1495 UNIVERSITY OF ABERDEEN | CELEBRATING 525 YEARS 1495 – 2020

ABERDEEN 2040

Revision – Week 2

Data Mining & Visualisation

Lecture 16

2025

Today…

- Exam-style questions that cover the past week's lectures
- We will walk through each one

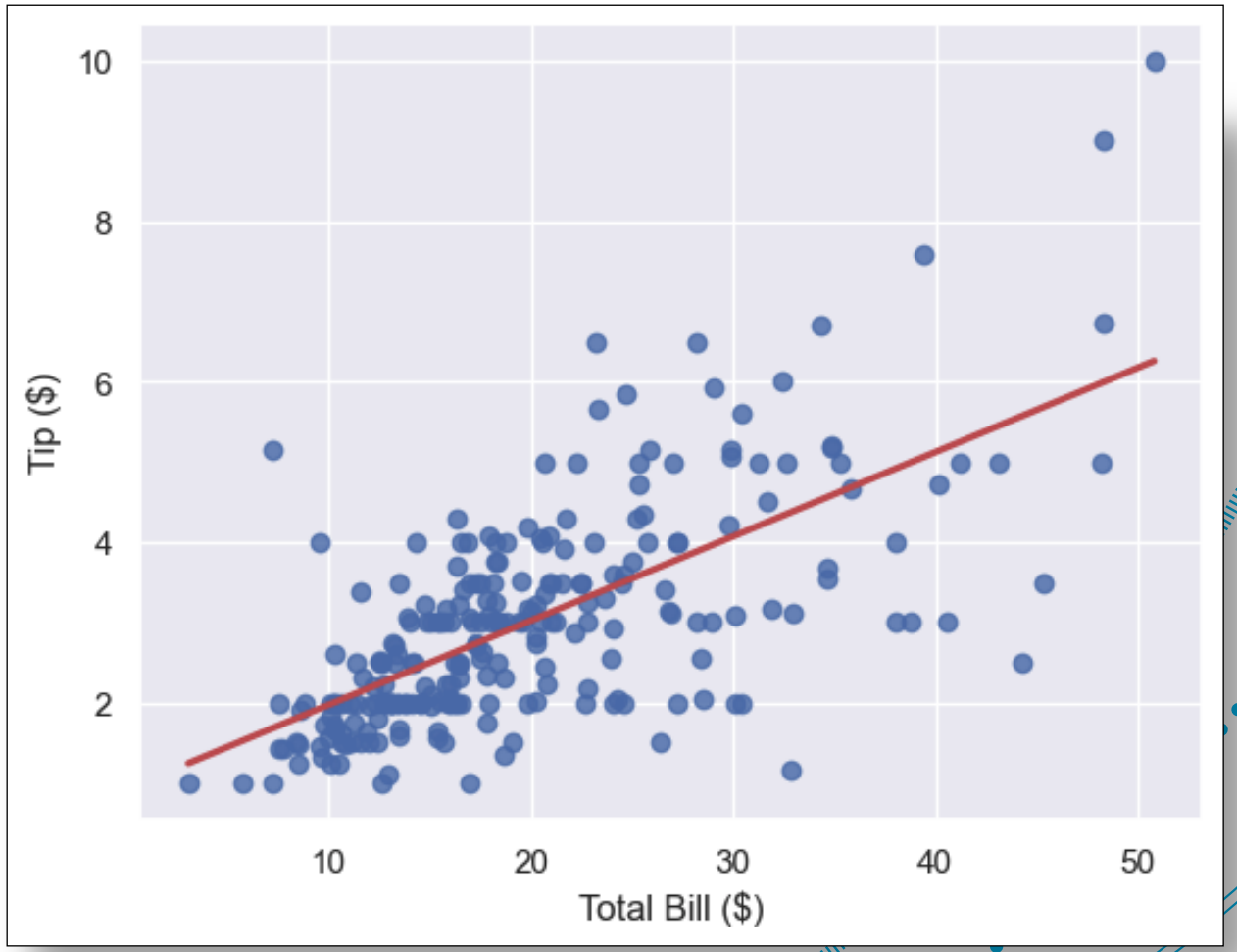
Regression

Regression Vs Classification Problems

- Note that when we discuss supervised learning, there are two broad types of problems that tend to come up:
- **Regression problems**, where our DV is quantitative
- **Classification problems**, where our DV is categorical
- **Note** that for both regression and classification problems, our IV(s) can be quantitative and/or categorical!

Regression Analysis

- Example: Do larger bills result in larger tips?
- A regression model estimates the function that most closely fits the data
- Note that today, we're just going to focus on **linear models** with **1 predictor**



Regression Model

- Recall the equation of a line:

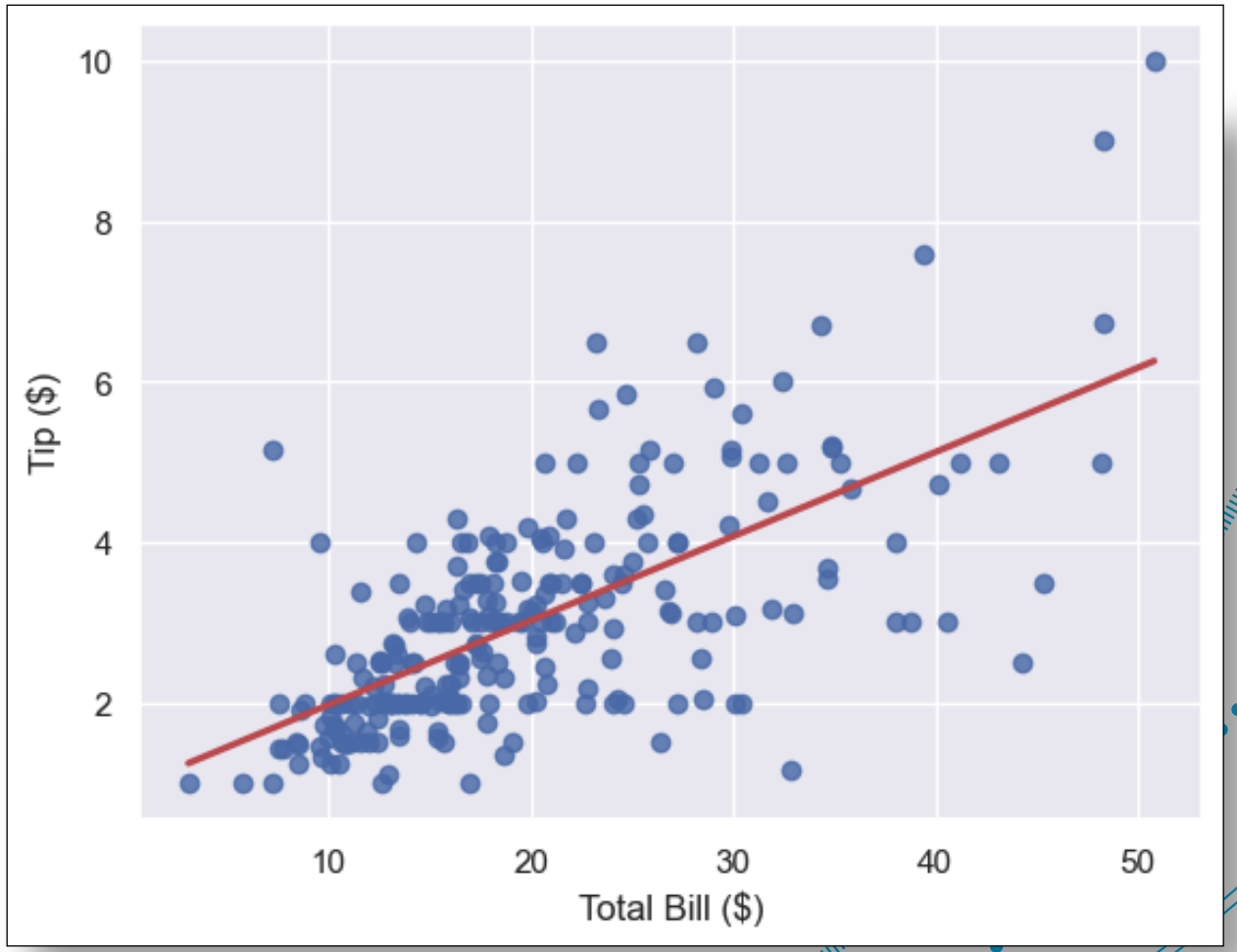
$$y = mx + c$$

Regression Model

Example: Do larger bills result in larger tips?

Intercept: $0.92

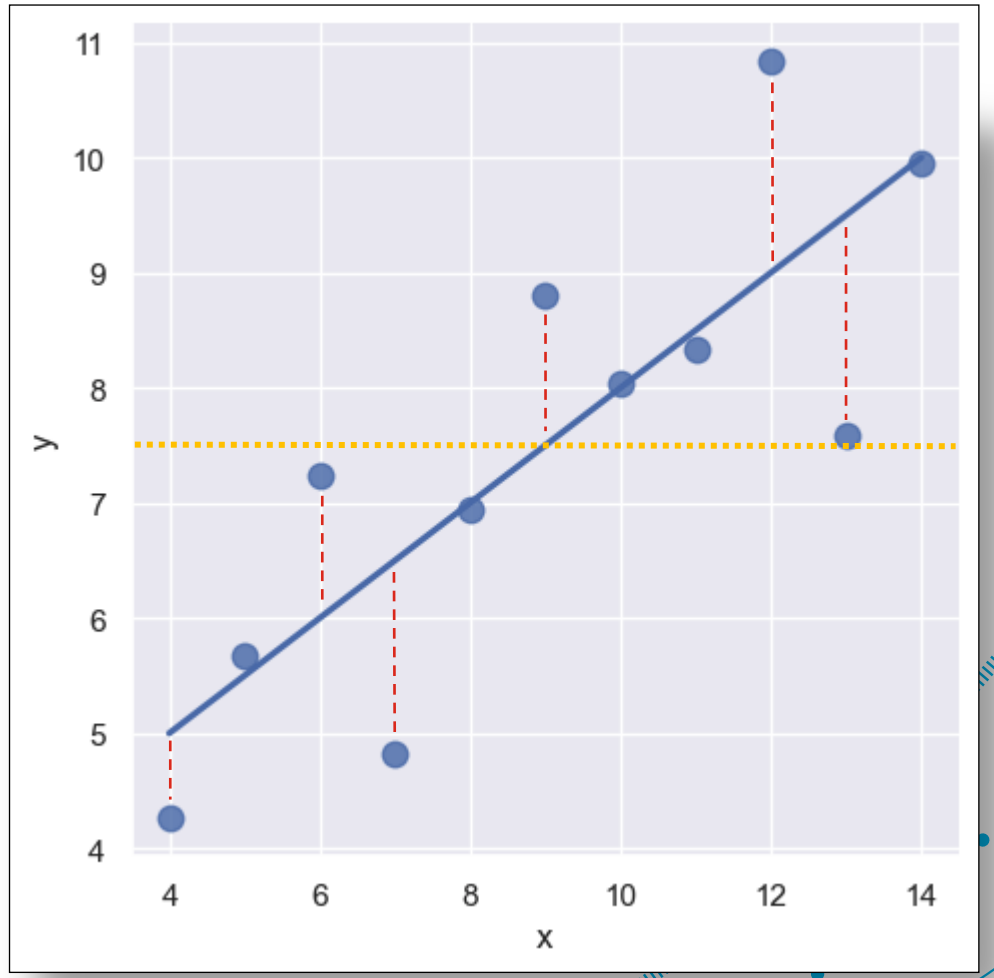
Gradient: 0.105



R^2 Value

- Recall the SSE calculation

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



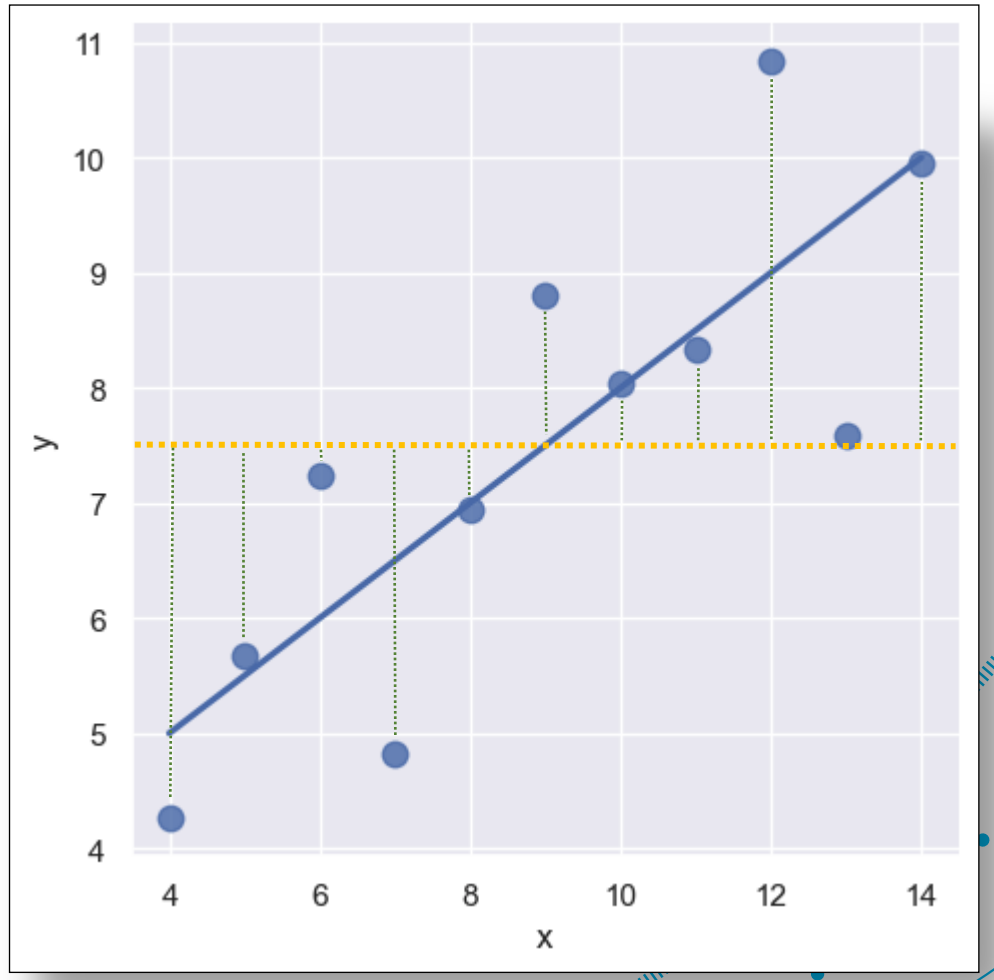
R^2 Value

- Recall the SSE calculation

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Let's say we calculate the Sum of Squares Total (SST), using the mean instead of ŷ

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



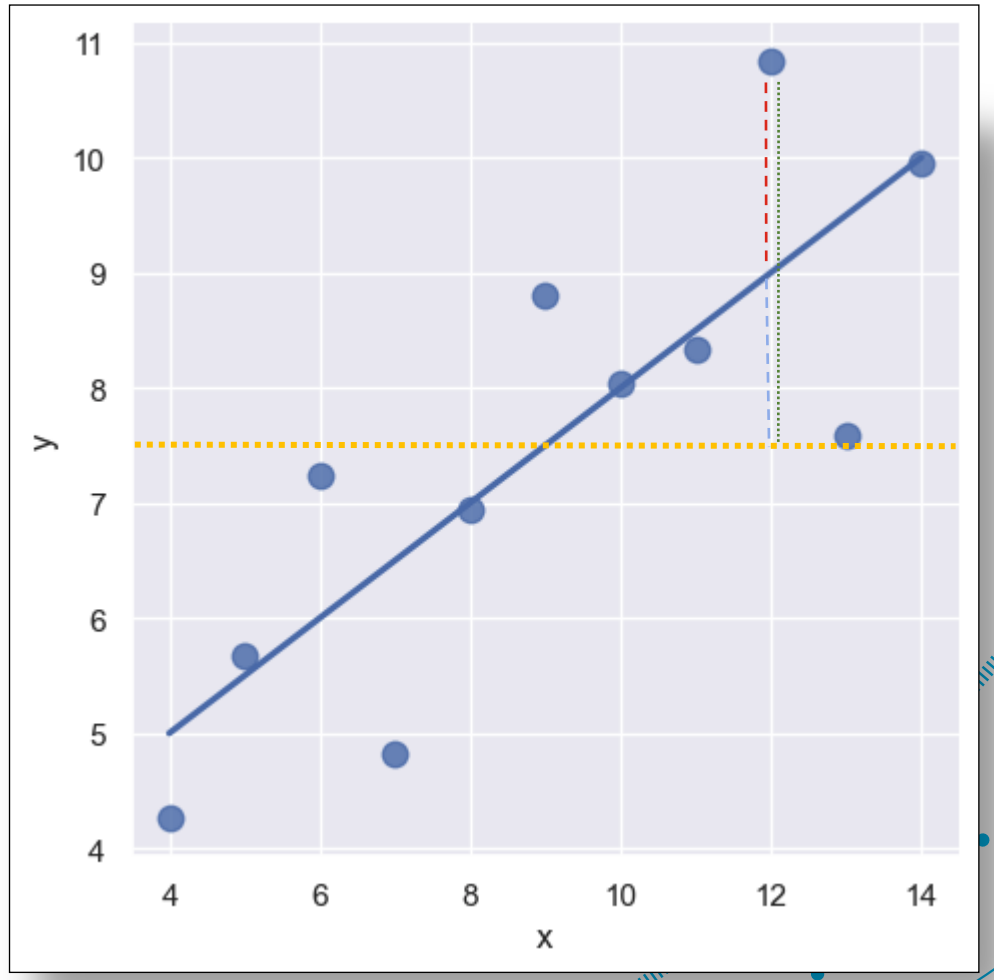
R^2 Value

- One way to think of this is we're differentiating between:

  explained variance (----),

  unexplained variance (----), and

  total variance (----)

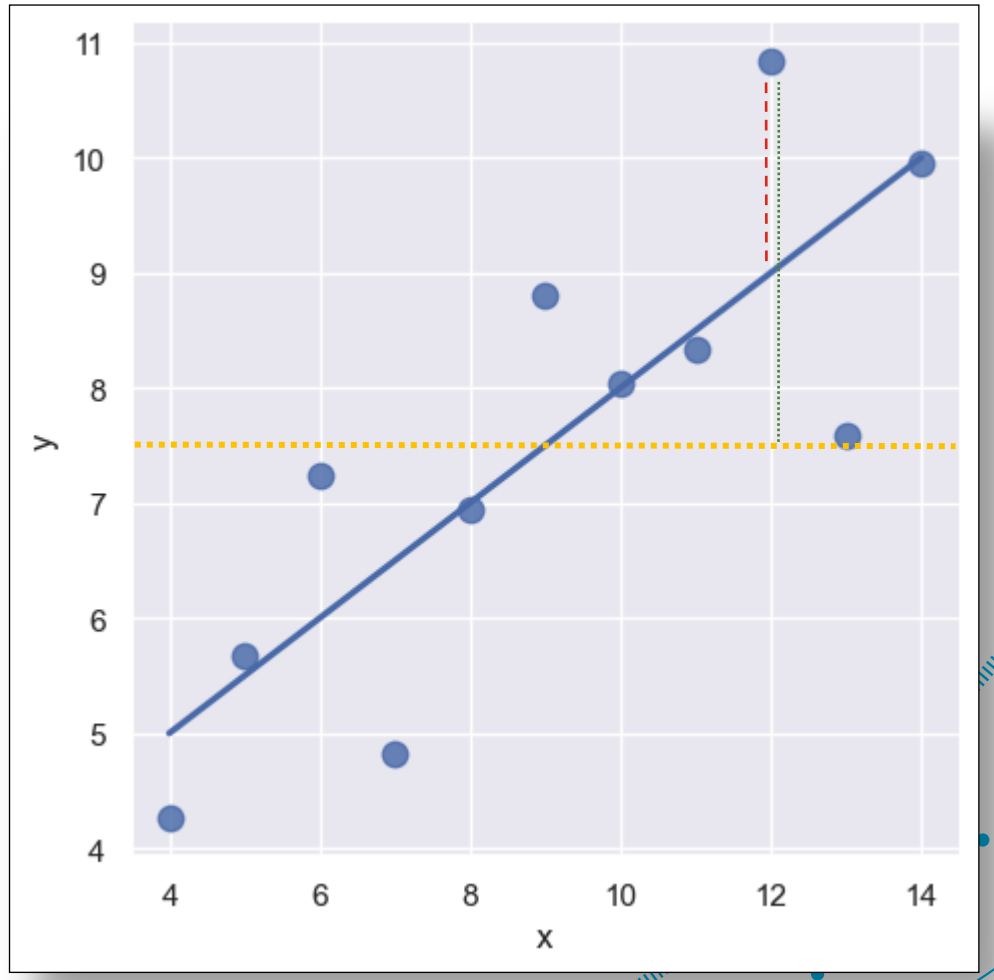


R^2 Value

- R^2 tells us what proportion of the total variance is explained by our regression model

$$R^2 = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$



Classification

Logistic Regression

- More specifically, we're looking to estimate:

$$Pr(y = 1|x_1, ...x_n)$$

- In other words, the probability that y is 1, given our IV(s)

- We typically estimate y = 1 whenever this probability > 0.5 (though we could technically choose a lower threshold)

Logistic Function

- To ensure that our probability is ($0 \leq Pr \leq 1$), we use the 'logistic function'

$$f(x) = \frac{1}{1+e^{-x}}$$

(Note, you don't need to know this)

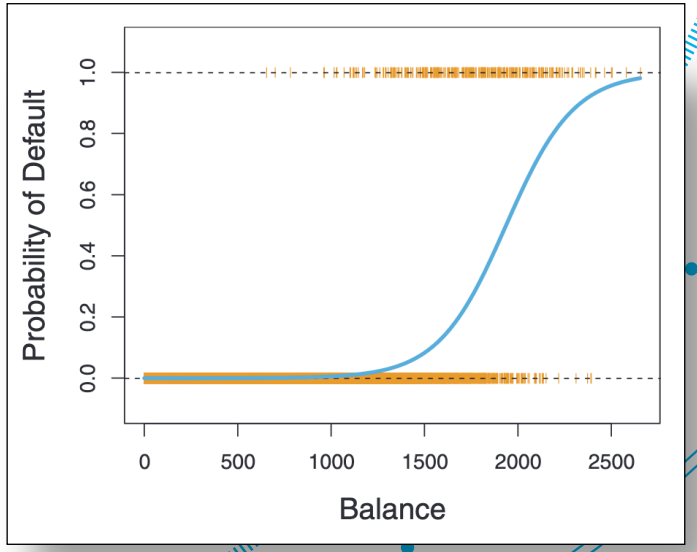
Logistic Function

- To ensure that our probability is ($0 \leq Pr \leq 1$), we use the 'logistic function'

$$f(x) = \frac{1}{1+e^{-x}}$$

(Note, you don't need to know this)

- This is what gives us a sigmoid (or S-shaped) curve (and logistic regression its name)



What Can We Do With Logistic Regression

- Similar to linear regression, with our logistic regression model:
  - We can use X values to predict the probability of Y being 1 (prediction)
  - We can also use X values to understand how they influence the Y value (interpretation)

Evaluating a Classification Model for Fit

- Variations of (pseudo) R^2 calculations exist for logistic regression, serving a similar function to that of R^2
  - i.e. quantify the variance explained by the model

- However, quite often what you'll see (especially in prediction contexts) is a *confusion matrix*

Confusion Matrix

- A confusion matrix is a table which reports these instances.
- It can also be used to calculate a number of different metrics.
- These include…

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

Confusion Matrix – Accuracy

Accuracy

= (TP + TN) / (TP + FP + FN + TN)

The proportion of all outcomes that were predicted correctly

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

Confusion Matrix – True Positive Rate

True Positive Rate

= TP / (TP + FN)

The proportion of {1} outcomes that were predicted correctly

Also known as **Sensitivity, Recall**

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

Confusion Matrix – True Negative Rate

True Negative Rate

= TN / (TN + FP)

The proportion of {0} outcomes that were predicted correctly

Also known as **Specificity**

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

Confusion Matrix – Positive Predictive Value

Positive Predictive Value

= TP / (TP + FP)

The proportion of outcomes that were predicted as {1} that were predicted correctly

Also known as **Precision**

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

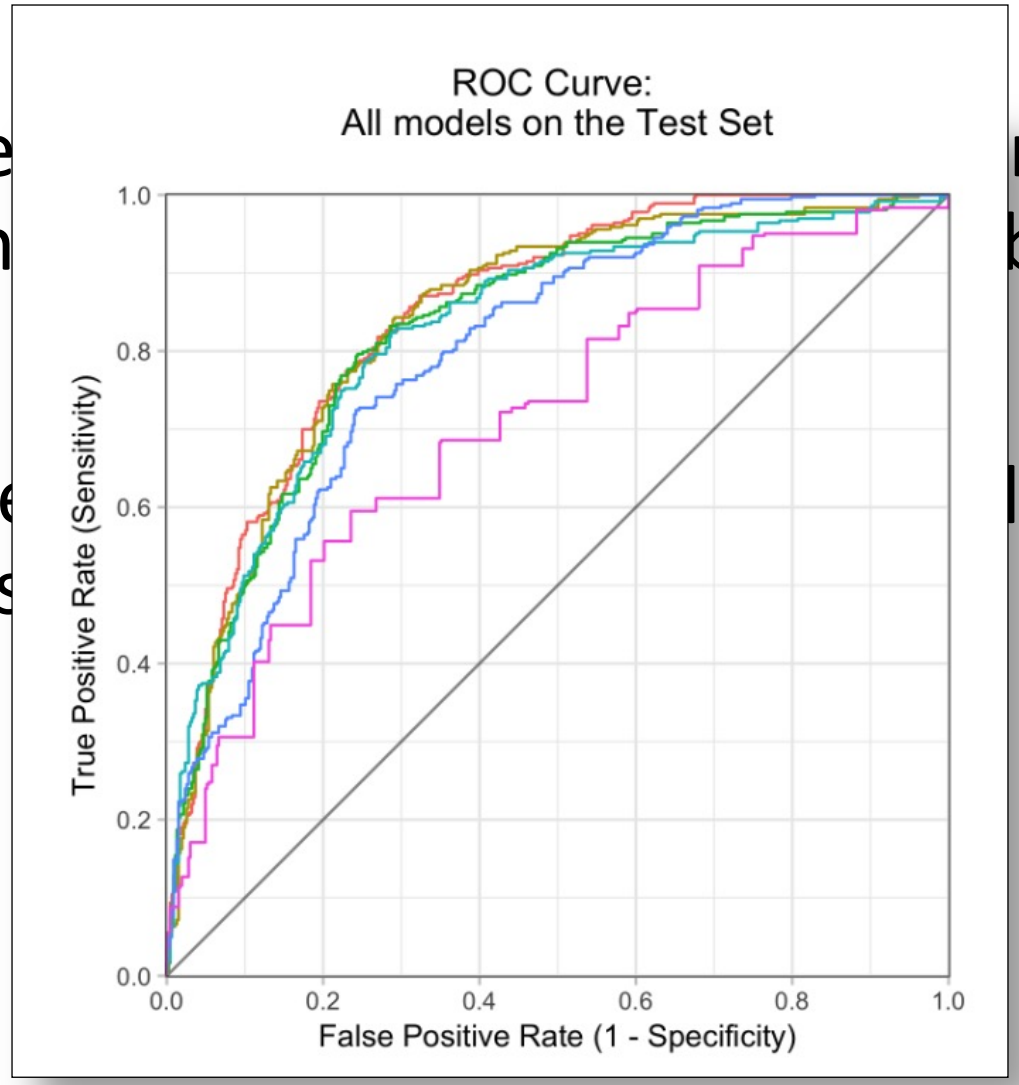
ROC Curve

- A ROC (Receiver operating characteristic) curve is a visual way to show the predictive performance of binary classifiers

- It plots the True Positive Rate against the False Positive Rate at different classification thresholds

ROC Curve

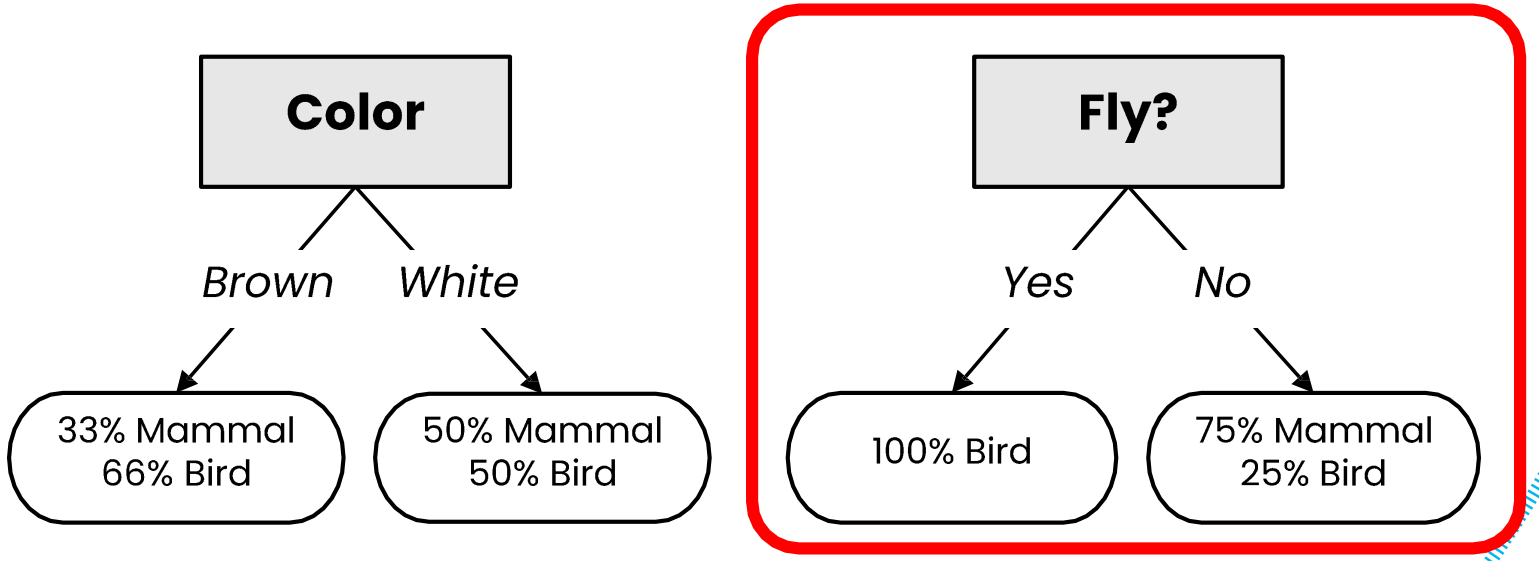
- A ROC (Receive rve is a visual way to show th binary classifiers
- It plots the True lse Positive Rate at different clas



Decision tree

What is a good attribute?

Does it fly?	Color	Class
No	Brown	**Mammal**
No	White	**Mammal**
Yes	Brown	**Bird**
Yes	White	**Bird**
No	White	**Mammal**
No	Brown	**Bird**
Yes	White	**Bird**



- Which attribute provides better splitting?
- Why?
  - Because the resulting subsets are more pure
  - Knowing the value of this attribute gives us more information about the label (the entropy of the subsets is lower)

Entropy

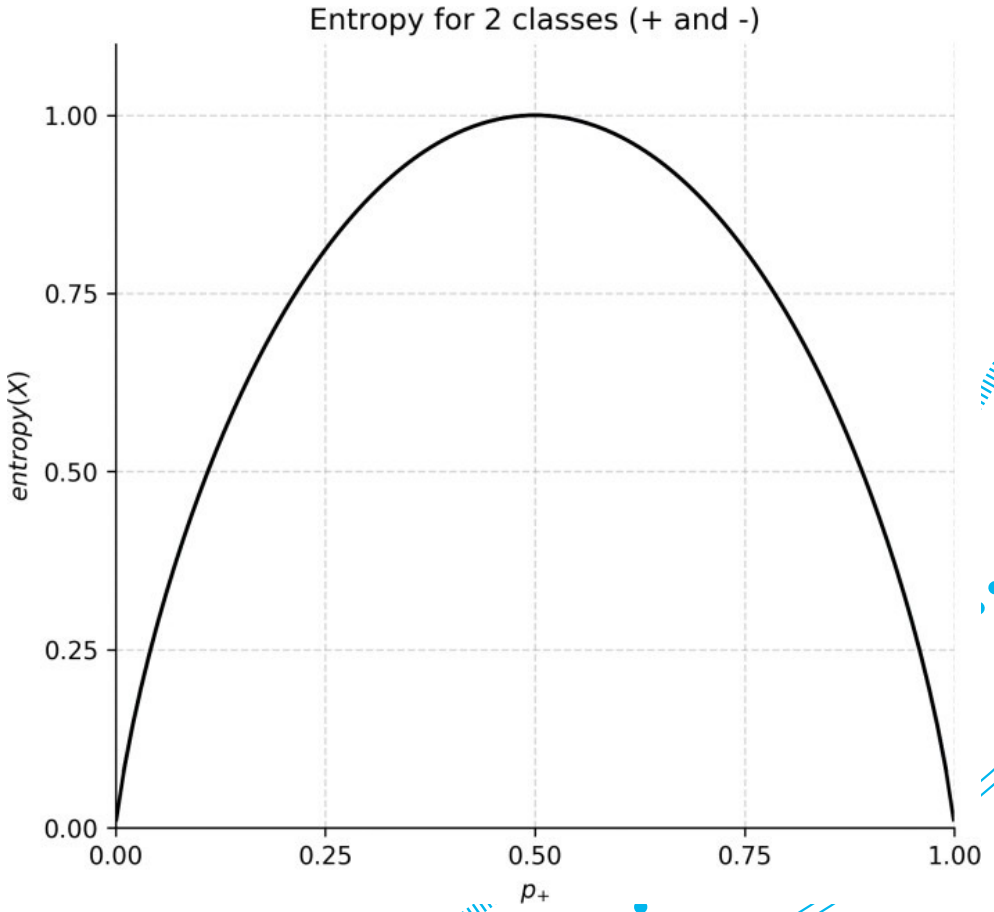
- Entropy measures the degree of randomness in data

- For a set of samples X with k classes:

$$entropy(X) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

where p_i is the proportion of elements of class i

- Lower entropy implies greater predictability!



How to calculate Entropy?

$$H(V) = -\sum_k P(v_k) \log_2 P(v_k)$$

- Example: If we had a total 10 data points in our dataset with 3 belonging to positive class and 7 belonging to negative class:

  -3/10 * log2 (3/10) – 7/10 * log2 (7/10) ≈ 0.876

- The Entropy is approximately 0.88 .

- ***High entropy means low level of purity.***

Information Gain

- The information gain of an attribute a is the expected reduction in entropy due to splitting on values of a:

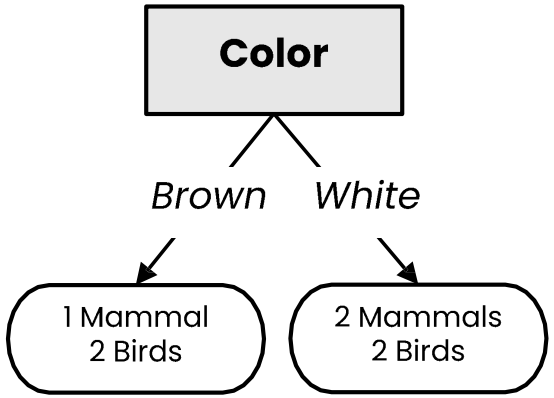
$$gain(X, a) = entropy(X) - \sum_{v \in Values(a)} \frac{|X_v|}{|X|} entropy(X_v)$$

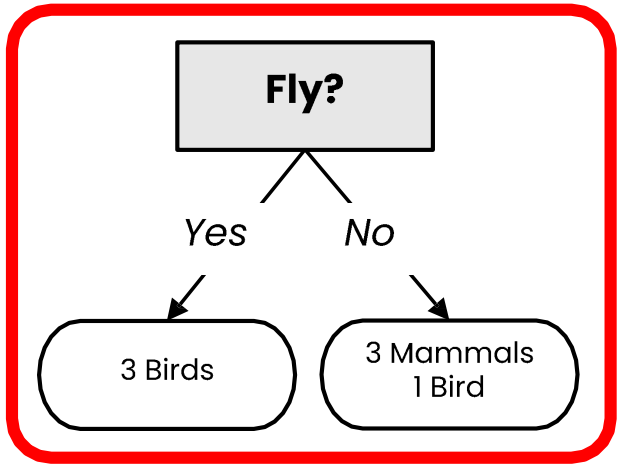
where X_v is the subset of X for which $a = v$

Best attribute = highest information gain

In practice, we compute $entropy(X)$ only once!

Does it fly?	Color	Class
No	Brown	**Mammal**
No	White	**Mammal**
Yes	Brown	**Bird**
Yes	White	**Bird**
No	White	**Mammal**
No	Brown	**Bird**
Yes	White	**Bird**





$$entropy\,(X) = -p_{\text{mammal}}\log_2 p_{\text{mammal}} - p_{\text{bird}}\log_2 p_{\text{bird}} = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} \approx 0.985$$

$$entropy\,(X_{color=brown}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} \approx 0.918 \qquad entropy\,(X_{color=white}) = 1$$

$$\boldsymbol{gain\,(X, color) = 0.985 - \frac{3}{7}\cdot 0.918 - \frac{4}{7}\cdot 1 \approx 0.020}$$

$$entropy\,(X_{fly=yes}) = 0 \qquad entropy\,(X_{fly=no}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} \approx 0.811$$

$$\boldsymbol{gain\,(X, fly) = 0.985 - \frac{3}{7}\cdot 0 - \frac{4}{7}\cdot 0.811 \approx 0.521}$$

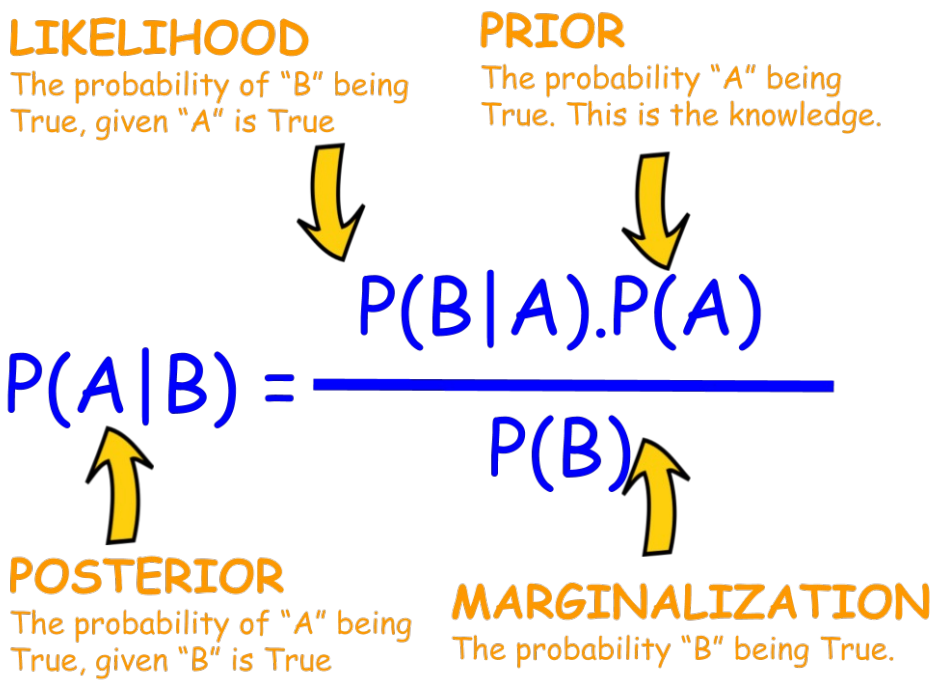
Naïve Bayes

Why is it called Naïve Bayes

- The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:
- **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.
    - Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- **Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:



Where:

- **P(A|B) is Posterior probability**: Probability of hypothesis A on the observed event B.
- **P(B|A) is Likelihood probability**: Probability of the data B given that the hypothesis A is true.
- **P(A) is Prior Probability**: Probability of hypothesis before observing the evidence. (regardless of the data)
- **P(B) is Marginal Probability**: Probability of Evidence. (regardless of the hypothesis)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example 2

- Example: Play Tennis

PlayTennis: training examples

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

Example 2

- Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=*No*
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=Yes) = 9/14 \quad P(\text{Play}=No) = 5/14$

Example 2

- Test Phase
  - Given a new instance,

    $\mathbf{x}'$=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
  - Look up tables

    P(Outlook=*Sunny*|Play=*Yes*) = 2/9

    P(Temperature=*Cool*|Play=*Yes*) = 3/9

    P(Huminity=*High*|Play=*Yes*) = 3/9

    P(Wind=*Strong*|Play=*Yes*) = 3/9

    P(Play=*Yes*) = 9/14

    P(Outlook=S*unny*|Play=*No*) = 3/5

    P(Temperature=*Cool*|Play==*No*) = 1/5

    P(Huminity=*High*|Play=*No*) = 4/5

    P(Wind=*Strong*|Play=*No*) = 3/5

    P(Play=*No*) = 5/14
  - MAP rule

    P(*Yes*|$\mathbf{x}'$): [P(*Sunny*|Yes)P(*Cool*|*Yes*)P(*High*|Yes)P(*Strong*|*Yes*)]P(Play=*Yes*) = 0.0053

    P(*No*|$\mathbf{x}'$): [P(*Sunny*|No) P(*Cool*|No)P(*High*|*No*)P(*Strong*|*No*)]P(Play=*No*) = 0.0206

    Given the fact P(*Yes*|$\mathbf{x}'$) < P(*No*|$\mathbf{x}'$), we label $\mathbf{x}'$ to be "*No*"