Day 2

Intro to Classification Algorithms

HPSI
Human Performance Systems, Inc.

Day 2 Outline

- Intro to Classifiers
- K-Nearest Neighbors
  - Algorithm details
  - Choosing ”k”
  - Pros/Cons
  - Exercise
- Probability
  - Probabilistic Classifiers
  - Probability Distributions
  - Exercise
  - Sample Spaces
  - Conditional Probability
  - Bayesian Inference
- Naïve Bayes Classifier
  - Algorithm Details
  - Exercise
  - Pros/Cons
- Logistic Regression
  - Algorithm Details
  - Exercise
- Model Comparison

Classification

- Definition
  - **Classification** is the process of identifying which class a new (unlabeled) observation belongs to
  - An algorithm that implements classification is known as a **Classifier**
  - Terminology across fields is quite varied
    - In **Statistics**, the properties of records are termed explanatory variables, *independent variables*, *regressors*, etc., and the categories to be predicted are known as *outcomes*, which are considered to be possible values of the *dependent variable*.
    - In **Machine Learning,** the records are more often referred to as *instances or observations*, the explanatory variables are termed *features* (grouped into a *feature vector*), and the possible categories to be predicted are *classes*.

Classification

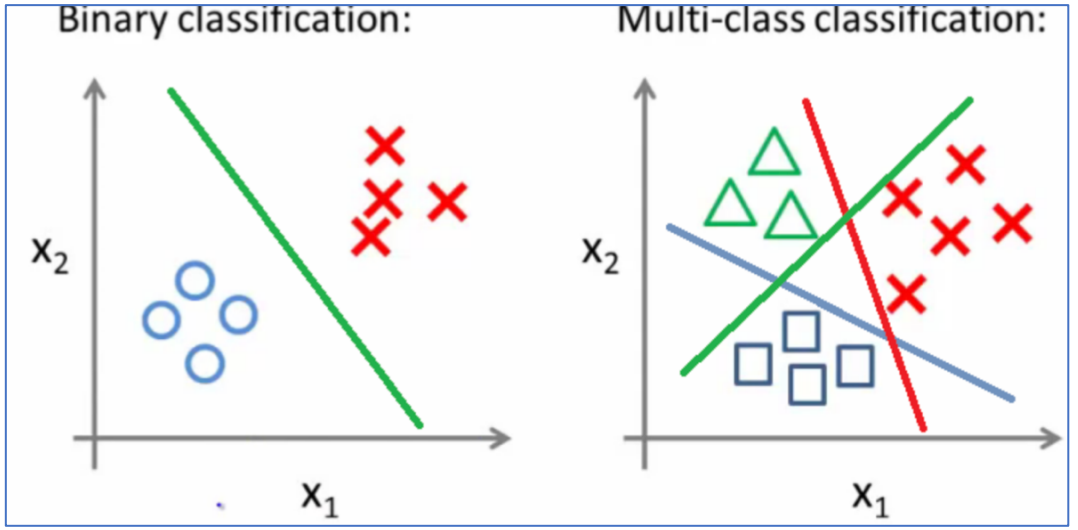
- **The Goal for us** today:
  - To build a few classification models
    - K-NN Classifier
    - Naïve Bayes Classifier
    - Logistic Regression Classifier
  - To gain some basic intuition for how the underlying algorithms behind each of these classifiers work
  - To understand Pros/Cons associated with each classifier

Classification

- **The Goal for the classifiers** today:
  - Given a set of labeled data, these classifiers should be able to:
    - Learn the relationship between features and labels
    - Take un-labeled data as input
    - Output corresponding labels for each input

Classification

- Types of Classification
  - Binary Classification
    - Only 2 Target Classes
  - Multiclass Classification
    - More than 2 Target Classes



Classifier Comparison

No-Free-Lunch

- Classifier performance depends greatly on the characteristics of the data to be classified
- There is no single classifier that always works best in all scenarios

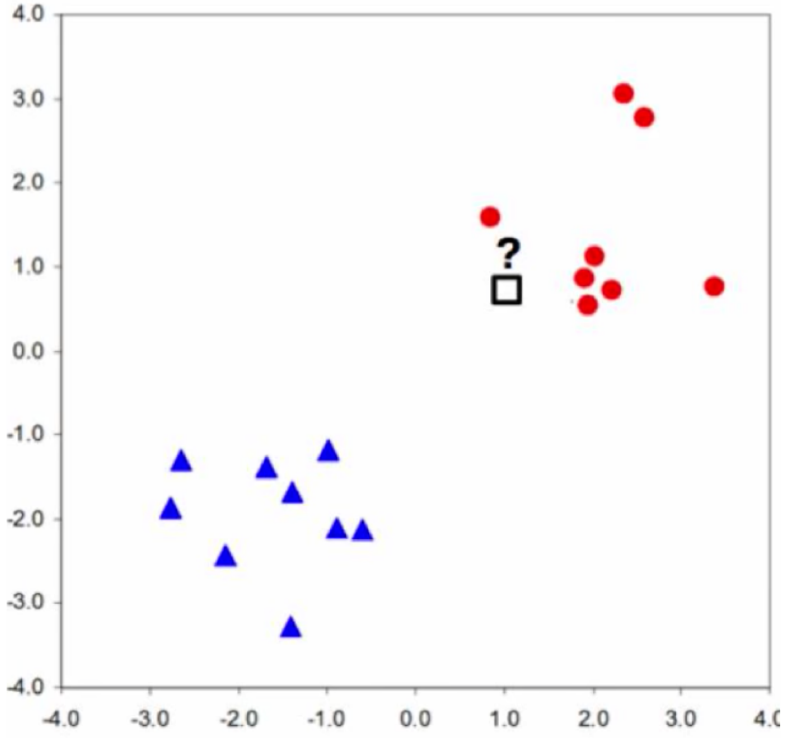
"Baseline" classifiers

- Usually used to compare against more complex algorithms
- Typically easier to explain/interpret

K-Nearest Neighbors

Intuition for k-NN

- Given this set of points
  - Two classes (Red & Blue)
  - Is this new point red or blue?
- What was your reasoning?
  - Did you compute the priors? (Naïve Bayes Classifier)
  - Did you look for a threshold and compute information gain? (Decision Tree Classifier)
  - Did you create a separating hyperplane and maximize the margin (SVM Classifier)
- You just noticed that it was closest to other red points (its neighbors)
- This is the intuition for kNN learning algorithm



HPSI
Human Performance Systems, Inc.

Simple approach for k-NN

Goal:

- Predict the label of a data point by:
  - Looking at the ‘k’ closest labeled data points (neighbors)
  - Taking a majority vote
- Memory-Based Learning
  - Also known as “case-based” or “example-based” learning
- Intuition behind memory-based learning
  - Similar inputs map to similar outputs
    - If true, we just have to define “similar”
    - Not all similarities created equal…

Memory-Based Learning

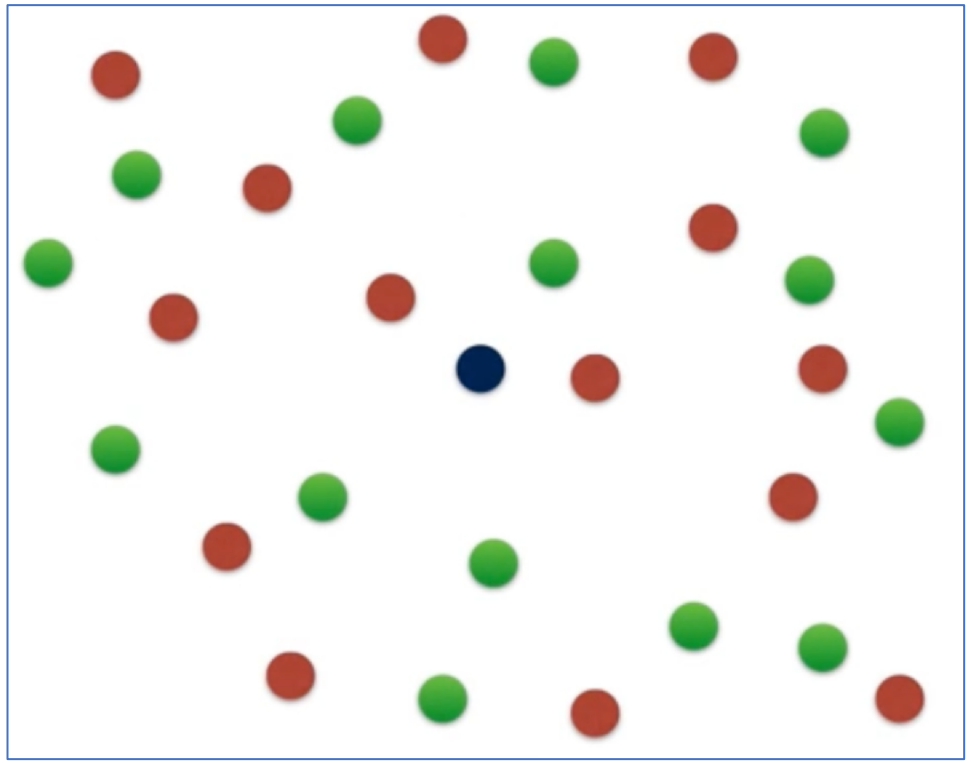
- How do we determine “similar”?
  - For instance, if we wanted to:
  - Predict Brent’s weight
    - Who are the similar people?
    - Similar age, height, waistline (in inches), bodyfat %…
  - Predict Brent’s IQ
    - Who are the similar people?
    - Similar occupation, writing style, undergraduate degree, …
  - How would you quantify a comparison for these two?
    - Need some metric…
      - Distance?

k-NN Approach

- Define a distance $d(x_1, x_2)$ between any 2 examples
  - More on this later…
  - Examples are just feature vectors…
  - We could just use Euclidean distance …
- Fitting/Training
  - Memorize/store the training examples for fast lookup
- Making Predictions
  - For each test input,
    - Calculate the distance from each training point
    - Find the “k” closest neighbors
    - Predicted class = majority vote of k_neighbors

k-NN Graphical Example

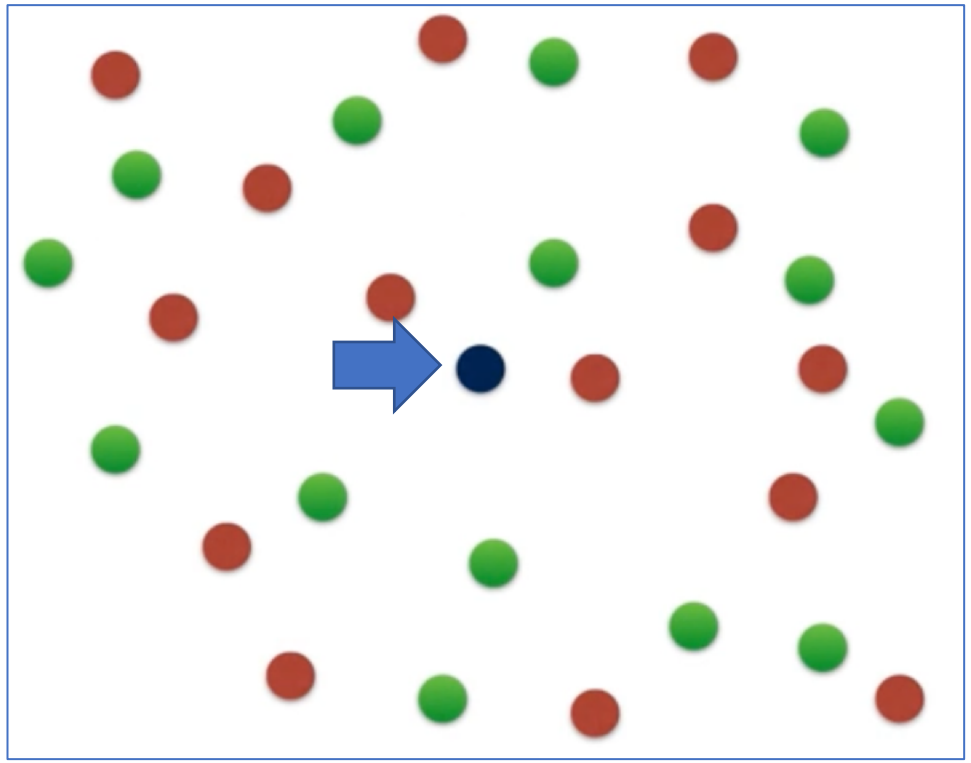
- Let's look at a simple graphical example
- Consider this two-dimensional dataset
  - 2-features (x1,x2)
  - 2-classes (Red, Green)



HPSI
Human Performance Systems, Inc.

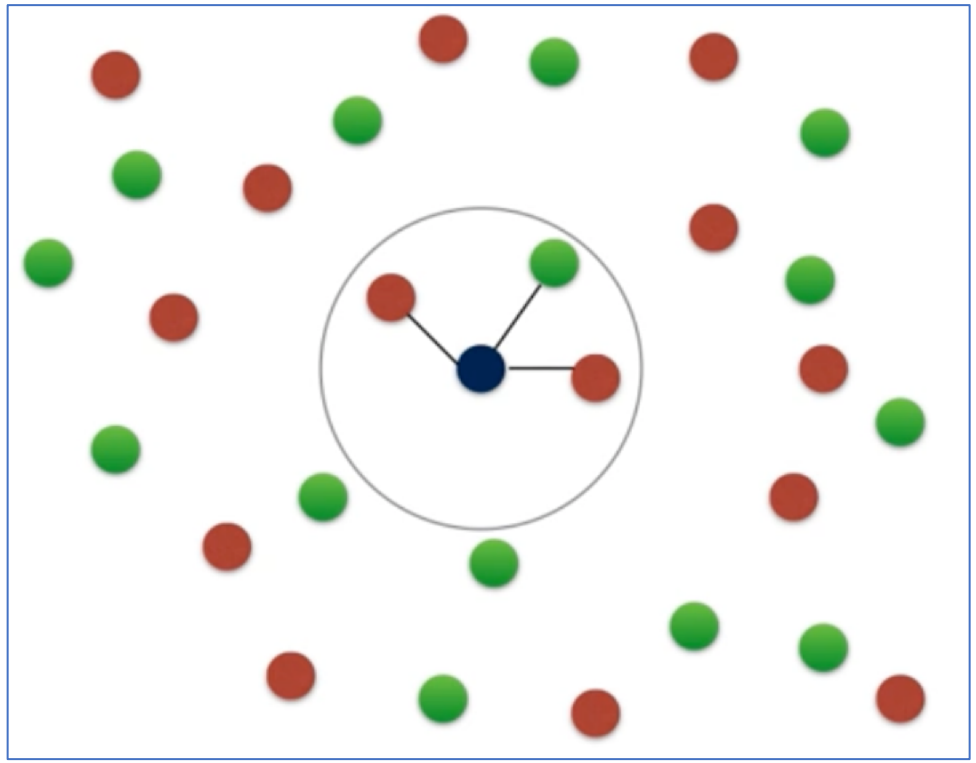
k-NN Graphical Example

- We want to Classify this point



k-NN Graphical Example

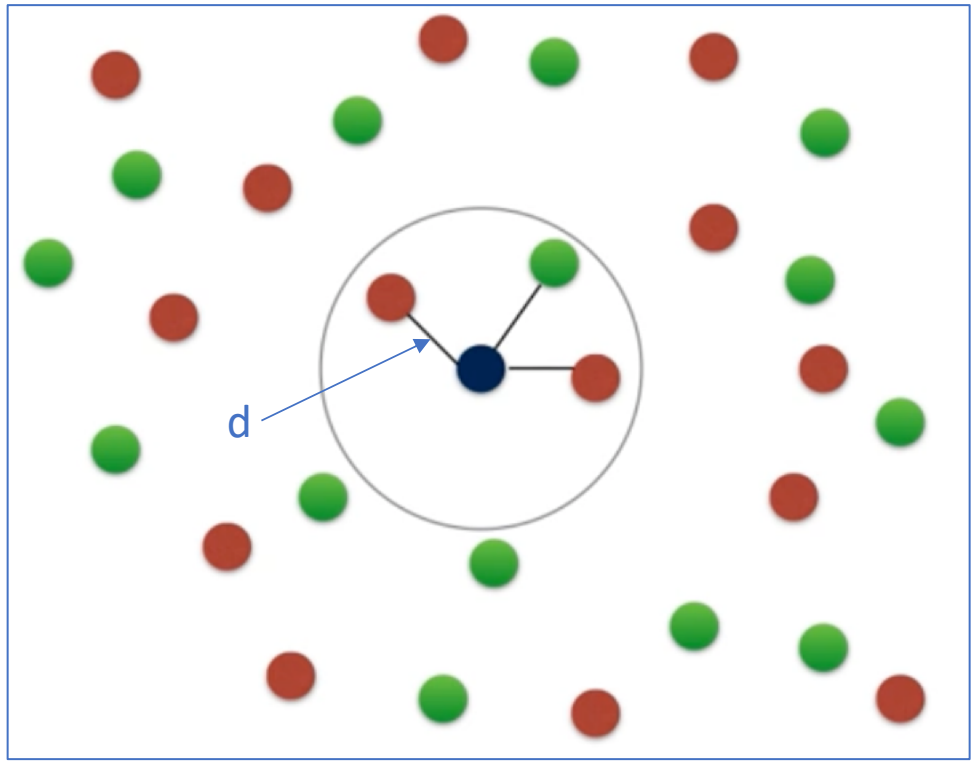
- We want to Classify this point
- If we consider k=3 neighbors



HPSI
Human Performance Systems, Inc.

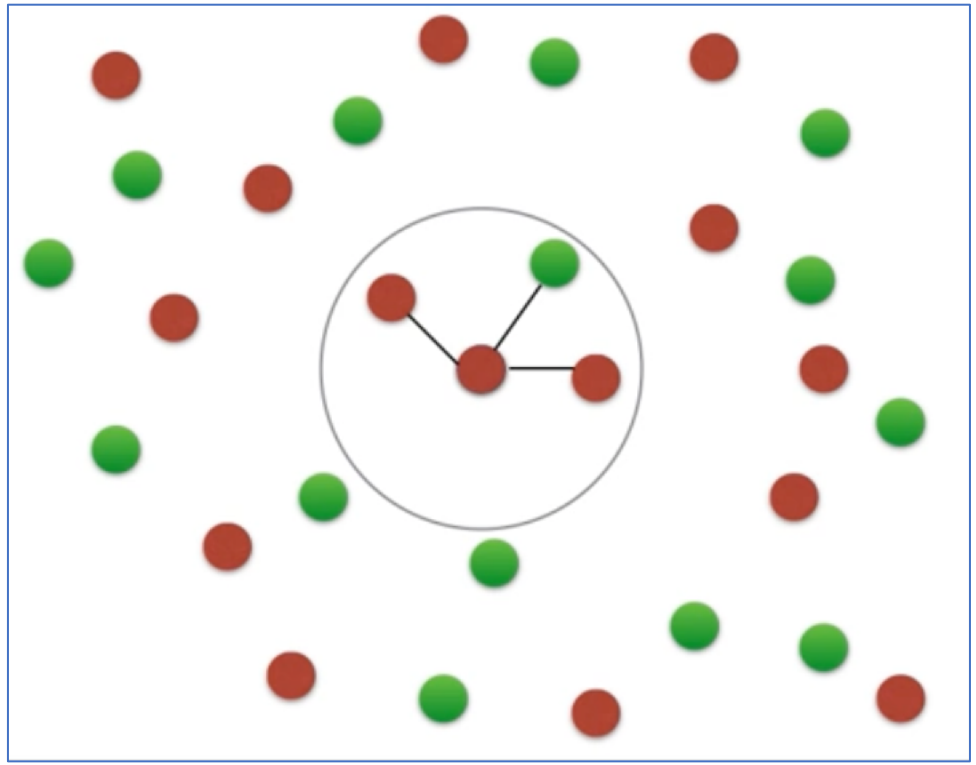
k-NN Graphical Example

- We want to Classify this point
- If we consider k=3 neighbors
  - Measured by some distance



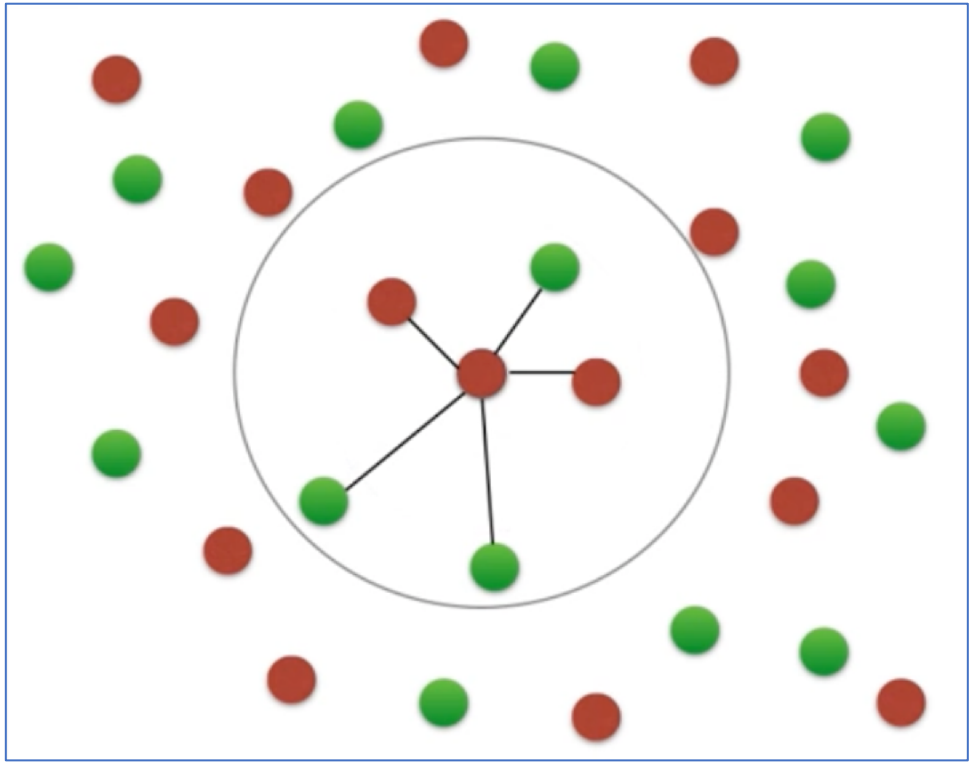
k-NN Graphical Example

- We want to Classify this point
- If we consider k=3 neighbors
  - Measured by some distance
  - The point is classified as Red



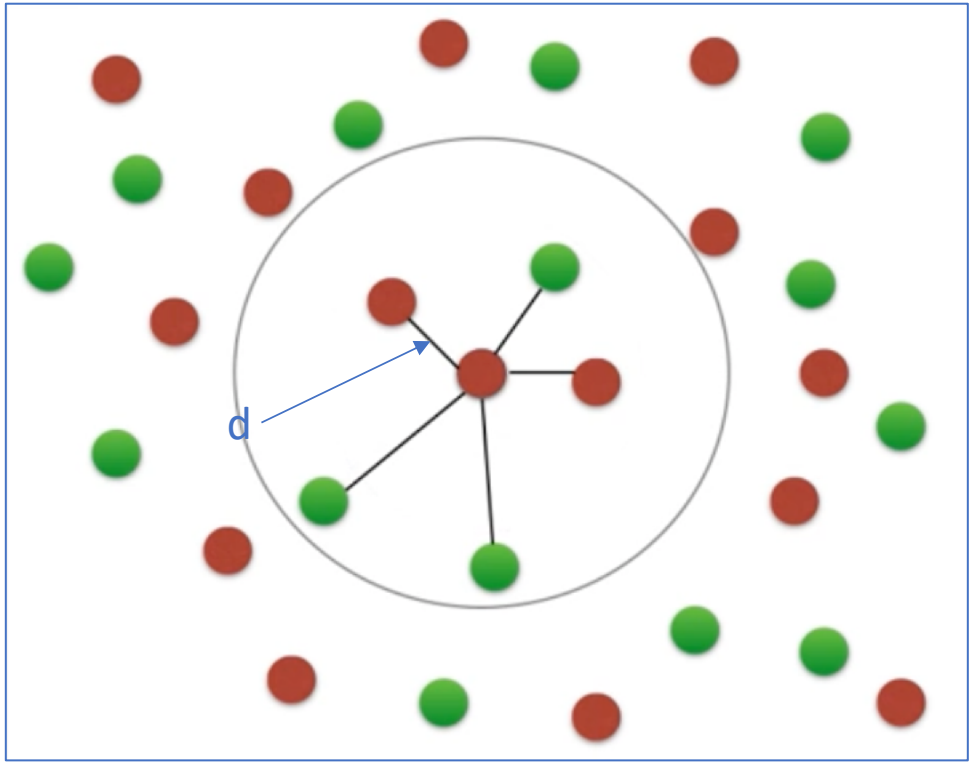
k-NN Graphical Example

- We want to Classify this point
- If we consider k=3 neighbors
  - Measured by some distance
  - The point is classified as Red
- If we consider k=5 neighbors



k-NN Graphical Example

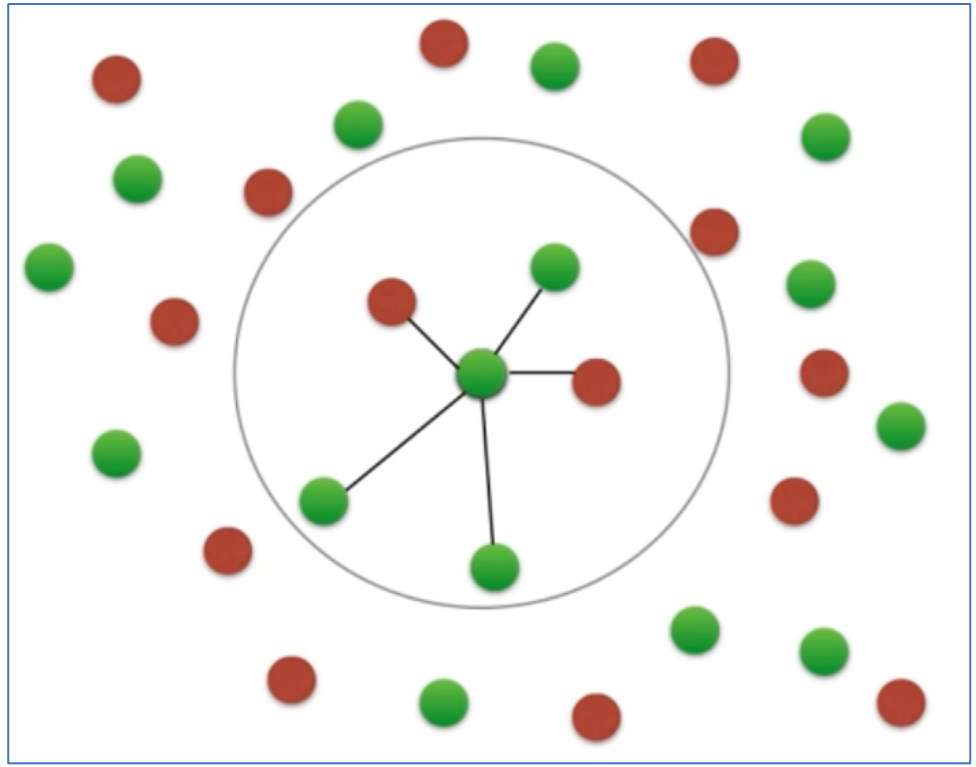
- We want to Classify this point
- If we consider k=3 neighbors
  - Measured by some distance
  - The point is classified as Red
- If we consider k=5 neighbors
  - Measured by some distance



HPSI
Human Performance Systems, Inc.

k-NN Graphical Example

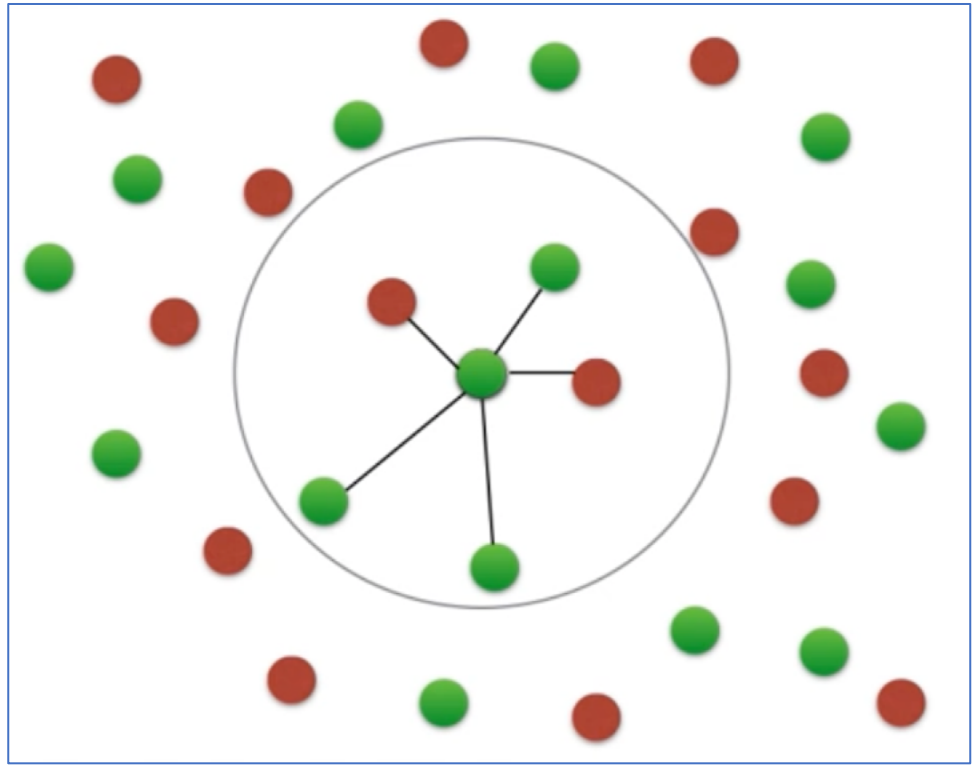
- We want to Classify this point
- If we consider k=3 neighbors
  - Measured by some distance
  - The point is classified as Red
- If we consider k=5 neighbors
  - Measured by some distance
  - The point is classified as Green



HPSI
Human Performance Systems, Inc.

k-NN Graphical Example

- We want to Classify this point
- If we consider k=3 neighbors
  - Measured by some distance
  - The point is classified as Red
- If we consider k=5 neighbors
  - Measured by some distance
  - The point is classified as Green
- So, k is a hyperparameter that we get to choose…



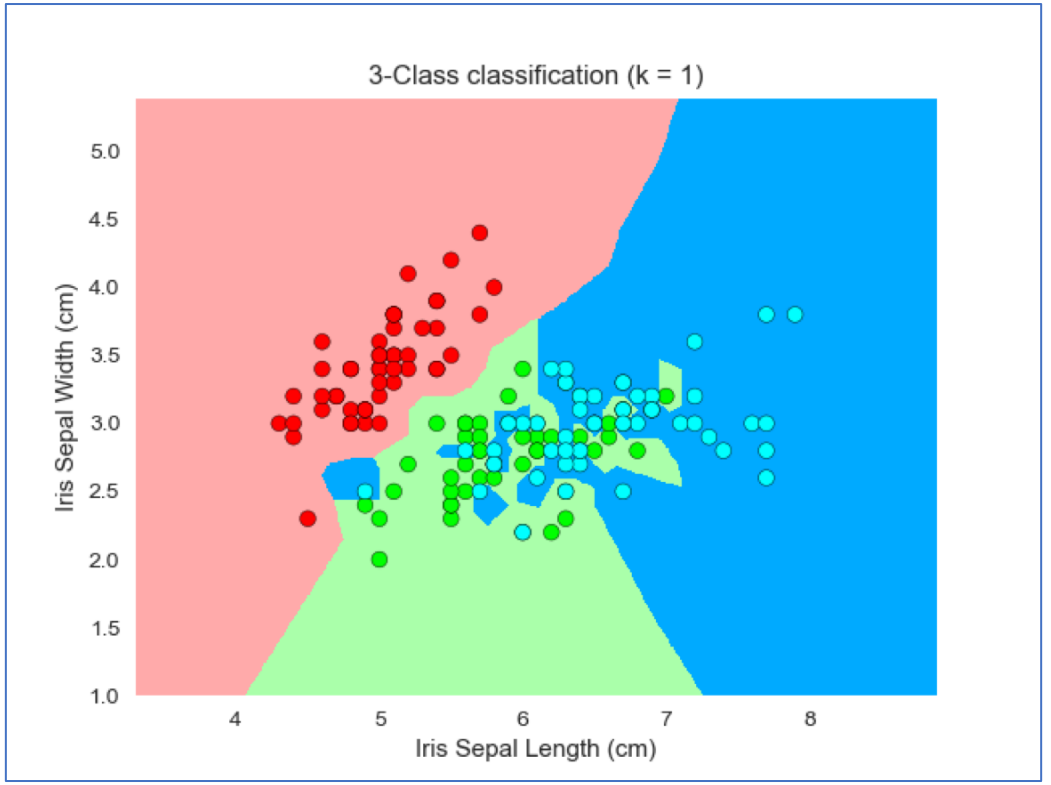
HPSI
Human Performance Systems, Inc.

kNN Decision Boundaries Example

- Consider this more complicated example
- Here we have a multiclass (R,G,B) dataset with significant overlap

kNN Decision Boundaries

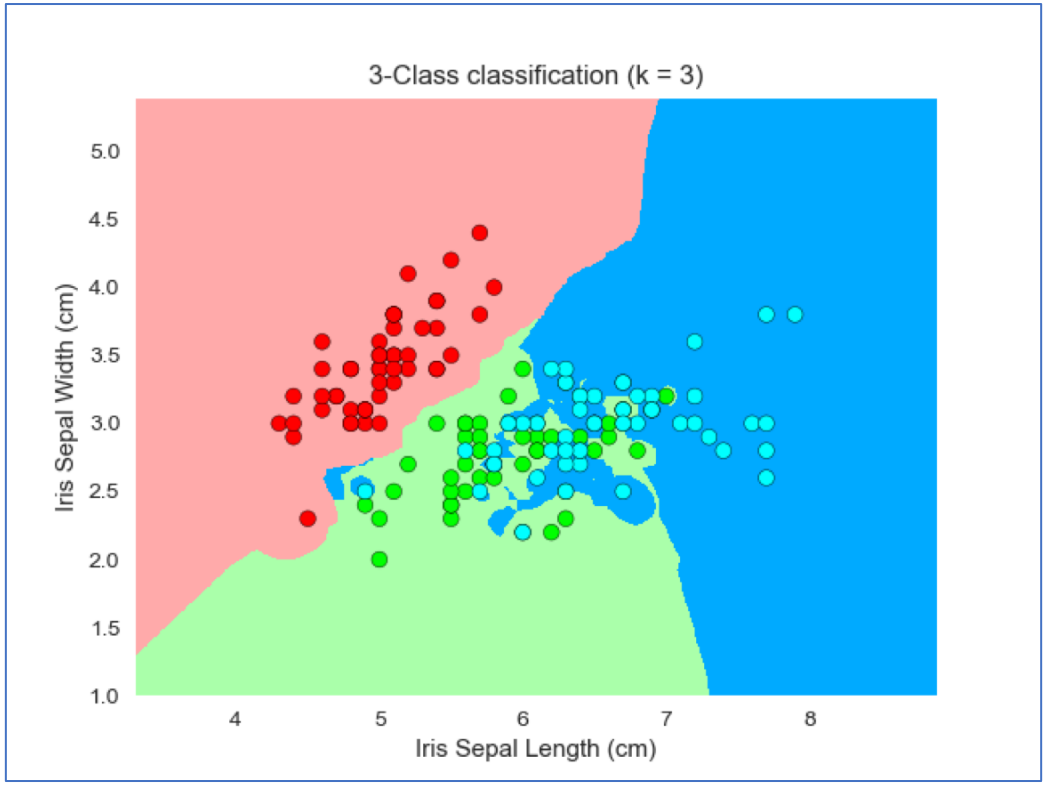
- k=1
- What if we increase k?



HPSI
Human Performance Systems, Inc.

kNN Decision Boundaries

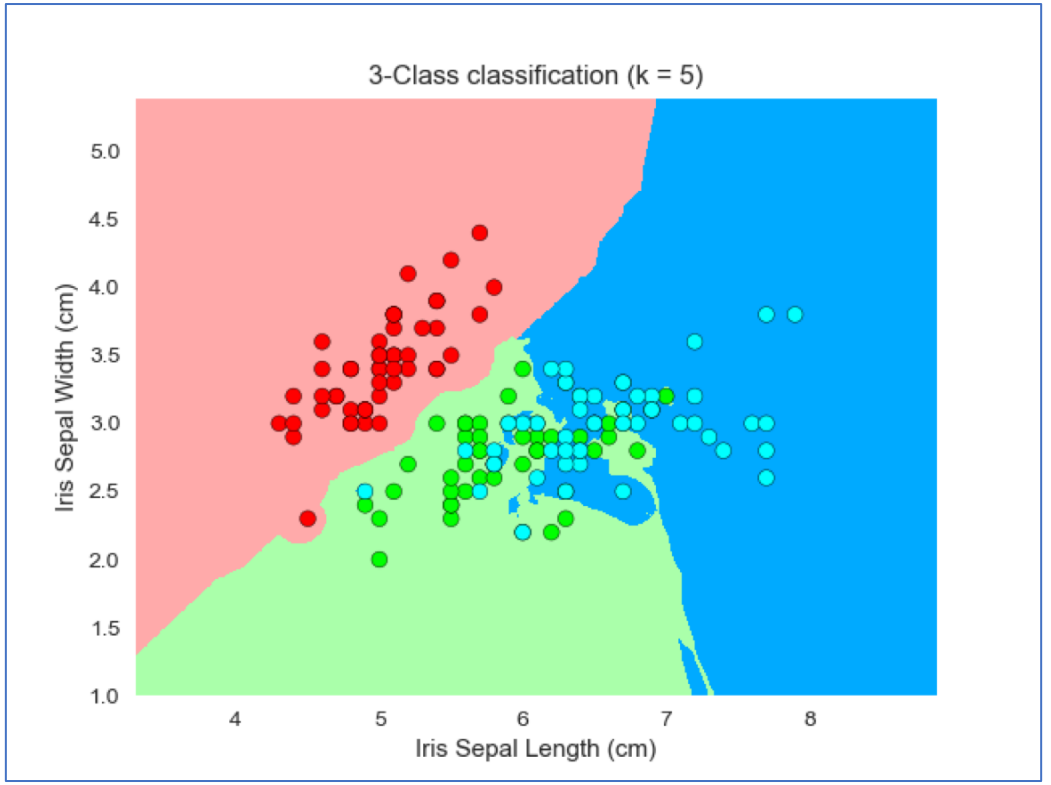
- k=3
- Increase Again?



HPSI
Human Performance Systems, Inc.

kNN Decision Boundaries

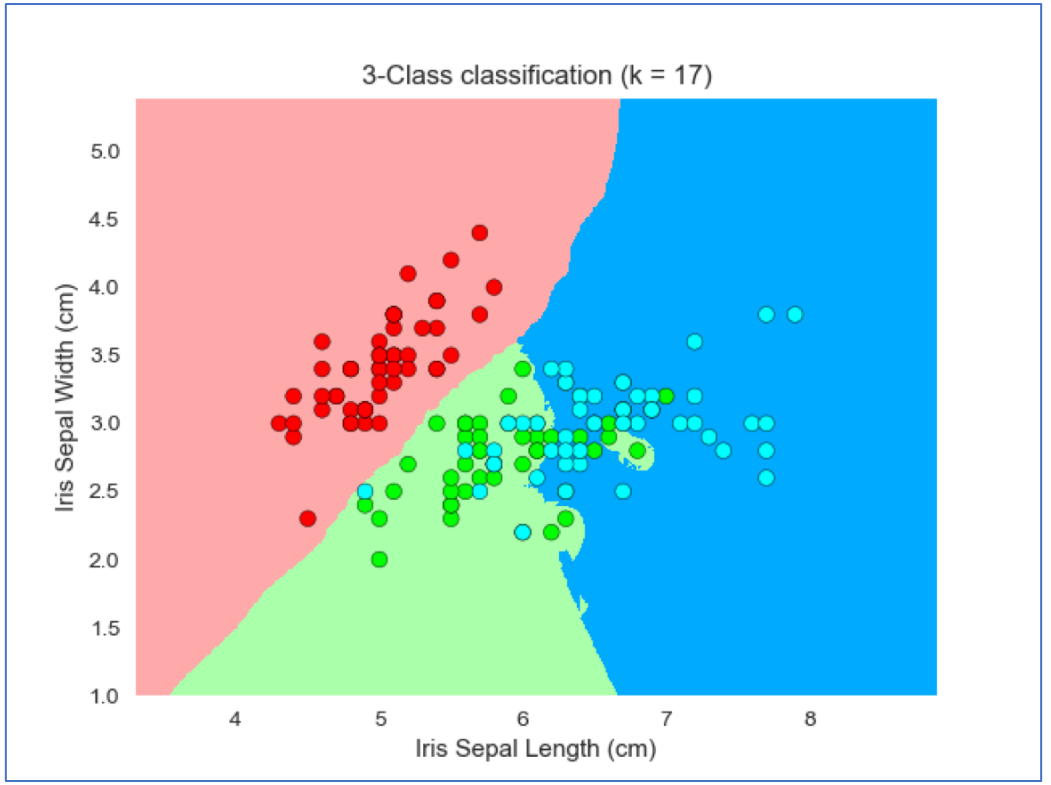
- k=5
- The decision boundary between R/G looks smoother…
- Let's keep increasing k



HPSI
Human Performance Systems, Inc.

kNN Decision Boundaries

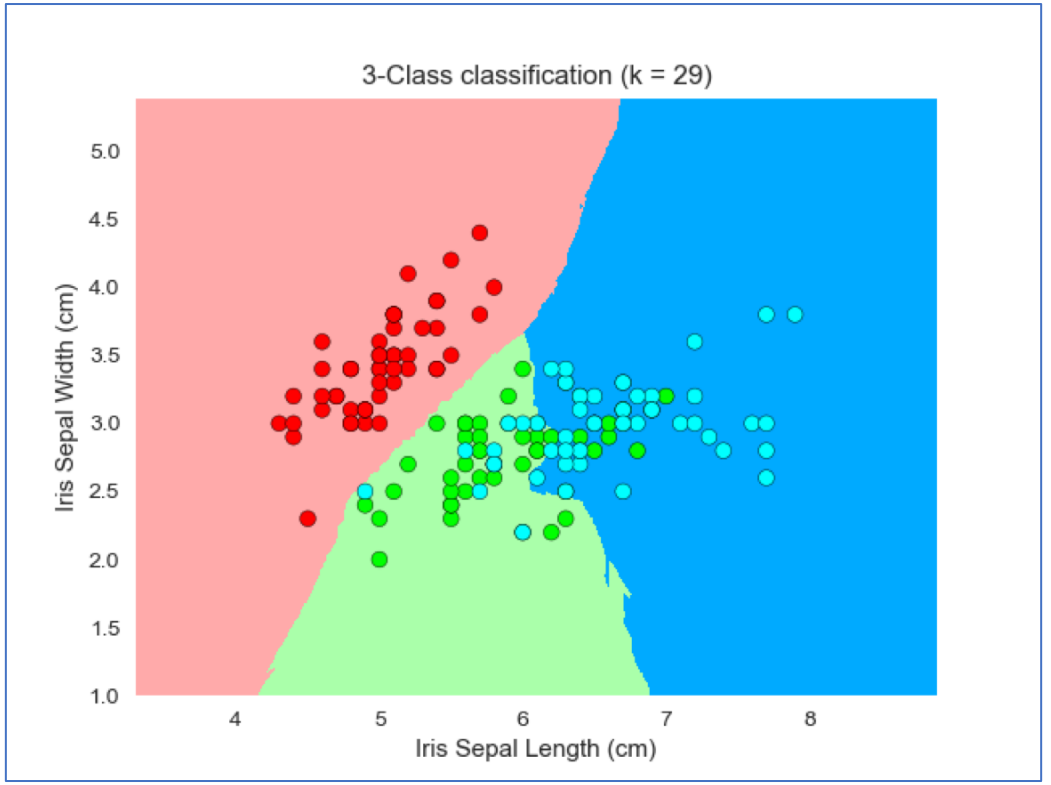
- K=17
- …smoother
- But we still have an "island"



HPSI
Human Performance Systems, Inc.

kNN Decision Boundaries

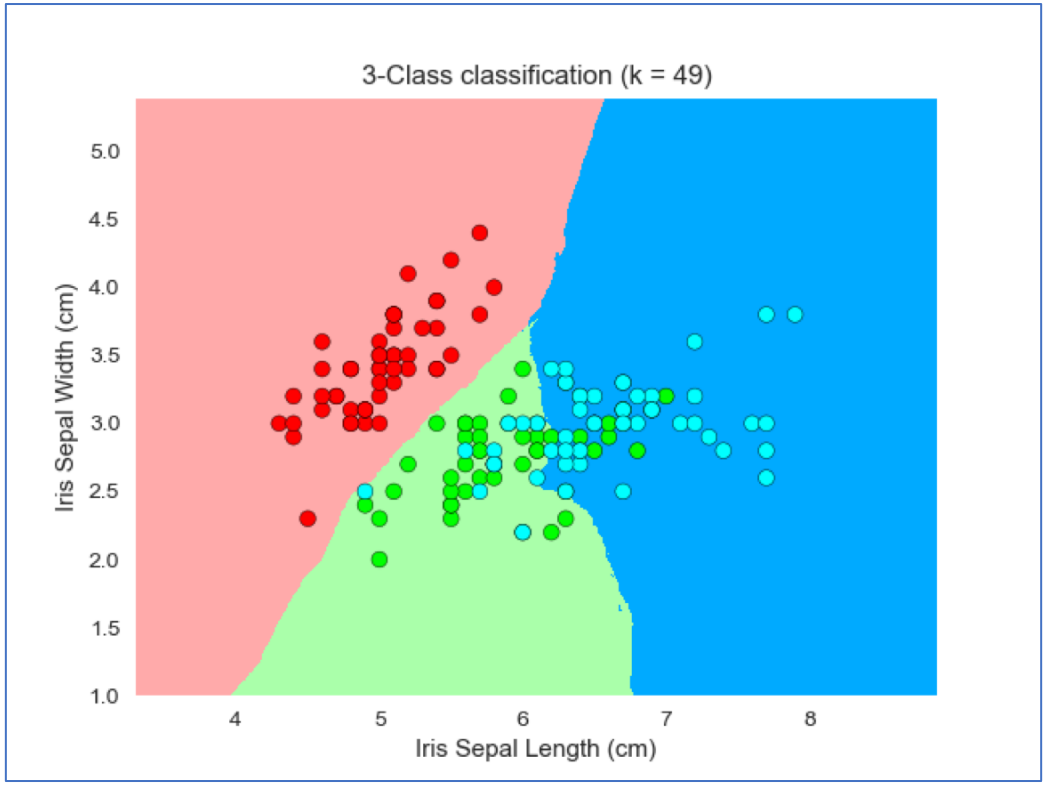
- K=29
- The “island” is gone



HPSI
Human Performance Systems, Inc.

kNN Decision Boundaries

- K=49
- …dang smooth
- So, do we just keep increasing k?
- How do we know what k to choose?



How to choose “k”

- Odd k (often 1, 3, or 5):
  - Avoids problem of breaking ties (there are other ways around this)
- Large k:
  - Less sensitive to noise (particularly class noise)
  - Better probability estimates for discrete classes
  - Larger training sets allow larger values of k
  - If you make k huge, every point will simply be classified as the most probable class (P(y))
- Small k:
  - Captures fine structure of problem space better
  - High variability in decision boundaries (more complex models)
  - May be necessary with small training sets
- It is dependent on your training data
- A good rule of thumb is $k = \sqrt{n}$ where n is the number of training examples
- Can plot a curve (see example)
- Can use grid-search

How to choose “distance”

- Recall the **norm** of a vector ***x*** is the length from the origin to ***x***
  - Used for measuring the size of a vector
  - Maps vectors to non-negative values
- L^p Norm
  - $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$
- L^2 Norm (Euclidean)
  - $\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2} = \sqrt{\sum_{i=1}^{n} |x_i|^2}$
- L^1 Norm (Manhattan/Taxicab)
  - $\|x\|_1 = \left(\sum_{i=1}^{n} |x_i|^1\right)^{1/1} = \sum_{i=1}^{n} |x_i|$

Distance Metrics

sklearn.neighbors.DistanceMetric

class sklearn.neighbors. **DistanceMetric**

DistanceMetric class

This class provides a uniform interface to fast distance metric functions. The various metrics can be accessed via the get_metric class method and the metric string identifier (see below). For example, to use the Euclidean distance:

```
>>> dist = DistanceMetric.get_metric('euclidean')
>>> X = [[0, 1, 2],
         [3, 4, 5]]
>>> dist.pairwise(X)
array([[ 0.        ,  5.19615242],
       [ 5.19615242,  0.        ]])
```

Available Metrics The following lists the string metric identifiers and the associated distance metric classes:

Metrics intended for real-valued vector spaces:

identifier	class name	args	distance function
“euclidean”	EuclideanDistance	•	`sqrt(sum((x - y)^2))`
“manhattan”	ManhattanDistance	•	`sum(|x - y|)`
“chebyshev”	ChebyshevDistance	•	`max(|x - y|)`
“minkowski”	MinkowskiDistance	p	`sum(|x - y|^p)^(1/p)`
“wminkowski”	WMinkowskiDistance	p, w	`sum(w * |x - y|^p)^(1/p)`
“seuclidean”	SEuclideanDistance	V	`sqrt(sum((x - y)^2 / V))`
“mahalanobis”	MahalanobisDistance	V or VI	`sqrt((x - y)' V^-1 (x - y))`

Distance Metrics

Metrics intended for integer-valued vector spaces: Though intended for integer-valued vectors, these are also valid metrics in the case of real-valued vectors.

identifier	class name	distance function
“hamming”	HammingDistance	`N_unequal(x, y) / N_tot`
“canberra”	CanberraDistance	`sum(|x - y| / (|x| + |y|))`
“braycurtis”	BrayCurtisDistance	`sum(|x - y|) / (sum(|x|) + sum(|y|))`

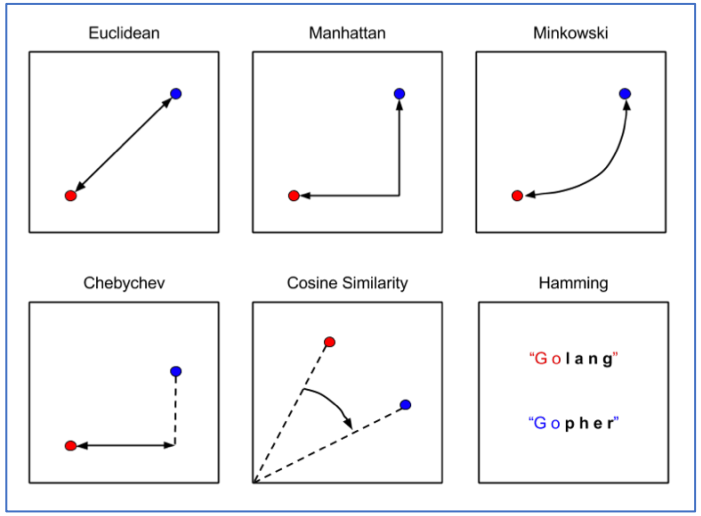
Metrics intended for boolean-valued vector spaces: Any nonzero entry is evaluated to “True”. In the listings below, the following abbreviations are used:

- N : number of dimensions
- NTT : number of dims in which both values are True
- NTF : number of dims in which the first value is True, second is False
- NFT : number of dims in which the first value is False, second is True
- NFF : number of dims in which both values are False
- NNEQ : number of non-equal dimensions, NNEQ = NTF + NFT
- NNZ : number of nonzero dimensions, NNZ = NTF + NFT + NTT

identifier	class name	distance function
“jaccard”	JaccardDistance	NNEQ / NNZ
“matching”	MatchingDistance	NNEQ / N
“dice”	DiceDistance	NNEQ / (NTT + NNZ)
“kulsinski”	KulsinskiDistance	(NNEQ + N - NTT) / (NNEQ + N)
“rogerstanimoto”	RogersTanimotoDistance	2 * NNEQ / (N + NNEQ)
“russellrao”	RussellRaoDistance	NNZ / N
“sokalmichener”	SokalMichenerDistance	2 * NNEQ / (N + NNEQ)
“sokalsneath”	SokalSneathDistance	NNEQ / (NNEQ + 0.5 * NTT)

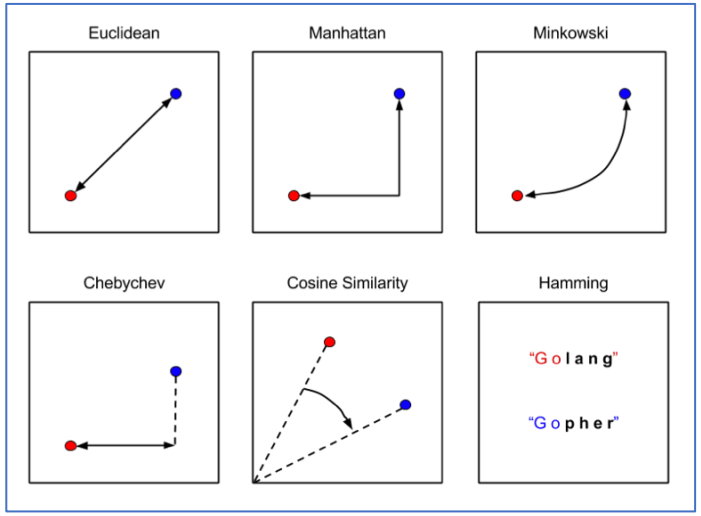
Distance Metric for Numeric Data

- **Euclidean Distance (L2-norm)**
  - The “standard”, “as the crow flies” – think Pythagorean theorem
  - Default for Numeric Attributes
  - Symmetric, spherical, treats all dimensions equally
  - Sensitive to extreme differences in single attribute
    - For example, if all training features are on a similar scale, but one feature is orders of magnitude larger
    - The large value of that feature is going to override the contribution of other features when computing distance
    - The value is squared



Distance Metric for Categorical Data

- **Hamming Distance**
  - Default for Categorical Attributes
  - Basically looks at each attribute and says “are they equal or not”
  - If they are equal, then distance is 0, if they are not equal, then distance is 1, so you basically count how many different attributes there are between the two instances



Distance Metrics Exercise

- Quick Exercise:
  - Consider this data set
    - 6 observations
    - 3 features (X1, X2, X3)
    - 2 Classes (Red, Green)
  - Suppose we want to use this data to make a prediction for a test point (X1=0, X2=0, X3=0)
  - Compute the Euclidean Distance between the test point and each observation
  - What is our prediction for k=1?
  - What is our prediction for k=3?

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

kNN Model Representation

- The model representation for KNN is the entire training dataset.
- It is as simple as that.
- KNN has no “model” other than storing the entire dataset, so there is no learning required.
- Efficient implementations can store the data using complex data structures (k-d trees) to make look-up during prediction efficient.

Preparing data for kNN

- General Best Practices
- Rescale Data
  - KNN performs much better if all of the data has the same scale.
    - Scale your data!
      - Min/Max
      - Standard Scaler
- Handle Missing Data
  - Missing data will mean that the distance between samples can not be calculated.
  - These samples should either be excluded (dropped) or the missing values should be imputed.
- Lower Dimensionality
  - KNN is best suited for lower dimensional data.
  - KNN can benefit from dimensionality reduction/feature selection (for ex: PCA) that reduces the dimensionality of the input feature space.

Advantages of k-NN

- Simple and fast to deploy
  - Little to no training time
- Easy to interpret/explain
- Naturally handles multiclass datasets
- Non-parametric
  - Does not assume any probability distributions on the input data
- Learns complex decision boundaries

Disadvantages of k-NN

- Storage of model takes a lot of disk space (contains entire training dataset)
- Curse of Dimensionality - often works best with 25 or fewer dimensions
  - There is little difference between the nearest and farthest neighbor in high dimensional data
- Computationally expensive predictions (large search problem to find nearest neighbors)
  - Might be impractical in industry settings
- Need to normalize - suffers from skewed class distributions
  - If one type of category occurs much more than another, classifying an input will be more biased towards that one category (dominates the majority vote since it is more likely to be neighbors with the input)

kNN Exercise

Naïve Bayes

Naïve Bayes

- Naive Bayes is a simple classification technique that relies on conditional probability, and predicts the most probable class given a set of inputs
- It is often used as a baseline for more complex models
- Can be thought of as a “Purely Statistical” model
- Naïve Bayes Classifiers are extremely fast and surprisingly accurate given their “naïve assumptions”
  - More on this later

Understanding Naïve Bayes Classifiers

- To be able to unpack the Naïve Bayes classifier definition, we need a good grasp on the following topics
  - Random Experiments
    - Sample Space
    - Event
    - Random Variable
  - Probability
    - Conditional Probability
    - Joint Probability
    - Marginal Probability
  - Statistical Independence
  - Mutual Exclusivity

Probability

- Suppose I have a coin that I am going to flip.
  - How likely is it to come up a head?
  - How likely is it to come up a tail?
  - How likely is it to come up an arm?
- Notice, that when we contemplate the likelihood of each outcome, we have in mind a set of all possible outcomes….
- Notice also that a single flip of a coin can result in only one outcome - It cannot be both heads and tails in a single flip
- We say that these outcomes are “mutually exclusive”

Probability

- Likelihood
  - When we compute the **likelihood** of an outcome, we first have in mind a set of all outcomes that are possible.
  - When this set exhausts all possible outcomes, and the outcomes are all mutually exclusive, this set is called the **sample space**.
- The sample space is determined by the measurement operation that we use to make an observation about the world.
  - There are often many things that we take for-granted about this measurement

Probability

- Consider the probability that a coin comes up heads when it is flipped.
  - If the coin is fair, it should come up heads in about 50% of the flips.
  - If the coin (or the flipping mechanism) is biased, then it will tend to come up heads more than or less than 50% of the flips.
  - The probability of coming up heads can be denoted with parameter label θ
  - For example:
    - A coin is fair when $\theta = 0.5$

Probability

- We can also consider our degree of belief that the coin is fair
- Maybe we know that the coin was recently manufactured by a government mint
  - High degree of belief that the coin is fair
- On the other hand, maybe the coin was manufactured by 'Acme Novelty Company'
  - High degree of belief that the coin is biased.

Probability

- We can denote the degree of belief about a parameter by $p(\theta)$
  - Minted Coin:
    - $p(\theta = 0.5) = 0.99$
  - Novelty Coin:
    - $p(\theta = 0.5) = 0.01$
    - $p(\theta = 0.9) = 0.99$

Probability

- Both “probability” of head or tail outcome, and “degree of belief” in biases refer to sample spaces.
- The sample space for flips of a coin consists of two possible outcomes:
  - {H,T}
- The sample space for coin bias consists of a continuum of possible values
  - $\{\theta = 0.0, \theta = 0.01, \theta = 0.02, \theta = 0.03, \ldots \theta = 1\}$
- When we flip a given coin, we are sampling from the first sample space, and when we grab a coin at random from a sack of coins (in which each coin may have a different bias) we are sampling from the second sample space

Probability

- Why should we care about coins?
- Because coin flips represent every real-life event that has a binary outcome….
- Any examples?
  - Heart surgery – successful or not
  - For a patient taking medicine – side effect or not
  - For a two-candidate election - whether one wins or not
  - …

Types of sample space

- Recall, what are the two kinds of sample space?
  - The sample space of the data (e.g. heads or tails)
  - The sample space of the parameters (e.g. bias)
- Each kind of sample space corresponds to a kind of probability
  - For data, the probability is over measureable outcomes that are “out there” in the world
  - For parameters, the probability is over unmeasurable beliefs that are “inside the head”

Probability as a mapping

- Either way, no matter whether the probability is “out there” in the world, or “inside the head”…
- A Probability is just a way of assigning numbers to a set of exhaustive and mutually exclusive events
- A “mapping” of possibilities to numbers

Probability Definition

- A probability needs to satisfy three properties (Kolmogorov, 1956):
  - A probability must be nonnegative
  - The sum of the probabilities across all events in the entire sample space must be 1
  - For any two mutually exclusive events, the probability that one or the other occurs is the sum of their individual probabilities

Probability Distributions

- A probability distribution is simply a list of all possible events and their corresponding probabilities
- There are two kinds of probability distributions
  - Discrete Distribution:
    - Probability of heads or tails
  - Continuous Distribution:
    - Probabilities of people's heights

Discrete Probability Distribution

- When the sample space consists of discrete outcomes (e.g., heads or tails), the probability distribution is a list of probabilities of the outcomes
- The probability of a discrete outcome is called a **probability mass**
- The sum of the probability masses across the sample space must be 1
- Figure 2.1 (see next page) shows the probability masses of four suspects in the Holmes example

Discrete Probability Example

- Example
  - Consider the simple experiment of tossing a coin three times. Let X = number of times the coin comes up heads. The 8 possible elementary events and the corresponding values for X are:

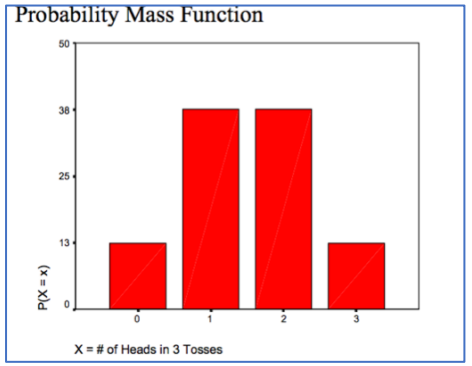
Elementary Event	Count of Heads (X)
TTT	0
TTH	1
THT	1
HTT	1
THH	2
HTH	2
HHT	2
HHH	3

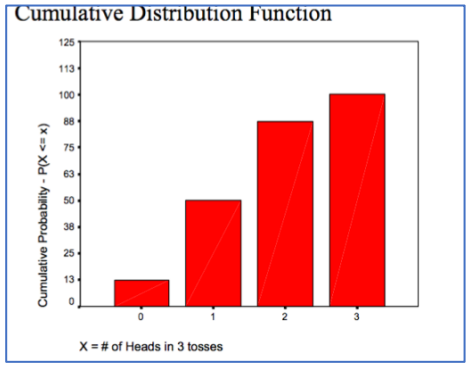
Discrete Probability Example

- Example
  - Therefore, the probability distribution for the number of heads occurring in three coin tosses is

Count of Heads (X)	P(X)	P(X<=X)
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	1

$$P(x) = \begin{cases} 1/8 & if\ x = 0 \\ 3/8 & if\ x = 1, 2 \\ 1/8 & if\ x = 3 \\ 0 & Otherwise \end{cases}$$





HPSI
Human Performance Systems, Inc.

Continuous Probability Distribution – Probability Density

- When the sample space consists of continuous outcomes (ex: people's heights) we cannot use probability mass for a specific outcome.
- Why not?
  - Because the probability mass for a specific outcome will be zero
  - In other words, the probability of someone's height being exactly 67.2141390842076153…
- Instead, we can:
  - Discretize the space into a finite set of mutually exclusive and exhaustive intervals
  - Calculate the probability mass in each interval
  - Use the ratio of probability mass to interval width
  - This ratio is called the **Probability Density**

Probability Density

- The top panel of this figure shows the discretized intervals and probability mass in each interval
  - The probability of a discrete outcome, such as the probability of falling into an interval on a continuous scale, is referred to as Probability Mass
  - $\frac{Probability\ Mass}{Interval\ Width} = Probability\ Density$
  - The amount of "stuff" per unit of space that it takes up
- The second panel shows the probability density
- The third panel shows the narrower intervals and probability mass in each interval
- The bottom panel shows the probability density corresponding to the more narrow intervals
- Generally, the skinnier the intervals are, the more accurate the probability density is

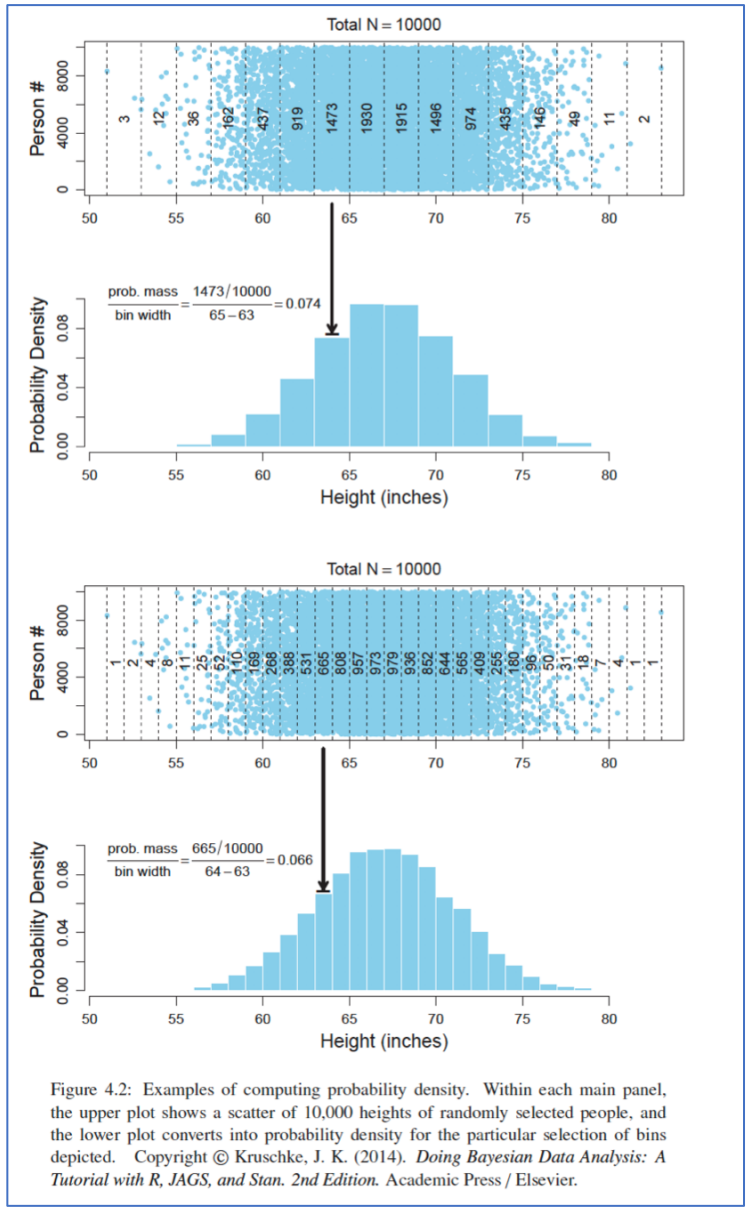


Figure 4.2: Examples of computing probability density. Within each main panel, the upper plot shows a scatter of 10,000 heights of randomly selected people, and the lower plot converts into probability density for the particular selection of bins depicted. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Probability Density

- While probability mass cannot exceed 1, probability densities can
- The upper panel of this figure shows that most of the probability mass is concentrated around 84
- Consequently, the probability density near 84 exceeds 1.0, as shown in the lower panel
- This simply means that there is a high concentration of probability mass relative to the width of the interval

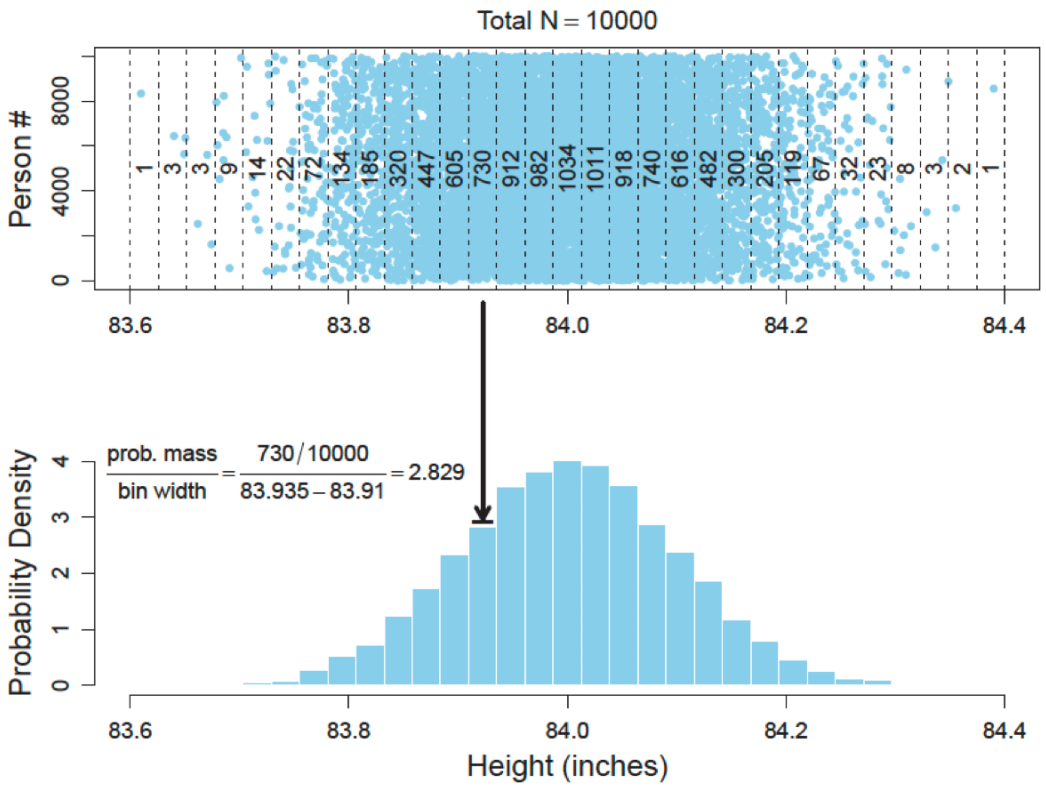


Figure 4.3: Example of probability density greater than 1.0. Here, all the probability mass is concentrated into a small region of the scale, and therefore the density can be high at some values of the scale. The annotated calculation of density uses rounded interval limits for display. (For this example, we can imagine that the points refer to manufactured doors instead of people, and therefore the *y*-axis of the top panel should be labelled "Door" instead of "Person.") Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

HPSI
Human Performance Systems, Inc.

Properties of Probability Density Functions

- We need to define some notations first
- Let:
  - x be the continuous variable
  - Δx be the width of an interval on x
  - i be an index for the intervals
  - $[x_i, x_i + \Delta x]$ be the interval between x_i and $x_i + \Delta x$
  - $P([x_i, x_i + \Delta x])$be the probability mass of the ith interval
- Then the sum of those probability masses must be 1:

$$\sum_i P([x_i, x_i + \Delta x]) = 1$$

- We can rewrite the equation above in terms of the density of each interval, by dividing and multiplying by x:

$$\sum_i \frac{\Delta x * P([x_i, x_i + \Delta x])}{\Delta x} = 1$$

Properties of Probability Density Functions

- In the limit, as the interval width becomes infinitesimal, we denote:
  - Summation as $\int$ instead of $\sum$
  - The width of the interval around x as dx instead of Δx
  - The probability density in the infinitesimal interval around x as $\mathrm{p}(x)$
- Then, the previous equation (in terms of density) can be rewritten as:

$$\sum_i \frac{\Delta x * P([x_i, x_i + \Delta x])}{\Delta x} = 1 \Rightarrow \int dx\, p(x) = 1$$

- We use $p(x)$ to represent the probability mass when x is discrete
- Thus, what $p(x)$ represents depends on the context
  - Is x discrete or continuous?

The Normal Probability Density Functions

- Perhaps the most famous probability density function is the normal distribution, also known as the Gaussian distribution
- The probability density function of normal distribution is

$$\mathrm{p}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right)}$$

- Recall, what are σ and μ? what do they control?
- An example of the probability density is shown in the figure where the x axis is divided into a dense comb of small intervals
- The figure also shows that the area under the curve is, in fact, 1

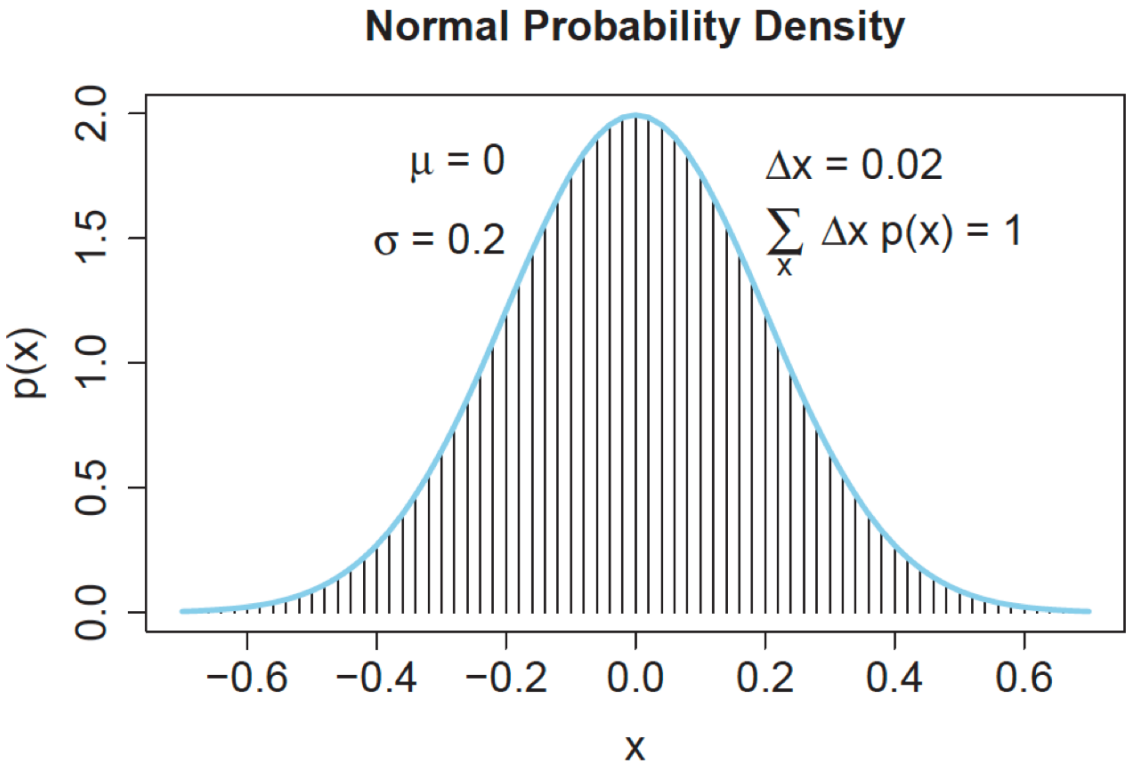


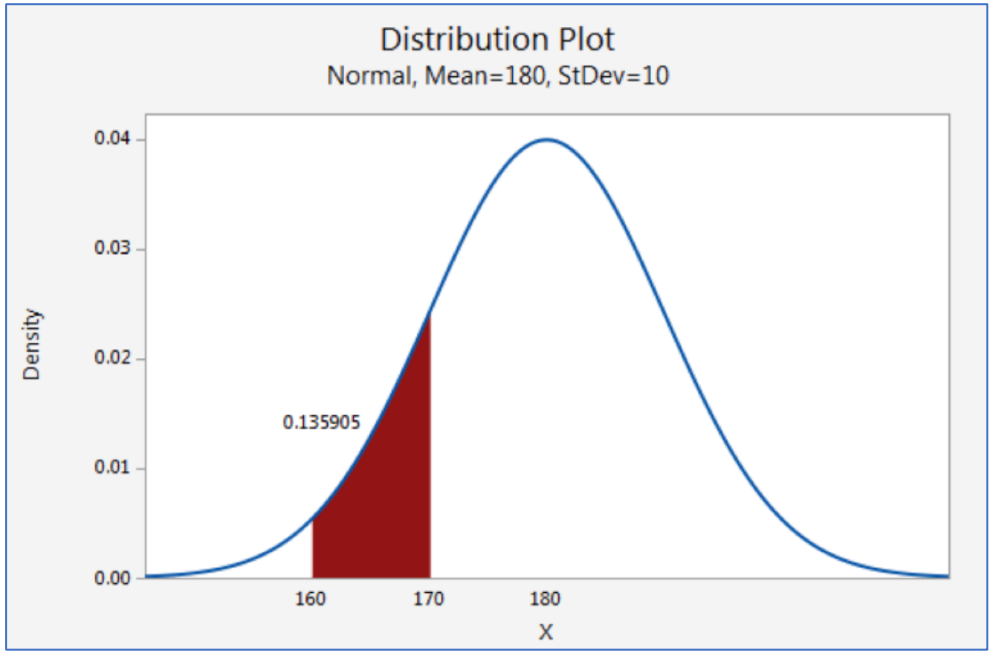
Figure 4.4: A normal probability density function, shown with a comb of narrow intervals. The integral is approximated by summing the width times height of each interval. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

HPSI
Human Performance Systems, Inc.

Example - Continuous Normal Distribution

- Example of the continuous distribution of weights
  - The continuous normal distribution can describe the distribution of weight of adult males.
  - For example, you can calculate the probability that a man weighs between 160 and 170 pounds.
  - The area of this range is 0.136; therefore, the probability that a randomly selected man weighs between 160 and 170 pounds is 13.6%.
  - The entire area under the curve equals 1.0

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$



Bayes Rule (Just a preview)

- So why did we need to learn all of this probability nonsense?

$$P(C|X)=$$

Bayes Rule (Just a preview)

- Bayes rule is merely the mathematical relation between the prior allocations of credibility and the posterior reallocation of credibility (conditional on data)

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

Conditional Probability – Order Matters

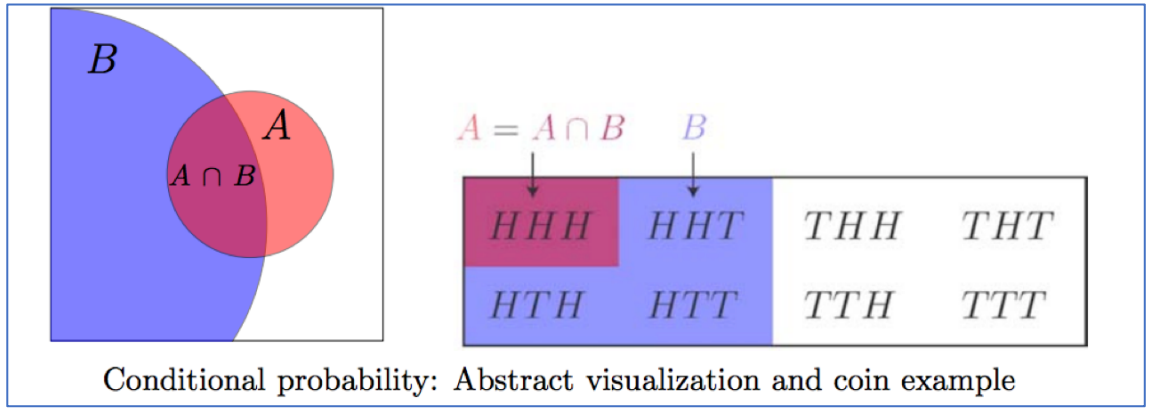
- Why?
- $\mathrm{P}(cute|puppy) \neq \mathrm{P}(puppy|cute)$

Conditional Probability Example

- In terms of **Sample Space**….
- Coin Toss Example:
  - Toss a fair coin 3 times
  - What is the probability of 3 heads?
    - Answer:
      - $Sample\ Space\ = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
      - All outcomes are equally likely (if the coin is fair)
      - $P(HHH) = \frac{1}{8}$
  - Suppose we are told that the first toss was heads
  - Given this information, how should we compute the **probability of Flipping "Heads" three times in a row**
    - Answer:
      - We have a new (reduced) $Sample\ Space\ = \{HHH, HHT, HTH, HTT\}$
      - All outcomes are still equally likely (the coin is still fair)
      - $P(HHH) = \frac{1}{4}$

Conditional Probability Example

- We can visualize the conditional probability as follows
  - Think of $P(A)$ as the proportion of the area of the whole sample space taken up by A
  - For $P(A|B)$ we restrict our attention to B
  - $P(A|B)$ is the proportion of B taken up by A
    - $P(A|B) = \frac{P(A \cap B)}{P(B)}$



Conditional probability: Abstract visualization and coin example

Joint Probability and Marginal Probability

- This table shows the probabilities of various combinations of people's eye/hair color
- Each entry indicates the **joint probability** of particular combinations of eye color (e) and hair color (h), denoted by $p(e, h)$
- The right margin of the table shows the probabilities of the eye colors overall, collapsed across hair colors
- Such probabilities are called **marginal probability**, denoted by $p(e)$:

$$p(e) = \sum_{h} p(e, h)$$

- The marginal probabilities of the hair colors, $p(h)$, are indicated on the lower margin of the table:

$$p(h) = \sum_{e} p(e, h)$$

Table 4.1: Proportions of combinations of hair color and eye color. Some rows or columns may not sum exactly to their displayed marginals because of rounding error from the original data. Data adapted from Snee (1974). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Eye Color	Hair Color				Marginal
	Black	Brunette	Red	Blond	(Eye Color)
Brown	.11	.20	.04	.01	.37
Blue	.03	.14	.03	.16	.36
Hazel	.03	.09	.02	.02	.16
Green	.01	.05	.02	.03	.11
Marginal (Hair Color)	.18	.48	.12	.21	1.0

Conditional Probability

- Conditional Probability
  - $P(e|h)$ is the probability of the occurrence of event e, given that h occurred is given as:
    - $P(e|h) = \frac{P(e \cap h)}{P(h)} = \frac{P(e,h)}{P(h)}$
  - Answers the question:
    - How does the probability of an event change if we have extra information?

Table 4.2: Example of conditional probability. Of the blue-eyed people in Table 4.1, what proportion have hair color h? Each cell shows $p(h|\text{blue}) = p(\text{blue}, h)/p(\text{blue})$ rounded to two decimal points. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

	Hair Color				Marginal
Eye Color	Black	Brunette	Red	Blond	(Eye Color)
Blue	.03/.36 = .08	.14/.36 = .39	.03/.36 = .08	.16/.36 = .45	.36/.36 = 1.0

Statistical Independence Example

- Independent Events
  - If we want to calculate the joint probability of two independent events, we can simply multiply each probability together to get the joint probability
  - "Joint Distribution" = "Product Distribution"
  - $\mathrm{P}(x, y) = P(x) * P(y)$
- For Example:
  - Probability of tossing a coin and getting "Heads":

$$P(Heads) = P(x) = \frac{1}{2}$$

  - Probability of rolling a dice and getting "3":

$$P(Roll\ \text{“3”}) = P(y) = \frac{1}{6}$$

$$P(Heads) * P(Roll\ \text{“3”}) = \mathrm{P}(x, y) = \frac{1}{2} * \frac{1}{6} = \frac{1}{12}$$

Bayesian Inference

➢The Naïve Bayes Classifier relies on three things:

1. Independence assumption
2. The notion of conditional probability
3. **Bayesian Inference -** a method of statistical inference in which Bayes Theorem is used to update the probability for a hypothesis as more evidence becomes available.

Bayesian Inference Example

- We observe that the sidewalk is wet
- What are the possible causes?

Bayesian Inference Example

- We observe that the sidewalk is wet
- What are the possible causes?
  - It rained recently
  - Sprinkler
  - Broken water main (pipe)
  - Person spilled a drink
  - Dog marked his territory

Bayesian Inference Example

- Based on information that we have, we have some notion that certain probabilities are greater than others
- For example:
  - P(recent rain) > P(sprinkler)
  - P(recent rain) > P(spilled drink)
- Bayesian inference incorporates previous knowledge (prior probabilities)

Bayesian Inference Example

- Observation A:
  - Suppose we *observe* that the sidewalk is wet, in addition to the grass, the trees, the street, and the parked cars
  - How do the probabilities change given this new information?
  - P(recent rain | Observation A) ↑
  - P(spilled drink | Observation A) ↓

Bayesian Inference Example

- Observation B:
  - Now suppose that instead we *observe* that the sidewalk is wet, but it is localized to a small area next to an empty water bottle
  - How do the probabilities change given this new information?
  - P(recent rain | Observation B) ↓
  - P(spilled drink | Observation B) ↑
- This leads us in to “Bayes Rule”

Bayesian Inference – Sherlock Holmes Example

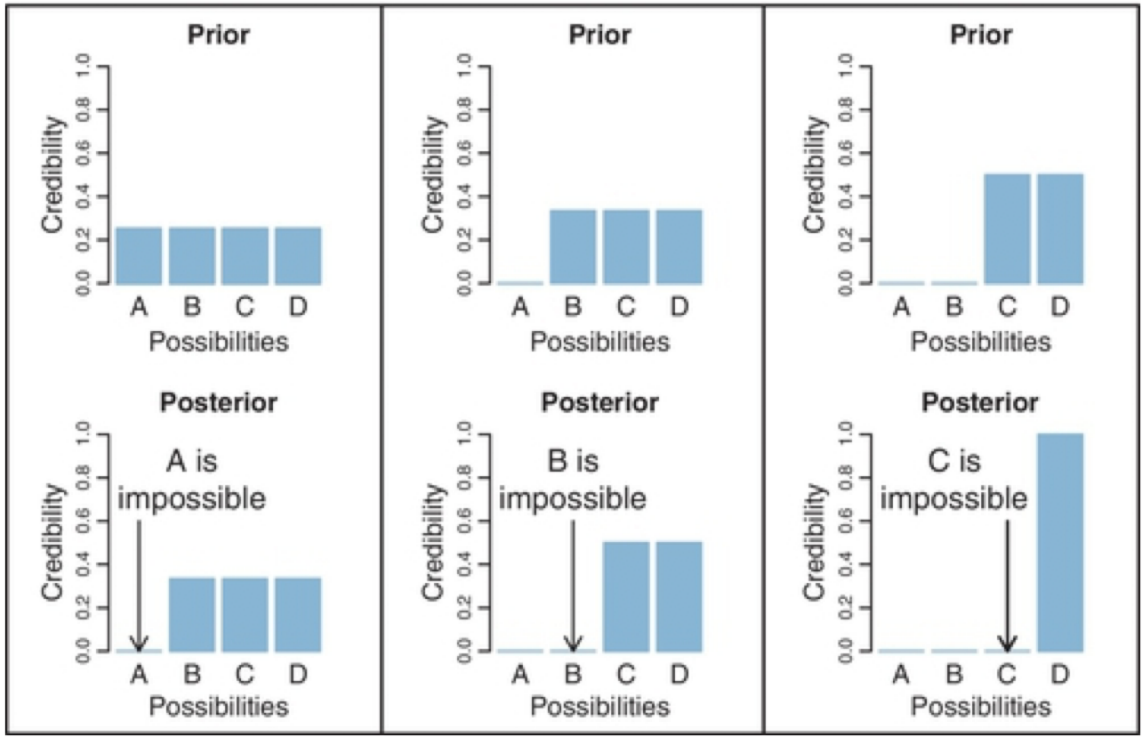


FIGURE 2.1 The upper-left graph shows the credibilities of the four possible causes for an outcome. The causes, labeled A, B, C, and D, are mutually exclusive and exhaust all possibilities. The causes happen to be equally credible at the outset; hence all have prior credibility of 0.25. The lower-left graph shows the credibilities when one cause is learned to be impossible. The resulting posterior distribution is used as the prior distribution in the middle column, where another cause is learned to be impossible. The posterior distribution from the middle column is used as the prior distribution for the right column. The remaining possible cause is fully implicated by Bayesian reallocation of credibility.

Bayes Rule Proof

- Recall, based on the definition of conditional probability, what are $p(\theta|y)$ and $p(y|\theta)$?

(1) $$p(\theta|y) = \frac{p(\theta,y)}{p(y)}$$

(2) $$p(y|\theta) = \frac{p(\theta,y)}{p(\theta)}$$

Bayes Rule Proof

- Recall, based on the definition of conditional probability, what are $p(\theta|y)$ and $p(y|\theta)$?

(1) $$p(\theta|y) = \frac{p(\theta,y)}{p(y)}$$

(2) $$p(y|\theta) = \frac{p(\theta,y)}{p(\theta)}$$

- We can rewrite (2) as:

(3) $$p(\theta, y) = p(y|\theta) * p(\theta)$$

Bayes Rule Proof

- Recall, based on the definition of conditional probability, what are $p(\theta|y)$ and $p(y|\theta)$?

(1) $$p(\theta|y) = \frac{p(\theta,y)}{p(y)}$$

(2) $$p(y|\theta) = \frac{p(\theta,y)}{p(\theta)}$$

- We can rewrite (2) as:

(3) $$p(\theta,y) = p(y|\theta) * p(\theta)$$

- Substituting into (1) above:

(4) $$p(\theta|y) = \frac{p(y|\theta)*p(\theta)}{p(y)}$$

Bayes Rule Proof

- Recall, based on the definition of conditional probability, what are $p(\theta|y)$ and $p(y|\theta)$?

(1) $$p(\theta|y) = \frac{p(\theta,y)}{p(y)}$$

(2) $$p(y|\theta) = \frac{p(\theta,y)}{p(\theta)}$$

- We can rewrite (2) as:

(3) $$p(\theta,y) = p(y|\theta) * p(\theta)$$

- Substituting into (1) above:

(4) $$p(\theta|y) = \frac{p(y|\theta)*p(\theta)}{p(y)}$$

Bayes Rule!

Applying Bayes Rule to an example

- Consider trying to diagnose a rare disease …
  - Suppose that in the general population, the probability of having the disease is only one in a thousand
  - Let's say that we developed a test (y) for the disease that has a **hit rate of 99%**
  - **This means that if a person has the disease, then the test result is positive 99% of the time**
  - We also have a false alarm rate (FPR) of 5%
  - Suppose a person tests positive
  - What is the probability of having the disease?

Applying Bayes Rule to an example

- Suppose the test result is positive. What is the probability of having the disease?
- What do we already know?
  - We already know the prior probability and the likelihood
    - $p(\theta)$
    - $p(y|\theta)$
- What do we want to know?
  - We want to know the posterior probability
    - $p(\theta|y)$
- Let's draw this out…

Applying Bayesian Inference to parameters/data

- Naïve Bayes is a machine learning method that can be used to predict the likelihood that an event will occur given evidence that is present in your data
- The Naïve Bayes Classifier relies on Bayesian inference at its core
- Makes two “Naïve” assumptions over attributes

Naïve Bayes Classifier Assumptions

- Fundamental **assumptions**:
  - Each feature makes an **independent** and **equal** contribution to the outcome
    - We assume that no pair of features are dependent on one another in any way (complete independence)
      - Temperature has nothing to do with humidity, and has no effect on whether or not it is windy
    - Each feature is given the same weight/importance
      - We assume that none of the attributes is irrelevant, and that they are all contributing equally to the outcome.

Naïve Bayes Classifier

- To illustrate the inner workings of the Naïve Bayes Classifier, we will consider an example:
  - We have recorded weather features about the last 14 times that we played golf.
  - We also recorded the result of whether the conditions were favorable or not.
  - We will demonstrate how a Naïve Bayes Classifier can be used to determine whether or not the a specific set of weather conditions supports playing golf.

Naïve Bayes Classifier Example

- Weather Features:
  - Outlook
    - (Rainy, Overcast, Sunny)
  - Temperature
    - (Hot, Mild, Cool)
  - Humidity
    - (High, Normal)
  - Windy
    - (False, True)
- Weather Dependent Variable:
  - Play Golf
    - (YES, NO)

	outlook	temperature	humidity	windy	play_golf
0	Sunny	Hot	High	False	NO
1	Sunny	Hot	High	True	NO
2	Overcast	Hot	High	False	YES
3	Rainy	Mild	High	False	YES
4	Rainy	Cool	Normal	False	YES
5	Rainy	Cool	Normal	True	NO
6	Overcast	Cool	Normal	True	YES
7	Sunny	Mild	High	False	NO
8	Sunny	Cool	Normal	False	YES
9	Rainy	Mild	Normal	False	YES
10	Sunny	Mild	Normal	True	YES
11	Overcast	Mild	High	True	YES
12	Overcast	Hot	Normal	False	YES
13	Rainy	Mild	High	True	NO

Naïve Bayes Classifier

How does the Naïve Bayes Classifier work?

- Step 1:
  - Convert the data set into frequency tables
  - Use the frequency tables to calculate likelihood tables
- Step 2:
  - Use the product rule to obtain a joint conditional probability for the attributes
- Step 3:
  - Use Bayes Rule to calculate the posterior probability for each class variable
  - Once this has been done for all classes, output the class with the highest probability

Naïve Bayes Classifier Example

- Step 1:
  - Create a frequency table for the Class:
    - P(C) or P(CLASS)
    - P(YES) = 9/14

	outlook	temperature	humidity	windy	play_golf
0	Sunny	Hot	High	False	NO
1	Sunny	Hot	High	True	NO
2	Overcast	Hot	High	False	YES
3	Rainy	Mild	High	False	YES
4	Rainy	Cool	Normal	False	YES
5	Rainy	Cool	Normal	True	NO
6	Overcast	Cool	Normal	True	YES
7	Sunny	Mild	High	False	NO
8	Sunny	Cool	Normal	False	YES
9	Rainy	Mild	Normal	False	YES
10	Sunny	Mild	Normal	True	YES
11	Overcast	Mild	High	True	YES
12	Overcast	Hot	Normal	False	YES
13	Rainy	Mild	High	True	NO

Naïve Bayes Classifier Example

- Step 1:
  - Create a frequency table for the Class:
    - P(C) or P(CLASS)
    - P(YES) = 9/14
    - P(NO) = 5/14

	outlook	temperature	humidity	windy	play_golf
0	Sunny	Hot	High	False	NO
1	Sunny	Hot	High	True	NO
2	Overcast	Hot	High	False	YES
3	Rainy	Mild	High	False	YES
4	Rainy	Cool	Normal	False	YES
5	Rainy	Cool	Normal	True	NO
6	Overcast	Cool	Normal	True	YES
7	Sunny	Mild	High	False	NO
8	Sunny	Cool	Normal	False	YES
9	Rainy	Mild	Normal	False	YES
10	Sunny	Mild	Normal	True	YES
11	Overcast	Mild	High	True	YES
12	Overcast	Hot	Normal	False	YES
13	Rainy	Mild	High	True	NO

Naïve Bayes Classifier Example

- Step 1:
  - Create a frequency table for the Class:
    - P(C) or P(CLASS)
    - P(YES) = 9/14
    - P(NO) = 5/14
- Class Frequency Table:

Play	
YES	9
NO	5
Total	**14**

	outlook	temperature	humidity	windy	play_golf
0	Sunny	Hot	High	False	NO
1	Sunny	Hot	High	True	NO
2	Overcast	Hot	High	False	YES
3	Rainy	Mild	High	False	YES
4	Rainy	Cool	Normal	False	YES
5	Rainy	Cool	Normal	True	NO
6	Overcast	Cool	Normal	True	YES
7	Sunny	Mild	High	False	NO
8	Sunny	Cool	Normal	False	YES
9	Rainy	Mild	Normal	False	YES
10	Sunny	Mild	Normal	True	YES
11	Overcast	Mild	High	True	YES
12	Overcast	Hot	Normal	False	YES
13	Rainy	Mild	High	True	NO

How does the Naïve Bayes Classifier work?

- Step 1:
  - Convert the data set into frequency tables
  - Use the frequency tables to calculate likelihood tables
- Step 2:
  - Use the product rule to obtain a joint conditional probability for the attributes
- Step 3:
  - Use Bayes Rule to calculate the posterior probability for each class variable
  - Once this has been done for all classes, output the class with the highest probability

Naïve Bayes Classifier Example

- Step 1:
  - P(C) or P(CLASS)
  - P(YES) = 9/14 = 0.643
  - P(NO) = 5/14 = 0.357
- The Class Frequency Table can be used to create the Class Likelihood Table by dividing each class frequency by the total (relative probability)
- Likelihood Table:

Play		P(YES)
YES	9	9/14 = 0.643
NO	5	5/14 = 0.357
Total	**14**	**14/14 = 100%**

	outlook	temperature	humidity	windy	play_golf
0	Sunny	Hot	High	False	NO
1	Sunny	Hot	High	True	NO
2	Overcast	Hot	High	False	YES
3	Rainy	Mild	High	False	YES
4	Rainy	Cool	Normal	False	YES
5	Rainy	Cool	Normal	True	NO
6	Overcast	Cool	Normal	True	YES
7	Sunny	Mild	High	False	NO
8	Sunny	Cool	Normal	False	YES
9	Rainy	Mild	Normal	False	YES
10	Sunny	Mild	Normal	True	YES
11	Overcast	Mild	High	True	YES
12	Overcast	Hot	Normal	False	YES
13	Rainy	Mild	High	True	NO

Naïve Bayes Classifier Example

- The next step is to repeat this process for each weather feature and compute their corresponding Feature Frequency Tables
- These Frequency tables can then be used to create Likelihood tables as previously demonstrated for each feature in our dataset

	outlook	temperature	humidity	windy	play_golf
0	Sunny	Hot	High	False	NO
1	Sunny	Hot	High	True	NO
2	Overcast	Hot	High	False	YES
3	Rainy	Mild	High	False	YES
4	Rainy	Cool	Normal	False	YES
5	Rainy	Cool	Normal	True	NO
6	Overcast	Cool	Normal	True	YES
7	Sunny	Mild	High	False	NO
8	Sunny	Cool	Normal	False	YES
9	Rainy	Mild	Normal	False	YES
10	Sunny	Mild	Normal	True	YES
11	Overcast	Mild	High	True	YES
12	Overcast	Hot	Normal	False	YES
13	Rainy	Mild	High	True	NO

Naïve Bayes Classifier Example

➢ **Outlook** Frequency and Likelihood

Outlook				
	YES	NO	P(YES)	P(NO)
Sunny	2	2	2/9 = 0.222	2/5 = 0.4
Overcast	4	0	4/9 = 0.444	0/5 = 0
Rainy	3	2	3/9 = 0.33	2/5 = 0.4
Total	9	5	9/9 = 1	5/5 = 1

	outlook	temperature	humidity	windy	play_golf
0	Sunny	Hot	High	False	NO
1	Sunny	Hot	High	True	NO
2	Overcast	Hot	High	False	YES
3	Rainy	Mild	High	False	YES
4	Rainy	Cool	Normal	False	YES
5	Rainy	Cool	Normal	True	NO
6	Overcast	Cool	Normal	True	YES
7	Sunny	Mild	High	False	NO
8	Sunny	Cool	Normal	False	YES
9	Rainy	Mild	Normal	False	YES
10	Sunny	Mild	Normal	True	YES
11	Overcast	Mild	High	True	YES
12	Overcast	Hot	Normal	False	YES
13	Rainy	Mild	High	True	NO

Naïve Bayes Classifier Example

➢ **Temperature** Frequency and Likelihood

Temperature				
	YES	NO	P(YES)	P(NO)
Hot	2	2	2/9 = 0.22	2/5 = 0.4
Mild	4	2	4/9 = 0.44	2/5 = 0.4
Cool	3	1	3/9 = 0.33	1/5 = 0.2
Total	9	5	9/9 = 1	5/5 = 1

	outlook	temperature	humidity	windy	play_golf
0	Sunny	Hot	High	False	NO
1	Sunny	Hot	High	True	NO
2	Overcast	Hot	High	False	YES
3	Rainy	Mild	High	False	YES
4	Rainy	Cool	Normal	False	YES
5	Rainy	Cool	Normal	True	NO
6	Overcast	Cool	Normal	True	YES
7	Sunny	Mild	High	False	NO
8	Sunny	Cool	Normal	False	YES
9	Rainy	Mild	Normal	False	YES
10	Sunny	Mild	Normal	True	YES
11	Overcast	Mild	High	True	YES
12	Overcast	Hot	Normal	False	YES
13	Rainy	Mild	High	True	NO

Naïve Bayes Classifier Example

➢ **Humidity** Frequency and Likelihood

Humidity				
	YES	NO	P(YES)	P(NO)
High	3	4	3/9 = 0.33	4/5 = 0.8
Normal	6	1	6/9 = 0.66	1/5 = 0.2
Total	9	5	9/9 = 1	5/5 = 1

	outlook	temperature	humidity	windy	play_golf
0	Sunny	Hot	High	False	NO
1	Sunny	Hot	High	True	NO
2	Overcast	Hot	High	False	YES
3	Rainy	Mild	High	False	YES
4	Rainy	Cool	Normal	False	YES
5	Rainy	Cool	Normal	True	NO
6	Overcast	Cool	Normal	True	YES
7	Sunny	Mild	High	False	NO
8	Sunny	Cool	Normal	False	YES
9	Rainy	Mild	Normal	False	YES
10	Sunny	Mild	Normal	True	YES
11	Overcast	Mild	High	True	YES
12	Overcast	Hot	Normal	False	YES
13	Rainy	Mild	High	True	NO

HPSI
Human Performance Systems, Inc.

Naïve Bayes Classifier Example

➢ **Windy** Frequency and Likelihood

Windy				
	YES	NO	P(YES)	P(NO)
FALSE	6	2	6/9 = 0.66	2/5 = 0.4
TRUE	3	3	3/9 = 0.33	3/5 = 0.6
Total	9	5	9/9 = 1	5/5 = 1

	outlook	temperature	humidity	windy	play_golf
0	Sunny	Hot	High	False	NO
1	Sunny	Hot	High	True	NO
2	Overcast	Hot	High	False	YES
3	Rainy	Mild	High	False	YES
4	Rainy	Cool	Normal	False	YES
5	Rainy	Cool	Normal	True	NO
6	Overcast	Cool	Normal	True	YES
7	Sunny	Mild	High	False	NO
8	Sunny	Cool	Normal	False	YES
9	Rainy	Mild	Normal	False	YES
10	Sunny	Mild	Normal	True	YES
11	Overcast	Mild	High	True	YES
12	Overcast	Hot	Normal	False	YES
13	Rainy	Mild	High	True	NO

HPSI
Human Performance Systems, Inc.

Bayes Rule applied to data

- Bayes rule is merely the mathematical relation between the prior allocations of credibility and the posterior reallocation of credibility (conditional on data)

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

Bayes Rule Explained

Posterior Probability of class (C) given predictor (X)

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

This represents the probability of C being true, provided X is true

Bayes Rule Explained

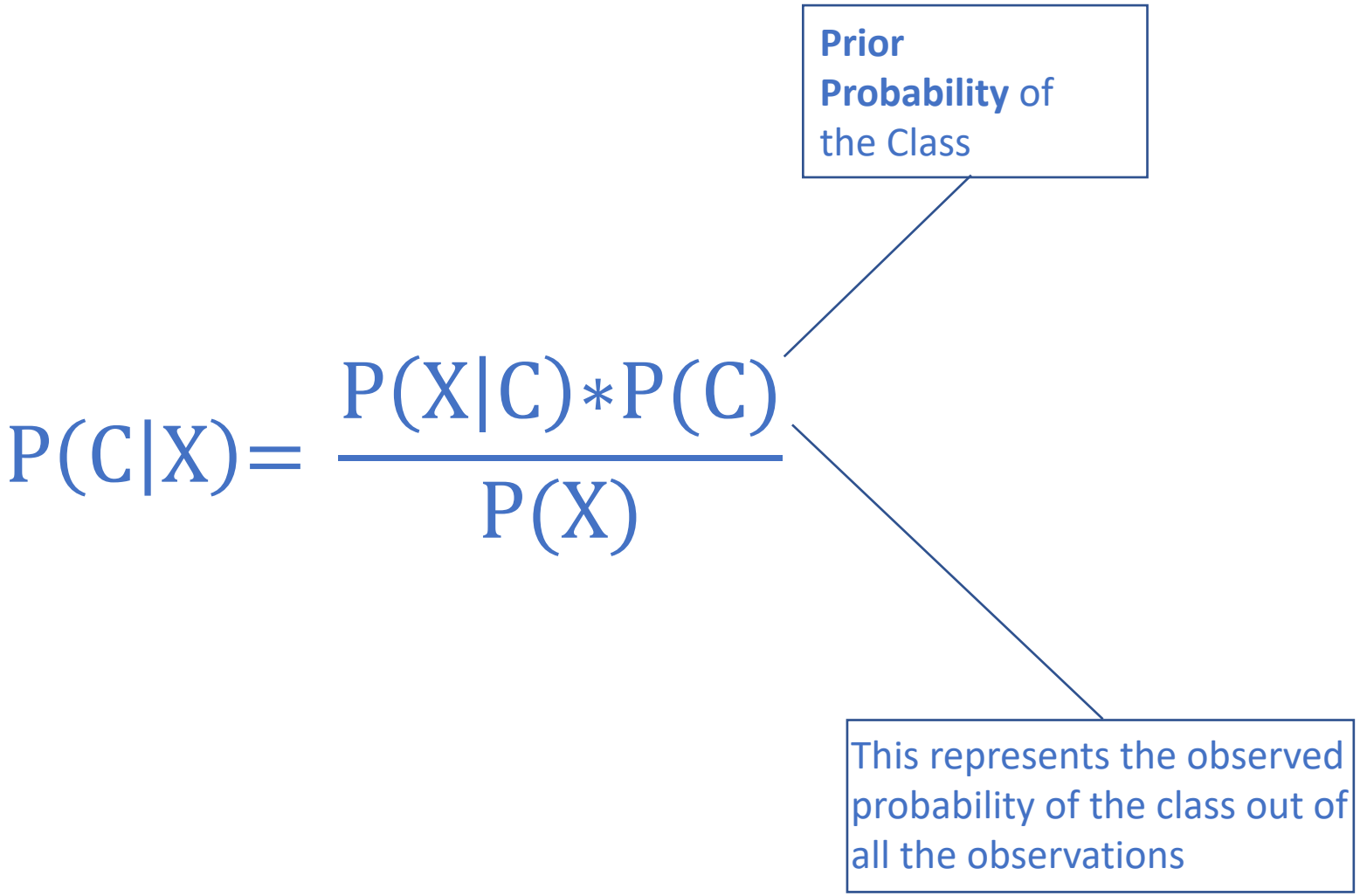
Likelihood - the conditional probability of the predictor-given the class

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

This represents the probability of X being true provided C is true

HPSI
Human Performance Systems, Inc.

Bayes Rule Explained



HPSI
Human Performance Systems, Inc.

Bayes Rule Explained

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

Prior Probability of the Predictor

This represents the observed probability of the predictor out of all the observations

How does the Naïve Bayes Classifier work?

- Step 1:
  - Convert the data set into frequency tables
  - Use the frequency tables to calculate likelihood tables
- Step 2:
  - Use the product rule to obtain a joint conditional probability for the attributes
- Step 3:
  - Use Bayes Rule to calculate the posterior probability for each class variable
  - Once this has been done for all classes, output the class with the highest probability

Bayes Rule Explained

Prior Probability of the Predictor can be estimated directly by multiplying the individual relative frequencies of each predictor due to the naive independence assumption

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

Bayes Rule Explained

Prior Probability of the Predictor can be estimated directly by multiplying the individual relative frequencies of each predictor (due to the naive independence assumption)

Key Idea: the "naive assumption" allows us to multiply the probabilities

$$P(C|X) = \frac{[P(X_1|C) * P(X_2|C) * P(X_3|C) \ldots * P(X_n|C)] * P(C)}{P(X)}$$

HPSI
Human Performance Systems, Inc.

Naïve Bayes Classifier Example

Lets consider the following case, and try to predict whether we should golf or not

Outlook	Temperature	Humidity	Windy?	Play?
X_1= Sunny	X_2 = Cool	X_3 = High	X_4 = TRUE	YES

$$P(YES|X) = \frac{[P(Sunny|YES) * P(Cool|YES) * P(High|YES) * P(TRUE|YES)] * P(YES)}{P(X)}$$

Naïve Bayes Classifier Example

Outlook	Temperature	Humidity	Windy?	Play?
Sunny	Cool	High	TRUE	YES

- Outlook = Sunny
  - P(Sunny|YES)=0.22
  - P(Sunny|NO)=0.6

Outlook				
	YES	NO	P(YES)	P(NO)
Sunny	2	3	**2/9 = 0.222**	**3/5 = 0.6**
Overcast	4	0	4/9 = 0.444	0/5 = 0
Rainy	3	2	3/9 = 0.33	2/5 = 0.4
Total	9	5	9/9 = 1	5/5 = 1

Naïve Bayes Classifier Example

Outlook	Temperature	Humidity	Windy?	Play?
Sunny	Cool	High	TRUE	YES

- Temperature = Cool
  - P(Cool|YES) = 0.33
  - P(Cool|NO) = 0.2

Temperature				
	YES	NO	P(YES)	P(NO)
Hot	2	2	2/9 = 0.22	2/5 = 0.4
Mild	4	2	4/9 = 0.44	2/5 = 0.4
Cool	3	1	**3/9 = 0.33**	**1/5 = 0.2**
Total	9	5	9/9 = 1	5/5 = 1

Naïve Bayes Classifier Example

Outlook	Temperature	Humidity	Windy?	Play?
Sunny	Cool	High	TRUE	YES

- Humidity = High
  - P(High|YES) = 0.33
  - P(High|NO) = 0.8

Humidity				
	YES	NO	P(YES)	P(NO)
High	3	4	**3/9 = 0.33**	**4/5 = 0.8**
Normal	6	1	6/9 = 0.66	1/5 = 0.2
Total	9	5	9/9 = 1	5/5 = 1

Naïve Bayes Classifier Example

Outlook	Temperature	Humidity	Windy?	Play?
Sunny	Cool	High	TRUE	YES

- Windy = True
  - P(TRUE|YES) = 0.33
  - P(TRUE|NO) = 0.6

Windy				
	YES	NO	P(YES)	P(NO)
FALSE	6	2	6/9 = 0.66	2/5 = 0.4
TRUE	3	3	**3/9 = 0.33**	**3/5 = 0.6**
Total	9	5	9/9 = 1	5/5 = 1

Naïve Bayes Classifier Example

Outlook	Temperature	Humidity	Windy?	Play?
Sunny	Cool	High	TRUE	YES

- Play = YES
  - P(C) = P(YES) = 0.64
  - P(C) = P(NO) = 0.36

Play		P(YES)
YES	9	**9/14 = 0.64**
NO	5	**5/14 = 0.36**
Total	14	14/14 = 100%

How does the Naïve Bayes Classifier work?

- Step 1:
  - Convert the data set into frequency tables
  - Use the frequency tables to calculate likelihood tables
- Step 2:
  - Use the product rule to obtain a joint conditional probability for the attributes
- Step 3:
  - Use Bayes Rule to calculate the posterior probability for each class variable
  - Once this has been done for all classes, output the class with the highest probability

Naïve Bayes Classifier Example

Lets consider the following case, and try to predict whether we can golf or not

	Outlook	Temperature	Humidity	Windy?
X	X_1= Sunny	X_2 = Cool	X_3 = High	X_4 = TRUE
P(X|YES)	0.22	0.33	0.33	0.33
P(X|NO)	0.6	0.2	0.8	0.6

	Play?
C	YES
P(YES)	0.64
P(NO)	0.36

$$P(YES|X) = \frac{[P(Sunny|YES) * P(Cool|YES) * P(High|YES) * P(TRUE|YES)] * P(YES)}{P(X)}$$

Naïve Bayes Classifier Example

Lets consider the following case, and try to predict whether we can golf or not

	Outlook	Temperature	Humidity	Windy?
X	X_1= Sunny	X_2 = Cool	X_3 = High	X_4 = TRUE
P(X|YES)	0.22	0.33	0.33	0.33
P(X|NO)	0.6	0.2	0.8	0.6

	Play?
C	YES
P(YES)	0.64
P(NO)	0.36

$$P(YES|X) = \frac{[0.22 * 0.33 * 0.33 * 0.33] * 0.64}{P(X)}$$

HPSI
Human Performance Systems, Inc.

Naïve Bayes Classifier Example

Lets consider the following case, and try to predict whether we can golf or not

	Outlook	Temperature	Humidity	Windy?
X	X_1= Sunny	X_2 = Cool	X_3 = High	X_4 = TRUE
P(X|YES)	0.22	0.33	0.33	0.33
P(X|NO)	0.6	0.2	0.8	0.6

	Play?
C	YES
P(YES)	0.64
P(NO)	0.36

$$P(YES|X) = \frac{0.0053}{P(X)}$$

Naïve Bayes Classifier Example

Lets consider the following case, and try to predict whether we can golf or not

	Outlook	Temperature	Humidity	Windy?
X	X_1= Sunny	X_2 = Cool	X_3 = High	X_4 = TRUE
P(X|YES)	0.22	0.33	0.33	0.33
P(X|NO)	0.6	0.2	0.8	0.6

	Play?
C	YES
P(YES)	0.64
P(NO)	0.36

$$P(NO|X) = \frac{[P(Sunny|NO) * P(Cool|NO) * P(High|NO) * P(TRUE|NO)] * P(NO)}{P(X)}$$

HPSI
Human Performance Systems, Inc.

Naïve Bayes Classifier Example

Lets consider the following case, and try to predict whether we can golf or not

	Outlook	Temperature	Humidity	Windy?
X	X_1= Sunny	X_2 = Cool	X_3 = High	X_4 = TRUE
P(X|YES)	0.22	0.33	0.33	0.33
P(X|NO)	0.6	0.2	0.8	0.6

	Play?
C	YES
P(YES)	0.64
P(NO)	0.36

$$P(NO|X) = \frac{[0.6 * 0.2 * 0.8 * 0.6] * 0.36}{P(X)}$$

Naïve Bayes Classifier Example

Lets consider the following case, and try to predict whether we can golf or not

	Outlook	Temperature	Humidity	Windy?
X	X_1= Sunny	X_2 = Cool	X_3 = High	X_4 = TRUE
P(X|YES)	0.22	0.33	0.33	0.33
P(X|NO)	0.6	0.2	0.8	0.6

	Play?
C	YES
P(YES)	0.64
P(NO)	0.36

$$P(NO|X) = \frac{0.0206}{P(X)}$$

Naïve Bayes Classifier Example

Lets consider the following case, and try to predict whether we can golf or not

	Outlook	Temperature	Humidity	Windy?
X	X_1= Sunny	X_2 = Cool	X_3 = High	X_4 = TRUE
P(X|YES)	0.22	0.33	0.33	0.33
P(X|NO)	0.6	0.2	0.8	0.6

	Play?
C	YES
P(YES)	0.64
P(NO)	0.36

Likelihood of playing golf under these conditions:

$$P(YES|X) = \frac{0.0053}{P(X)}$$

Likelihood of *NOT* playing golf under these conditions:

$$P(NO|X) = \frac{0.0206}{P(X)}$$

Naïve Bayes Classifier Example

Lets consider the following case, and try to predict whether we can golf or not

	Outlook	Temperature	Humidity	Windy?
X	X_1= Sunny	X_2 = Cool	X_3 = High	X_4 = TRUE
P(X|YES)	0.22	0.33	0.33	0.33
P(X|NO)	0.6	0.2	0.8	0.6

	Play?
C	YES
P(YES)	0.64
P(NO)	0.36

Likelihood of playing golf under these conditions:

$$P(YES|X) = \frac{0.0053}{P(X)}$$

Probability of Play = YES:

$$P(YES|X) = \frac{0.0053}{(0.0053 + 0.0206)} = 20.5\%$$

Likelihood of *NOT* playing golf under these conditions:

$$P(NO|X) = \frac{0.0206}{P(X)}$$

Probability of Play = NO:

$$P(NO|X) = \frac{0.0206}{(0.0053 + 0.0206)} = 79.5\%$$

Naïve Bayes Classifier Example

Lets consider the following case, and try to predict whether we can golf or not

	Outlook	Temperature	Humidity	Windy?
X	X_1= Sunny	X_2 = Cool	X_3 = High	X_4 = TRUE
P(X|YES)	0.22	0.33	0.33	0.33
P(X|NO)	0.6	0.2	0.8	0.6

	Play?
C	YES
P(YES)	0.64
P(NO)	0.36

Likelihood of playing golf under these conditions:

$$P(YES|X) = \frac{0.0053}{P(X)}$$

Probability of Play = YES:

$$P(YES|X) = \frac{0.0053}{(0.0053 + 0.0206)} = 20.5\%$$

Likelihood of *NOT* playing golf under these conditions:

$$P(NO|X) = \frac{0.0206}{P(X)}$$

Probability of Play = NO:

$$P(NO|X) = \frac{0.0206}{(0.0053 + 0.0206)} = 79.5\%$$

HPSI
Human Performance Systems, Inc.

Naïve Bayes Classifier Results

The Naïve Bayes Classifier predicts that we should **not** play golf under the conditions in question

Outlook	Temperature	Humidity	Windy?	Play?
Sunny	Cool	High	TRUE	YES

Naïve Bayes Exercise

Logistic Regression

Logistic Regression – a few points

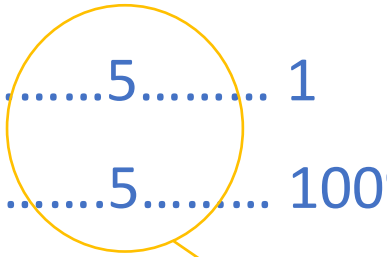
- Logistic Regression is also very popular for establishing a baseline
- Here are a few of the traits that make it popular:
  - Like Naïve Bayes, Logistic Regression is very fast
  - It's highly interpretable
  - Doesn't require features to be scaled
  - Doesn't require any tuning
  - Easy to regularize
  - Outputs predicted probabilities

Logistic Regression – a few points

- Although Logistic Regression is popular, it is often misunderstood.
- Here are a few common questions that people have when they think of logistic regression:
  - Why is it called "logistic regression" if it's used for **classification**?
  - Why is it called a li**near model**?
  - How do you interpret the **model coefficients**?

Probability Reminder

- Probability
  - Measure of **Likelihood**, meaning how **likely** is something going to happen, how likely is something true
  - Expressed as a number between 0-1
  - 0 ……5…… 1
  - 0……5…… 100%



In the case of a **binary outcome** we have some midpoint where it's equally probable/likely

ODDS

- Probability
  - Another measure of probability is **ODDS**
  - $Probability = \frac{one\ outcome}{\boldsymbol{ALL\ possible}\ outcome}$
  - $Odds = \frac{one\ outcome}{\boldsymbol{ALL\ OTHER}\ outcomes}$

Back to the coinflip

- Flip a coin
  - $P(Heads) = \frac{one\ outcome}{\textbf{ALL possible}\ outcome} = \frac{1}{1+1} = \frac{H}{H+T} = 0.5$
  - $O(Heads) = \frac{one\ outcome}{\textbf{ALL OTHER}\ outcome} = \frac{1}{1} = \frac{H}{T} = 1$
  - Ratio Representation
    - Odds can be represented as (1:1)
    - We say the odds are **even** because these numbers are the same

Rolling a Die

- Roll a Die
  - We are interested in rolling a 3
  - $P(roll = 3) = \frac{one\ outcome}{\boldsymbol{ALL\ possible}\ outcome} = \frac{\{3\}}{\{1,2,3,4,5,6\}} = \frac{1}{6} = 0.1667$
  - $O(roll = 3) = \frac{one\ outcome}{\boldsymbol{ALL\ OTHER}\ outcome} = \frac{\{3\}}{\{1,2,4,5,6\}} = \frac{1}{5} = 0.2$
  - Ratio Representation
    - Odds can be represented as (0.2:1)
    - We say the odds are **less than even**

Rolling a Die

- Roll a Die
  - Now say we are interested in rolling <=5
  - $P(roll \leq 5) = \frac{one\ outcome}{\boldsymbol{ALL\ possible}\ outcome} = \frac{\{5,4,3,2,1\}}{\{6,5,4,3,2,1\}} = \frac{5}{6} = 0.83$
  - $O(roll \leq 5) = \frac{one\ outcome}{\boldsymbol{ALL\ OTHER}\ outcome} = \frac{\{5,4,3,2,1\}}{\{6\}} = \frac{5}{1} = 5$
  - Ratio Representation
    - Odds can be represented as (5:1)
    - We say the odds are **better than even** (in your favor)

Ranges for Odds and Probability

$$0 \leq Odds \leq \infty$$

$$0 \leq Probability \leq 1$$

Odds Ratio Example

- Smokers, non-smokers study
- Interested in studying smoking/lung cancer, 2 groups
- Group 1 (Case Group) – Have lung cancer
  - 8 Smokers
  - 2 Non-smokers
- Group 2 (Control Group) – Don't have lung cancer
  - 4 Smokers
  - 6 Non-smokers
- How can we compare the odds?
- Conclusion:
  - The Case group was 6 times more likely to be a smoker
  - Making no claims about causation
  - Simply that people who have lung cancer are 6 times more likely to have smoked

	Case	Control
Smokers	8	4
Non-Smokers	2	6

$$Odds\ of\ Smoking, Case\ Group = \frac{8}{2} = 4$$

$$Odds\ of\ Smoking, Control\ Group = \frac{4}{6} = 0.66$$

$$Odds\ Ratio = \frac{Odds\ Case\ Group}{Odds\ Control\ Group} = \frac{4}{0.66} = 6$$

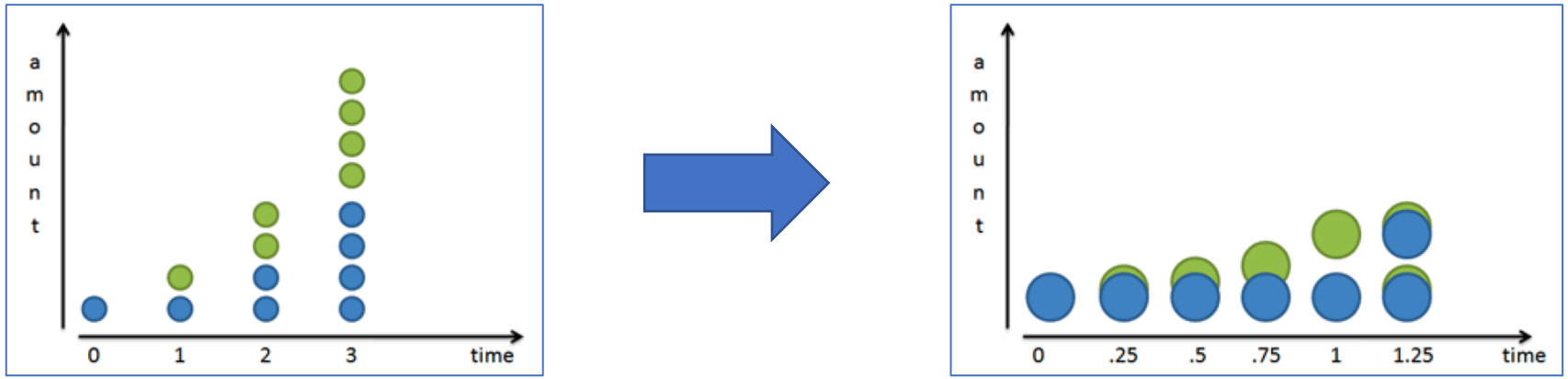
- From Wikipedia:
  - *"The natural logarithm, formerly known as the hyperbolic logarithm, is the logarithm to the base e, where e is an irrational constant approximately equal to 2.718281828459."*
  - Describing *e* as "a constant approximately 2.71828…" is like calling pi "an irrational number, approximately equal to 3.1415…". Sure, it's true, but you completely missed the point.

e

- A Better way to think about π/ e
  - π is the ratio between circumference and diameter shared by all circles. It is a fundamental ratio inherent in all circles and therefore impacts any calculation of circumference, area, volume, and surface area for circles, spheres, cylinders, and so on. (not to mention the trigonometric functions derived from circles (sin, cos, tan).
  - e is the base rate of growth shared by all continually growing processes. e lets you take a simple growth rate (where all change happens at the end of the year) and find the impact of compound, continuous growth
  - e shows up whenever systems grow exponentially and continuously: population, radioactive decay, interest calculations, and more.
  - Just like every number can be considered a scaled version of 1 (the base unit), every circle can be considered a scaled version of the unit circle (radius 1), and every rate of growth can be considered a scaled version of e (unit growth, perfectly compounded).

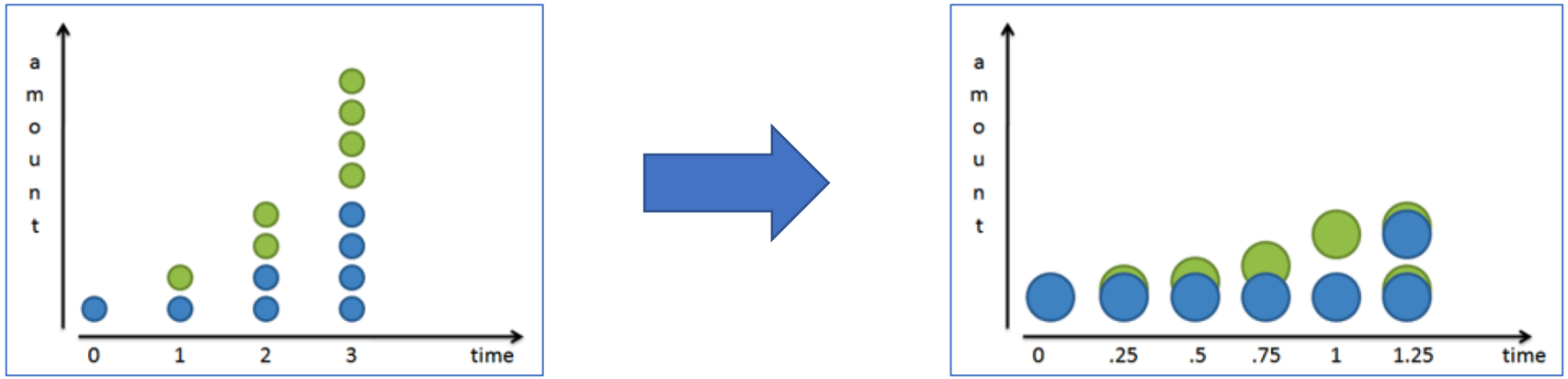
Exponential Growth

- *Understanding Exponential Growth*
  - Let's look at a basic system that doubles after an amount of time. For example,
  - Bacteria can split and "doubles" every 24 hours
  - We get twice as many noodles when we fold them in half.
  - Your money doubles every year if you get 100% return
- It looks like this:



The secret to *e*

- *The big secret: e merges rate and time*
  - This is wild, e^x can mean two things:
  - x is the number of times we multiply a growth rate: 100% growth for 3 years is e^3
  - x is the growth rate itself: 300% growth for one year is e^3



The secret to *e*

- *The big secret: e merges rate and time*
  - This is wild, e^x can mean two things:
  - x is the number of times we multiply a growth rate: 100% growth for 3 years is e^3
  - x is the growth rate itself: 300% growth for one year is e^3
- And now you know why it's "e", and not pi or some other number: e raised to "r*t" gives you the growth impact of rate r and time t.

$$growth = e^x = e^{rt}$$

Demystifying the Natural Logarithm (ln)

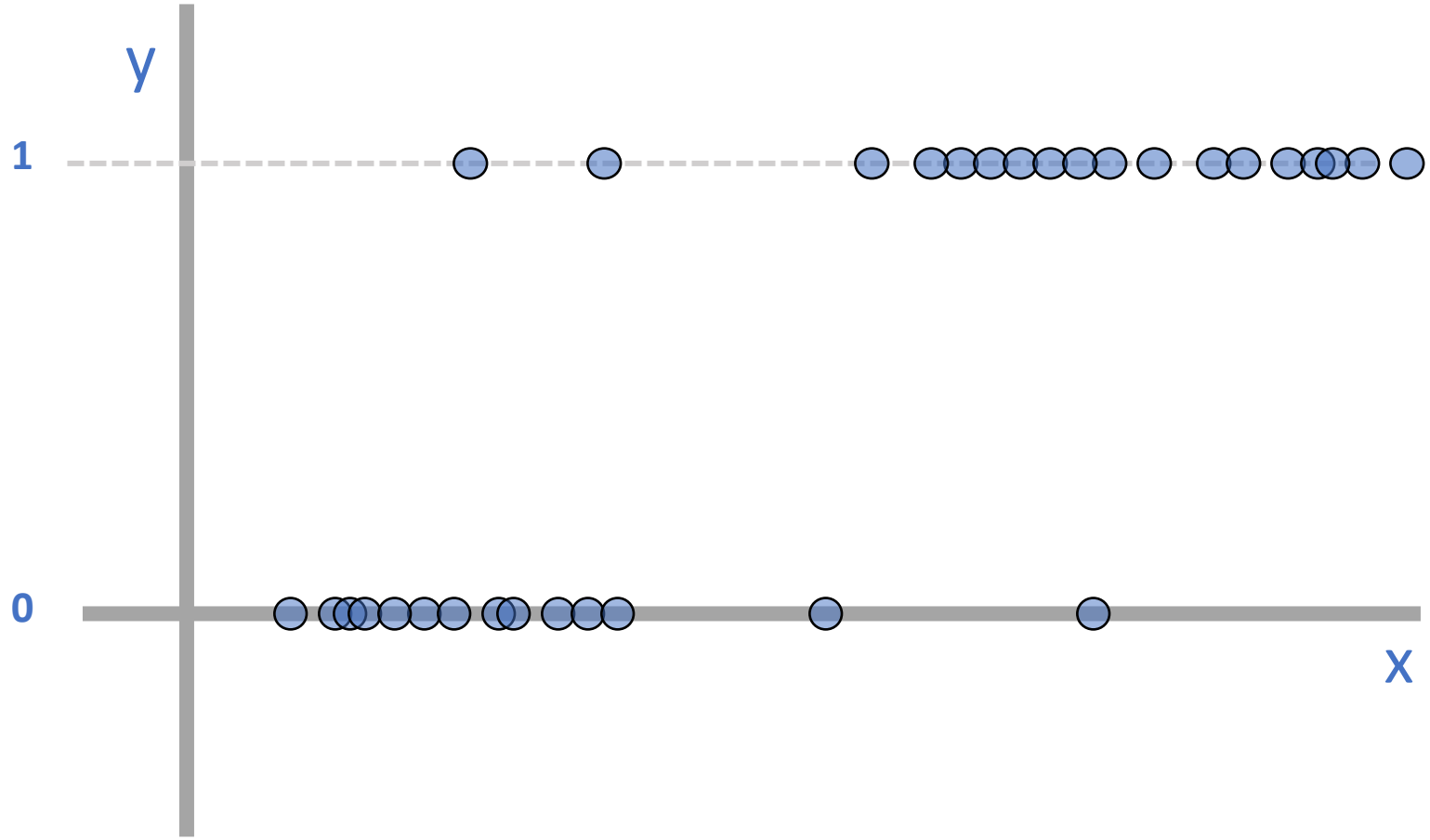
- After understanding the exponential function, the next component required for logistic regression is the natural log
- If you look-up the Wikipedia definition, it's "the inverse of e^x"
- e and the Natural Log are twins:
  - e^x is the amount of continuous growth after a certain amount of time.
  - Natural Log (ln) is the amount of time needed to reach a certain level of continuous growth

Logistic Regression

- When we think of linear regression, we usually think of predicting a numeric quantity
- Logistic Regression is somewhat misleading in that it serves primarily as a binary classification model
- Categorical response (y) with 2 levels (binary: 0 and 1)
  - Passing or failing a test
  - Surviving a plane crash or not
  - Hospitalisation required or not
  - Diagnosis of diabetes (yes / no)
  - Labelling (over/under some threshold)
- Predictor variables (x_i) can take on any form: binary, categorical, and/or continuous

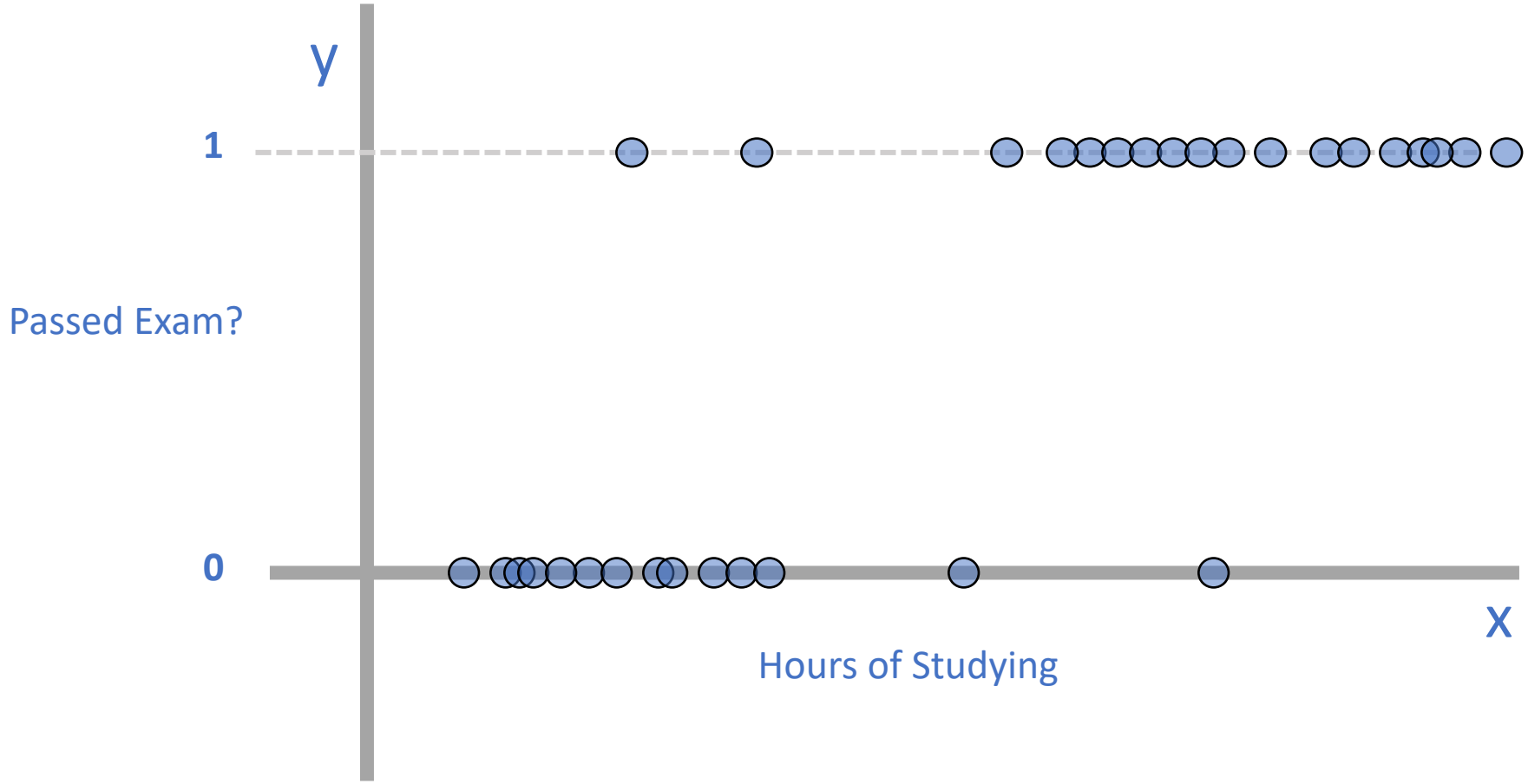
Logistic Regression Classification

Consider this set of binary data

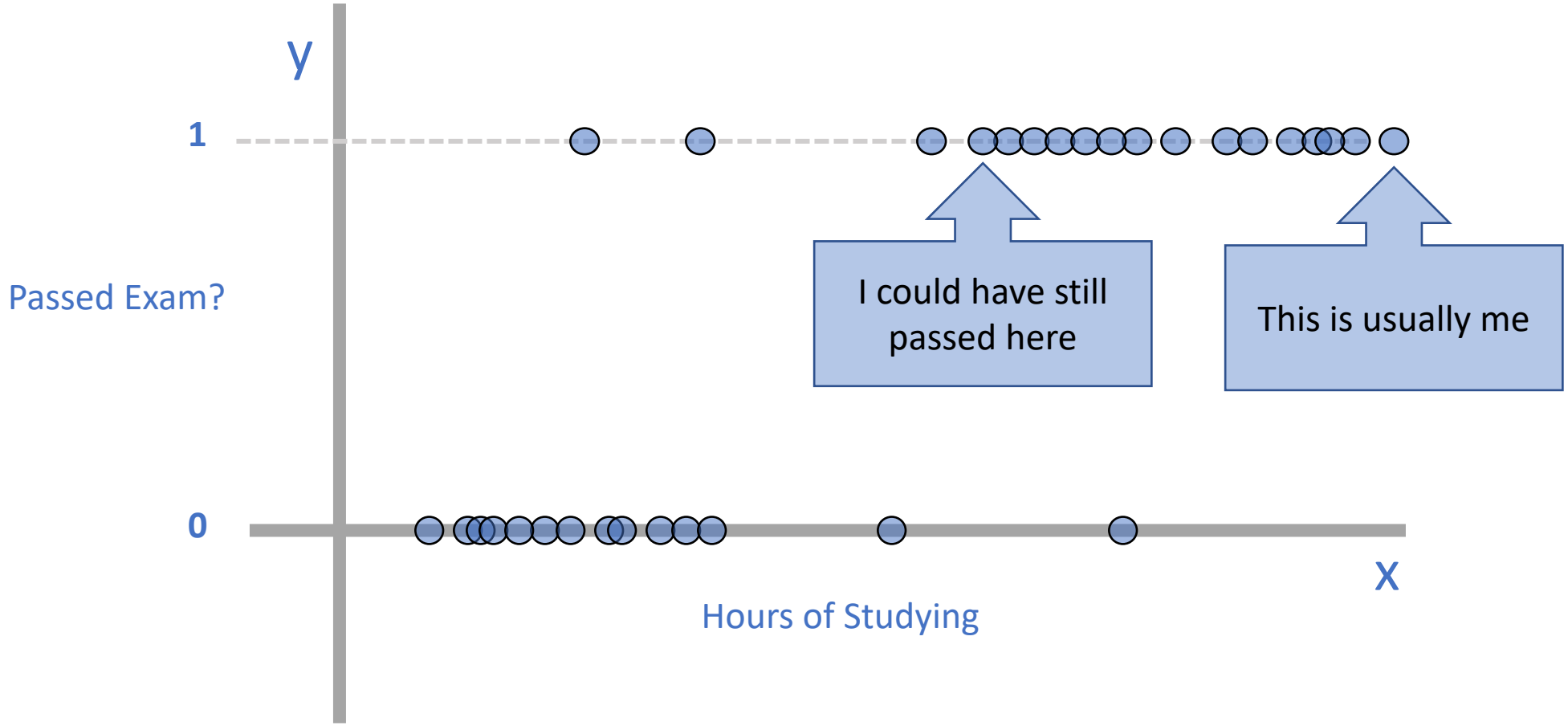


HPSI
Human Performance Systems, Inc.

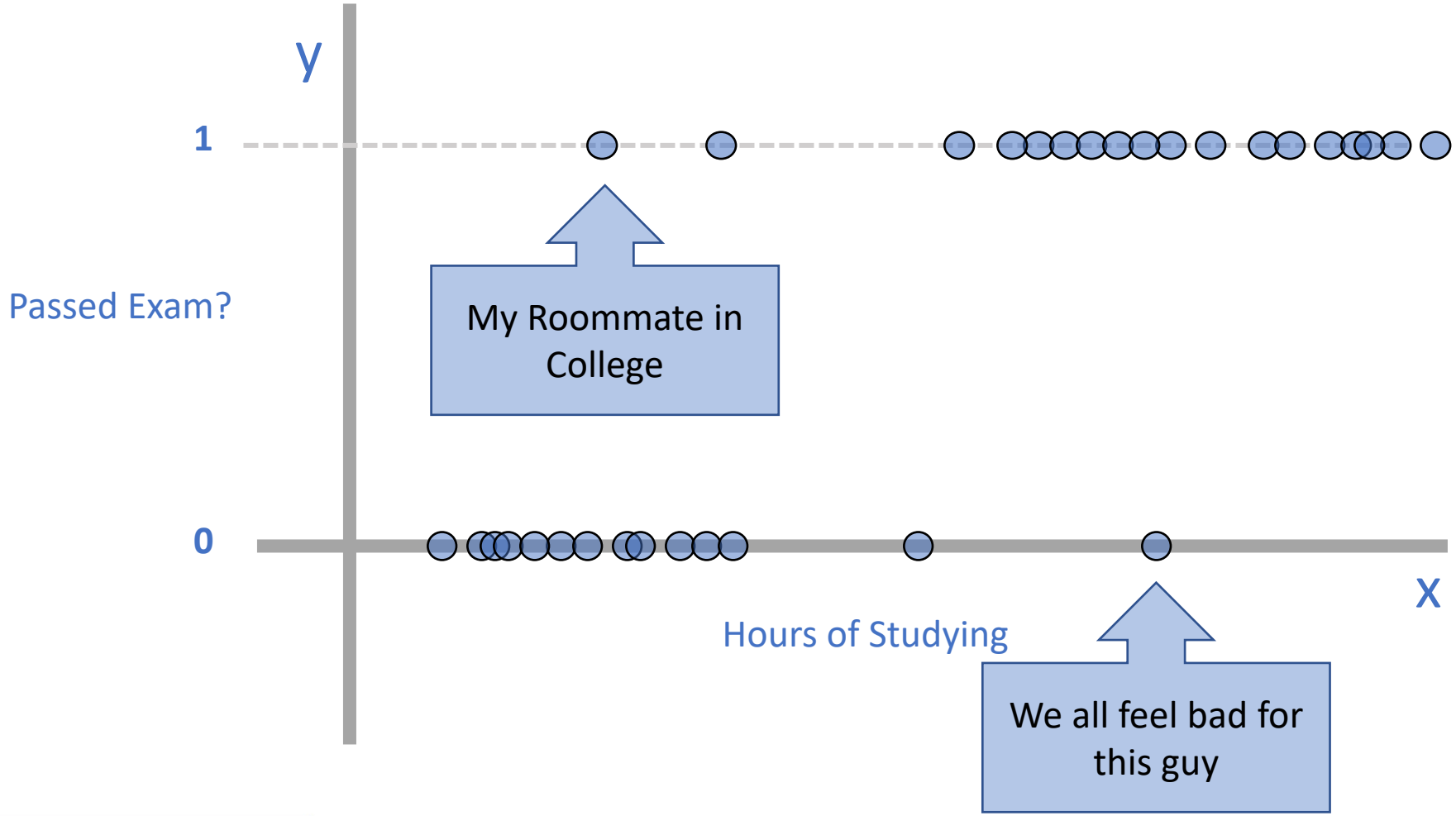
Consider this set of binary data



Consider this set of binary data

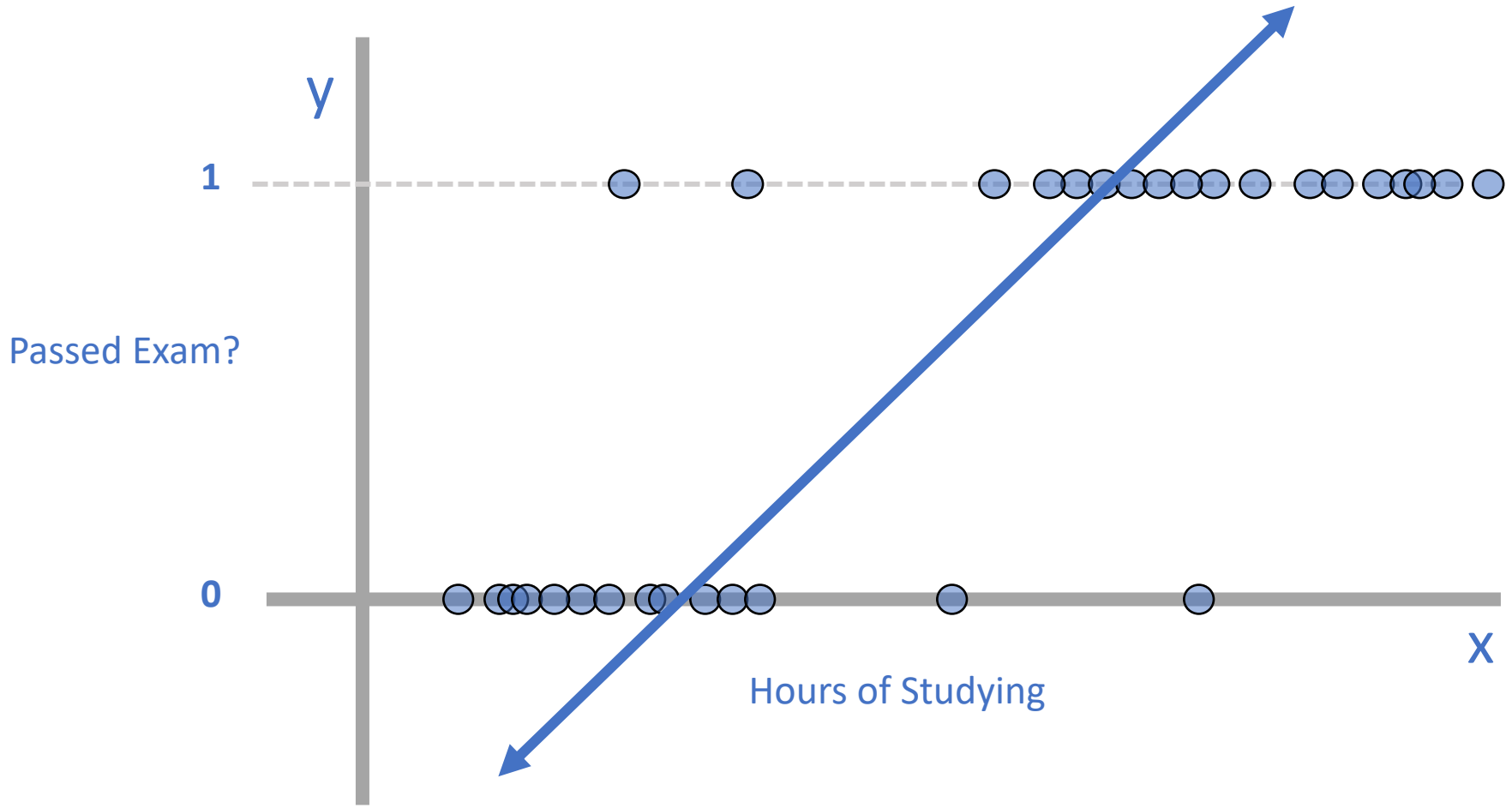


Consider this set of binary data



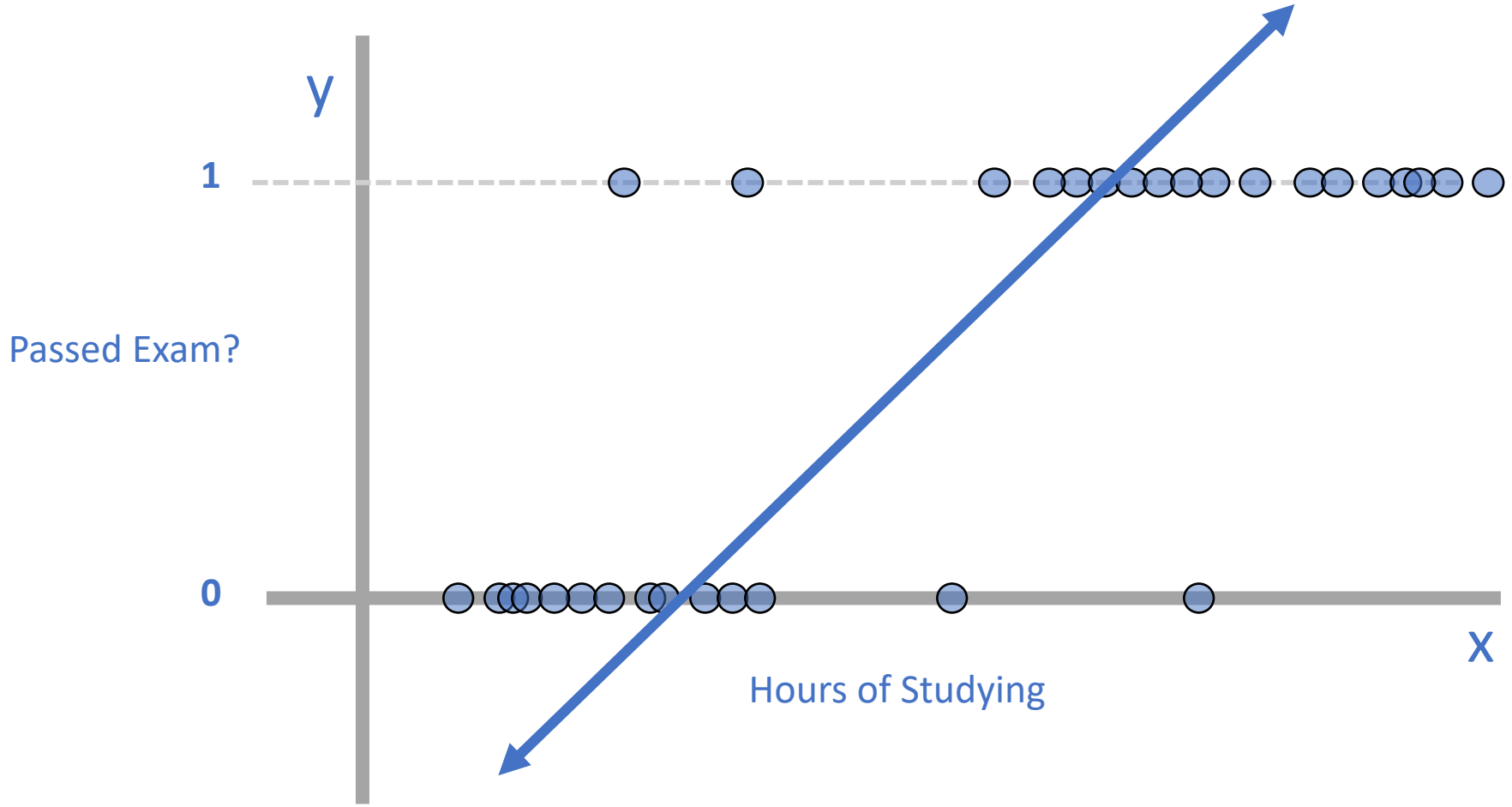
Logistic Regression Classification

- Linear Model? – Aside from being binary, there's nothing special about (y)



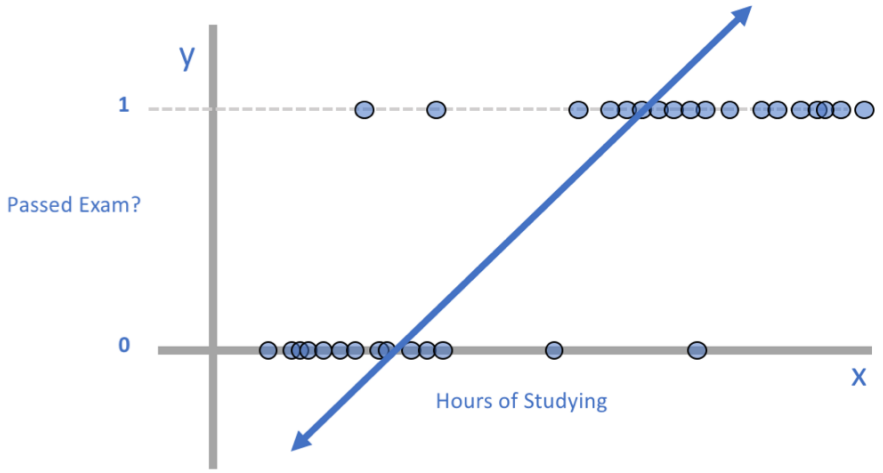
Logistic Regression Classification

- The value of "Passed Exam" is higher if a student studies more



Logistic Regression Classification

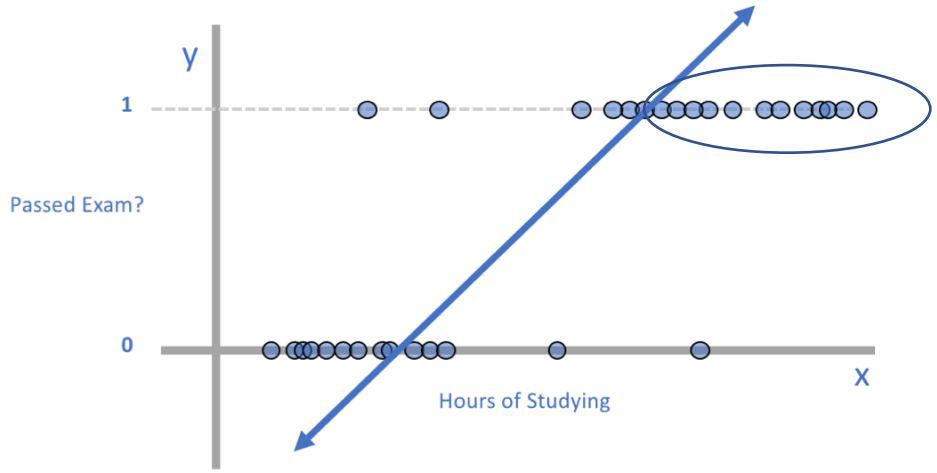
- Linear Model
  - $Pass = \beta_0 + \beta_1 hours\ of\ studying + \varepsilon$
- Problem
  - We want to see what makes the dependent variable change from a 0 to a 1
  - This can also be interpreted as what increases the likelihood of passing, or $P(pass = 1)$ which we can simply denote as p.
  - We should then be able to re-write the linear model as
  - $P(pass = 1) = p = \beta_0 + \beta_1 hours\ of\ studying + \varepsilon$
  - Every additional hour of studying increases the probability of passing by $x\%$



Logistic Regression Classification

$$P(pass = 1) = \beta_0 + \beta_1 hours\ of\ studying + \varepsilon$$

- Recall
  - Probabilities are bounded by $0 \leq p \leq 1$
  - If we were to look at students who studied this many hours, our model would predict a probability greater than 1

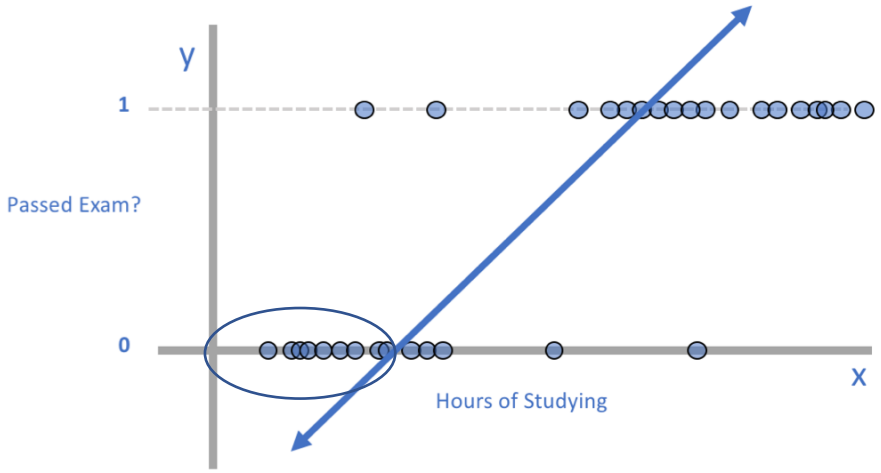


HPSI
Human Performance Systems, Inc.

Logistic Regression Classification

$$P(pass = 1) = \beta_0 + \beta_1 hours\ of\ studying + \varepsilon$$

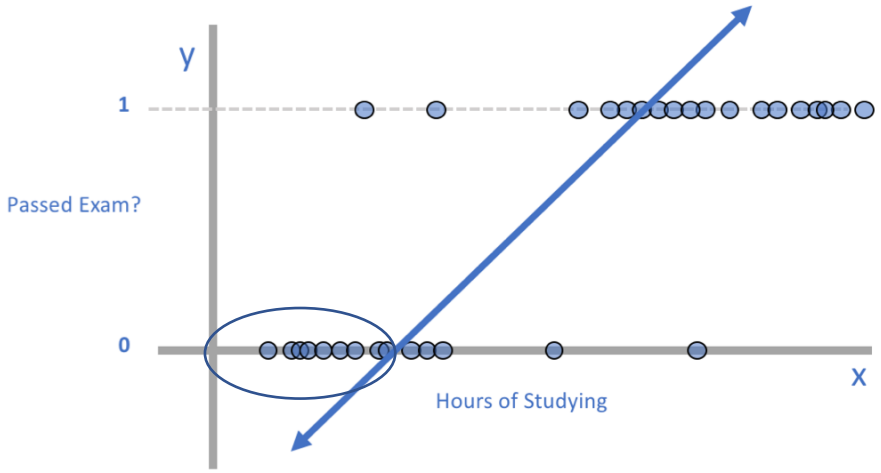
- Recall
  - Probabilities are bounded by $0 \leq p \leq 1$
  - If we were to look at students who studied this many hours, our model would predict a probability greater than 1
  - Likewise, here our model would predict a probability less than 0



Logistic Regression Classification

$$P(pass = 1) = \beta_0 + \beta_1 hours\ of\ studying + \varepsilon$$

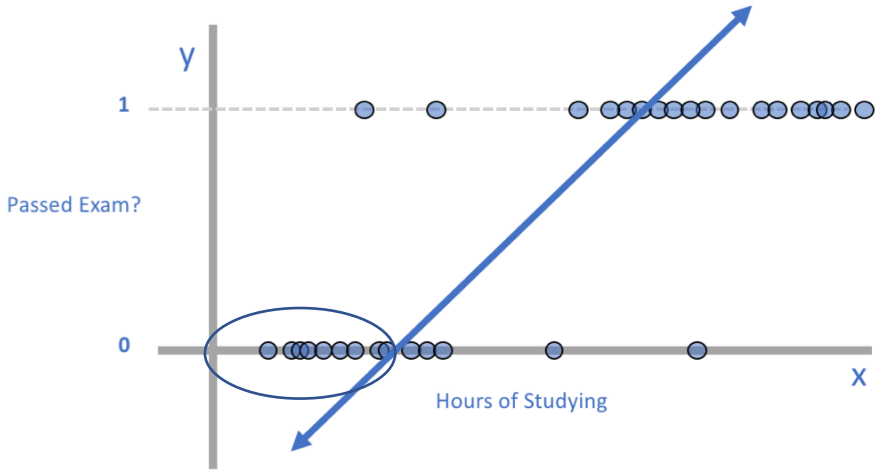
- Recall
  - Probabilities are bounded by $0 \leq p \leq 1$
  - If we were to look at students who studied this many hours, our model would predict a probability greater than 1
  - Likewise, here our model would predict a probability less than 0
- In addition to violating the laws of probability, our model would not maintain normally distributed residuals (which is another requirement of a linear regression model)



Logistic Regression Classification

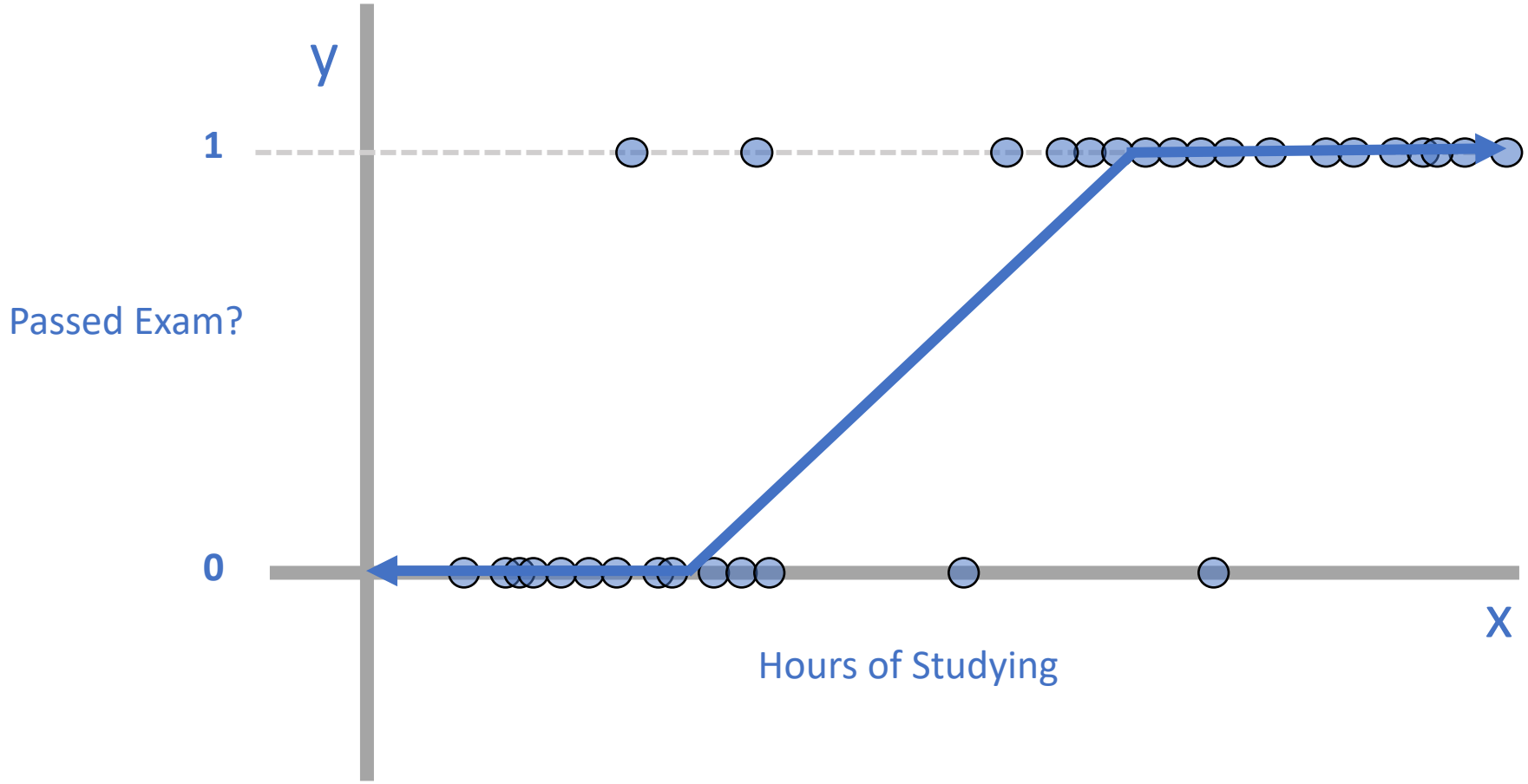
$$P(pass = 1) = \beta_0 + \beta_1 hours\ of\ studying + \ \varepsilon$$

- Recall
  - Probabilities are bounded by $0 \leq p \leq 1$
  - If we were to look at students who studied this many hours, our model would predict a probability greater than 1
  - Likewise, here our model would predict a probability less than 0
- In addition to violating the laws of probability, our model would not maintain normally distributed residuals (which is another requirement of a linear regression model)
- What can we do to fix this?



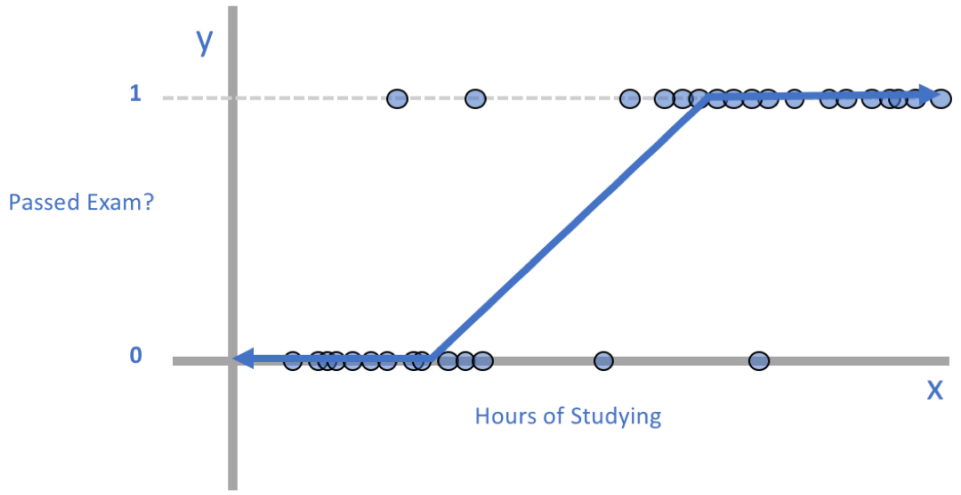
Logistic Regression Classification

- We could cap the probabilities at 0 and 1



Logistic Regression Classification

- Fixing the prior approach
  - We need to somehow constrain p such that $0 \leq p \leq 1$
  - We know $P(pass) = f(hours\ of\ studying)$ but the linear function didn't work
  - Let's try to develop a new function $f(hours\ of\ studying)$ that satisfies this criteria



Logistic Regression Classification

- **Two Constraints**
  1. **It must always be positive** since $0 \leq p(pass)$
     - $|x|$?
     - x^2 ?
     - What about $p(pass) = e^{\beta_0 + \beta_1 * hours\ of\ studying}$?
     - This works, but there are times when it would be greater than 1
  2. **It must always be less than 1** $(p(pass) \leq 1)$
     - If you think about proportions, any number that is divided by a number slightly greater than it will give us a number smaller than 1
     - What if we just add 1 to the denominator?
     - $p(pass) = \frac{(e^{\beta_0 + \beta_1 * hours\ of\ studying})}{(e^{\beta_0 + \beta_1 * hours\ of\ studying}) + 1}$
     - Note that we could have added any small number (ε) and this condition would have been met, but we use 1 for reasons that will become clear shortly

Logistic Regression Classification

- The previous expression:

$$p(pass) = p = \frac{(e^{\beta_0 + \beta_1 * hours\ of\ studying})}{(e^{\beta_0 + \beta_1 * hours\ of\ studying}) + 1}$$

- After applying some algebra, can be re-written as:

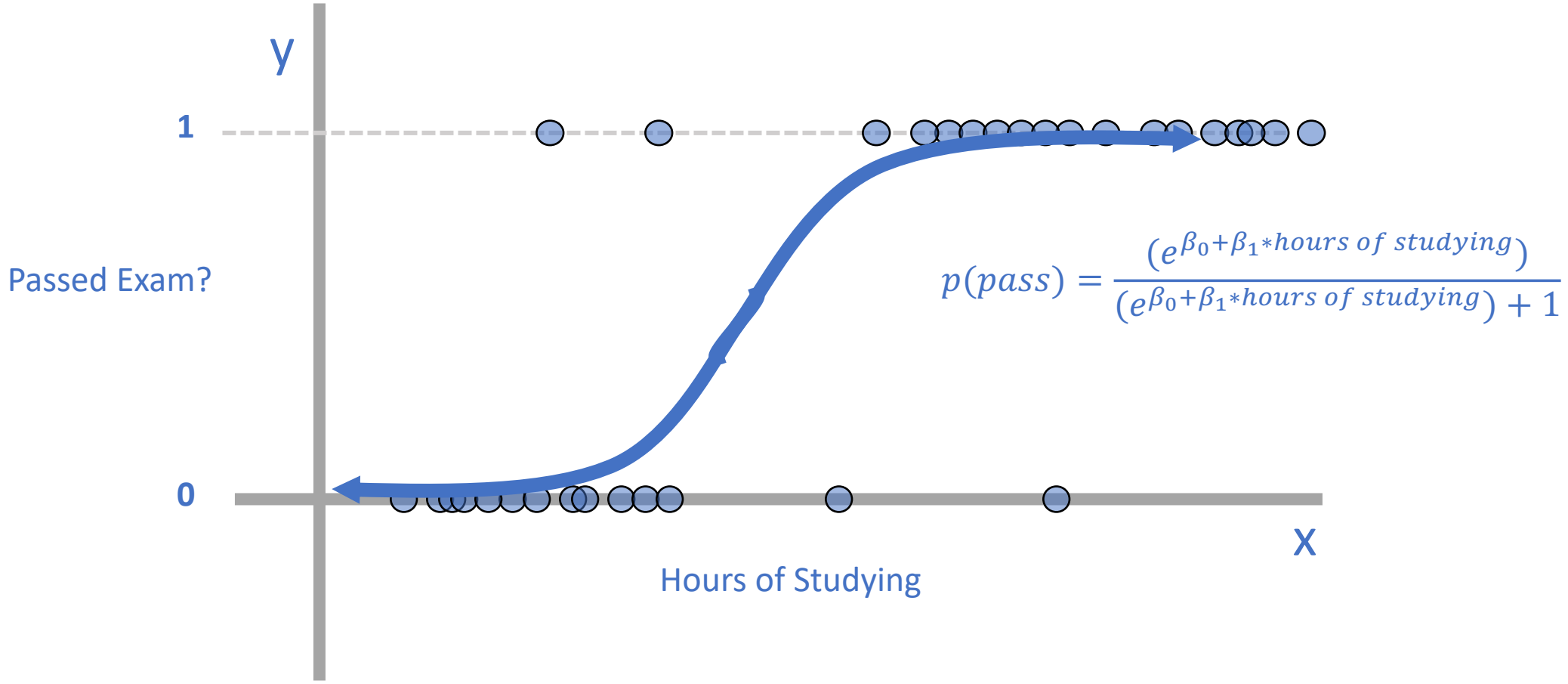
$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 * hours\ of\ studying$$

Logistic Regression Classification

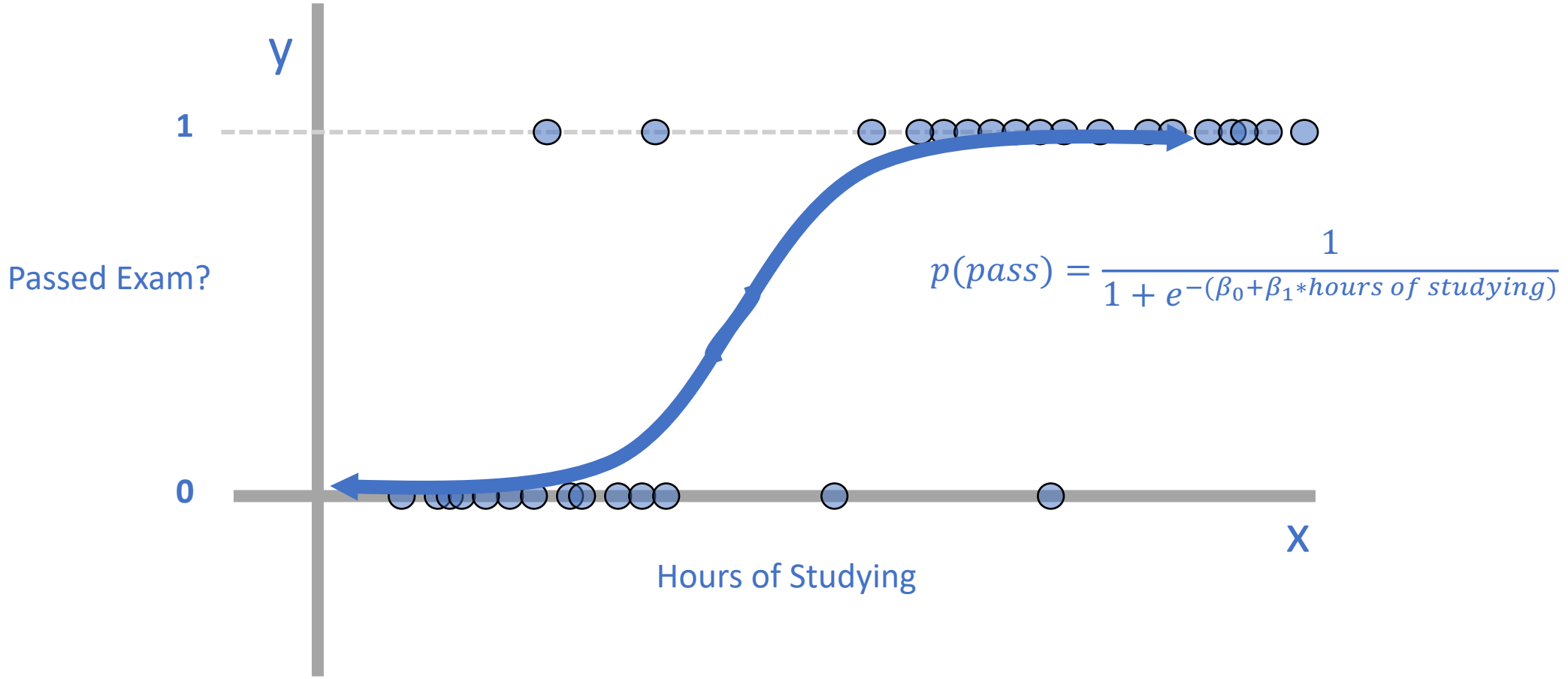
$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * hours\ of\ studying$$

- Does this look familiar?
  - Yes!
  - It's in the form of a standard linear model
- So, even though the probability of a student passing is not a linear function of study-hours, the simple transformation is a linear function of study-hours
- This is the equation used in logistic regression

Logistic Regression

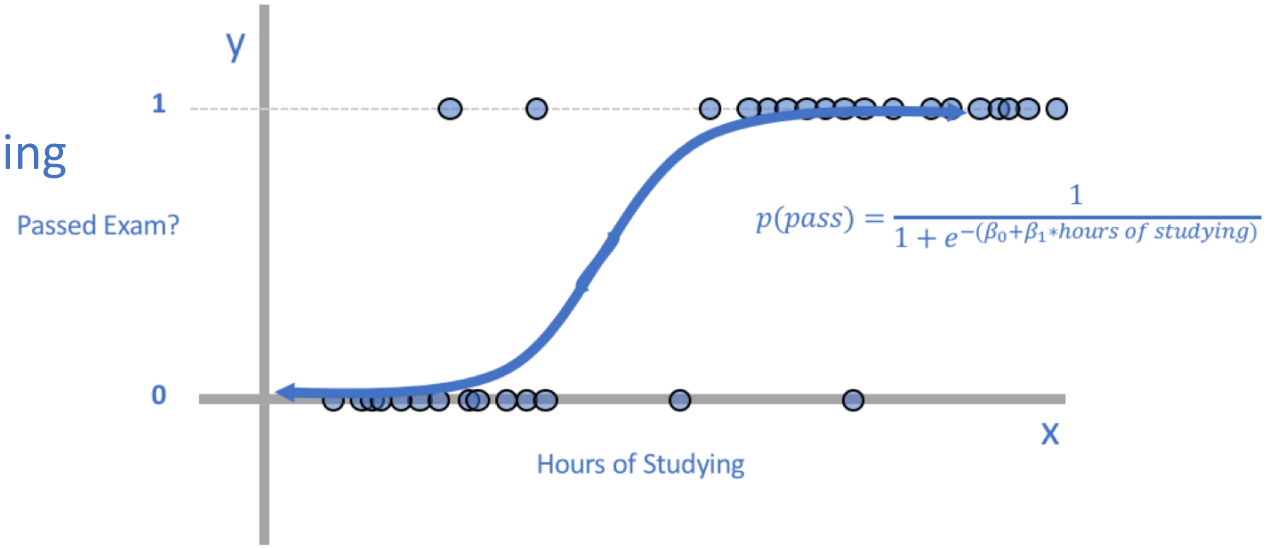


Logistic Regression

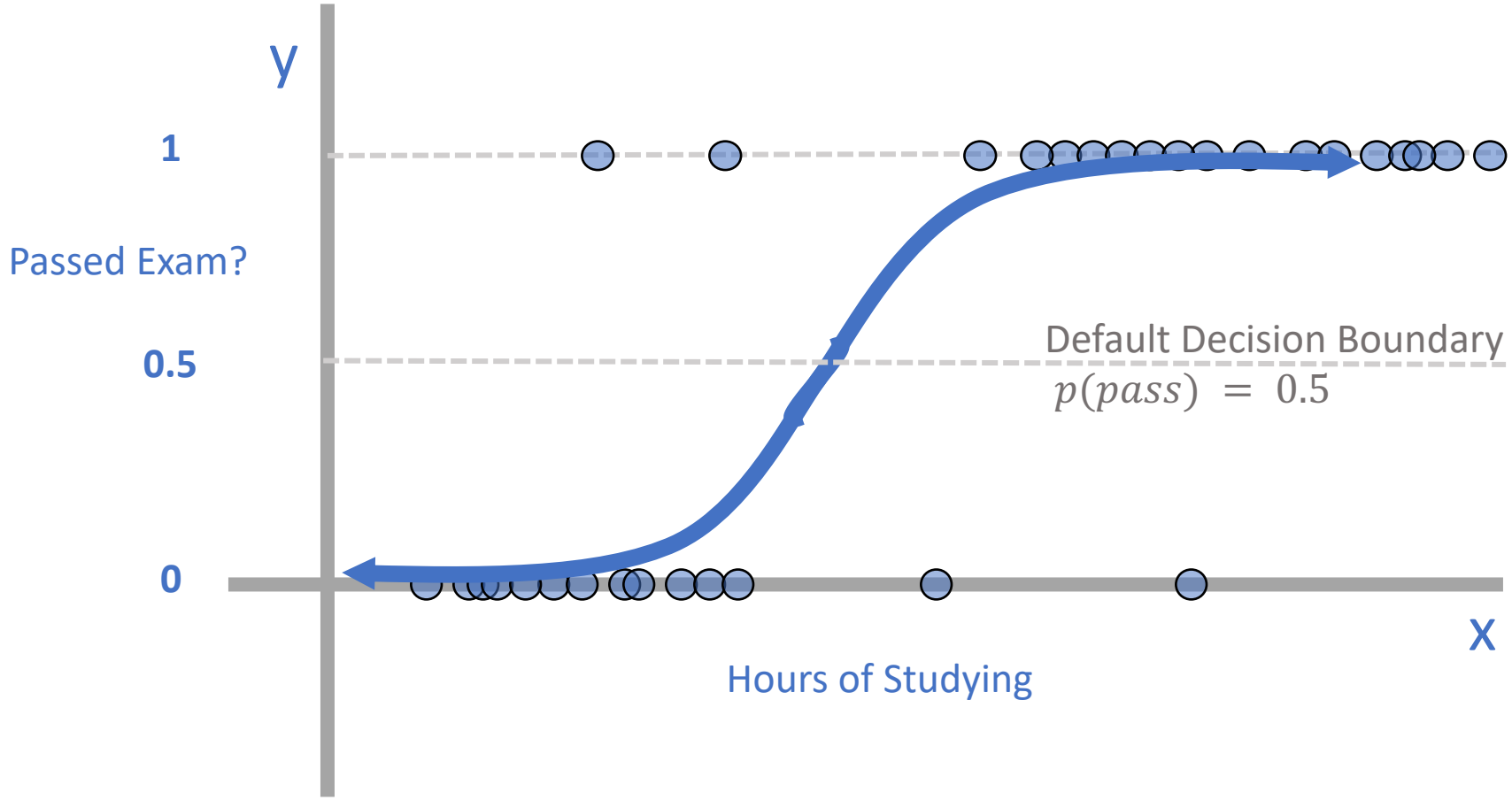


Logistic Regression Classification

- Note that the probability of passing is now between 0 and 1
  - $0 \leq p \leq 1$
- We now have a continuous function
- As study-hours approach 0, the probability of passing goes (asymptotically) to zero
- As study-hours approach infinity, probability of passing goes (asymptotically) to 1

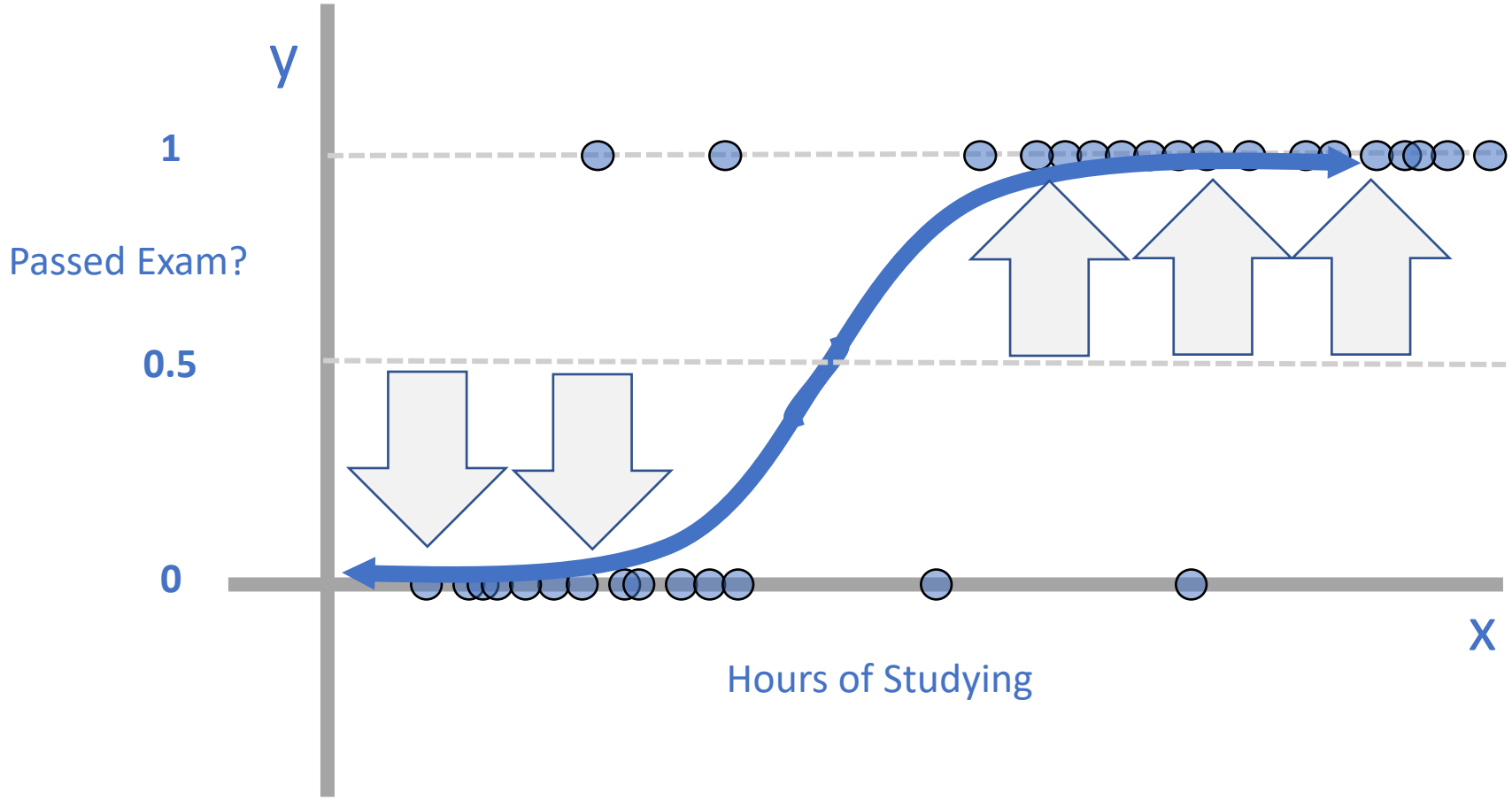


Decision Boundary



Logistic Regression Classification

Decision Boundary



Interpreting Binary Logistic Regression Model Output

- Interpreting the coefficients from a Logistic Regression model is different from a Linear Regression model
- Consider these results from a fitted logistic regression model
  - $\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 * study\_hrs$
  - $\beta_0 = -4.077$
  - $\beta_1 = 1.5046$
  - $\ln\left(\frac{p}{1-p}\right) = -4.077 + 1.5046 * study\_hrs$
  - For every additional unit increase in $study\_hrs$, $\ln\left(\frac{p}{1-p}\right)$ increases by 1.5046 units?
  - But what does that mean?

Interpreting Binary Logistic Regression Model Output

- If we have

$$y^* = logit = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * study\_hrs$$

- y^* is called the "logit" function
  - A logit is defined as the natural log of the odds
- Exponentiating both sides of equation results in:

$$\frac{p}{1-p} = e^{(\beta_0 + \beta_1 * study\_hrs)}$$

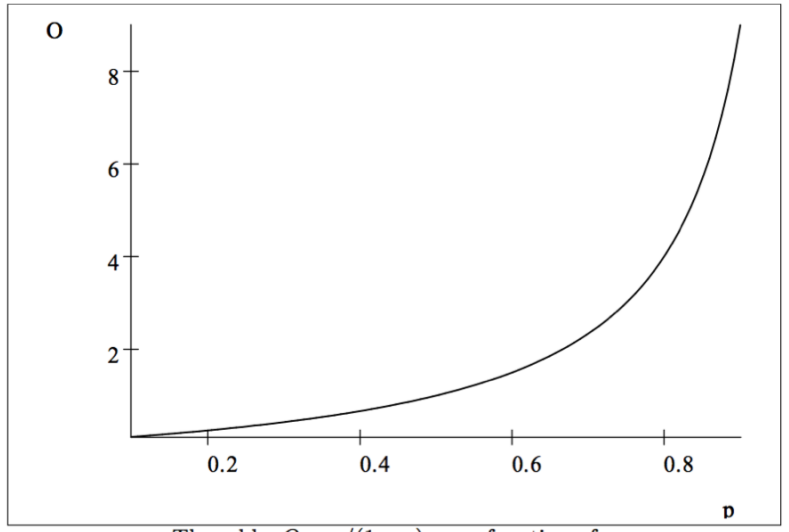
- We can exponentiate each of the coefficients to generate their corresponding odds ratios

Interpreting Binary Logistic Regression Model Output

- Odds

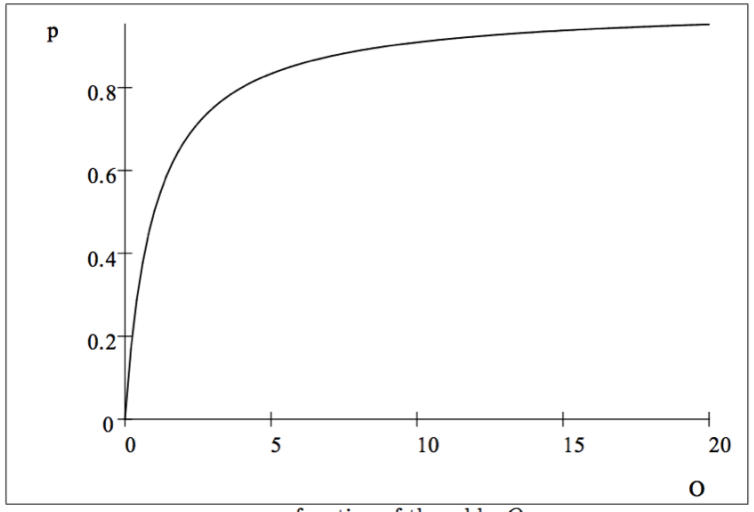
$$Odds(O) = \frac{P}{1 - P}$$

  - The odds ratio tells you how a unit increase or decrease in the corresponding input variable affects the odds of passing the test
    - When the odds ratio is greater than 1, it describes a positive relationship
    - An odds ratio less than 1 implies a negative relationship



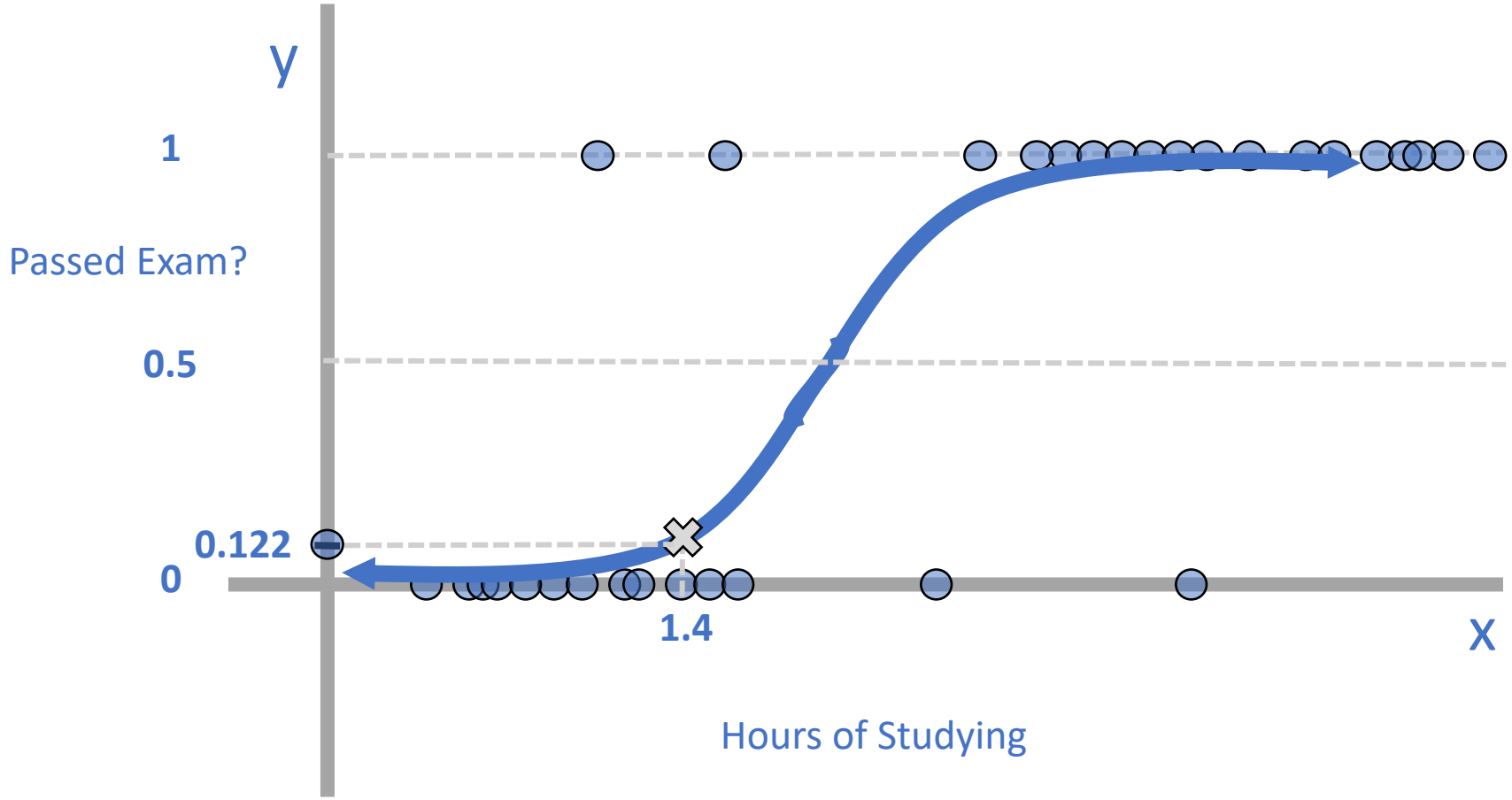
The odds, $O = p/(1-p)$, as a function of p

Note that $0 \leq O$, and O is undefined for $p = 1$. Solving $O = \frac{p}{1-p}$ for p, $p = \frac{O}{(O+1)}$



p as a function of the odds, O

Visualizing the Logistic Regression Model



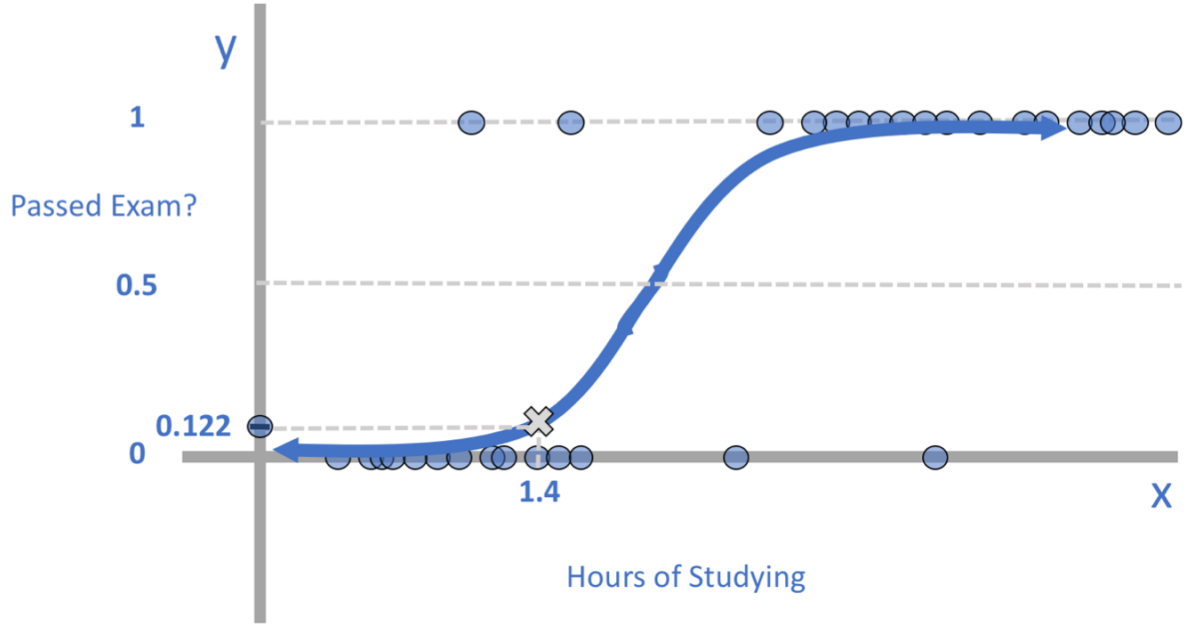
Logistic Regression Classification

- If a student only prepares 1.4 hours for the test
  - $study\_hours$=1.4
  - $\beta_0 = -4.077$
  - $\beta_1 = 1.5046$

$$p(pass) = \frac{\left(e^{\beta_0+\beta_1*studyhrs}\right)}{\left(e^{\beta_0+\beta_1*studyhrs}\right)+1}$$

$$p(pass) = \frac{\left(e^{-4.077+1.5046*1.4}\right)}{\left(e^{-4.077+1.5046*1.4}\right)+1} = 0.122$$

- Our model predicts a probability of passing of 12.2%
- This falls below the 0.5 threshold and results in a prediction of failing the exam



Logistic Regression Example

References

1. *Locally Weighted Learning* by Atkeson, Moore, Schaal
2. *Tuning Locally Weighted Learning* by Schaal, Atkeson, Moore
3. Kruschke, J. K. (2014). Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.
4. Machine Learning, Neural and Statistical Classification, Editors: D. Michie, D.J. Spiegelhalter, C.C. Taylor
5. http://www.dataschool.io/guide-to-logistic-regression/