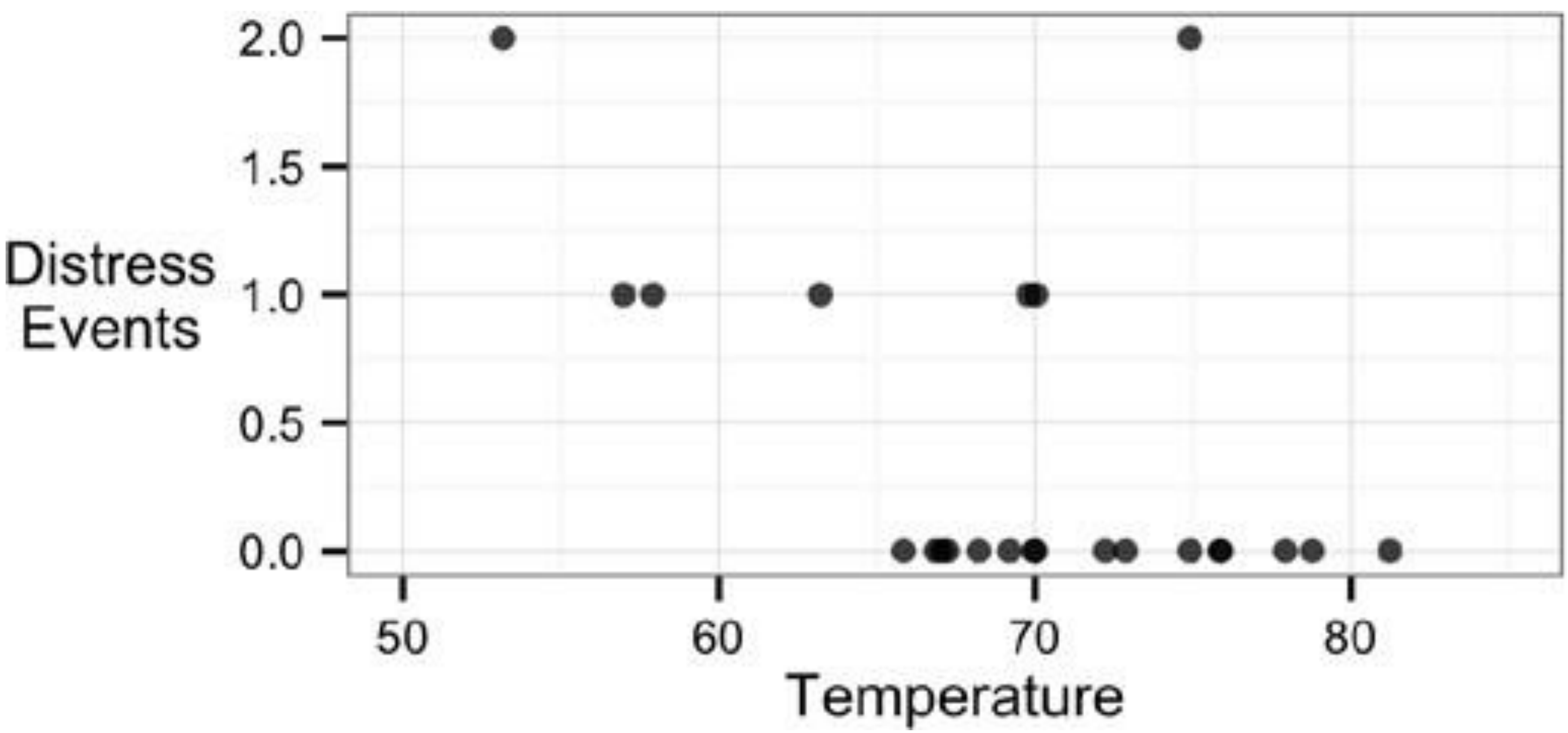
Simple Linear Regression.

- Defines the relationship between a dependent variable and a single independent predictor variable using a line defined by an equation in the following form:

$$y = \alpha + \beta x$$

- **intercept, *α*** **(alpha) → the line crosses the *y* axis.**
- **slope, *β*** **(beta) → change in *y* given an increase of *x*.**

Scatterplot shows a plot of primary O-ring distresses detected for the previous 23 launches, as compared to the temperature at launch:



Simple Linear Regression – Example(cntd).

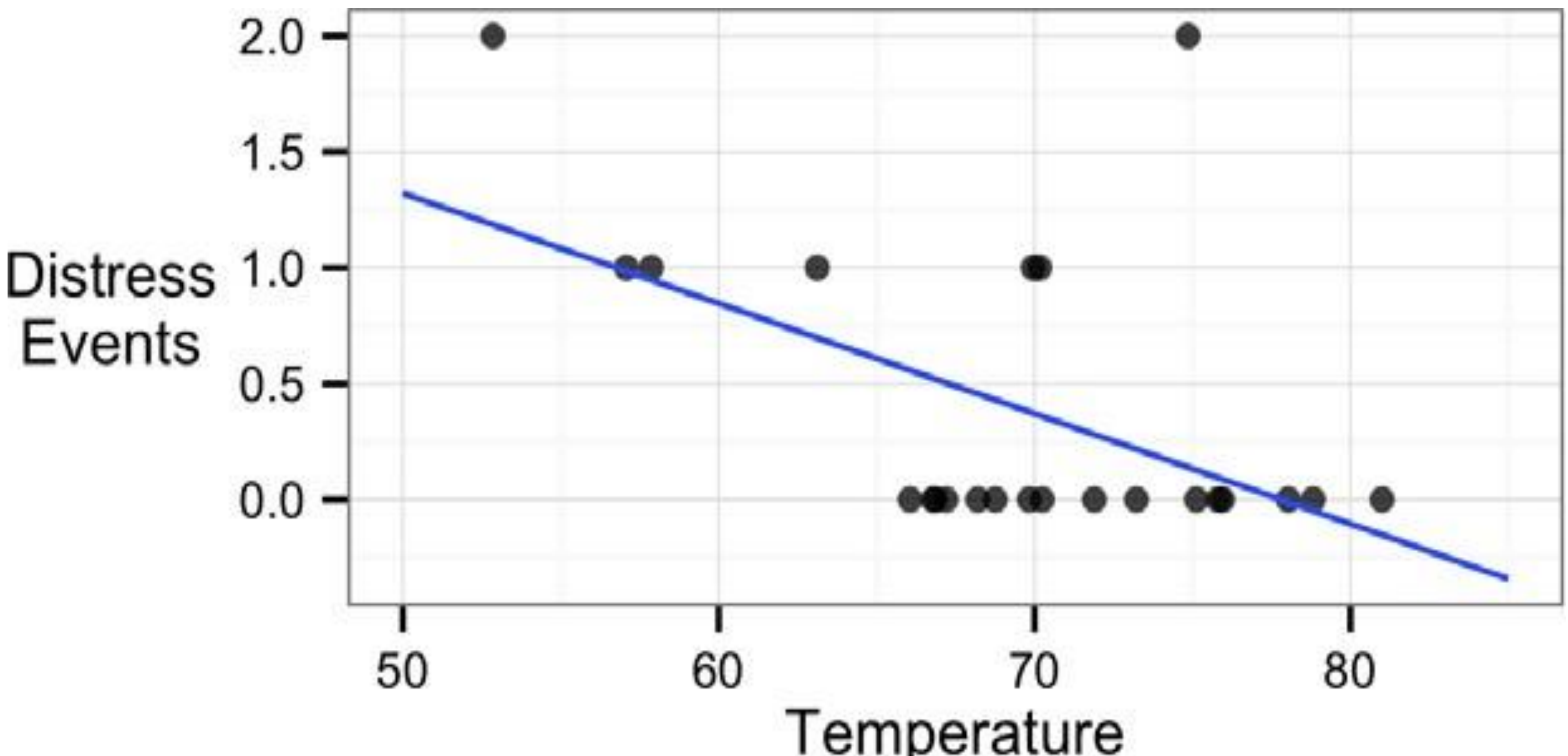
- estimated regression parameters in the equation for the shuttle launch data are:
- a = 3.70 , and
- b = -0.048.

Hence, the full linear equation is:

- $y = 3.70 - 0.048x$.

plot the line on the scatterplot:

- Ss

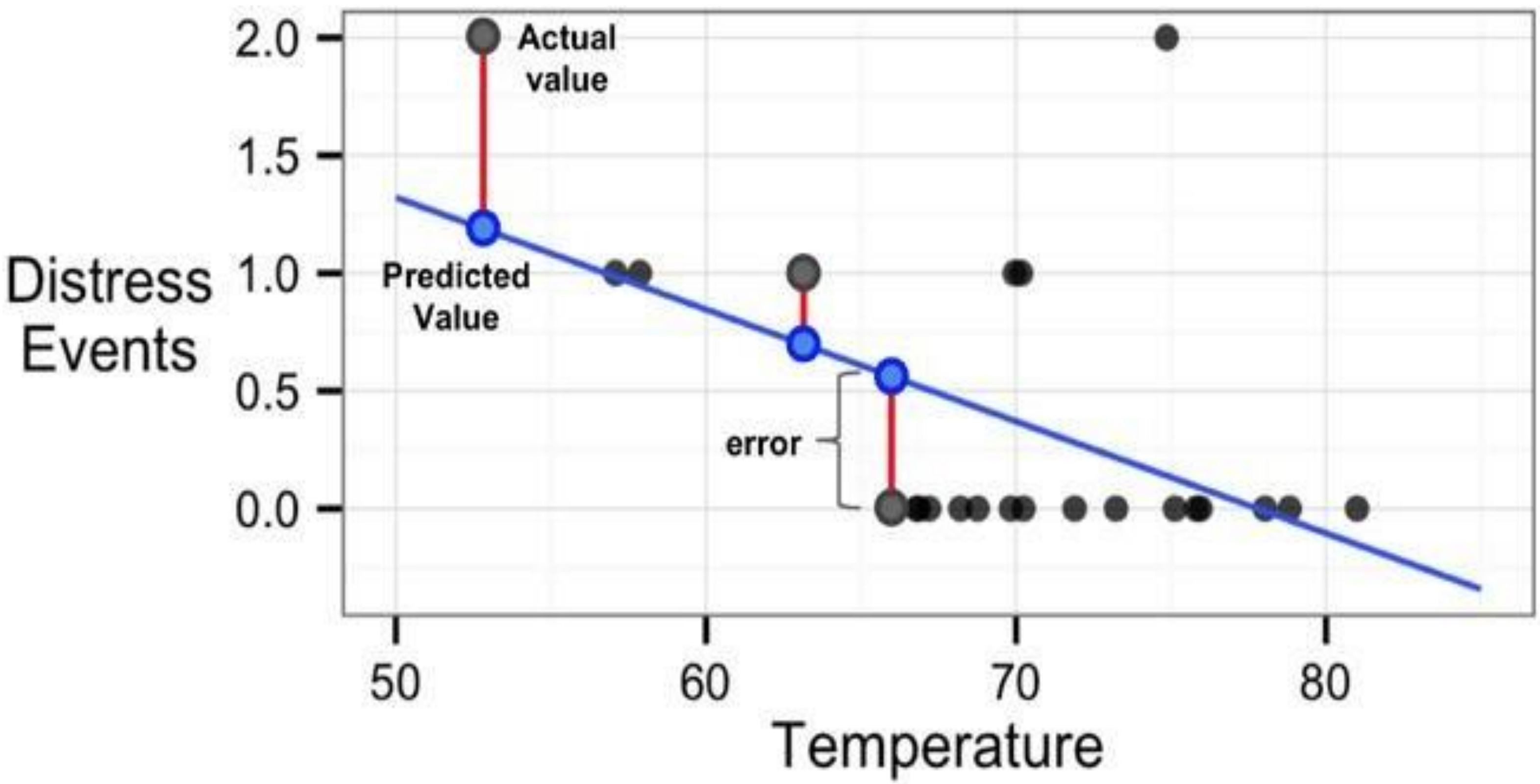


Ordinary Least Squares (OLS) Estimation

- In order to determine the optimal estimates of α and β, an estimation method known as Ordinary Least Squares (OLS) was used.
- In OLS regression, the slope and intercept are chosen so that they minimize the sum of the squared errors;
  - that is, the vertical distance between the predicted *y* value and the actual *y* value.
- These errors are known as residuals.

Ordinary Least Squares (OLS) Estimation(cntd.)

- **Errors are known as residuals, and are illustrated for several points in the following diagram:**



Ordinary Least Squares (OLS) Estimation(cntd.)

- In mathematical terms, the goal of OLS regression can be expressed as the task of minimizing the following equation:

$$\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

- equation defines:
- *e* (error) → difference between the actual *y* value and the predicted *y* value.
- The error values are squared and summed across all the points in the data.

Ordinary Least Squares (OLS) Estimation(cntd.)

- **The solution for *a* depends on the value of *b*. It can be obtained using the following formula:**

$$a = \bar{y} - b\bar{x}$$

Ordinary Least Squares (OLS) Estimation(cntd.)

- value of *b* that results in the minimum squared error is:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Ordinary Least Squares (OLS) Estimation(cntd.)

- If we break this equation apart into its component pieces, we can simplify it a bit.
- The denominator for *b* should look familiar;
- it is very similar to the variance of *x* . (*Var(x)*).
- the variance involves finding the average squared deviation from the mean of *x*.
- This can be expressed as:

Ordinary Least Squares (OLS) Estimation(cntd.)

- variance involves finding the average squared deviation from the mean of *x*.
- This can be expressed as:

$$\mathrm{Var}(x) = \frac{\sum (x_i - \bar{x})^2}{n}$$

Ordinary Least Squares (OLS) Estimation(cntd.)

- The numerator(b) involves taking the sum of each data point's deviation from the mean *x* value, multiplied by that point's deviation away from the mean *y* value.
- This is similar to the covariance function for *x* and *y*, denoted as *Cov(x, y)*.
- The covariance formula is:

$$\mathrm{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Ordinary Least Squares (OLS) Estimation(cntd.)

- If we divide the covariance function by the variance function, the *n* terms get cancelled and we can rewrite the formula for *b* as:

$$b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

- Given this restatement, it is easy to calculate the value of *b* using built-in R functions.
- Let's apply it to the rocket launch data to estimate the regression line.

- Assume that our shuttle launch data is stored in a data frame named launch, the independent variable x is temperature, and the dependent variable y is distress_ct.
- We can then use R's cov() and var() functions to estimate b:

```
b <- cov(launch$temperature, launch$distress_ct) /
        var(launch$temperature)
b
1] -0.04753968
```

- **We can estimate *a* using the mean() function:**

```
a <- mean(launch$distress_ct) - b * mean(launch$temperature)
a
L] 3.698413
```

Correlations

- The correlation between two variables is:
  - A number that indicates how closely their relationship follows a straight line.
- Without additional qualification, correlation typically refers to Pearson's correlation coefficient;
  - developed mathematician Karl Pearson.
- The correlation ranges between -1 and +1.
- The extreme values → a perfectly linear relationship;
- a correlation close to zero → the absence of a linear relationship.

Correlations(cntd.)

- **Pearson's Correlation - Formula :**

$$\rho_{x,y} = \mathrm{Corr}(x, y) = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

Correlations(cntd.)

- Using this formula, we can calculate the correlation between the launch temperature and the number of O-ring distress events.
- Recall that the covariance function is cov() and the standard deviation function is sd().
- store the result in r, a letter that is commonly used to indicate the estimated correlation:
  - r= cov() / sd().

Correlations(cntd.)

- The correlation between the temperature and the number of distressed O-rings is -0.51.
- The negative correlation → increases in temperature are related to decreases in the number of distressed O-rings.
- Ie; low temperature launch could be problematic.
- The correlation → relative strength of the relationship between temperature and O-ring distress.
- Because -0.51 is halfway to the maximum negative correlation of -1, → moderately strong negative linear association.

Interpretation of correlation strength:

- "weak" → values between 0.1 and 0.3;
- "moderate" → 0.3 to 0.5, and
- "strong" → values above 0.5
  - (these also apply to similar ranges of negative correlations).
- Often, the correlation must be interpreted in context.
- For data involving human beings, a correlation of 0.5 may be considered extremely high;
- for data generated by mechanical processes, a correlation of 0.5 may be weak.

Multiple Linear Regression.

- More than one independent variable.
- Extension of simple linear regression.
- The goal in both cases is similar:
  - Find values of beta coefficients that minimize the prediction error of a linear equation.
- The key difference is :
  - There are additional terms for additional independent variables.

Multiple Linear Regression(cntd).

- General form of Multiple regression :

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

- An error term (*epsilon* - €) has been added → predictions are not perfect.
- € → residual term

Multiple Linear Regression(cntd).

- A coefficient is provided for each feature.
- This allows each feature to have a separate estimated effect on the value of *y*.
- Ie; *y* changes by the amount *βi* for each unit increase in *xi*.
- when All independent variables = zero.
  - Expected value of *y* = *α* (Intercept)

Multiple Linear Regression(cntd).

- **Since the intercept *α* is really no different than any other regression parameter, it is also sometimes denoted as *ß0* (pronounced beta-naught), as shown in the following equation:**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

Multiple Linear Regression(cntd).

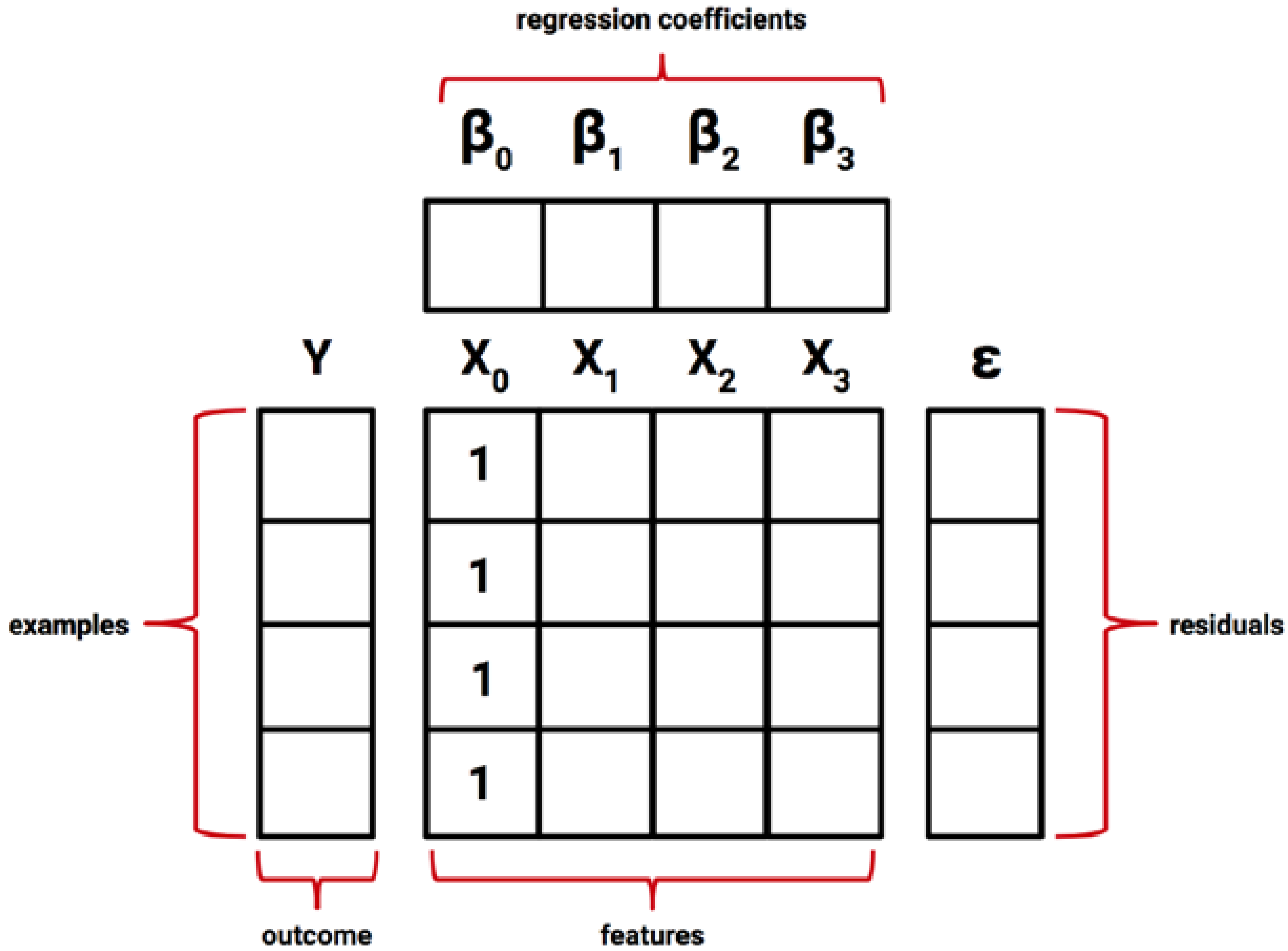
- **Intercept is unrelated to any of the independent *x* variables.**
- **Imagine *β0* as if it were being multiplied by a term *x0,* which is a constant with the value 1:**

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

Multiple Linear Regression(cntd).

- To estimate the values of the regression parameters:
  - Each observed value of the dependent variable *y* must be related to the observed values of the independent *x* variables using the regression equation in the previous form.
- The following figure illustrates this structure:

Multiple Linear Regression(cntd).



Multiple Linear Regression(cntd).

- The many rows and columns of data illustrated in the preceding figure can be described in a condensed formulation using "bold font" matrix notation to indicate that each of the terms represents multiple values:

$$\mathbf{Y} = \beta \mathbf{X} + \varepsilon$$

Multiple Linear Regression(cntd).

- The dependent variable is now a vector, Y, with a row for every example.
- The independent variables have been combined into a matrix, X, with a column for each feature.

  +
- an additional column of '1' values for the intercept term.
- Each column has a row for every example.
- The regression coefficients β and residual errors ε are also now vectors.

Multiple Linear Regression(cntd).

- The goal is now to solve for β, the vector of regression coefficients that minimizes the sum of the squared errors between the predicted and actual Y values.
- Finding the optimal solution requires the use of matrix algebra;
- the best estimate of the vector β can be computed as:

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

Multiple Linear Regression(cntd).

$$\hat{\beta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

- **This solution uses a pair of matrix operations:**
- **T → Transpose of matrix X;**
- **Negative exponent → Matrix inverse.**
- **Using R's built-in matrix operations, we can implement a simple multiple regression learner.**

- **Function to the shuttle launch data.**
- **As shown in the following code, the dataset includes three features and the distress count (distress_ct), which is the outcome of interest:**

- str(launch)
- 'data.frame': 23 obs. of 4 variables:
- $ distress_ct : int 0 1 0 0 0 0 0 0 1 1 ...
- $ temperature : int 66 70 69 68 67 72 73 70 57 63 ...
- $ field_check_pressure: int 50 50 50 50 50 50 100 100 200 ...
- $ flight_num : int 1 2 3 4 5 6 7 8 9 10 ...

strengths and weaknesses of multiple linear regression

Strengths	Weaknesses
• By far the most common approach for modeling numeric data • Can be adapted to model almost any modeling task • Provides estimates of both the strength and size of the relationships among features and the outcome	• Makes strong assumptions about the data • The model's form must be specified by the user in advance • Does not handle missing data • Only works with numeric features, so categorical data requires extra processing • Requires some knowledge of statistics to understand the model

Activation Functions