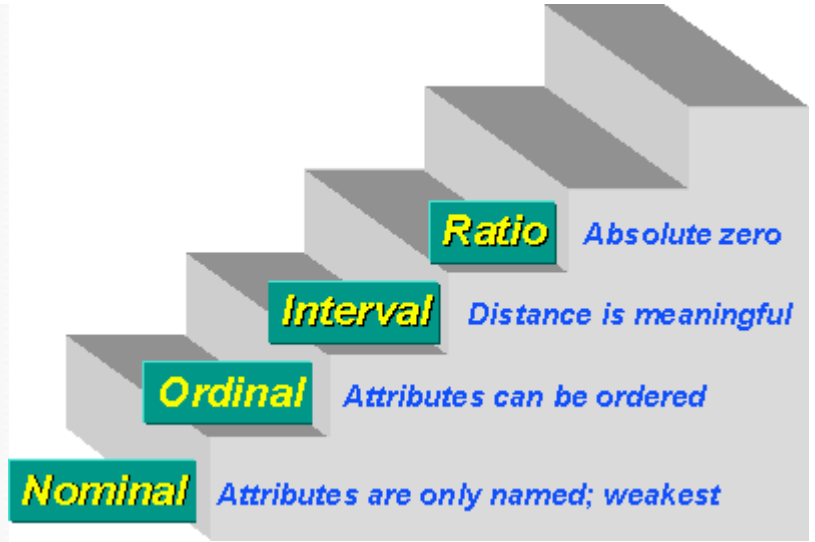
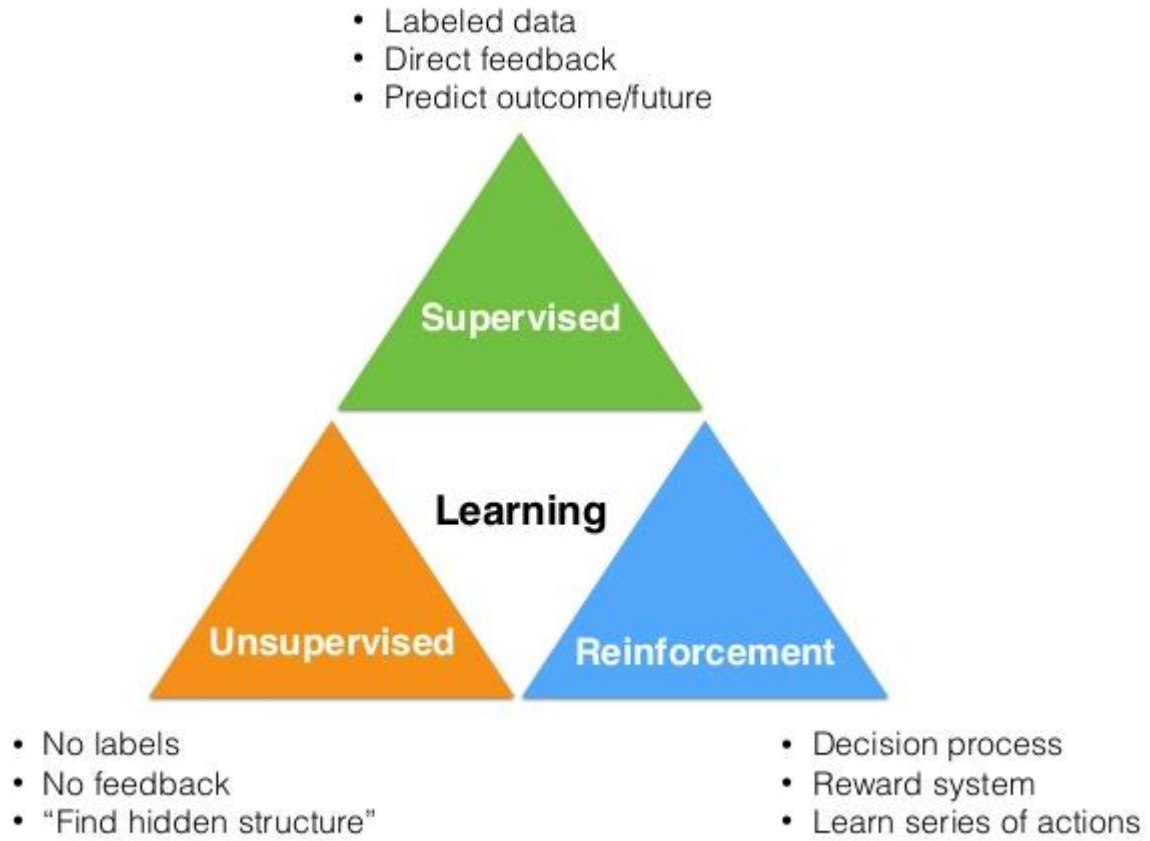
Levels of measurement

		EXAMPLES:
Ratio Data	Differences between measurements, true zero exists	Height, Age, Weekly Food Spending
Interval Data	Differences between measurements but no true zero	Temperature in Fahrenheit, Standardized exam score
Ordinal Data	Ordered Categories (rankings, order, or scaling)	Service quality rating, Student letter grades
Nominal Data	Categories (no ordering or direction)	Marital status, Type of car owned



Machine learning triangle



Machine Learning Use Cases

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
 Banking	**Predict credit worthiness of credit card holders**: Build a machine learning model to look for delinquency attributes by providing it with data on delinquent and non-delinquent customers	**Segment customers by behavioral characteristics**: Survey prospects and customers to develop multiple segments using clustering	**Create a 'next best offer' model for the call center group**: Build a predictive model that learns over time as users accept or reject offers made by the sales staff
 Healthcare	**Predict patient readmission rates**: Build a regression model by providing data on the patients' treatment regime and readmissions to show variables that best correlate with readmissions	**Categorize MRI data by normal or abnormal images**: Use deep learning techniques to build a model that learns different features of images to recognize different patterns	**Allocate scarce medical resources to handle different types of ER cases**: Build a Markov Decision Process that learns treatment strategies for each type of ER case
 Retail	**Analyze products customers buy together**: Build a supervised learning model to identify frequent item sets and association rules from transactional data	**Recommend products to customers based on past purchases**: Build a collaborative filtering model based on past purchases by "customers like them"	**Reduce excess stock with dynamic pricing**: Build a dynamic pricing model that adjusts the price based on customer response to offers

ML Lingo

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering

Data Mining vs Machine Learning vs Deep Learning

Data Mining

Data mining can be considered a superset of many different methods to extract insights from data. It might involve traditional statistical methods and machine learning. Data mining applies methods from many different areas to identify previously unknown patterns from data. This can include statistical algorithms, machine learning, text analytics, time series analysis and other areas of analytics. Data mining also includes the study and practice of data storage and data manipulation.

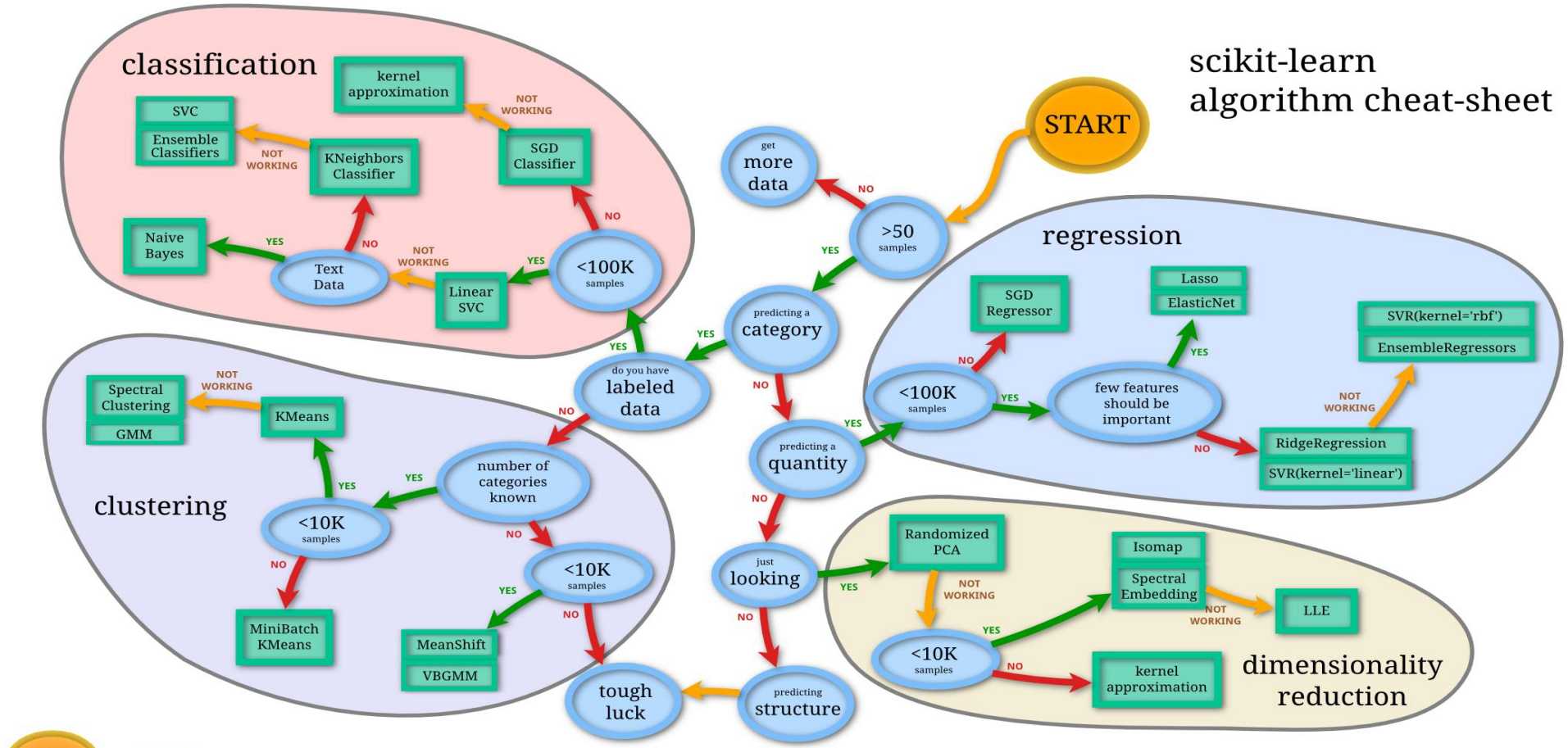
Machine Learning

The main difference with machine learning is that just like statistical models, the goal is to understand the structure of the data – fit theoretical distributions to the data that are well understood. So, with statistical models there is a theory behind the model that is mathematically proven, but this requires that data meets certain strong assumptions too. Machine learning has developed based on the ability to use computers to probe the data for structure, even if we do not have a theory of what that structure looks like. The test for a machine learning model is a validation error on new data, not a theoretical test that proves a null hypothesis. Because machine learning often uses an iterative approach to learn from data, the learning can be easily automated. Passes are run through the data until a robust pattern is found.

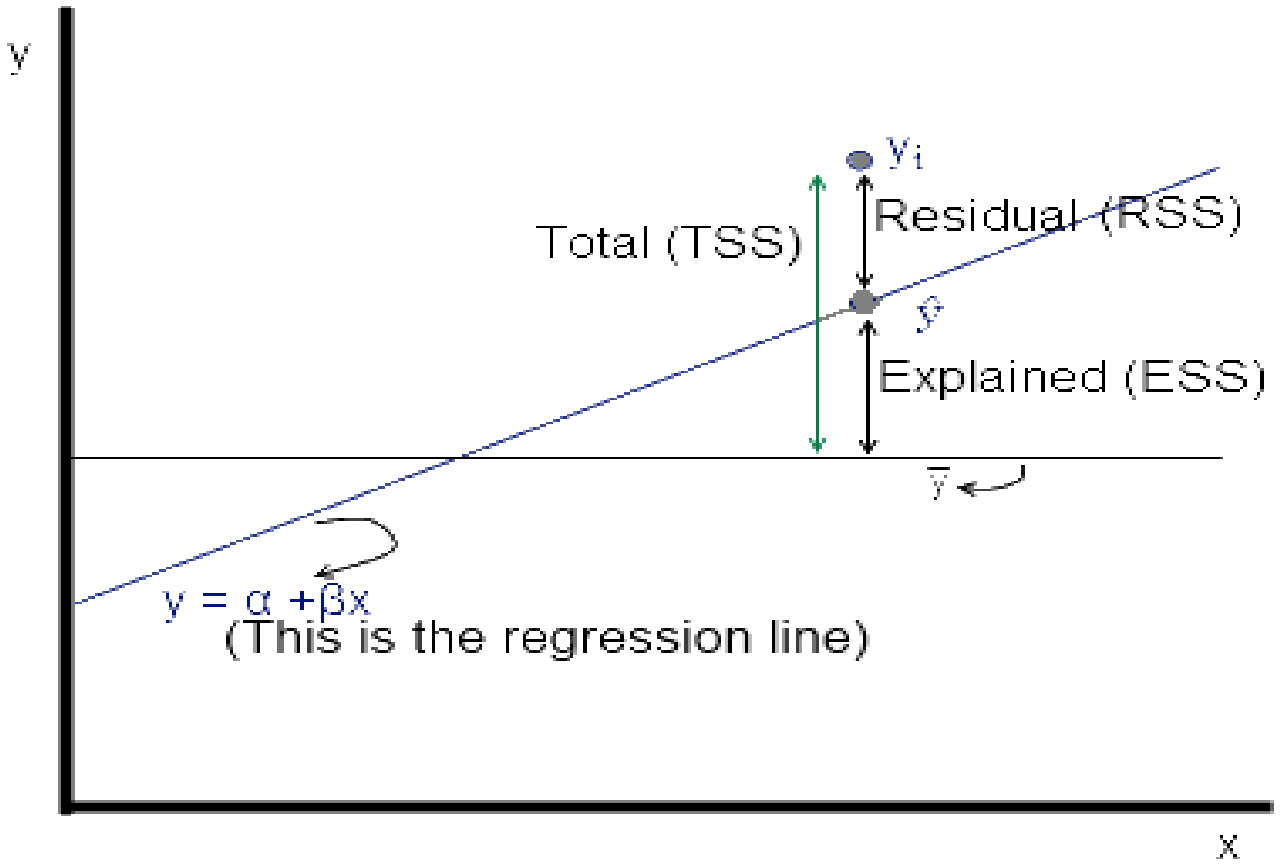
Deep learning

Deep learning combines advances in computing power and special types of neural networks to learn complicated patterns in large amounts of data. Deep learning techniques are currently state of the art for identifying objects in images and words in sounds. Researchers are now looking to apply these successes in pattern recognition to more complex tasks such as automatic language translation, medical diagnoses and numerous other important social and business problems.

ML cheat sheet



Regression graph [R2 = ESS/TSS]



$\hat{y}$ is the predicted value of y given x, using the equation $y = \alpha + \beta x$.

y_i is the actual observed value of y

$\bar{y}$ is the mean of y.

The distances that RSS, ESS and TSS represent are shown in the diagram to the left - but remember that the actual calculations are squares of these distances.

$$TSS = \Sigma (y_i - \bar{y})^2$$

$$RSS = \Sigma (y_i - \hat{y})^2$$

$$ESS = \Sigma (\hat{y} - \bar{y})^2$$

Degrees of Freedom

Degrees of freedom of an estimate is **the number of independent pieces of information that went into calculating the estimate**. It's not quite the same as the number of items in the sample. In order to get the df for the estimate, you have to subtract 1 from the number of items. Let's say you were finding the mean weight loss for a low-carb diet. You could use 4 people, giving 3 degrees of freedom (4 – 1 = 3), or you could use one hundred people with df = 99.

		Softdrink Choice		
		Coke	*Pepsi*	Total
Gender	*Male*	19	6	25
	Female	10	15	25
	Total	29	21	50

Which test what to use

Select A Statistical Test

- Hypothesis tests to find relationships between project Y and potential X's

		Y: Continuous	Y: Discrete
X	Continuous	Simple Linear Regression	Logistic Regression
X	Discrete	2 Sample t-Test *(Compare Means of two samples)* ANOVA *(Compare means of multiple samples)* Homgeneity of Variance *(Compare variances)*	Chi-Square Test

Cars dataset

```r
set.seed(122)
speed.c = scale(cars$speed, center=TRUE, scale=FALSE)
mod1 = lm(formula = dist ~ speed.c, data = cars)
summary(mod1)
```

```
##
## Call:
## lm(formula = dist ~ speed.c, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.9800     2.1750  19.761  < 2e-16 ***
## speed.c       3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Output of regression

- Formula Call
- As you can see, the first item shown in the output is the formula R used to fit the data. Note the simplicity in the syntax: the formula just needs the predictor (speed) and the target/response variable (dist), together with the data being used (cars).
- Residuals
- The next item in the model output talks about the residuals. Residuals are essentially the difference between the actual observed response values (distance to stop dist in our case) and the response values that the model predicted. The Residuals section of the model output breaks it down into 5 summary points. When assessing how well the model fit the data, you should look for a symmetrical distribution across these points on the mean value zero (0). In our example, we can see that the distribution of the residuals do not appear to be strongly symmetrical. That means that the model predicts certain points that fall far away from the actual observed points. We could take this further consider plotting the residuals to see whether this normally distributed, etc. but will skip this for this example.
- Coefficients
- The next section in the model output talks about the coefficients of the model. Theoretically, in simple linear regression, the coefficients are two unknown constants that represent the intercept and slope terms in the linear model. If we wanted to predict the Distance required for a car to stop given its speed, we would get a training set and produce estimates of the coefficients to then use it in the model formula. Ultimately, the analyst wants to find an intercept and a slope such that the resulting fitted line is as close as possible to the 50 data points in our data set.
- Coefficient - Estimate
- The coefficient Estimate contains two rows; the first one is the intercept. The intercept, in our example, is essentially the expected value of the distance required for a car to stop when we consider the average speed of all cars in the dataset. In other words, it takes an average car in our dataset 42.98 feet to come to a stop. The second row in the Coefficients is the slope, or in our example, the effect speed has in distance required for a car to stop. The slope term in our model is saying that for every 1 mph increase in the speed of a car, the required distance to stop goes up by 3.9324088 feet.
- Coefficient - Standard Error
- The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable. We'd ideally want a lower number relative to its coefficients. In our example, we've previously determined that for every 1 mph increase in the speed of a car, the required distance to stop goes up by 3.9324088 feet. The Standard Error can be used to compute an estimate of the expected difference in case we ran the model again and again. In other words, we can say that the required distance for a car to stop can vary by 0.4155128 feet. The Standard Errors can also be used to compute confidence intervals and to statistically test the hypothesis of the existence of a relationship between speed and distance required to stop.

Output of regression

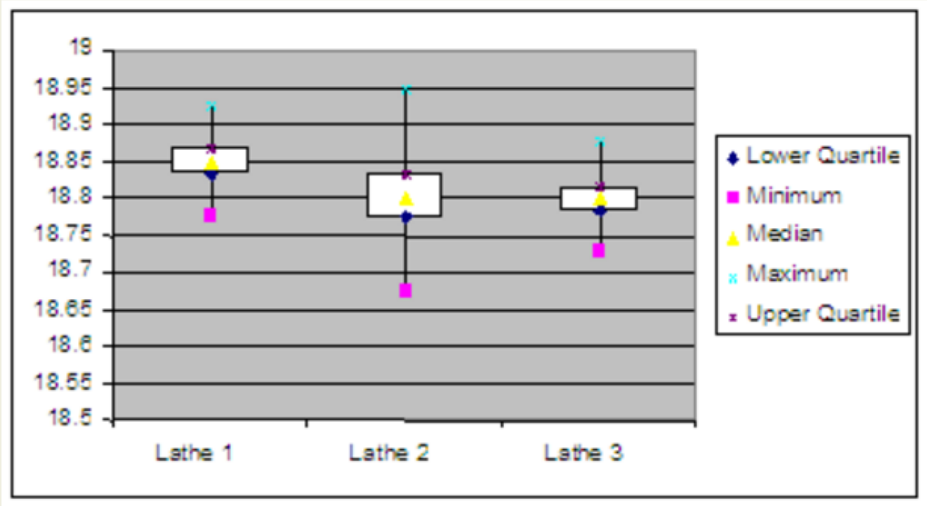
- Coefficient - t value
- The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between speed and distance exist. In our example, the t-statistic values are relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists. In general, t-values are also used to compute p-values.
- Coefficient - Pr(>|t|)
- The Pr(>|t|) acronym found in the model output relates to the probability of observing any value equal or larger than |t|. A small p-value indicates that it is unlikely we will observe a relationship between the predictor (speed) and response (dist) variables due to chance. Typically, a p-value of 5% or less is a good cut-off point. In our model example, the p-values are very close to zero. Note the 'signif. Codes' associated to each estimate. Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between speed and distance.
- Residual Standard Error
- Residual Standard Error is measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term E. Due to the presence of this error term, we are not capable of perfectly predicting our response variable (dist) from the predictor (speed) one. The Residual Standard Error is the average amount that the response (dist) will deviate from the true regression line. In our example, the actual distance required to stop can deviate from the true regression line by approximately 15.3795867 feet, on average. In other words, given that the mean distance for all cars to stop is 42.98 and that the Residual Standard Error is 15.3795867, we can say that the percentage error is (any prediction would still be off by) 35.78%. It's also worth noting that the Residual Standard Error was calculated with 48 degrees of freedom. Simplistically, degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account these parameters (restriction). In our case, we had 50 data points and two parameters (intercept and slope).
- Multiple R-squared, Adjusted R-squared
- The R-squared statistic (R2R2) provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. The R2R2 is a measure of the linear relationship between our predictor variable (speed) and our response / target variable (dist). It always lies between 0 and 1 (i.e.: a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable). In our example, the R2R2 we get is 0.6510794. Or roughly 65% of the variance found in the response variable (dist) can be explained by the predictor variable (speed). Step back and think: If you were able to choose any metric to predict distance required for a car to stop, would speed be one and would it be an important one that could help explain how distance would vary based on speed? I guess it's easy to see that the answer would almost certainly be a yes. That why we get a relatively strong R2R2. Nevertheless, it's hard to define what level of R2R2 is appropriate to claim the model fits well. Essentially, it will vary with the application and the domain studied.
- A side note: In multiple regression settings, the R2R2 will always increase as more variables are included in the model. That's why the adjusted R2R2 is the preferred measure as it adjusts for the number of variables considered.
- F-Statistic
- F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is. However, how much larger the F-statistic needs to be depends on both the number of data points and the number of predictors. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis (H0 : There is no relationship between speed and distance). The reverse is true as if the number of data points is small, a large F-statistic is required to be able to ascertain that there may be a relationship between predictor and response variables. In our example the F-statistic is 89.5671065 which is relatively larger than 1 given the size of our data.

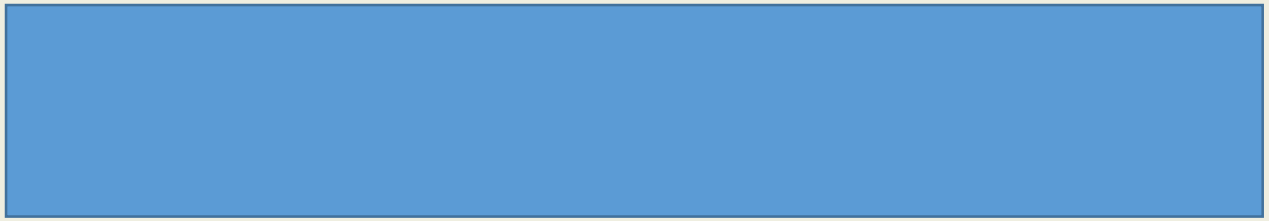
Box plot

Box and Whisker Plot Example

Suppose you wanted to compare the performance of three lathes responsible for the rough turning of a motor shaft. The design specification is 18.85 +/- 0.1 mm.

Diameter measurements from a sample of shafts taken from each roughing lathe are displayed in a box and whisker plot.





A bag contains 9 white and 6 black balls. What is the probability of selecting
(i) 2 white balls
(ii) 3 white balls
(iii) 4 black balls
(iv) 1 white and 3 black balls
(v) 4 white and 5 black balls
Solution: -
We know that the combination formula for selecting r items from n is
nCr= n!/r!(n-r)!
Also we know that probability of an event A is
P(A) = Number of favorable outcomes/ Total number of outcomes
(i) We have to find the probability of selecting 2 white balls.
So favorable cases will be obtained when we select 2 balls from 9 white balls. This can be done in 9C2 ways.
Since there are 9 + 6 = 15 balls, the total number of outcomes will be obtained by selecting 2 balls from the 15 balls. This can be done in 15C2 ways.
So the required probability is **9C2/15C2**
(ii)We have to find the probability of selecting 3 white balls.
So favorable cases will be obtained when we select 3 balls from 9 white balls. This can be done in 9C3 ways.
Since there are 9 + 6 = 15 balls, the total number of outcomes will be obtained by selecting 3 balls from the 15 balls. This can be done in 15C3 ways.
So the required probability is **9C3/15C3**
(iii) We have to find the probability of selecting 4 blackballs.
So favorable cases will be obtained when we select 4 balls from 6 black balls. This can be done in 6C4 ways.
Since there are 9 + 6 = 15 balls, the total number of outcomes will be obtained by selecting 4 balls from the 15 balls. This can be done in 15C4 ways.
So the required probability is **6C4/15C4**
(iv) We have to find the probability of selecting 1 white and 3 black balls.
So favorable cases will be obtained when we select 1 white from the 9 white balls and 3 black balls from the 6 black balls. This can be done in 9C1 x 6C3 ways.
Since there are 9 + 6 = 15 balls, the total number of outcomes will be obtained by selecting 4 balls from the 15 balls. This can be done in 15C4 ways.
So the required probability is **9C1 x 6C3 /15C4**
(v) We have to find the probability of selecting 4 white and 5 black balls.
So favorable cases will be obtained when we select 4 white from the 9 white balls and 3 black balls from the 5 black balls. This can be done in 9C4 x 6C5 ways.
Since there are 9 + 6 = 15 balls, the total number of outcomes will be obtained by selecting 9 balls from the 15 balls. This can be done in 15C9 ways.
So the required probability is **9C4 x 6C5 /15C9**

More probability

3. Given two urns, suppose urn I contains 4 black and 7 white balls. Urn II contains 3 black, 1 white, and 4 yellow balls. Select an urn and then select a ball.
   a) What is the probability that you obtain a black ball? (**Answer: 65/176**)

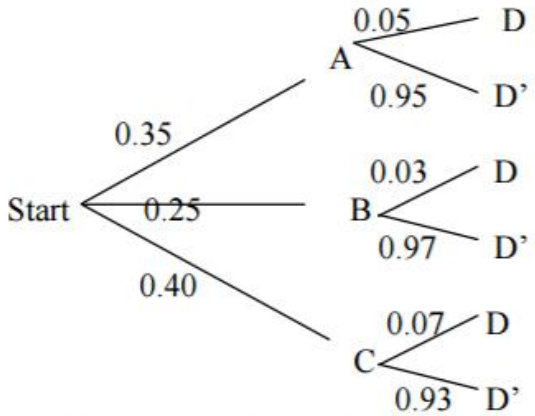
0.5*choose(4,1)/choose(11,1) + 0.5*choose(3,1)/choose(8,1)

Bayes theorem & Conditional probability

1. **In New York State, 48% of all teenagers own a skateboard and 39% of all teenagers own a skateboard and roller blades. What is the probability that a teenager owns roller blades given that the teenager owns a skateboard?**

81.25

Example 1. A company produces 1,000 refrigerators a week at three plants. Plant A produces 350 refrigerators a week, plant B produces 250 refrigerators a week, and plant C produces 400 refrigerators a week. Production records indicate that 5% of the refrigerators at plant A will be defective, 3% of those produced at plant B will be defective, and 7% of those produced at plant C will be defective. All refrigerators are shipped to a central warehouse. If a refrigerator at the warehouse is found to be defective, what is the probability that it was produced a) at plant A? b) at plant B? c) at plant C?
We consider D as defective and D' as non defective.



We now answer all questions from the tree diagram.

a) $$P(A \mid D) = \frac{P(A \cap D)}{P(D)} = \frac{0.35(0.05)}{0.35(0.05)+0.25(0.03)+0.40(0.07)} = \frac{1}{3}$$

You now try to find b) $$P(B \mid D) = \frac{P(B \cap D)}{P(D)} = \frac{0.25(0.03)}{0.35(0.05)+0.25(0.03)+0.40(0.07)} = ?$$

c) $$P(C \mid D) = \frac{P(C \cap D)}{P(D)} = \frac{0.40(0.07)}{0.35(0.05)+0.25(0.03)+0.40(0.07)} = ?$$

Form Null and alternate hypothesis

- I expect the average recovery period to be less than 8.2 weeks.
- *Put half a pack of Mentos into a 2-Liter Diet Coke bottle and check for explosion*
- *Bomb on Plane*

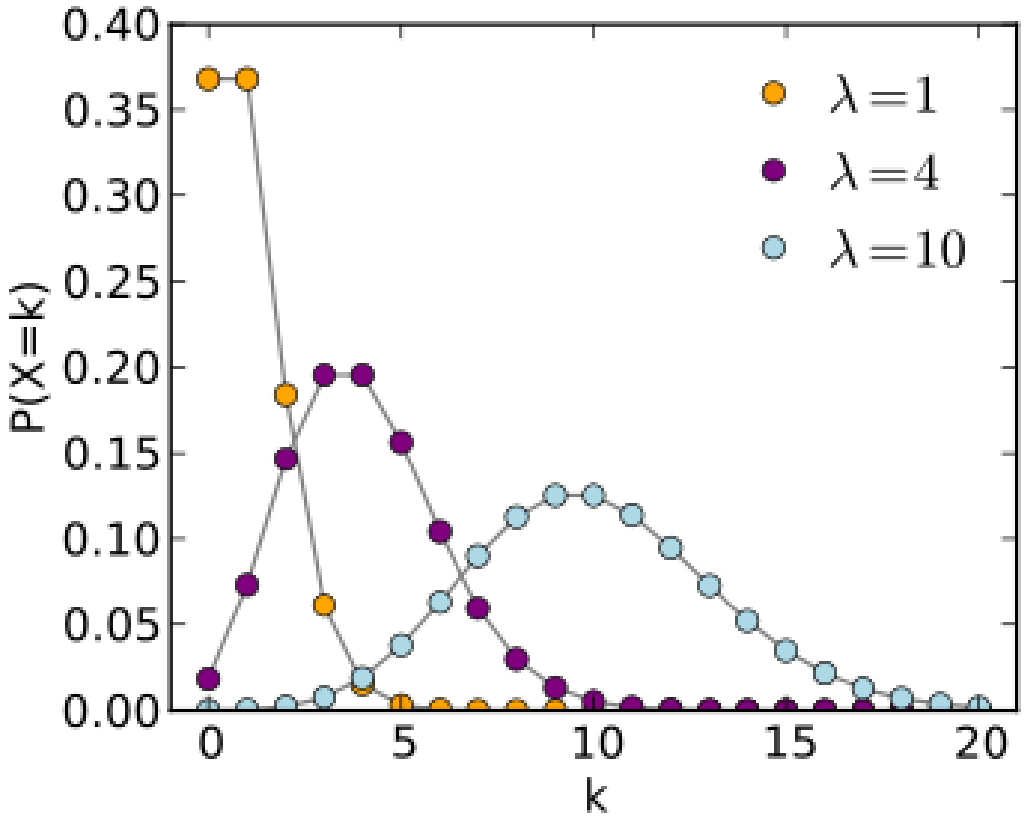
Poisson distribution example

- **Problem**
- If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.
- **Solution**
- The probability of having *sixteen or less* cars crossing the bridge in a particular minute is given by the function ppois.
- > ppois(16, lambda=12) # lower tail
  [1] 0.89871
- Hence the probability of having seventeen or more cars crossing the bridge in a minute is in the *upper tail* of the probability density function.
- > ppois(16, lambda=12, lower=FALSE) # upper tail
  [1] 0.10129

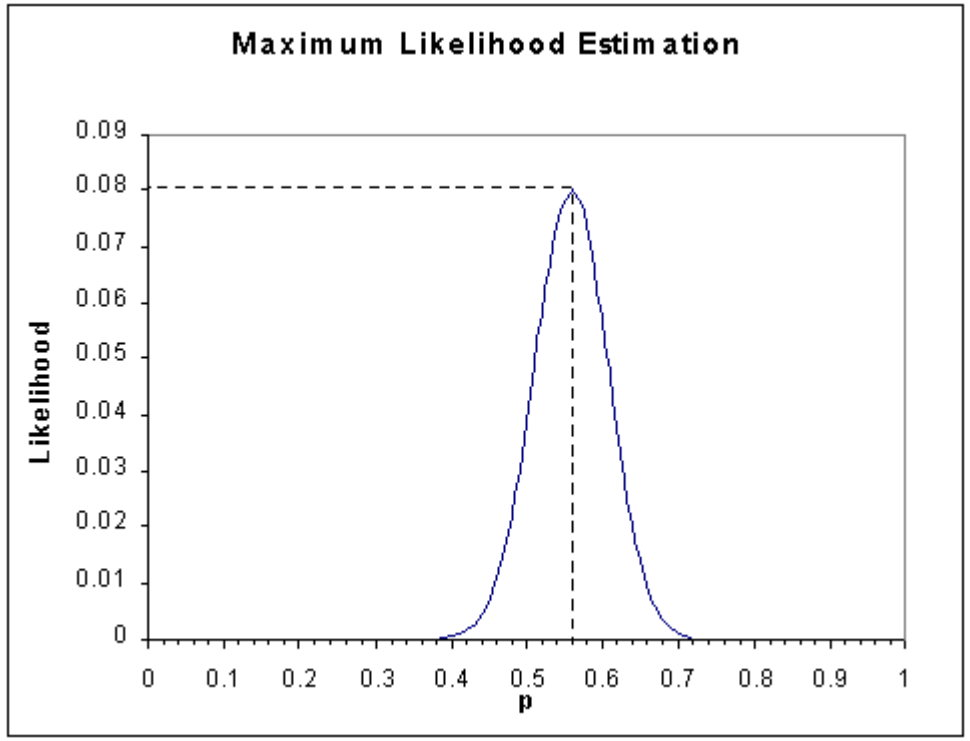
Poisson distribution

- Any business process where lines are a matter of fact — this means:
- Emergency or Doctor’s offices
- Restaurants
- Server Load in a network environment
- Fulfillment/Distribution Center or Warehousing
- Project Management
- Call Centers
- Software Engineering
- etc…

Lambda is the average number of events per interval



Maximum likelihood estimation



Probability

Knowing parameters -> Prediction of outcome

Likelihood

Observation of data -> Estimation of parameters

Normal / Gaussian problem

Problem

Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

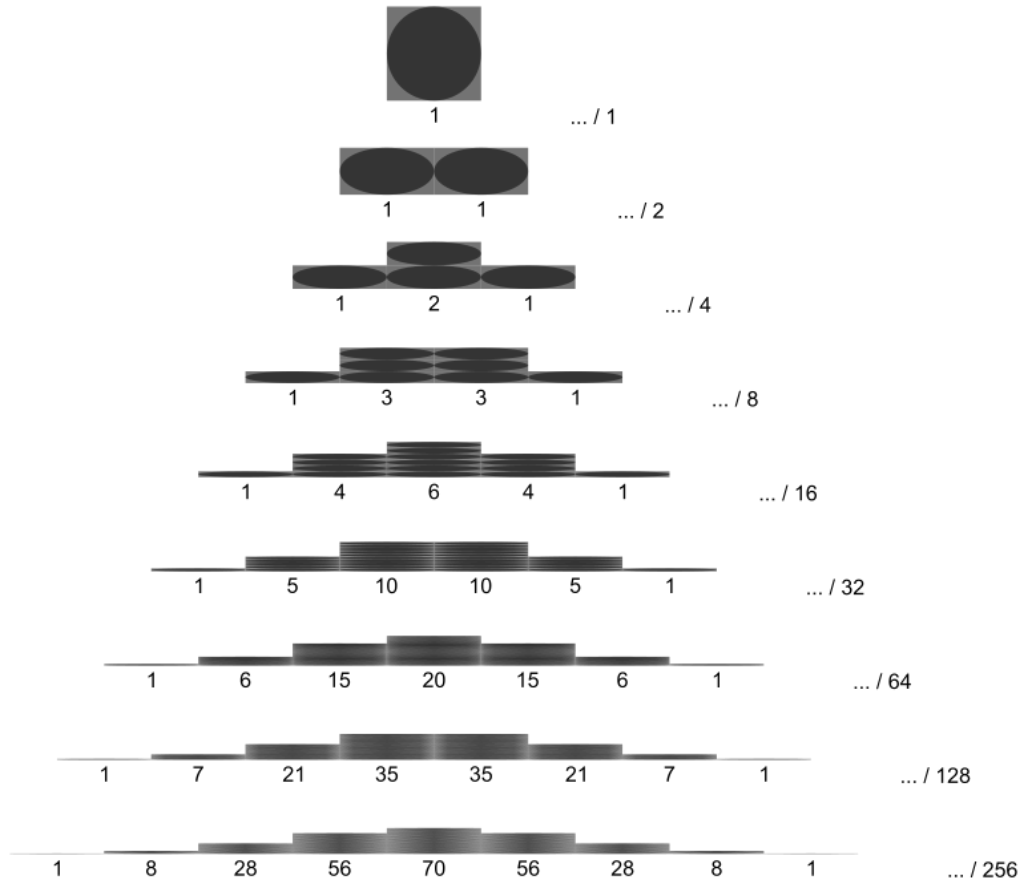
Solution

We apply the function pnorm of the normal distribution with mean 72 and standard deviation 15.2. Since we are looking for the percentage of students scoring higher than 84, we are interested in the *upper tail* of the normal distribution.

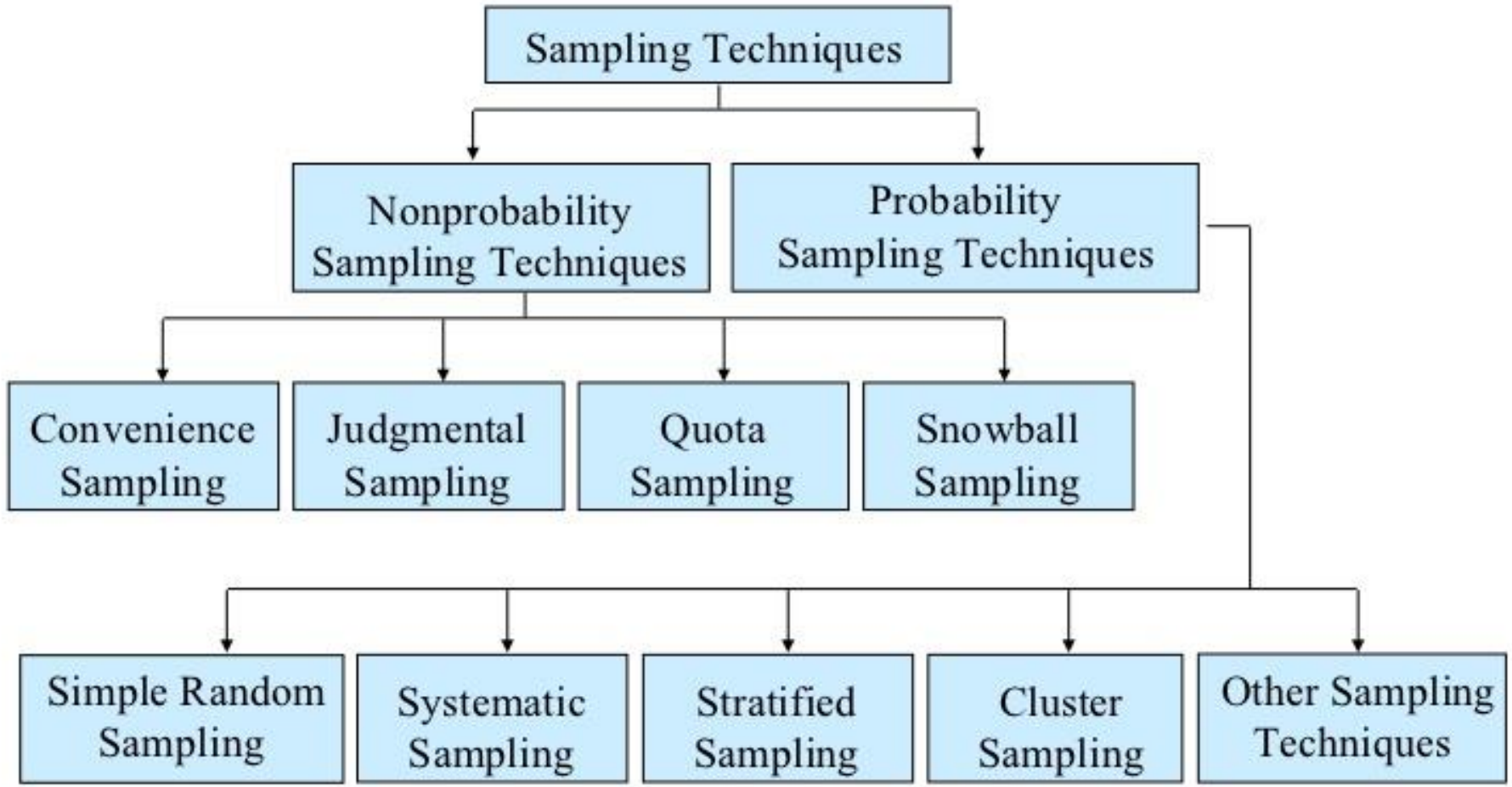
```
> pnorm(84, mean=72, sd=15.2, lower.tail=FALSE)
[1] 0.21492
```

Answer

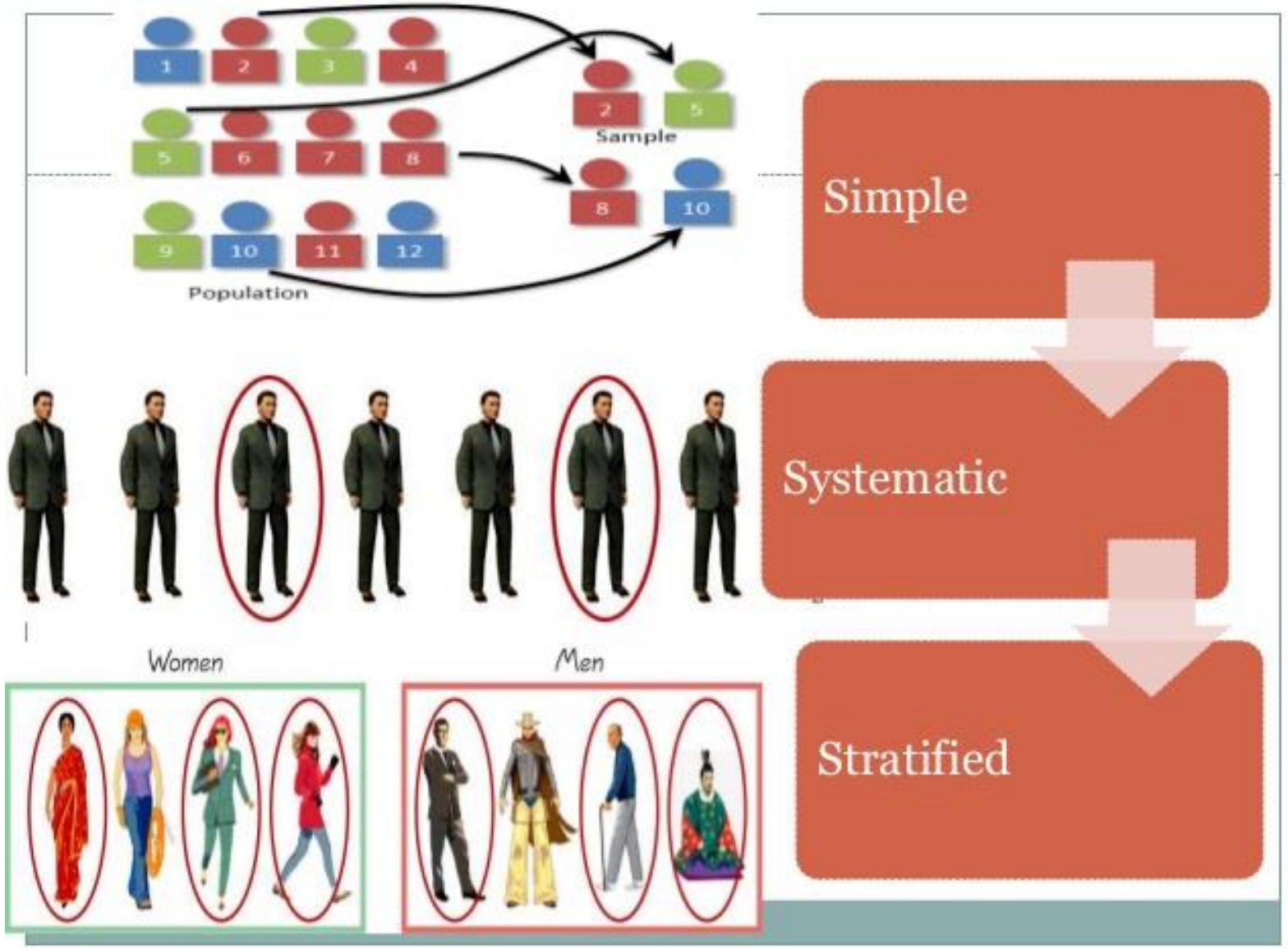
The percentage of students scoring 84 or more in the college entrance exam is 21.5%.


1 ... / 1
1 1 ... / 2
1 2 1 ... / 4
1 3 3 1 ... / 8
1 4 6 4 1 ... / 16
1 5 10 10 5 1 ... / 32
1 6 15 20 15 6 1 ... / 64
1 7 21 35 35 21 7 1 ... / 128
1 8 28 56 70 56 28 8 1 ... / 256

Classification of Sampling Techniques



Simple vs systematic



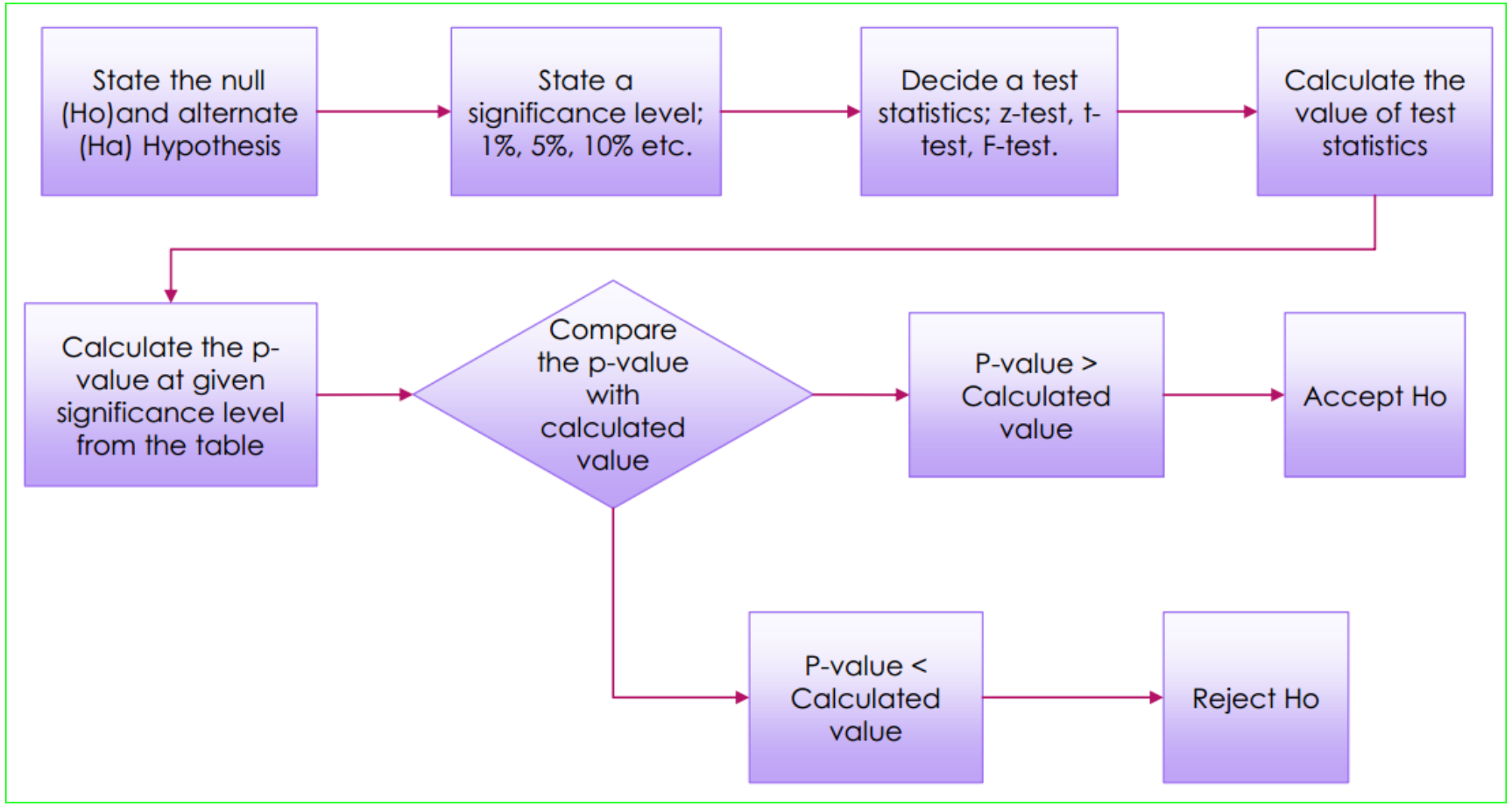
Stratified vs Cluster sampling

Stratified Sampling Vs Cluster Sampling

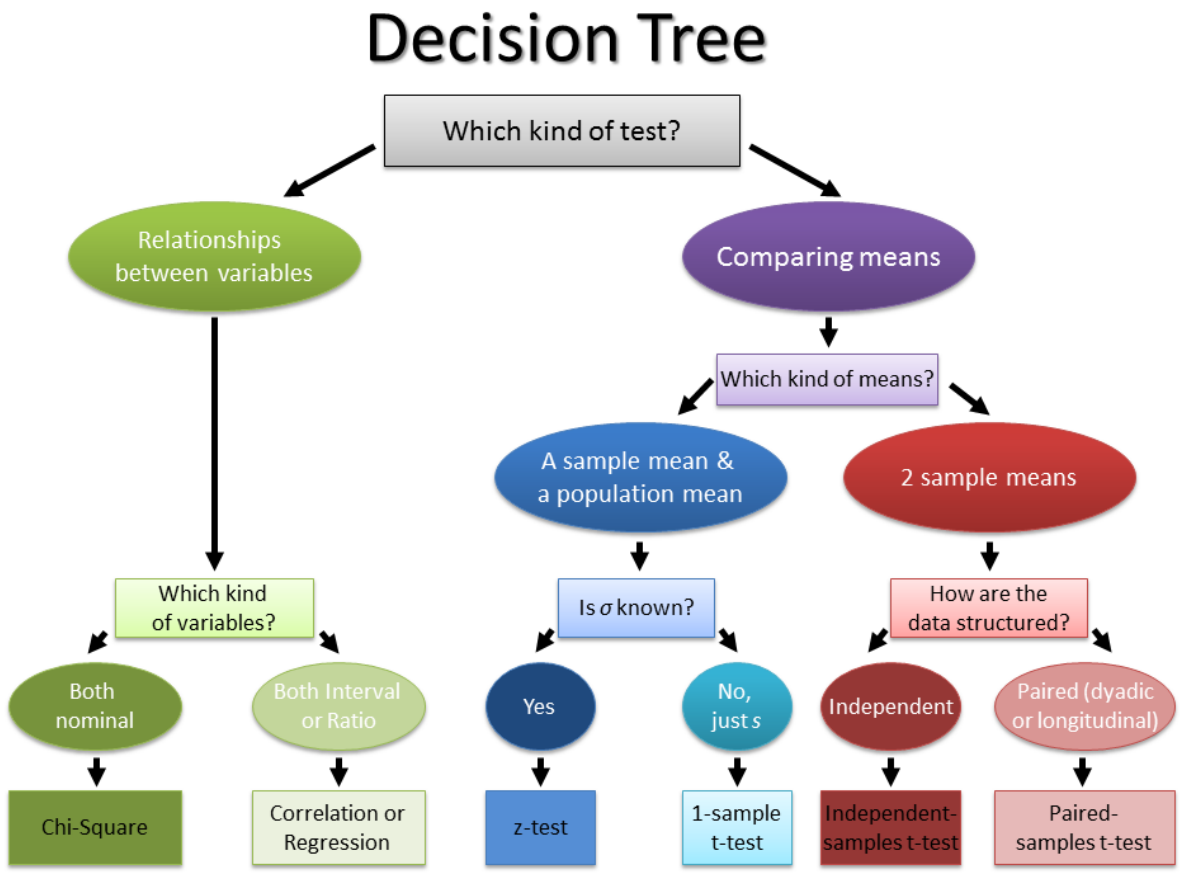
Find outliers

- **Find the outliers, if any, for the following data set:**
- **10.2, 14.1, 14.4. 14.4, 14.4, 14.5, 14.5, 14.6, 14.7, 14.7, 14.7, 14.9, 15.1, 15.9, 16.4**
- To find out if there are any outliers, I first have to find the IQR. There are fifteen data points, so the median will be at position (15 + 1) ÷ 2 = 8. Then Q_2 = 14.6. There are seven data points on either side of the median, so Q_1 is the fourth value in the list and Q_3 is the twelfth: Q_1 = 14.4 and Q_3 = 14.9. Then IQR = 14.9 – 14.4 = 0.5.
- Outliers will be any points below $Q_1 – 1.5 \times IQR = 14.4 – 0.75 = 13.65$ or above $Q_3 + 1.5 \times IQR = 14.9 + 0.75 = 15.65$.
- **Then the outliers are at 10.2, 15.9, and 16.4.**

Steps …


State the null (Ho)and alternate (Ha) Hypothesis
State a significance level; 1%, 5%, 10% etc.
Decide a test statistics; z-test, t-test, F-test.
Calculate the value of test statistics
Calculate the p-value at given significance level from the table
Compare the p-value with calculated value
P-value > Calculated value
Accept Ho
P-value < Calculated value
Reject Ho

Decision tree in inferential stats.



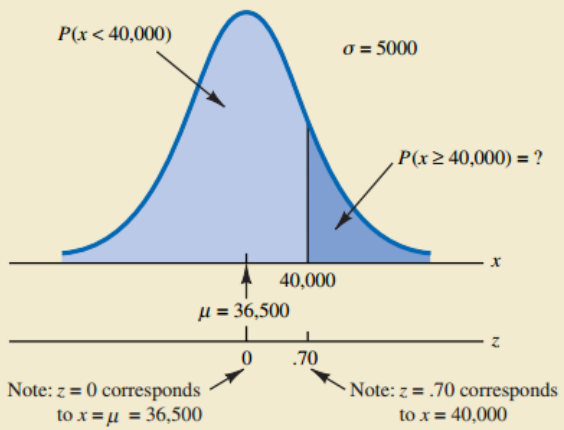
Normal distribution [239 Anderson Sweeny]

Grear Tire Company Problem

We turn now to an application of the normal probability distribution. Suppose the Grear Tire Company developed a new steel-belted radial tire to be sold through a national chain of discount stores. Because the tire is a new product, Grear's managers believe that the mileage guarantee offered with the tire will be an important factor in the acceptance of the product. Before finalizing the tire mileage guarantee policy, Grear's managers want probability information about x = number of miles the tires will last.

From actual road tests with the tires, Grear's engineering group estimated that the mean tire mileage is $\mu = 36{,}500$ miles and that the standard deviation is $\sigma = 5000$. In addition, the data collected indicate that a normal distribution is a reasonable assumption. What percentage of the tires can be expected to last more than 40,000 miles? In other words, what is the probability that the tire mileage, x, will exceed 40,000? This question can be answered by finding the area of the darkly shaded region in Figure 6.6.

FIGURE 6.6 GREAR TIRE COMPANY MILEAGE DISTRIBUTION



A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr?

2. Let x be the random variable that represents the speed of cars. x has μ = 90 and σ = 10. We have to find the probability that x is higher than 100 or P(x > 100)

For x = 100 , z = (100 - 90) / 10 = 1

P(x > 90) = P(z >, 1) = [total area] - [area to the left of z = 1]

= 1 - 0.8413 = 0.1587

The probability that a car selected at a random has a speed greater than 100 km/hr is equal to 0.1587

Chi-square

Example

In the built-in data set survey, the **Smoke** column records the students smoking habit, while the **Exer** column records their exercise level. The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".

We can tally the students smoking habit against the exercise level with the table function in R. The result is called the **contingency table** of the two variables.

```
> library(MASS)       # load the MASS package
> tbl = table(survey$Smoke, survey$Exer)
> tbl                 # the contingency table

       Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
```

Problem

Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

Solution

We apply the chisq.test function to the contingency table tbl, and found the p-value to be 0.4828.

Solution

We apply the chisq.test function to the contingency table tbl, and found the p-value to be 0.4828.

```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  table(survey$Smoke, survey$Exer)
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(table(survey$Smoke, survey$Exer)) :
  Chi-squared approximation may be incorrect
```

Answer

As the p-value 0.4828 is greater than the .05 significance level, we do not reject the null hypothesis that the smoking habit is independent of the exercise level of the students.

Z – test

Consider the following exercise: The policy of a particular bank branch is that its ATMs must be stocked with enough cash to satisfy customers making withdrawals over an entire weekend. At this branch the expected (i.e., population) average amount of money withdrawn from ATM machines per customer transaction over the weekend is \$160 with an expected (i.e., population) standard deviation of \$30. Suppose that a random sample of 36 customer transactions is examined and it is observed that the sample mean withdrawal is \$172. Note that the standard deviation that we are to use for this problem did *not* come from the sample. Therefore, this will be a **Z**-test, not a ***t***-test. For this problem, we have $\sigma = 30$, $n = 36$, and $\overline{x} = 172$.

(a)

State the null and alternative hypothesis.
This is not very clear, but apparently we are to check if the average *exceeds* \$160, which would mean the ATMs are not "stocked with enough cash." This is what we will try to show, so it should go in the alternative hypothesis. Therefore the hypotheses would be:

$$H_0 : \mu \leq 160$$

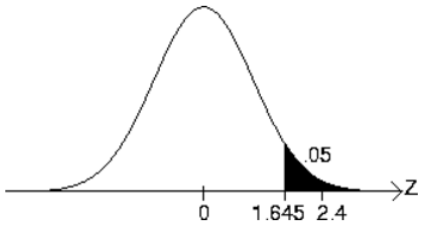
$$H_1 : \mu > 160$$

(b)

At the .05 level of significance, using the critical value approach to hypothesis testing, is there enough evidence to believe that the true average withdrawal is greater than \$160?
Let us get the test statistic:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{172 - 160}{30/\sqrt{36}} = 2.4$$

Set up the rejection region by drawing a Z-curve and shade the last 5% of the right tail. We need the Z critical value associated with this area, which is $Z_\alpha = Z_{.05} = 1.645$. Use the `invNorm` function with .95 as the argument, since, as your hand-drawn curve should clearly show, the area from $-\infty$ to $Z_{.05}$ is $1 - \alpha = 1 - .05 = .95$.



The test statistic falls into the rejection region, i.e., 2.4 > 1.645, therefore we reject H_0. Yes, there is enough evidence that the average is more than \$160.

Z-test

Blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw cornstarch will have a positive or negative effect on blood glucose levels. A sample of 30 patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect.

Step 1: State the null hypothesis: H_0:μ=100

Step 2: State the alternate hypothesis: H_1:≠100

Step 3: State your alpha level. We'll use 0.05 for this example. As this is a two-tailed test, split the alpha into two. 0.05/2=0.025

Step 4: Find the z-score associated with your alpha level. You're looking for the area in *one tail only*. A z-score for 0.75(1-0.025=0.975) is 1.96. As this is a two-tailed test, you would also be considering the left tail (z=1.96)

Step 5: Find the test statistic using this formula:

$$Z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

z=(140-100)/(15/√30)=14.60.

Step 6: If Step 5 is less than -1.96 or greater than 1.96 (Step 3), reject the null hypothesis. In this case, it is greater, so you *can* reject the null.

Paired Z-test!

Formula: $$z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two samples, Δ is the hypothesized difference between the population means (0 if testing for equal means), σ_1 and σ_2 are the standard deviations of the two populations, and n_1 and n_2 are the sizes of the two samples.

The amount of a certain trace element in blood is known to vary with a standard deviation of 14.1 ppm (parts per million) for male blood donors and 9.5 ppm for female donors. Random samples of 75 male and 50 female donors yield concentration means of 28 and 33 ppm, respectively. What is the likelihood that the population means of concentrations of the element are the same for men and women?

Null hypothesis: $H_0: \mu_1 = \mu_2$

or $H_0: \mu_1 - \mu_2 = 0$

alternative hypothesis: $H_a: \mu_1 \neq \mu_2$

or: $H_a: \mu_1 - \mu_2 \neq 0$ $$z = \frac{28 - 33 - 0}{\sqrt{\frac{14.1^2}{75} + \frac{9.5^2}{50}}} = \frac{-5}{\sqrt{2.65 + 1.81}} = -2.37$$

The computed *z*-value is negative because the (larger) mean for females was subtracted from the (smaller) mean for males. But because the hypothesized difference between the populations is 0, the order of the samples in this computation is arbitrary—$\bar{x}_1$ could just as well have been the female sample mean and $\bar{x}_2$ the male sample mean, in which case *z* would be 2.37 instead of −2.37. An extreme *z*-score in either tail of the distribution (plus or minus) will lead to rejection of the null hypothesis of no difference.

The area of the standard normal curve corresponding to a *z*-score of −2.37 is 0.0089. Because this test is two-tailed, that figure is doubled to yield a probability of 0.0178 that the population means are the same. If the test had been conducted at a pre-specified significance level of $\alpha < 0.05$, the null hypothesis of equal means could be rejected. If the specified significance level had been the more conservative (more stringent) $\alpha < 0.01$, however, the null hypothesis could not be rejected.

In practice, the two-sample *z*-test is not used often, because the two population standard deviations σ_1 and σ_2 are usually unknown. Instead, sample standard deviations and the *t*-distribution are used.

T-test [qt(c(0.05), df=19)]

Consider the following exercise.
A manufacturer claims that the average capacity of a certain type of battery the company produces is at least 140 ampere-hours. An independent consumer protection agency wishes to test the credibility of the manufacturer's claim and measures the capacity of 20 batteries from a recently produced batch. The results, in ampere-hours, are as follows:

```
137.4  140.0  138.8  139.1  144.4  139.2  141.8  137.3  133.5  138.2
141.1  139.7  136.7  136.3  135.6  138.0  140.9  140.6  136.7  134.1
```

Using the .05 level of significance, is there enough evidence that the manufacturer's claim is being overstated?
The claim is that the average capacity is at least 140, and we will try to show that it is in fact less than 140. Thus the hypotheses here will be

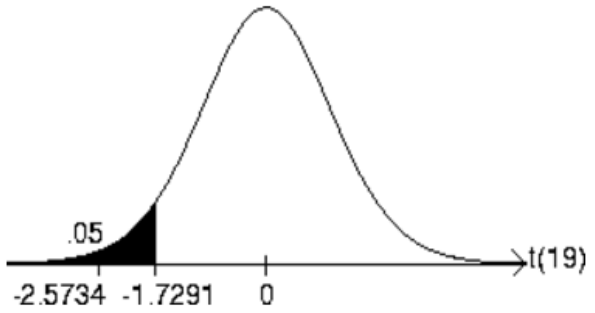
$$H_0 : \mu \geq 140$$

$$H_1 : \mu < 140$$

Now let us get the test statistic:

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{138.47 - 140}{2.6589/\sqrt{20}} = -2.5734$$

Set up the rejection region by drawing a *t*-curve and shade the leftmost 5% of the left tail.



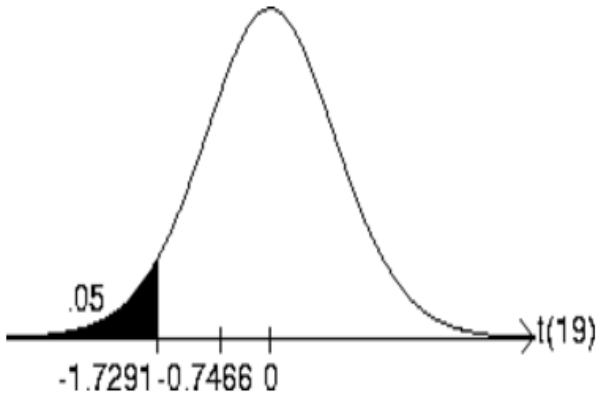
We need the *t* critical value associated with this, which is $-t_{\alpha}^{(n-1)} = -t_{.05}^{(19)} = -1.7291$. We get this *t* critical value with the EQUATION SOLVER, just like we did for the last problem.

T-test

What is your answer in (a) if the last two values are 146.7 and 144.1 instead of 136.7 and 134.1?

Obtain the summary statistics again. The mean and standard deviation will be different. The test statistic changes to:

$$t = \frac{\overline{X}-\mu}{s/\sqrt{n}} = \frac{139.47-140}{3.1749/\sqrt{20}} = -0.7466$$



The rejection region stays the same, so we see that this time, the test statistic does *not* fall into the rejection region, i.e., $-0.7466 \not< -1.7291$. Therefore, we do not reject H_0.

For the *p*-value, shade the area under the *t*-curve to the left of the test statistic, -0.7466.

2- sample T test

Example:

Consider the lifespan of 18 rats. 12 were fed a restricted calorie diet and lived an average of 700 days (standard deviation=21 days). The other 6 had unrestricted access to food and lived an average of 668 days (standard deviation=30 days). Does a restricted calorie diet increase the lifespan of rats (assume α=0.05)?

μ_1=700, s_1=21, n_1=12; μ_2=668, s_2=30, n_2=6

H_0: $\mu_1=\mu_2$

H_a: $\mu_1>\mu_2$ (because we are only asking if a restricted calorie diet increases lifespan)

We cannot assume that the variances of the two populations are equal because the different diets could also affect the variability in lifespan.

The t-statistic is: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)}} = \dfrac{700 - 668}{\sqrt{\dfrac{21^2}{12} + \dfrac{30^2}{6}}} = 2.342$

Degrees of freedom = $(n_1\text{-}1)+(n_2\text{-}1)$ = (12-1)+(6-1)=16

From the t-distribution table, the p-value falls between 0.01 and 0.02, so we do reject H_0. The restricted calorie diet does increase the lifespan of rats.

Paired T-test

Student	Pre-module score	Post-module score	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

Calculating the mean and standard deviation of the differences gives:

$\bar{d} = 2.05$ and $s_d = 2.837$. Therefore, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{2.837}{\sqrt{20}} = 0.634$

So, we have:

$$t = \frac{2.05}{0.634} = 3.231 \qquad \text{on 19 df}$$

Looking this up in tables gives $p = 0.004$. Therefore, there is strong evidence that, on average, the module does lead to improvements.

F-TEST: Below find study hours of 6 female and 5 male students [var.test(x, y) in R]

H15

	A	B	C
1	Female	Male	
2	26	23	
3	25	30	
4	43	18	
5	34	25	
6	18	28	
7	52		
8			
9			

981

as

Q: test the null hypothesis that the variances of two populations are equal

Result:

Important: be sure that the variance of Variable 1 is higher than the variance of Variable 2. This is the case, 160 > 21.7. If not, swap your data. As a result, Excel calculates the correct F value, which is the ratio of Variance 1 to Variance 2 (F = 160 / 21.7 = 7.373).

Conclusion: if F > F Critical one-tail, we reject the null hypothesis. This is the case, 7.373 > 6.256. Therefore, we reject the null hypothesis. The variances of the two populations are unequal.

whether
these two sample have same population variance or not

Mann-Whitney-Wilcoxon Test

we can decide whether the population distributions are identical *without* assuming them to follow the [normal distribution]

In particular, the gas mileage data for manual and automatic transmissions are independent.

Problem

Without assuming the data to have normal distribution, decide at .05 significance level if the gas mileage data of manual and automatic transmissions in mtcars have identical data distribution.

Solution

The null hypothesis is that the gas mileage data of manual and automatic transmissions are identical populations. To test the hypothesis, we apply the wilcox.test function to compare the independent samples. As the p-value turns out to be 0.001817, and is less than the .05 significance level, we reject the null hypothesis.

```
> wilcox.test(mpg ~ am, data=mtcars)

        Wilcoxon rank sum test with continuity correction

data:  mpg by am
W = 42, p-value = 0.001871
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(21.4, 18.7, 18.1, 14.3, 24.4, 22.8,  :
  cannot compute exact p-value with ties
```

Answer

At .05 significance level, we conclude that the gas mileage data of manual and automatic transmissions in mtcar are *nonidentical* populations.

Wilcoxon Signed-Rank Test

The Wilcoxon two-sample paired signed rank test is used to test the null hypothesis that the population median of the paired differences of the two samples is 0.

```
> library(MASS)          # load the MASS package
> head(immer)
  Loc Var    Y1    Y2
1  UF   M  81.0  80.7
2  UF   S 105.4  82.3
    .....
```

Problem

Without assuming the data to have normal distribution, test at .05 significance level if the barley yields of 1931 and 1932 in data set immer have identical data distributions.

Solution

The null hypothesis is that the barley yields of the two sample years are identical populations. To test the hypothesis, we apply the wilcox.test function to compare the matched samples. For the paired test, we set the "paired" argument as TRUE. As the p-value turns out to be 0.005318, and is less than the .05 significance level, we reject the null hypothesis.

```
> wilcox.test(immer$Y1, immer$Y2, paired=TRUE)

        Wilcoxon signed rank test with continuity correction

data:  immer$Y1 and immer$Y2
V = 368.5, p-value = 0.005318
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(immer$Y1, immer$Y2, paired = TRUE) :
  cannot compute exact p-value with ties
```

Answer

At .05 significance level, we conclude that the barley yields of 1931 and 1932 from the data set immer are *nonidentical* populations.

Kruskal wallis

A collection of data samples are **independent** if they come from unrelated populations and the samples do not affect each other. Using the **Kruskal-Wallis Test**, we can decide whether the population distributions are identical *without* assuming them to follow the normal distribution.

Example

In the built-in data set named **airquality**, the daily air quality measurements in New York, May to September 1973, are recorded. The ozone density are presented in the data frame column Ozone.

```
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
    .....
```

Problem

Without assuming the data to have normal distribution, test at .05 significance level if the monthly ozone density in New York has identical data distributions from May to September 1973.

Solution

The null hypothesis is that the monthly ozone density are identical populations. To test the hypothesis, we apply the kruskal.test function to compare the independent monthly data. The p-value turns out to be nearly zero (6.901e-06). Hence we reject the null hypothesis.

```
> kruskal.test(Ozone ~ Month, data = airquality)

        Kruskal-Wallis rank sum test

data:  Ozone by Month
Kruskal-Wallis chi-squared = 29.267, df = 4, p-value = 6.901e-06
```

Answer

At .05 significance level, we conclude that the monthly ozone density in New York from May to September 1973 are *nonidentical* populations.

Anova single factor in excel

Below you can find the salaries of people who have a degree in economics, medicine or history.

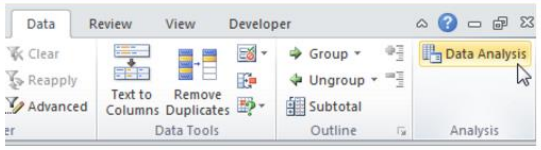
H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: at least one of the means is different.

H18 f_x

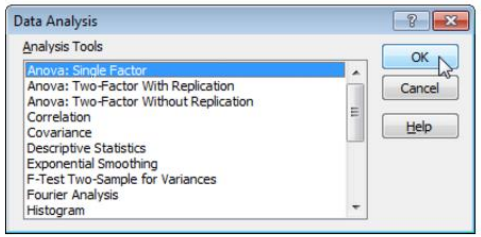
	A	B	C	D
1	economics	medicine	history	
2	42	69	35	
3	53	54	40	
4	49	58	53	
5	53	64	42	
6	43	64	50	
7	44	55	39	
8	45	56	55	
9	52		39	
10	54		40	
11				
12				

Anova excel



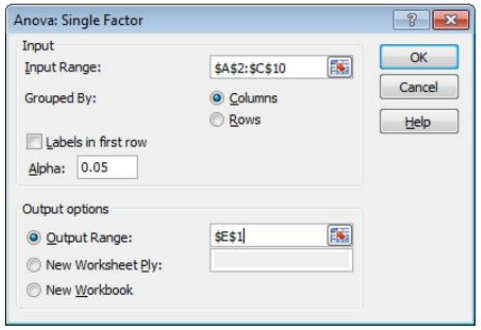
Note: can't find the Data Analysis button? Click here to load the Analysis ToolPak add-in.

2. Select Anova: Single Factor and click OK.



3. Click in the Input Range box and select the range A2:C10.

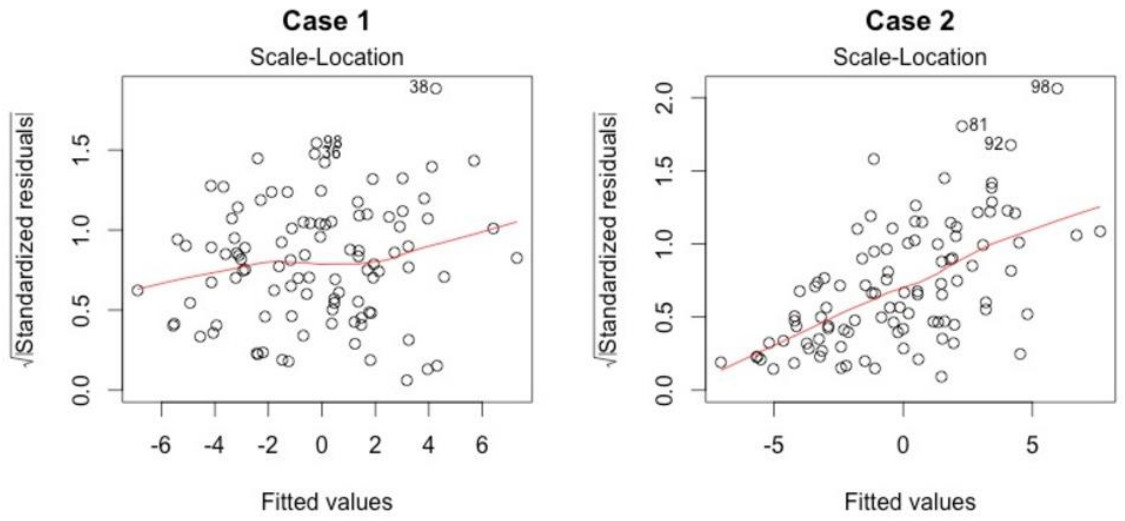
4. Click in the Output Range box and select cell E1.



Scale location

3. Scale-Location

It's also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points.



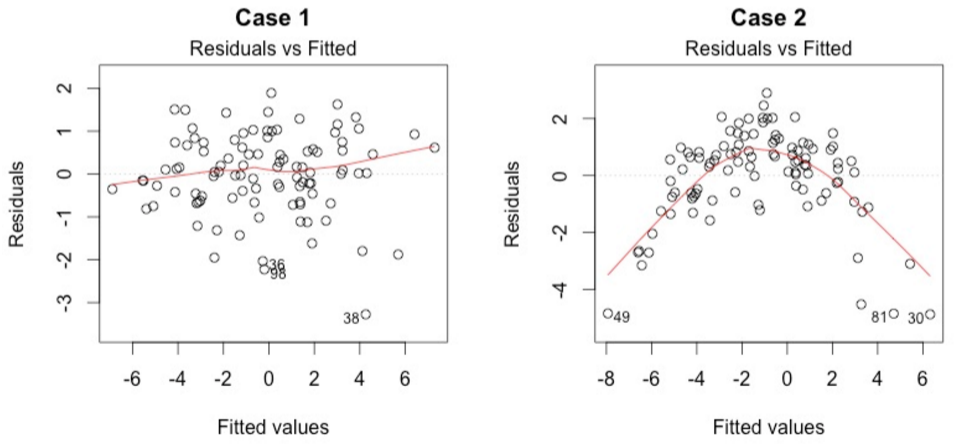
What do you think? In Case 1, the residuals appear randomly spread. Whereas, in Case 2, the residuals begin to spread wider along the x-axis as it passes around 5. Because the residuals spread wider and wider, the red smooth line is not horizontal and shows a steep angle in Case 2.

Residual vs fitted

1. Residuals vs Fitted

This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships.

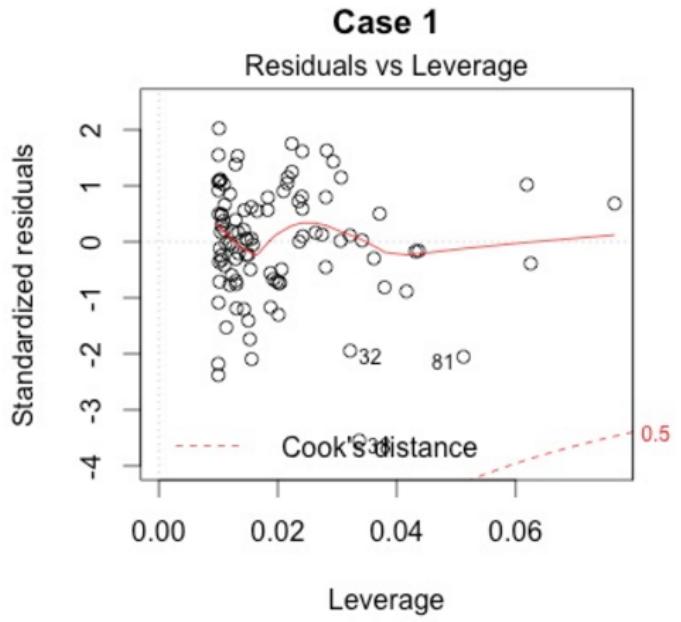
Let's look at residual plots from a 'good' model and a 'bad' model. The good model data are simulated in a way that meets the regression assumptions very well, while the bad model data are not.

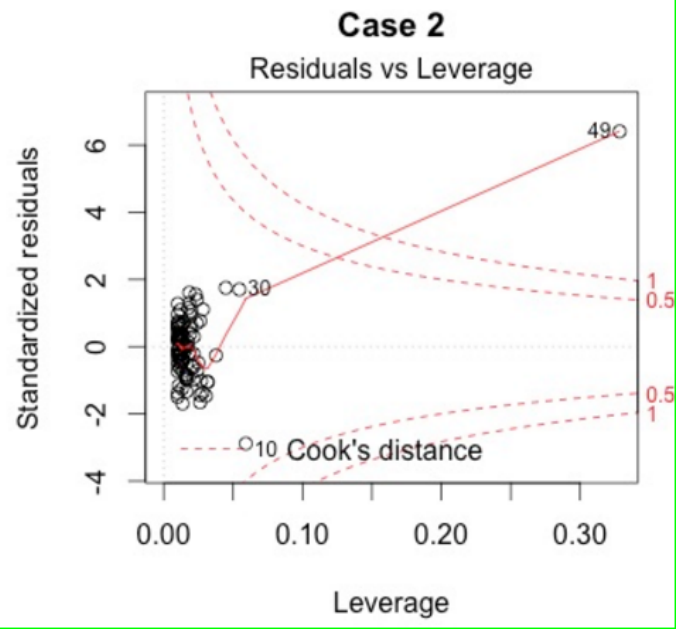


What do you think? Do you see differences between the two cases? I don't see any distinctive pattern in Case 1, but I see a parabola in Case 2, where the non-linear relationship was not explained by the model and was left out in the residuals.

Residual vs leverage

Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right corner o at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look fo cases outside of a dashed line, Cook's distance. When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered we exclude those cases.





2-way Anova – Stress levels

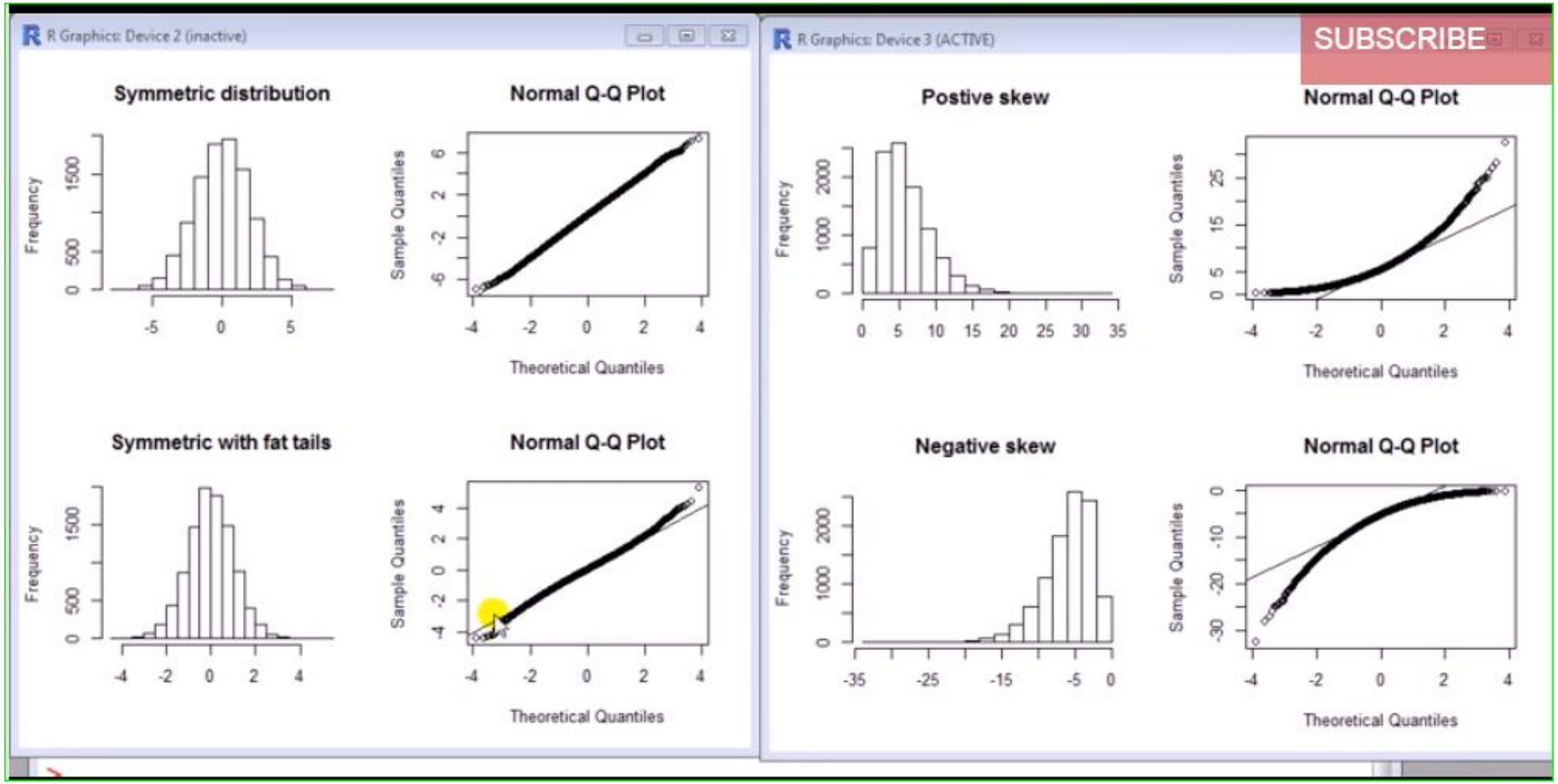
A	B	C	D
		Residence	
		Urban	Rural
	None	9	7
		6	5
		7	7
		8	8
		4	9
		7	8
		8	10
		7	8
		6	7
		7	8
Physical Activity Level	Consistent	1	3
		4	5
		5	7
		4	3
		3	3
		6	3
		5	3
		6	4
		7	6
		8	2

Independent groups of individuals, that are not being measured more than once. Hence 2 way Anova without replication.

1 vs 2-way Anova

BASIS FOR COMPARISON	ONE WAY ANOVA	TWO WAY ANOVA
Meaning	One way ANOVA is a hypothesis test, used to test the equality of three of more population means simultaneously using variance.	Two way ANOVA is a statistical technique wherein, the interaction between factors, influencing variable can be studied.
Independent Variable	One	Two
Compares	Three or more levels of one factor.	Effect of multiple level of two factors.
Number of Observation	Need not to be same in each group.	Need to be equal in each group.
Design of experiments	Need to satisfy only two principles.	All three principles needs to be satisfied.

QQ plot


R Graphics: Device 2 (inactive)
Symmetric distribution
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Symmetric with fat tails
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
R Graphics: Device 3 (ACTIVE)
SUBSCRIBE
Postive skew
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Negative skew
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

Association rule

Association Rule

An implication expression of the form X ® Y, where X and Y are item sets

Example: {Milk, Diaper} -> {Beer}

Here X is {Milk, Diaper] -> Y which is {Beer}

TID	Items
1	Chips, Milk
2	Chips, Diaper, Beer, Cornflakes
3	Milk, Diaper, Beer, Pepsi
4	Chips, Milk, Diaper, Beer
5	Chips, Milk, Diaper, pepsi

Support (s) = Fraction of transactions that contain both X and Y i.e. how often Milk, Diaper and Beer occur together in the transactions. Milk, Diaper and Beer occur in 2 out of total 5 transactions, hence support =2/5=0.4

Confidence (c) = Measures how often each item in Y appears in transactions that contain X

That is- How often beer occurs in the transactions which contain milk and diaper. Now milk and diaper are together in 3 transactions (TID=3, 4 and 5), and out of the 3, beer is present in 2 of them, hence confidence = 2/3 (No. of transactions with Milk, Diaper and Beer/No. of transactions with Milk and Beer) =0.67

Lift: The Lift of the rule is X=>Y is the confidence of the rule divided by the expected confidence, assuming that the item sets are independent.

Coming back to our Example-> Lift (X->Y) = confidence(X->Y) / support(Y)

=Support (X+Y)/Support (X)*Support (Y)

= 0.67 / (3/5)=0.67/0.60 = 1.1167

Now, Let us do a bit of Math here-> ((0.67-0.60)/0.60)*100=70/6=11.67 i.e. probability of finding beer in the transactions which have Milk and Diaper is greater than the normal probability of finding Beer in the above 5 transactions by 11.67%.

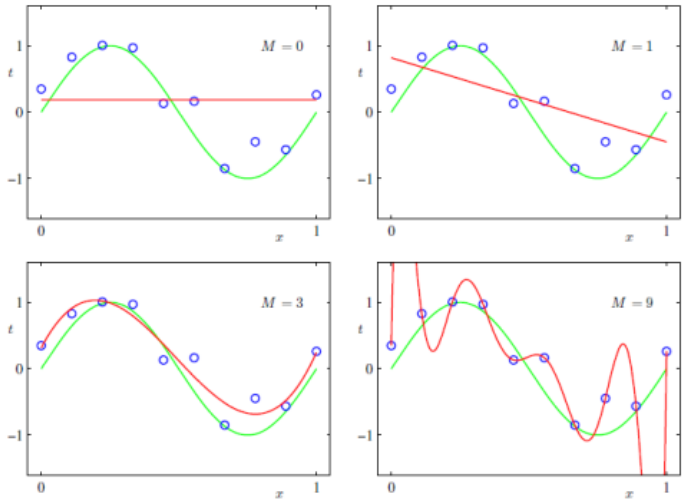
Triple trade off

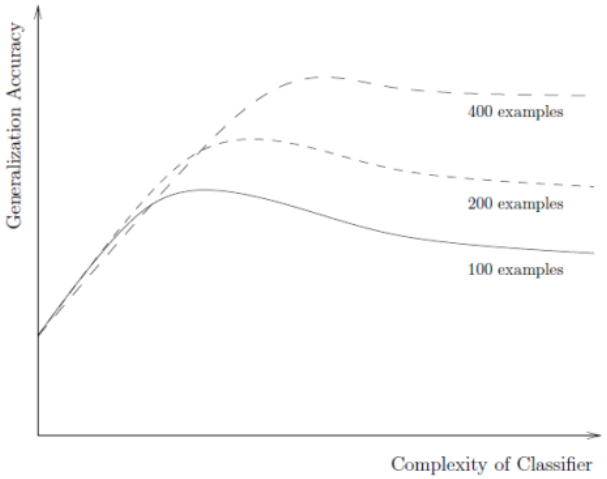
Diettrich 2003 ("Machine Learning") describes the **triple tradeoff** for empirical (supervised) learning. That is, there exists a tradeoff between:

1) The size or complexity of the learned classifier

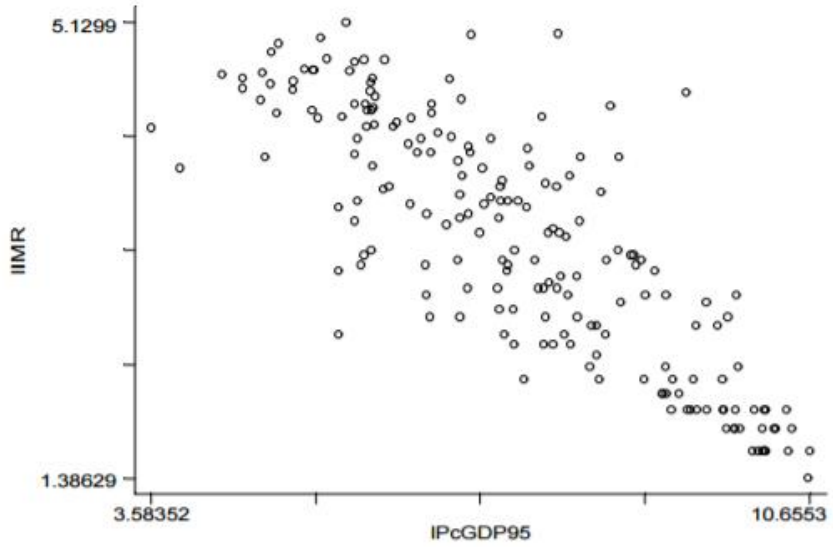
2) The amount of training data

3) The generalization accuracy on new examples





Log log




```
. regress lIMR lPcGDP95
  Source |       SS       df       MS              Number of obs =     194
---------+------------------------------           F(  1,   192) =  404.52
   Model |  131.035233     1  131.035233            Prob > F      =  0.0000
Residual |  62.1945021   192  .323929698           R-squared     =  0.6781
---------+------------------------------           Adj R-squared =  0.6765
   Total |  193.229735   193  1.00119034           Root MSE      =  .56915

------------------------------------------------------------------------------
    lIMR |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
lPcGDP95 |  -.4984531   .0247831   -20.113   0.000    -.5473352    -.449571
   _cons |   7.088676   .1908519    37.142   0.000     6.71224    7.465111
------------------------------------------------------------------------------
```

Linear vs Logistics

Problems with the Linear Approach

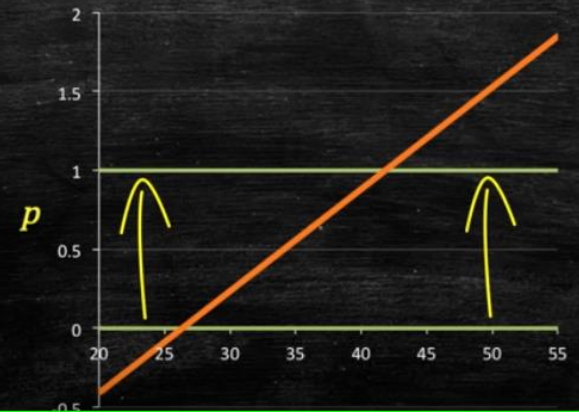
- Probabilities are bounded whereby $0 \leq p \leq 1$.
- The range of age in the data is such that $20 \leq age \leq 55$.
- The probability that a 35 year-old person subscribes is:

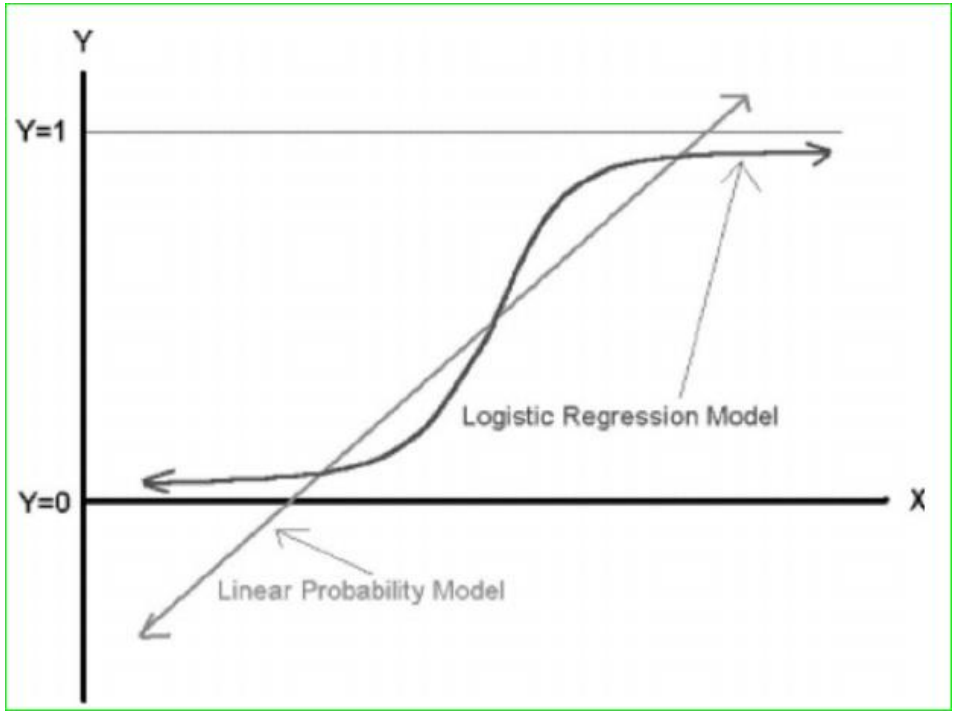
$$p = -1.700 + 0.064 \times 35 = 0.54$$

- What about people with 25 and 45 years of age?

$$p = -1.700 + 0.064 \times 25 = -0.09$$

Linear Model Plot





Linear vs Logistics

The Three Regression Types

a short guide

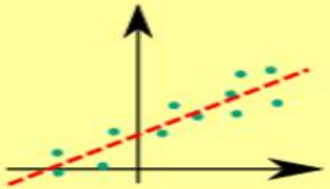
Generalized Linear Models (GLM) extend the ordinary linear regression and allow the response variable y to have an error distribution other than the normal distribution.

GLMs are:

a) Easy to understand
b) Simple to fit and interpret in any statistical package
c) Sufficient in a lot of practical applications

LINEAR REGRESSION

❶ Econometric modelling
❷ Marketing Mix Model
❸ Customer Lifetime Value



Continuous ⇒ Continuous

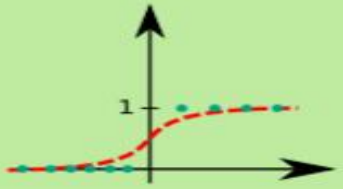
$$y = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$$

lm(y ~ x1 + x2, data)

1 unit increase in x increases y by α

LOGISTIC REGRESSION

❶ Customer Choice Model
❷ Click-through Rate
❸ Conversion Rate
❹ Credit Scoring



Continuous ⇒ True/False

$$y = \frac{1}{1 + e^{-z}}$$

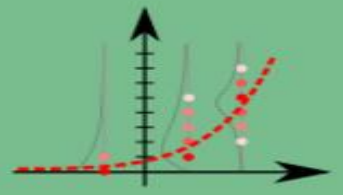
$$z = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$$

glm(y ~ x1 + x2, data, family=binomial( ))

1 unit increase in x increases log odds by α

POISSON REGRESSION

❶ Number of orders in lifetime
❷ Number of visits per user



Continuous ⇒ 0,1,2,...

$$y \sim Poisson(\lambda)$$

$$ln\lambda = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$$

glm(y ~ x1 + x2, data, family=poisson( ))

1 unit increase in x multiplies y by e^{α}

Linear vs Logistic

MARSIAN Technologies LLP EXPERIENCE PERFECTION www.marsiantech.com

Computational method

- Multiple regression uses the least-squares method to find the coefficients for the independent variables in the regression equation, i.e. it computed coefficients that minimized the residuals for all cases.
- Logistic regression uses maximum-likelihood estimation to compute the coefficients for the logistic regression equation. This method finds attempts to find coefficients that match the breakdown of cases on the dependent variable.

MARSIAN Technologies LLP EXPERIENCE PERFECTION www.marsiantech.com

Computational method…

- The overall measure of how will the model fits is given by the likelihood value, which is similar to the residual or error sum of squares value for multiple regression.

RSS

MARSIAN Technologies LLP EXPERIENCE PERFECTION www.marsiantech.com

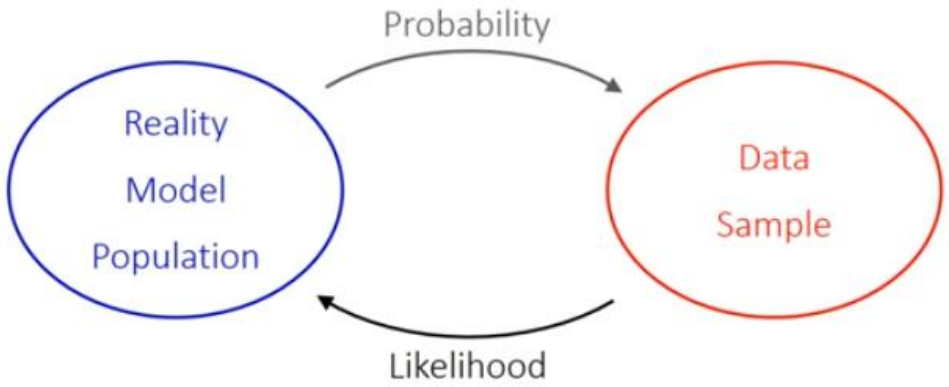
Assessing Predictors: The Wald Statistic

$$Wald\_stat = \frac{b}{SE_b}$$

- Similar to *t*-statistic in Regression.
- Tests the null hypothesis that $b = 0$.
- Is biased when b is large.
- Better to look at Likelihood-ratio statistics.

Max log likely hood

What is Likelihood?



Likelihood: given observed data, what is the chance that a given reality or model is true?

If you count 32% quartz grains, what is the chance that the true proportion is 0.2?

If you observe *x*, what is the best normal distribution (μ, σ)?

Maximum Likelihood Estimation

A procedure to:

1. Determine best model parameters (*reality*) that fit given data
2. Compare multiple models to determine the best fit to data

What it does:

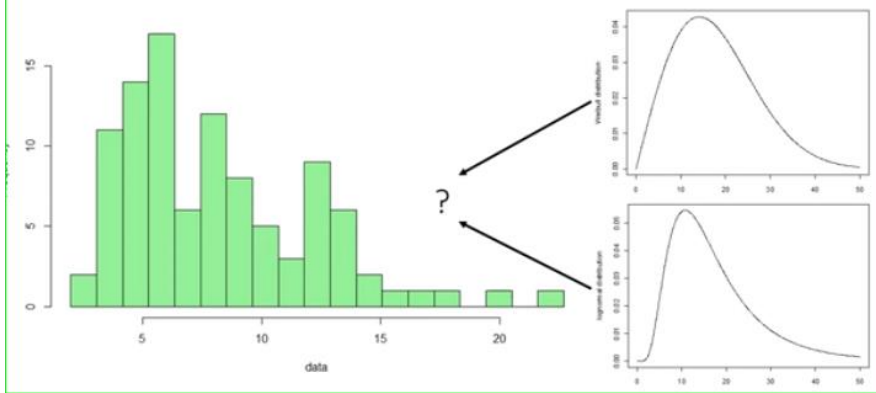
1. Maximizes log-likelihood function to estimate parameters
2. Uses information theory to compare model fits

Finding L

Maximum Likelihood Estimation – Example

What underlying distribution best describes this sample data?

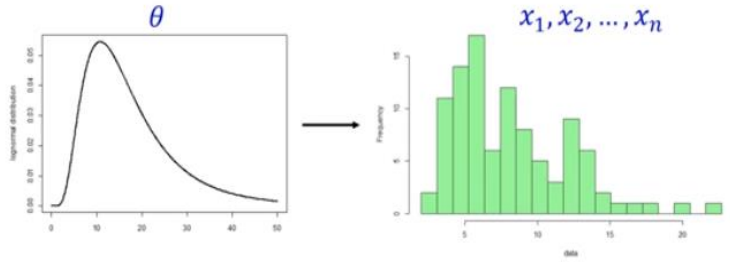
Is it best fit by statistical distribution A or distribution B?



Probability Density Function

Data come from a distribution with probability density function:

$f(x_1, x_2, \ldots, x_n \mid \theta)$ Probability of observing $x_i \ldots x_n$ given parameter(s) θ



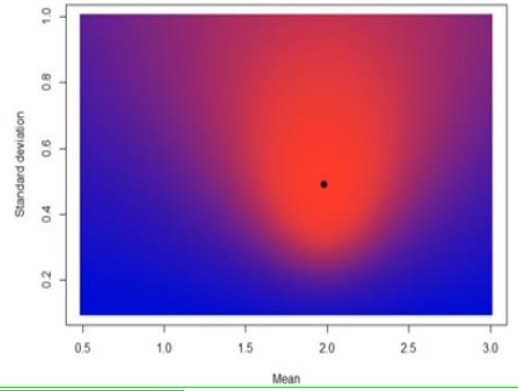
This is the same as:

$f(x_1 \mid \theta) \cdot f(x_2 \mid \theta) \cdot \cdots \cdot f(x_n \mid \theta)$ or $\prod f(x_i \mid \theta)$

Log-Likelihood Function

Maximization of the likelihood function is difficult in practice, so the method maximizes log-likelihood instead

$$L(\theta \mid x_i) = \prod f(x_i \mid \theta) \longrightarrow \ln L(\theta \mid x_i) = \sum \ln f(x_i \mid \theta)$$



Kullback-Leibler Information

Can use information theory techniques to quantify the distance between models f and g (each is a probability density function)

Model f

$$I(f,g) = \int f(x) ln \frac{f(x)}{g(x|\theta)} dx$$

Information lost when g used to approximate f

Model g over parameter space θ

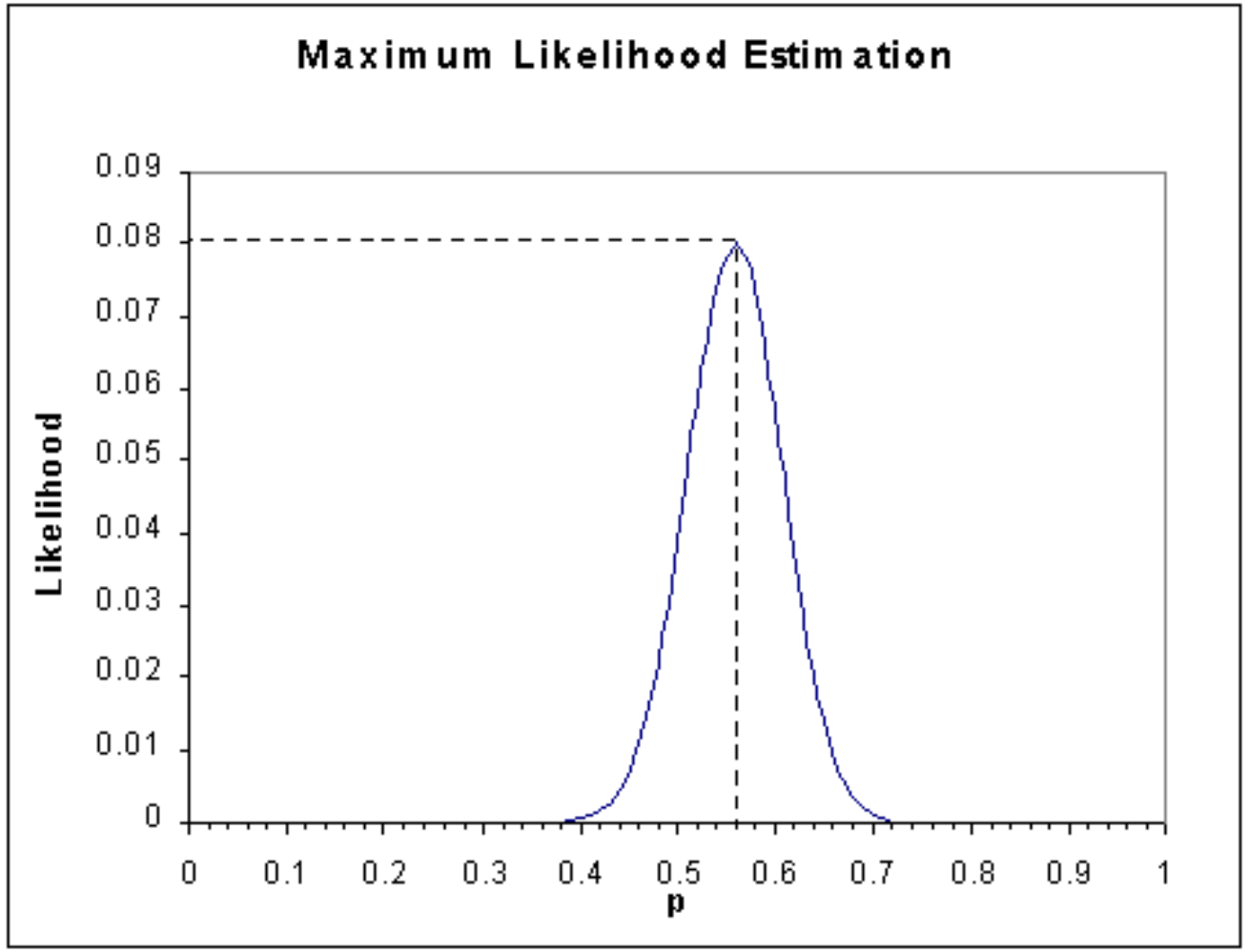
Model Selection with AIC

Hirotugu Akaike showed that K-L information could be estimated based on the maximum log-likelihood and created "an information criterion" (AIC) – later *Akaike information criterion*

$$\text{AIC} = -2 \ln\left[L(\hat{\theta}|x)\right] + 2K$$

Adding parameters K increases model fit (maximum log-likelihood)

Tossing a coin, n = 100, h = 56



Say we toss a coin 100 times and observe 56 heads and 44 tails. Instead of *assuming* that p is 0.5, we want to find the MLE for p. Then we want to ask whether or not this value differs significantly from 0.50.

How do we do this? We find the value for p that makes the observed data most likely.

As mentioned, the observed data are now fixed. They will be constants that are plugged into our binomial probability model :-

- n = 100 (total number of tosses)
- h = 56 (total number of heads)

Imagine that p was 0.5. Plugging this value into our probability model as follows :-

$$L(p = 0.5 \mid data) = \frac{100!}{56!44!} 0.5^{56} 0.5^{44} = 0.0389$$

But what if p was 0.52 instead?

$$L(p = 0.52 \mid data) = \frac{100!}{56!44!} 0.52^{56} 0.48^{44} = 0.0581$$

So from this we can conclude that p is more likely to be 0.52 than 0.5. We can tabulate the likelihood for different parameter values to find the maximum likelihood estimate of p:

p	L
0.48	0.0222
0.50	0.0389
0.52	0.0581
0.54	0.0739
0.56	0.0801
0.58	0.0738
0.60	0.0576
0.62	0.0378

Warnings AIC

Warnings about Model Selection and AIC

- Likelihood values (-2logL) are only relevant **in comparison** to other models with the same data.
- AIC also only relevant **in comparison** to other models with the same data (actual value of AIC is unimportant).
- AIC only chooses the best **of the candidate models** – it may not be a good model in an absolute sense if they are all bad!
- **Don't combine** model selection with hypothesis testing. The p value significance will be inflated because you are implicitly testing multiple hypotheses with model selection.

Over dispersion: It makes standard error too small

Overdispersion is an important concept in the analysis of discrete data. Many a time data admit more variability than expected under the assumed distribution. The greater variability than predicted by the generalized linear model random component reflects overdispersion. Overdispersion occurs because the mean and variance components of a GLM are related and depends on the same parameter that is being predicted through the independent vector.

Reasons for overdispersion

Overdispersion can be explained by

- variation among the success probabilities or

- correlation between the binary responses

Both reasons are the same, since variation leads to correlation and vice versa. But for interpretative reasons one explanation might be more reasonable than the other.

Decision tree

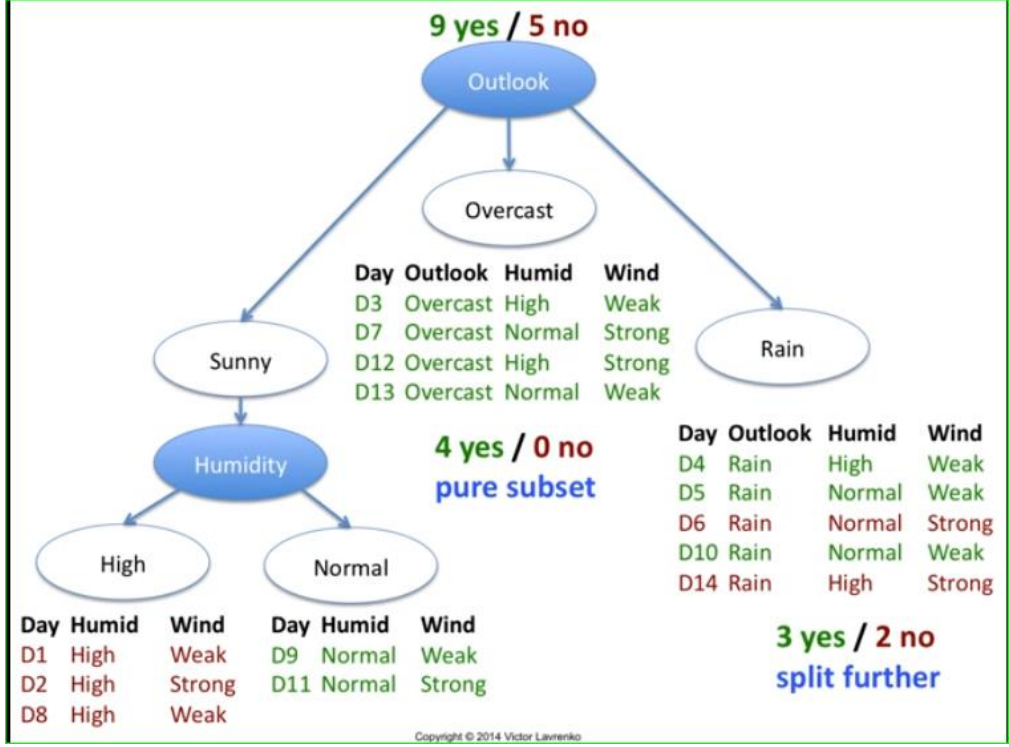
Predict if John will play tennis

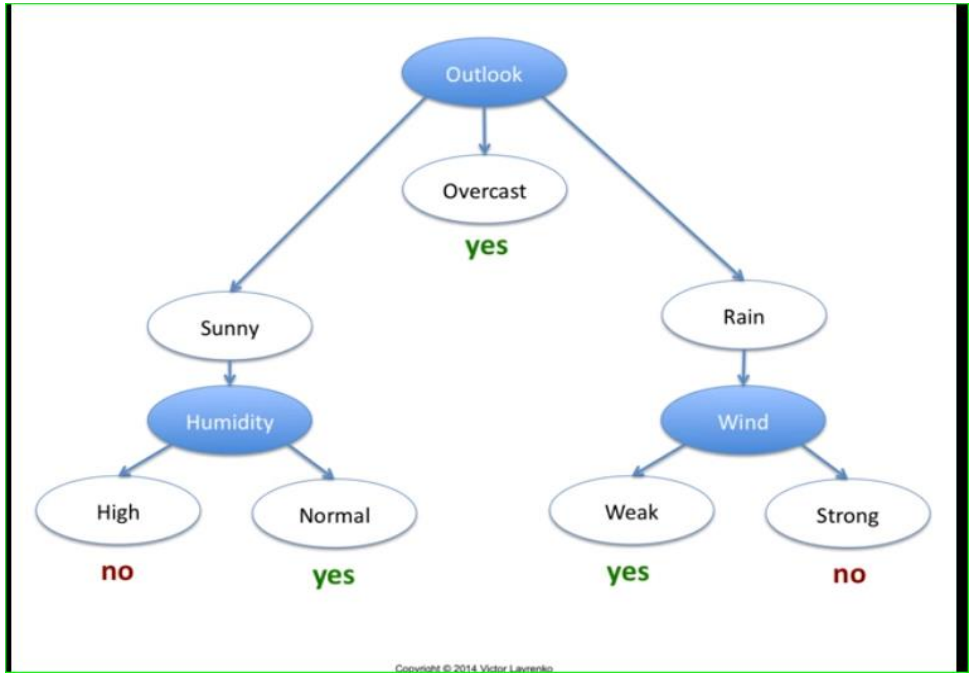
Training examples: **9 yes / 5 no**

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

New data:

D15	Rain	High	Weak	?





Split should be as certain as possible

Which attribute to split on?

- Want to measure "purity" of the split
  - more certain about Yes/No after the split
    - pure set (4 yes / 0 no) => completely certain (100%)
    - impure (3 yes / 3 no) => completely uncertain (50%)
  - can't use P("yes" | set):
    - must be symmetric: 4 yes / 0 no as pure as 0 yes / 4 no

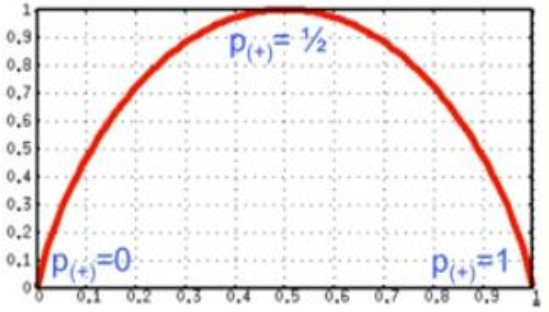
Entropy

- Entropy: $H(S) = - p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$ bits
  - S … subset of training examples
  - $p_{(+)}$ / $p_{(-)}$ … % of positive / negative examples in S
- Interpretation: assume item X belongs to S
  - how many bits need to tell if X positive or negative
- impure (3 yes / 3 no):

$$H(S) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1 \text{ bits}$$

- pure set (4 yes / 0 no):

$$H(S) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0 \text{ bits}$$



Info. Gain

Information Gain

- Want many items in pure sets
- Expected drop in entropy after split:

$$Gain(S,A) = H(S) - \sum_{V \in Values(A)} \frac{|S_V|}{|S|} H(S_V)$$

V … possible values of A
S … set of examples {X}
S_v … subset where $X_A = V$

- Mutual Information
  - between attribute A and class labels of S

Gain (S, Wind)
= H(S) – $^{8}/_{14}$ H(S_{weak}) – $^{6}/_{14}$ H(S_{weak})
= 0.94 – $^{8}/_{14}$ * 0.81 – $^{6}/_{14}$ * 1.0
= 0.049

$-\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$ **9 yes / 5 no**

H(S) = 0.94

6 yes / 2 no

$-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$

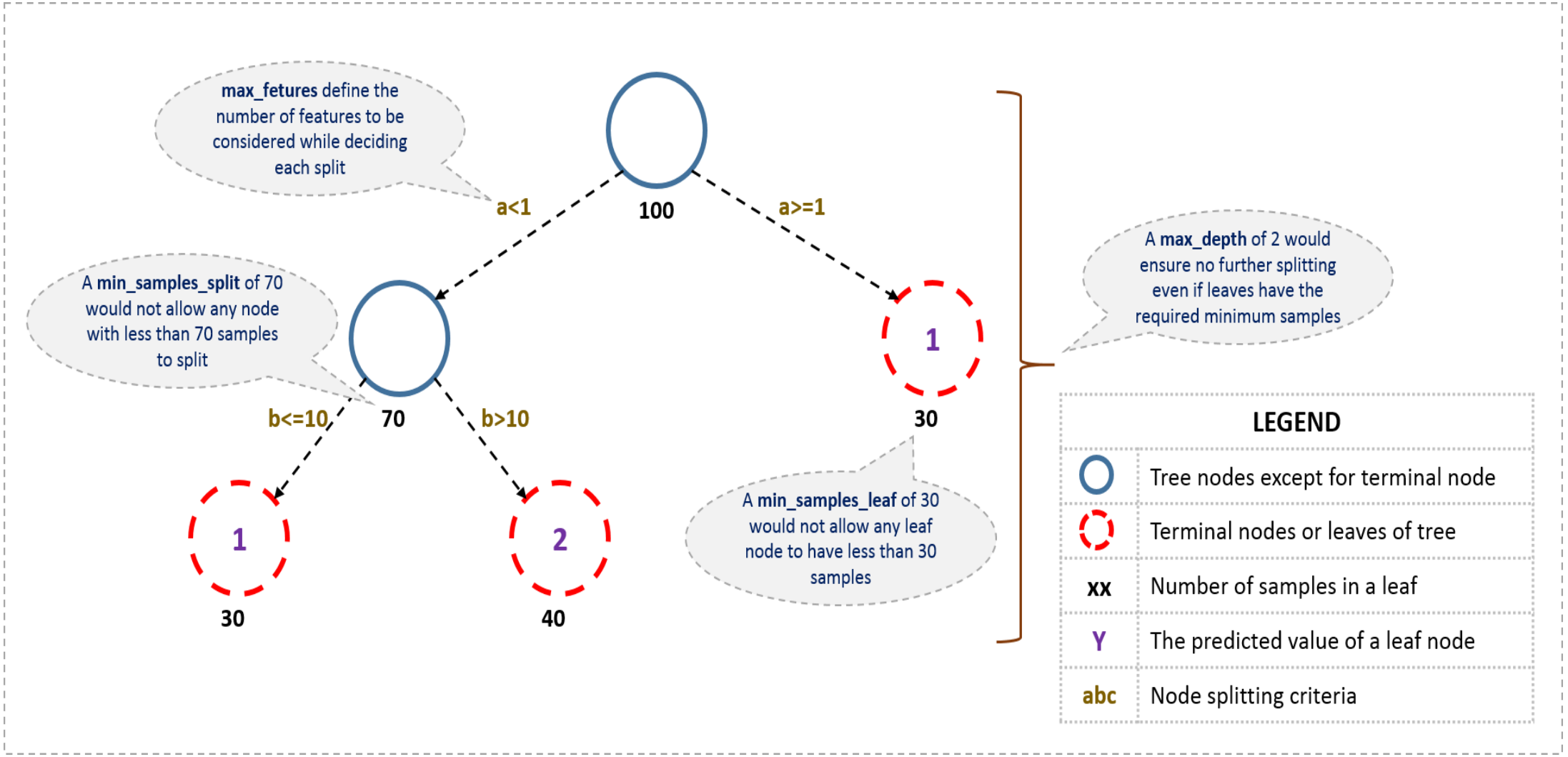
$H(S_{weak}) = 0.81$

3 yes / 3 no

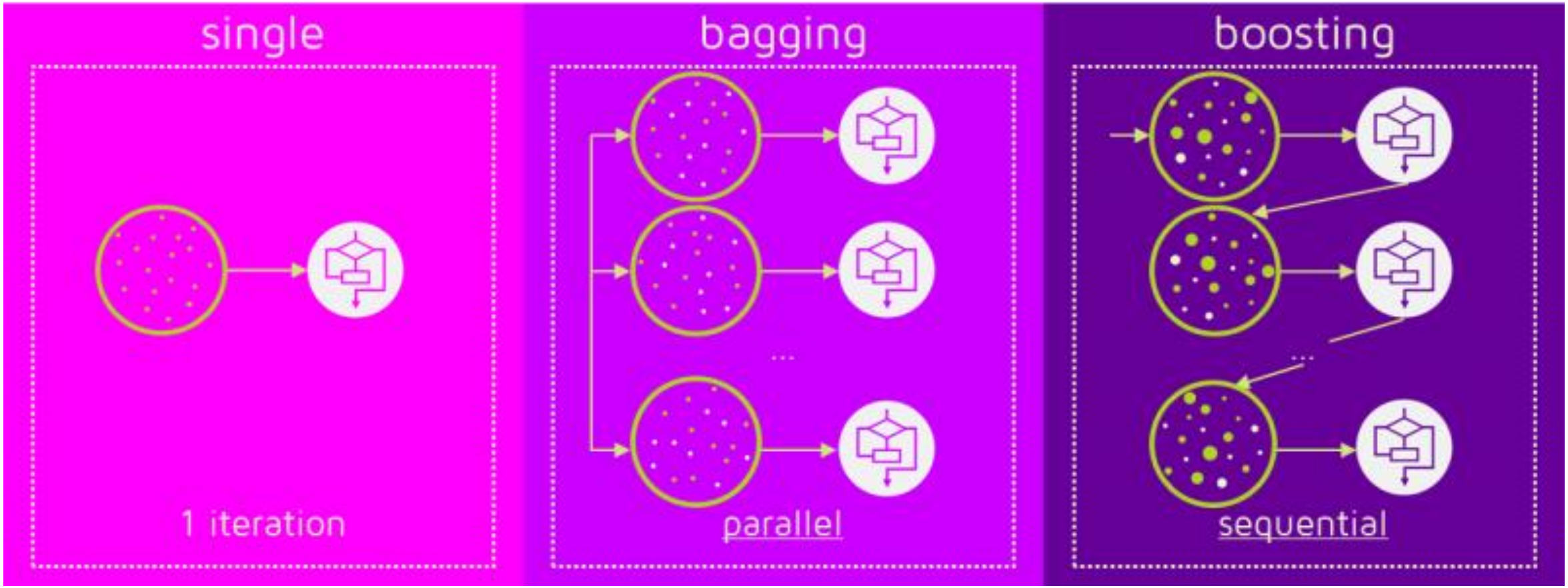
$-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$

$H(S_{strong}) = 1.0$

Setting Constraints on Tree Size



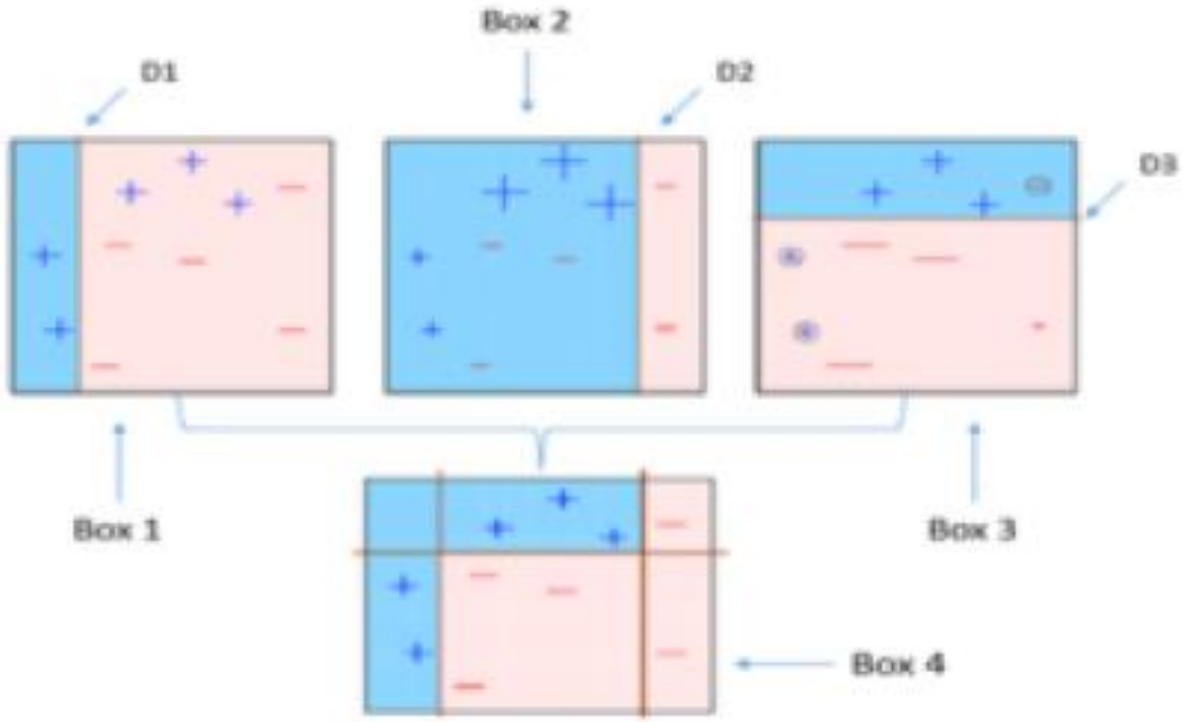
Bagging vs Boosting



Gradient boosting

- Boosting is a sequential technique which works on the principle of **ensemble**.
- It combines a set of **weak learners** and delivers improved prediction accuracy. At any instant t, the model outcomes are weighed based on the outcomes of previous instant t-1.
- The outcomes predicted correctly are given a lower weight and the ones miss-classified are weighted higher.
- This technique is followed for a classification problem while a similar technique is used for regression

Gradient boosting



Assume, you are given a previous model M to improve on. Currently you observe that the model has an accuracy of 80% (any metric). How do you go further about it?

One simple way is to build an entirely different model using new set of input variables and trying better ensemble learners. On the contrary, I have a much simpler way to suggest. It goes like this:

```
Y = M(x) + error
```

What if I am able to see that error is not a white noise but have same correlation with outcome(Y) value. What if we can develop a model on this error term? Like,

```
error = G(x) + error2
```

Probably, you'll see error rate will improve to a higher number, say 84%. Let's take another step and regress against error2.

```
error2 = H(x) + error3
```

Now we combine all these together :

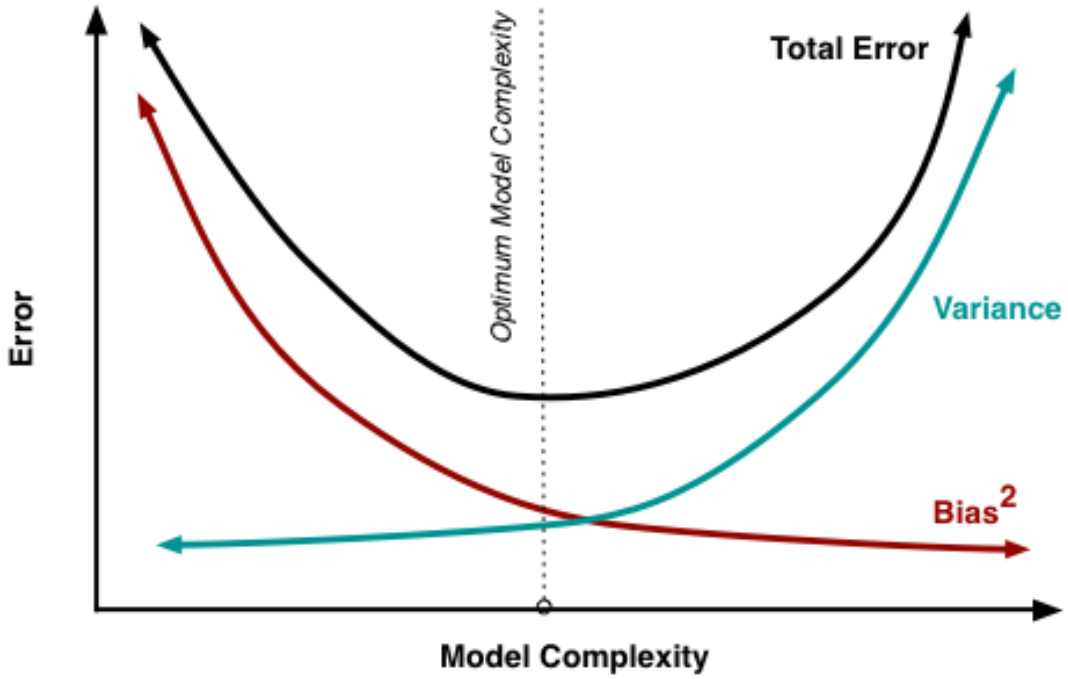
```
Y = M(x) + G(x) + H(x) + error3
```

This probably will have a accuracy of even more than 84%. What if I can find an optimal weights for each of the three learners,

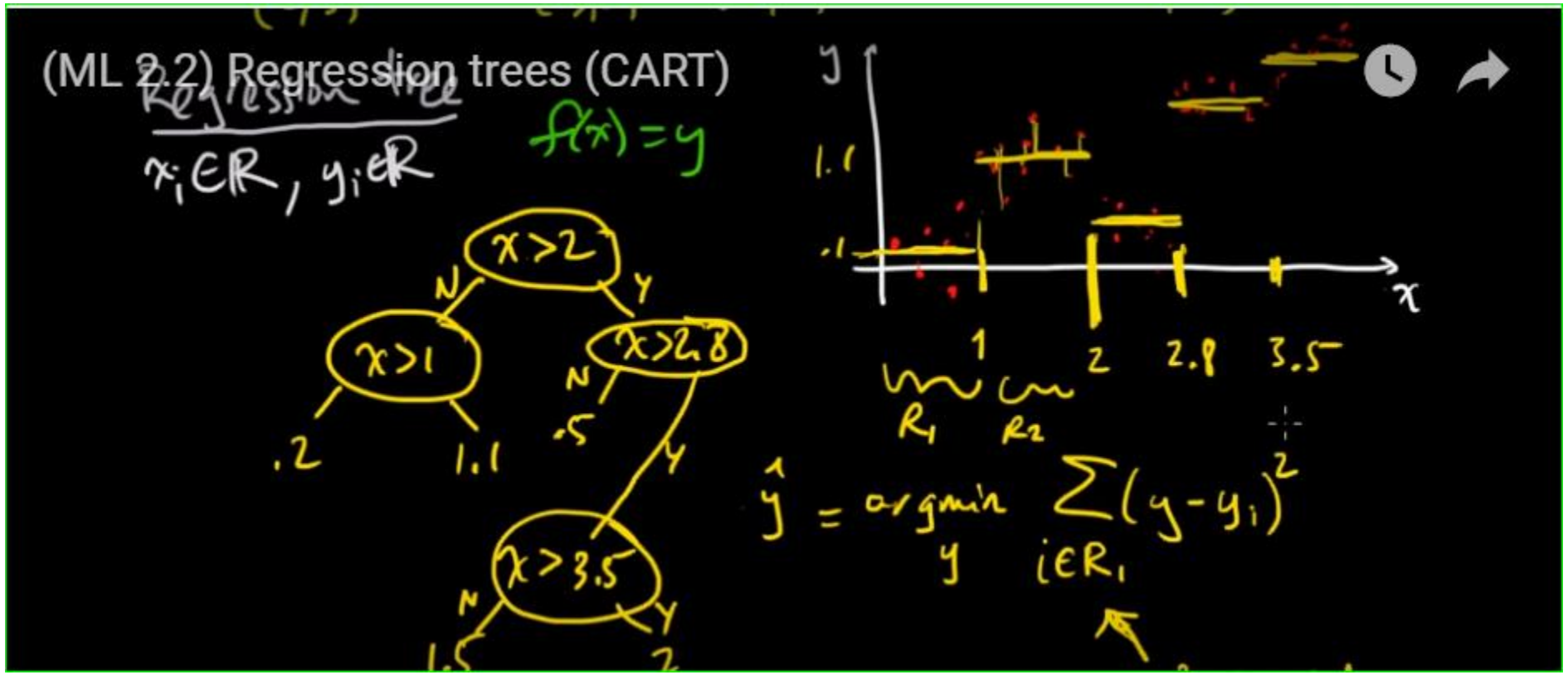
```
Y = alpha * M(x) + beta * G(x) + gamma * H(x) + error4
```

Optimum model complexity

Normally, as you increase the complexity of your model, you will see a reduction in prediction error due to lower bias in the model. As you continue to make your model more complex, you end up over-fitting your model and your model will start suffering from high variance.



CART (regression)



Missing value techniques

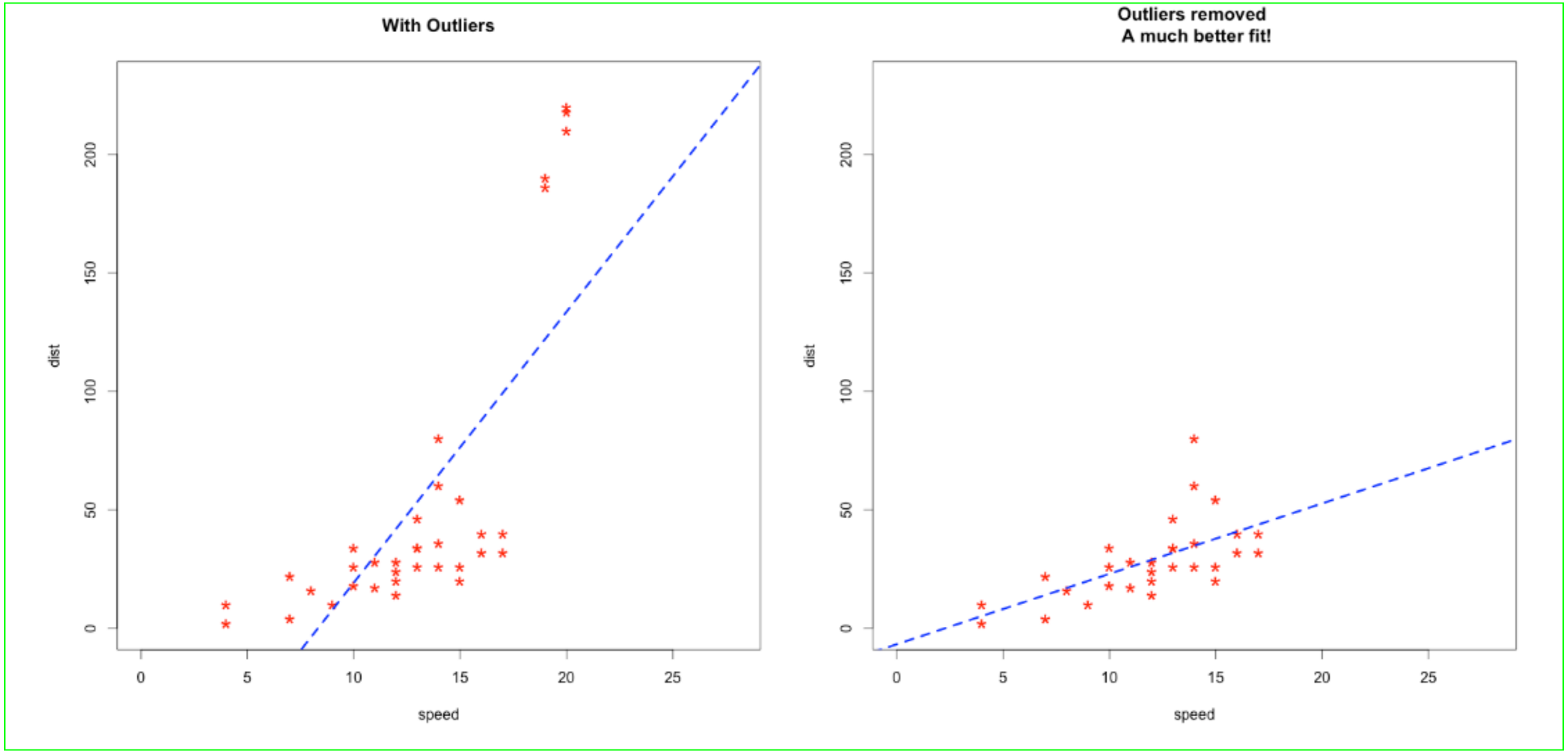
- DELETE observation or variable
  - Have sufficient data points, so the model doesn't lose power.
  - Not to introduce bias (meaning, disproportionate or non-representation of classes)
  - if by removing that one variable you can save many observations, then you are better off without that variable unless it is a really important predictor that makes a lot of business sense

- Hmisc
  - Imputation with mean / median / mode
  - Crude technique

- Knn
  - For every observation to be imputed, it identifies 'k' closest observations based on the euclidean distance and computes the weighted average (weighted based on distance) of these 'k' obs.
  - you could impute all the missing values in all variables with one call
  - be cautious not to include the response variable while imputing

- Mice
  - Multivariate Imputation by Chained Equations
  - 2-step using mice() build a random forest model and complete() to generate completed data

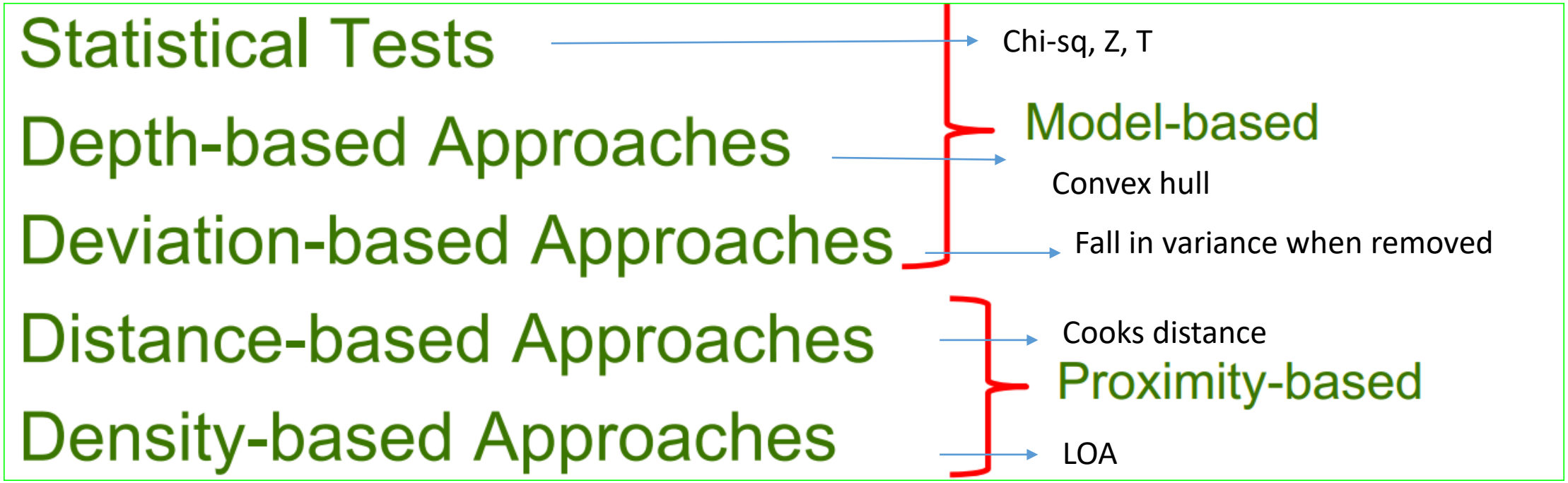
Dramatic impact of outliers



Applying outliers

- Sample applications of outlier detection
  - Fraud detection
    - Purchasing behavior of a credit card owner usually changes when the card is stolen
    - Abnormal buying patterns can characterize credit card abuse
  - Medicine
    - Unusual symptoms or test results may indicate potential health problems of a patient
    - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, …)
  - Public health
    - The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city
    - Whether an occurrence is abnormal depends on different aspects like frequency, spatial correlation, etc.

Outlier identification types



Treatment of outliers

1. Imputation

- Imputation with mean / median / mode.

2. Capping

- For missing values that lie outside the 1.5 * IQR limits, we could cap it by replacing those observations outside the lower limit with the value of 5th %ile and those that lie above the upper limit, with the value of 95th %ile.

3. Prediction

- outliers can be replaced with missing values (NA) and then can be predicted by considering them as a response variable.

Posterior probability

Suppose there is a mixed school having 60% boys and 40% girls as students. The girls wear trousers or skirts in equal numbers; the boys all wear trousers. An observer sees a (random) student from a distance; all the observer can see is that this student is wearing trousers. What is the probability this student is a girl? The correct answer can be computed using Bayes' theorem.

The event G is that the student observed is a girl, and the event T is that the student observed is wearing trousers. To compute the posterior probability $P(G|T)$, we first need to know:

- $P(G)$, or the probability that the student is a girl regardless of any other information. Since the observer sees a random student, meaning that all students have the same probability of being observed, and the percentage of girls among the students is 40%, this probability equals 0.4.
- $P(B)$, or the probability that the student is not a girl (i.e. a boy) regardless of any other information (B is the complementary event to G). This is 60%, or 0.6.
- $P(T|G)$, or the probability of the student wearing trousers given that the student is a girl. As they are as likely to wear skirts as trousers, this is 0.5.
- $P(T|B)$, or the probability of the student wearing trousers given that the student is a boy. This is given as 1.
- $P(T)$, or the probability of a (randomly selected) student wearing trousers regardless of any other information. Since $P(T) = P(T|G)P(G) + P(T|B)P(B)$ (via the law of total probability), this is $P(T) = 0.5 \times 0.4 + 1 \times 0.6 = 0.8$.

Given all this information, the **posterior probability** of the observer having spotted a girl given that the observed student is wearing trousers can be computed by substituting these values in the formula:

$$P(G|T) = \frac{P(T|G)P(G)}{P(T)} = \frac{0.5 \times 0.4}{0.8} = 0.25.$$

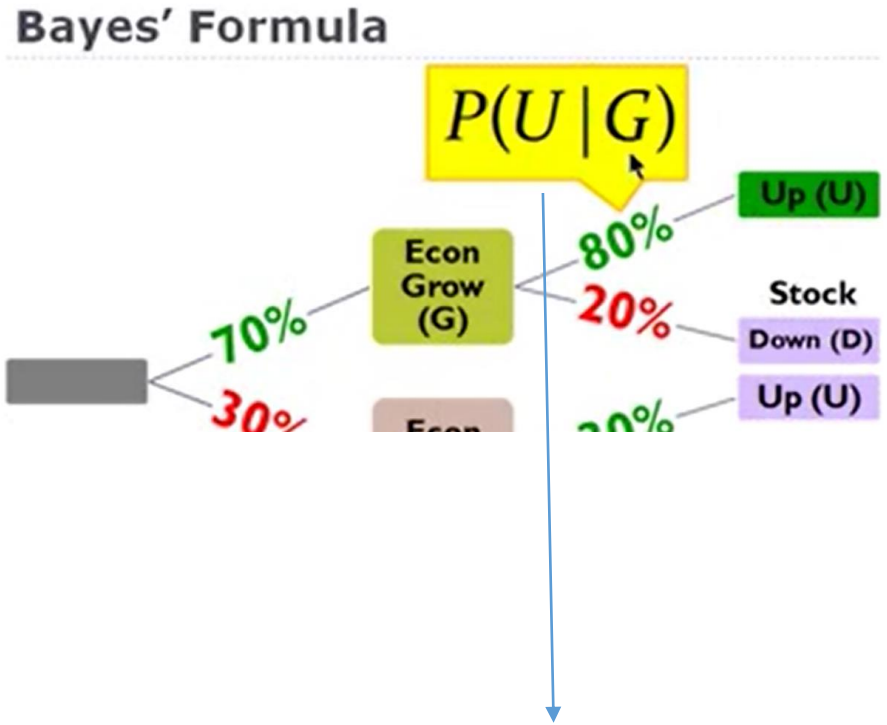
Bayes' Formula

Updates a probability given new information

$$P(G|U) = \frac{P(U|G)P(G)}{P(U)}$$

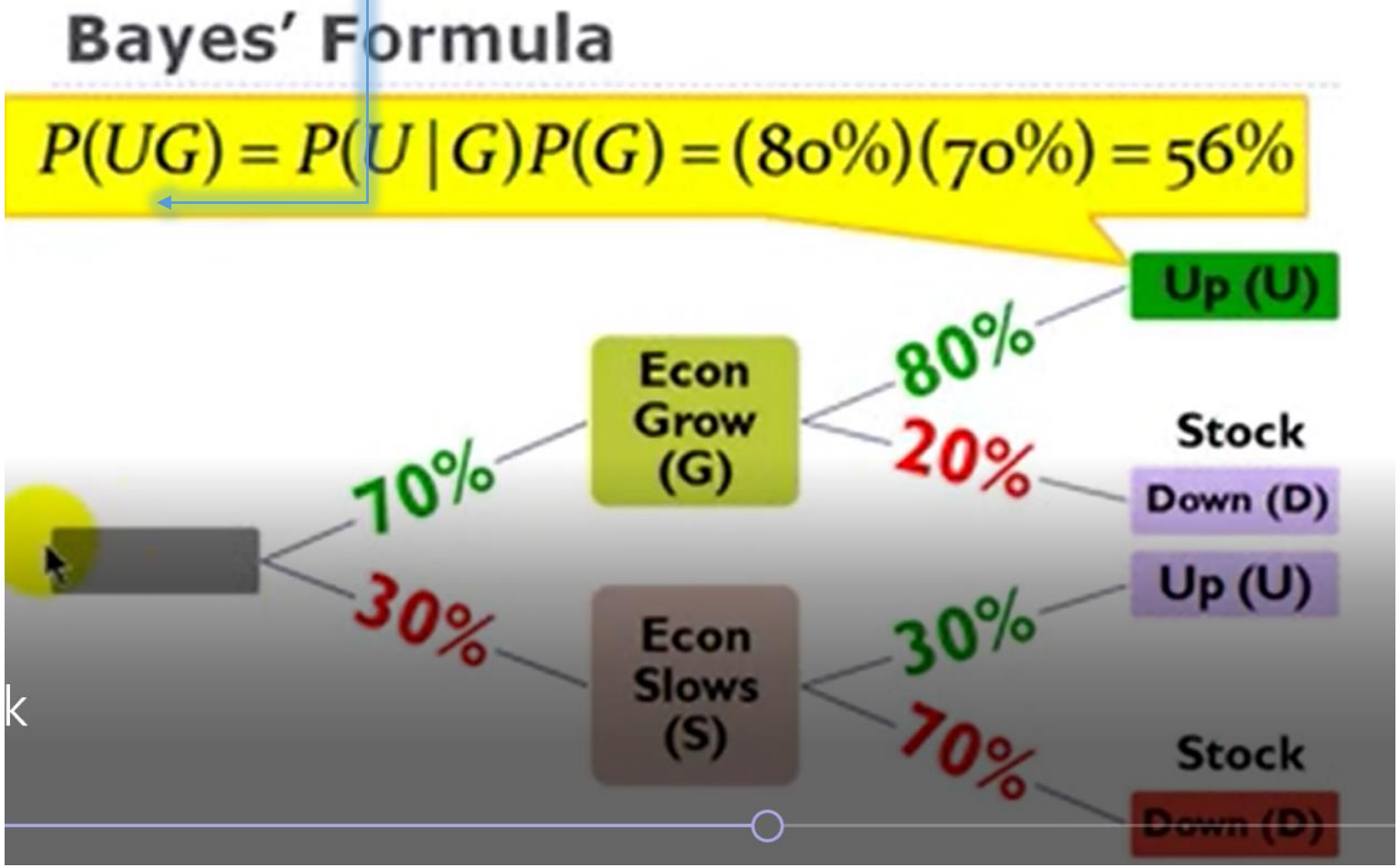
$$P(G|U) = \frac{P(U|G)P(G)}{P(U|G)+P(U|G')}$$

Bayes' Formula



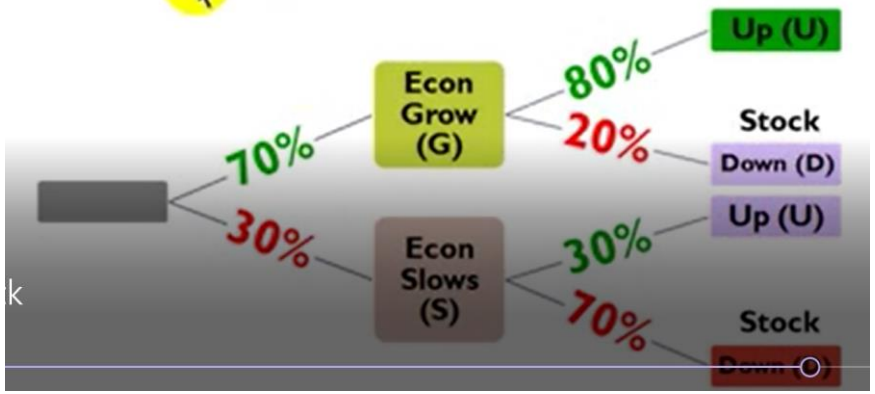
Conditional probability

Joint Probability



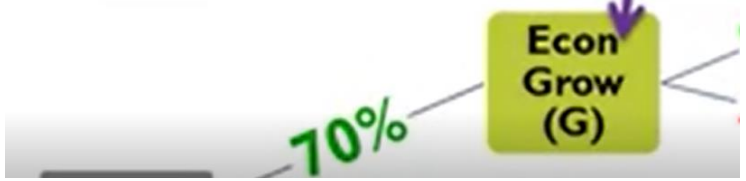
Bayes' Formula

$$P(G|U) = ?$$



Bayes' Formula

$$P(G|U) = \frac{P(U|G)P(G)}{P(U)}$$

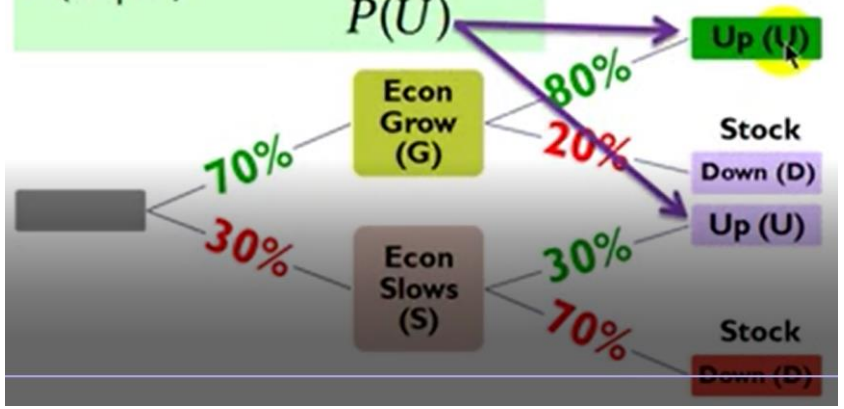


Bayes' Formula

$$P(G|U) = \frac{(80\%)(70\%)}{(80\%)(70\%)+(30\%)(30\%)}$$

Bayes' Formula

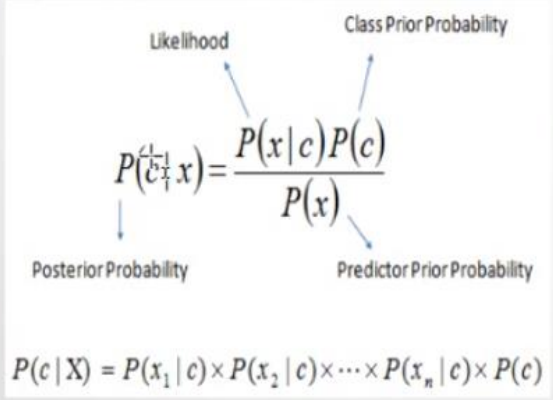
$$P(G|U) = \frac{P(U|G)P(G)}{P(U)}$$



Naïve Bayes

How it works

- Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c)
- NB classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute)
- P(c) is the prior probability of class
- P(x|c) is the likelihood which is the probability of predictor given class
- P(x) is the prior probability of predictor

Which one is the best predictor ?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Example

- The posterior probability can be calculated by first, constructing a frequency table for each attribute against the target
- Then, transforming the freq. tables to likelihood tables and finally using the Naive Bayesian equation to calculate the posterior probability for each class
- The class with the highest posterior probability is the outcome of prediction

$P(x|Yes) = P(Sunny|Yes) = 3/9 = 0.33$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(Yes) = 9/14 = 0.64$

Posterior Probability: $P(c|x) = P(Yes|Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$

Example

Frequency Tables

		Play Golf	
		Yes	No
Outlook	Sunny	3 2/9	2 2/5
	Overcast	4 4/9	0 0/5
	Rainy	2 2/9	3 3/5

		Play Golf	
		Yes	No
Temp.	Hot	2 2/9	2 2/5
	Mild	4 4/9	2 2/5
	Cool	3 3/9	1 1/5

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

		Play Golf	
		Yes	No
Windy	False	6 6/9	2 2/5
	True	3 3/9	3 3/5

The zero-frequency problem

- When an attribute value (Outlook=Overcast) doesn't occur with every class value (Play Golf=no)
- Add 1 to all the counts

Example

- Let's assume we have a day with:

  Outlook = Rainy

  Temp = Mild

  Humidity = Normal

  Windy = True

  Likelihood of Yes = P(Outlook=Rainy|Yes)*P(Temp=Mild|Yes)*P(Humidity=Normal|Yes)*P(Windy=True|Yes)*P(Yes) =

  2/9 * 4/9 * 6/9 * 3/9 * 9/14 = 0.014109347

  Likelihood of No = P(Outlook=Rainy|No)* P(Temp=Mild|No)*P(Humidity=Normal|No)*P(Windy=True|No)*P(No) =

  3/5 * 2/5 * 1/5 * 3/5 * 5/14 = 0.010285714

- Now we normalize:

  P(Yes) = 0.014109347/(0.014109347+0.010285714) = 0.578368999

  P(No) = 0.010285714/(0.014109347+0.010285714) = 0.421631001

Generative models in the context of Machine Learning [edit]

A generative algorithm models how the data was generated in order to categorize a signal. It asks the question: based on my generation assumptions, which category is most likely to generate this signal? A discriminative algorithm does not care about how the data was generated, it simply categorizes a given signal.

Suppose the input data is x and the set of labels for x is y. A generative model learns the joint probability distribution $p(x, y)$ while a discriminative model learns the conditional probability distribution $p(y|x)$ "probability of y given x".

Let's try to understand this with an example. Consider the following 4 data points: $(x, y) = \{(0, 0), (0, 0), (1, 0), (1, 1)\}$

For above data, $p(x, y)$ will be following:

	$y = 0$	$y = 1$
$x = 0$	$1/2$	0
$x = 1$	$1/4$	$1/4$

2/4

while $p(y|x)$ will be following:

	$y = 0$	$y = 1$
$x = 0$	1	0
$x = 1$	$1/2$	$1/2$

So discriminative algorithms try to learn $p(y|x)$ directly from the data and then try to classify data. On the other hand, generative algorithms try to learn $p(x, y)$ which can be transformed into $p(y|x)$ later to classify the data. One of the advantages of generative algorithms is that you can use $p(x, y)$ to generate new data similar to existing data. On the other hand, discriminative algorithms generally give better performance in classification tasks.

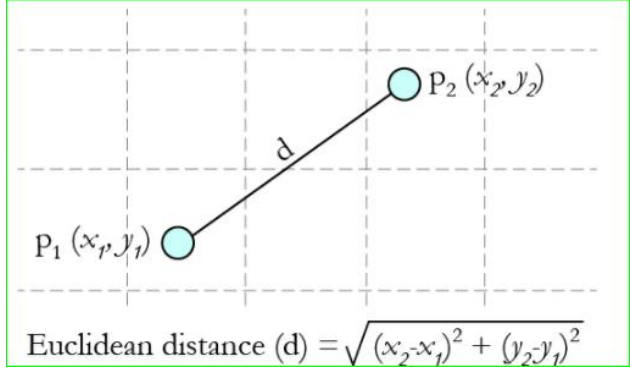
Generative: Naive Bayes, Latent Dirichlet Allocation, Probabilistic Context-Free Grammars, Hidden Markov Models

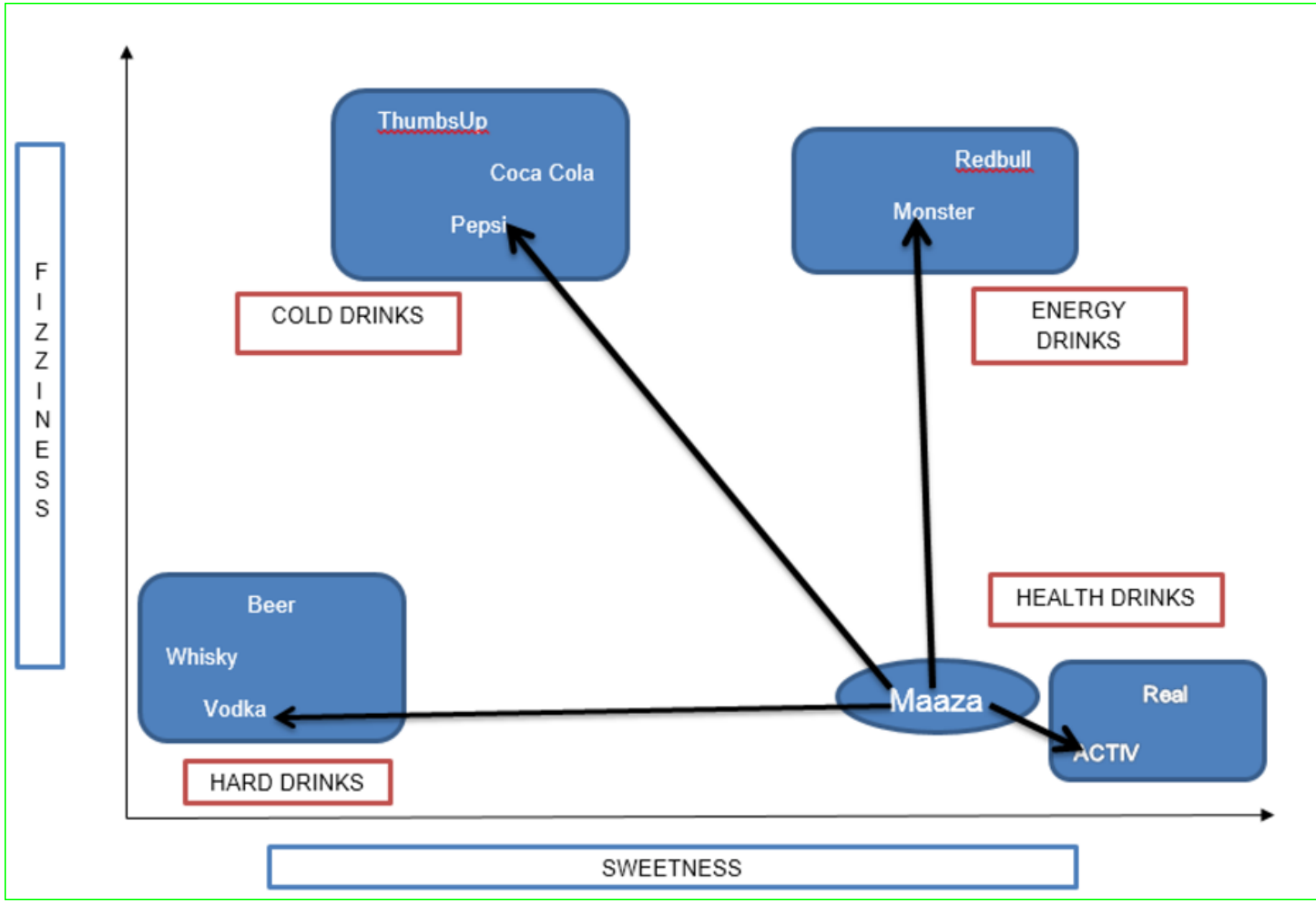
Discriminative: Logistic regression, Support Vector Machines, Maximum Entropy Markov Model, Conditional Random Fields, Neural Networks

KNN Example [Maaza (8,2) ?]

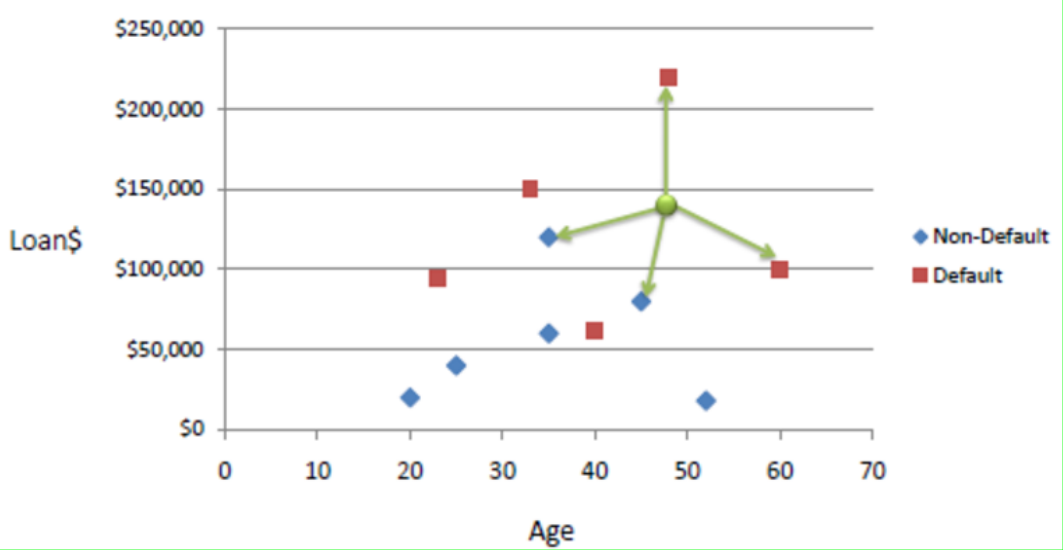
Ingredient	Sweetness	Fizziness	Type of Drink
Monster	8	8	Energy booster
ACTIV	9	1	Health drink
Pepsi	4	8	Cold drink
Vodka	2	1	Hard drink

Ingredient	Sweetness	Fizziness	Type of Drink	Distance to Maaza
Monster	7	8	Energy booster	6.08
ACTIV	9	1	Health drink	1.41
Pepsi	4	8	Cold drink	7.21
Vodka	2	1	Hard drink	6.08





KNN and Normalization



Age	Loan	Default	Distance	
25	$40,000	N	102000	
35	$60,000	N	82000	
45	$80,000	N	62000	
20	$20,000	N	122000	
35	$120,000	N	22000	2
52	$18,000	N	124000	
23	$95,000	Y	47000	
40	$62,000	Y	80000	
60	$100,000	Y	42000	3
48	$220,000	Y	78000	
33	$150,000	Y	8000	1
48	$142,000	?		

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Distance functions

Euclidean

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^{k} |x_i - y_i|$$

Minkowski

$$\left(\sum_{i=1}^{k} \left(|x_i - y_i| \right)^q \right)^{1/q}$$

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	0.61	?	

Standardized Variable

$$X_s = \frac{X - Min}{Max - Min}$$

k Nearest Neighbor Classification

- Strengths
  - Robust
  - Conceptually simple
  - Often works well
  - Powerful (arbitrary decision boundaries)
- Weaknesses
  - Performance is very dependent on the similarity measure used (and to a lesser extent on the number of neighbors k used)
  - Finding a good similarity measure can be difficult
  - Computationally expensive

K means

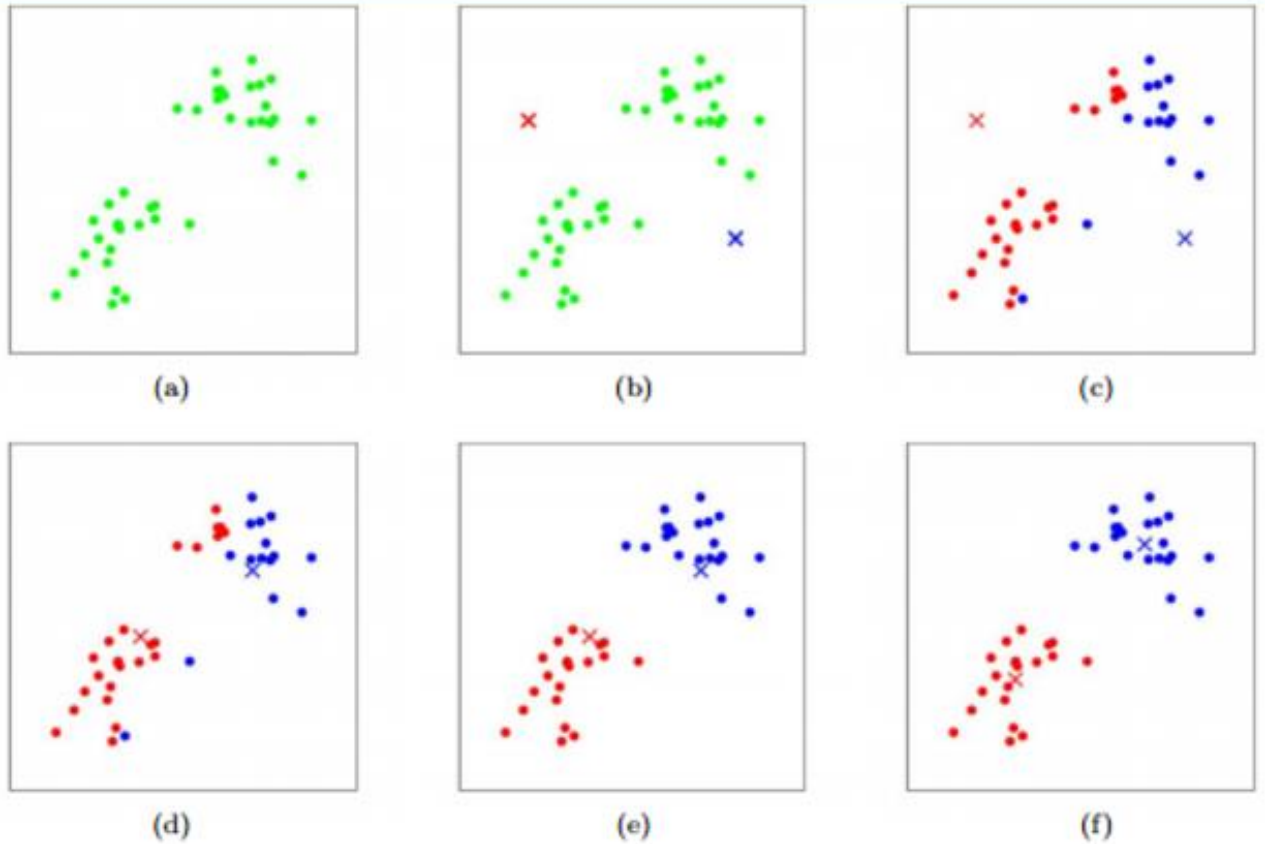


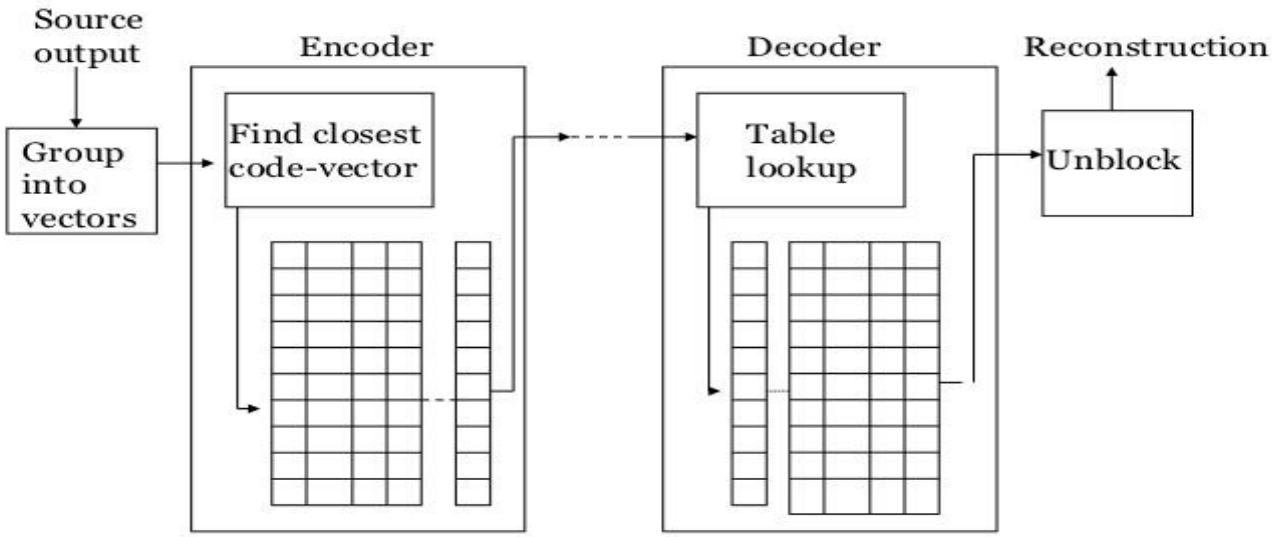
Figure 1: K-means algorithm. Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids. (c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid (shown by "painting" the training examples the same color as the cluster centroid to which is assigned); then we move each cluster centroid to the mean of the points assigned to it. Images courtesy of Michael Jordan.

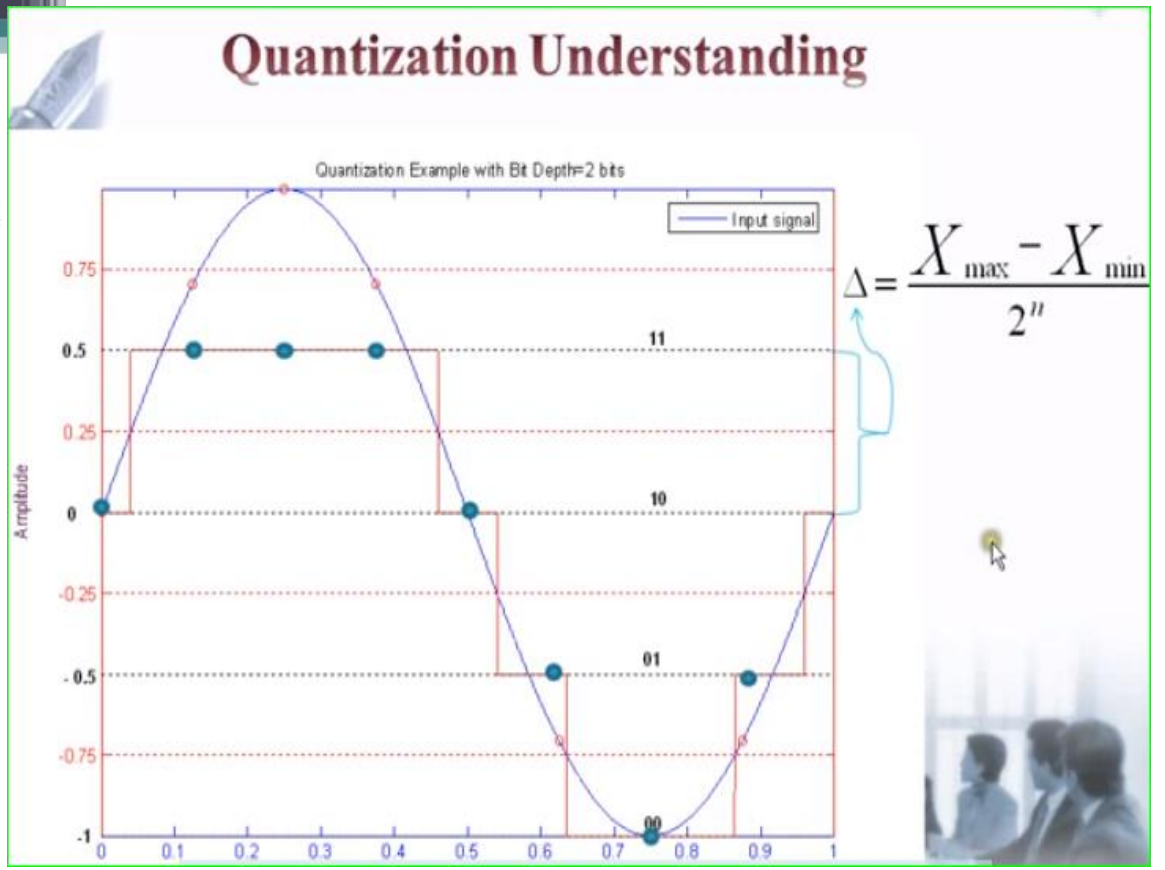
Quantization

The simplest way to quantize a signal is to choose the digital amplitude value closest to the original analog amplitude. This example shows the original analog signal (green), the quantized signal (black dots), the signal reconstructed from the quantized signal (yellow) and the difference between the original signal and the reconstructed signal (red). The difference between the original signal and the reconstructed signal is the quantization error and, in this simple quantization scheme, is a deterministic function of the input signal.

Color quantization

VECOR QUANTIZATION





Vector quantization (VQ) is a classical quantization technique from signal processing that allows the modeling of probability density functions by the distribution of prototype vectors. It was originally used for data compression. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means and some other clustering algorithms.

The density matching property of vector quantization is powerful, especially for identifying the density of large and high-dimensioned data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error. This is why VQ is suitable for lossy data compression. It can also be used for lossy data correction and density estimation.

K-means vs KNN

K-nearest neighbors is a classification algorithm, which is a subset of supervised learning.

K-means is a clustering algorithm, which is a subset of unsupervised learning.

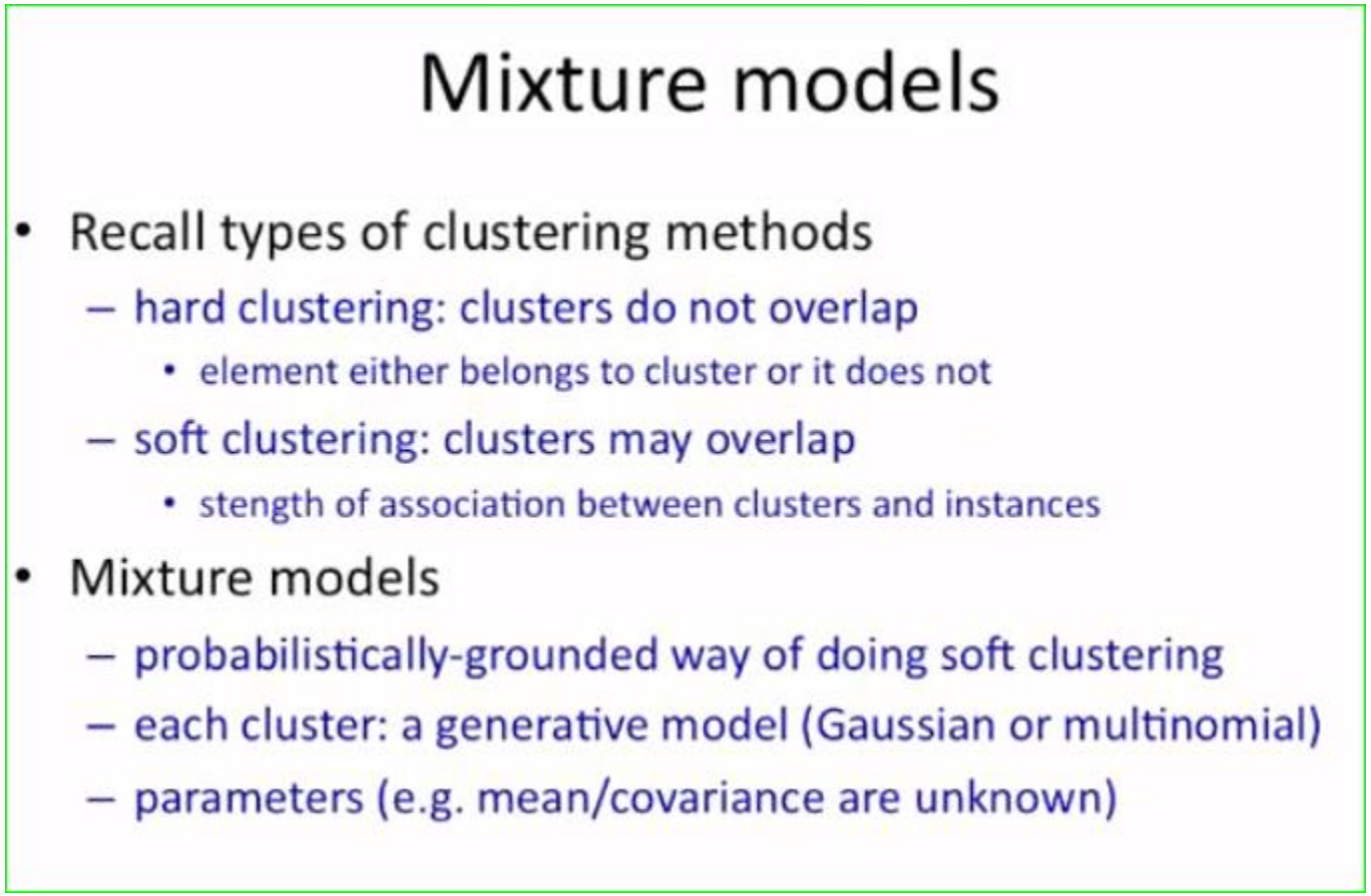
If I have a dataset of basketball players, their positions, and their measurements, and I want to assign positions to basketball players in a new dataset where I have measurements but no positions, I might use **k-nearest neighbors**.

On the other hand, if I have a dataset of basketball players who need to be grouped into k distinct groups based off of similarity, I might use **k-means.**

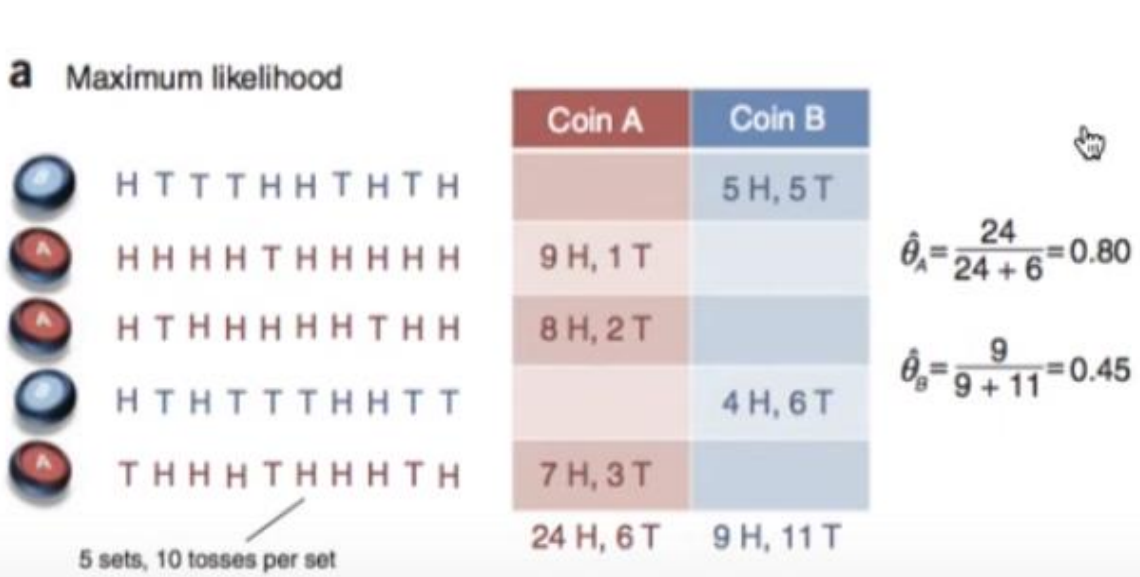
Correspondingly, the K in each case also mean different things! In **k-nearest neighbors**, the k represents the number of neighbors who have a vote in determining a new player's position. Take the example where k =3. If I have a new basketball player who needs a position, I take the 3 basketball players in my dataset with measurements closest to my new basketball player, and I have them vote on the position that I should assign the new player.

The k in **k-means** means the number of clusters I want to have in the end. If k = 5, I will have 5 clusters, or distinct groups, of basketball players after I run the algorithm on my dataset.

Mixture models



EM Bino-dist and likely-hood



One way to think about this is:

1. Assign random averages to both coins
2. For each of the 5 rounds of 10 coin tosses
   - Check the percentage of heads
   - Find the probability of it coming from each coin
   - Compute the expected number of heads: using that probability as a weight, multiply it by the number of heads
   - Record those numbers
   - Re-Compute new means for coin A and B.
3. With these new means go back to step 2.

So, We have:

$\theta_A = 0.6;\ \theta_B = 0.5$

1	H	T	T	T	H	H	T	H	T	H
2	H	H	H	H	H	T	H	H	H	H
3	H	T	H	H	H	H	H	T	H	H
4	H	T	H	T	T	T	H	H	T	T
5	T	H	H	H	T	H	H	H	T	H

Let's take the first round: $\frac{5}{10}$ heads and $\frac{5}{10}$ tails.

likelihood of "A"$= p_A(h)^h(1 - p_A(h))^{10-h} = 0.0007962624$

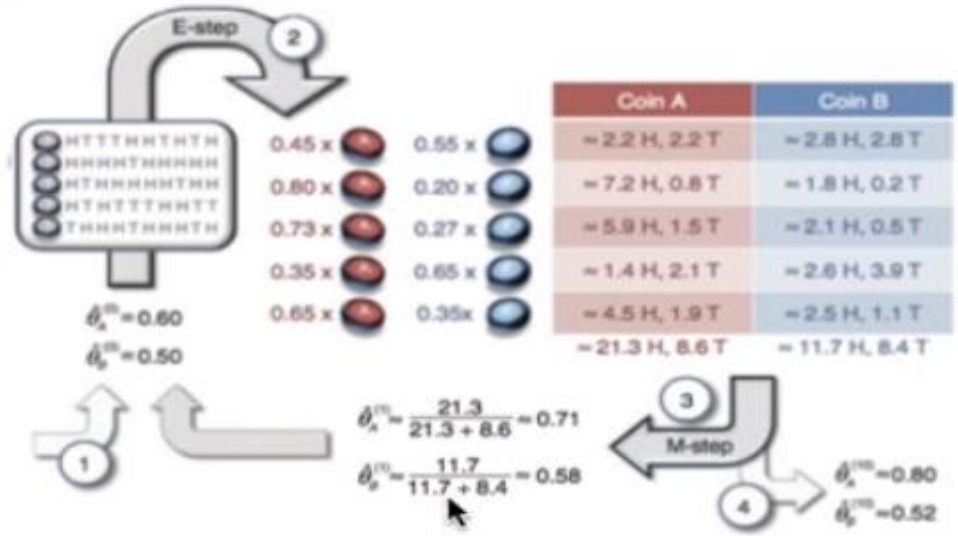
likelihood of "B"$= p_B(h)^h(1 - p_B(h))^{10-h} = 0.0009765625$

Normalizing I get probabilities: 0.45 and 0.55

So, We have:

$\theta_A = 0.6;\ \theta_B = 0.5$

b Expectation maximization



Repeat E-Step and M-Step until convergence

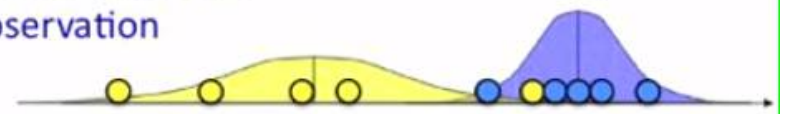
EM

Mixture models in 1-d

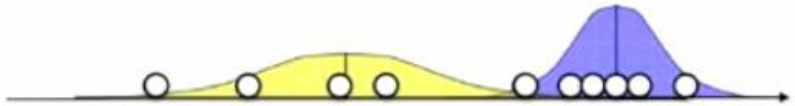
- Observations $x_1 \ldots x_n$
  - K=2 Gaussians with unknown μ, σ^2
  - estimation trivial if we know the source of each observation

$$\mu_b = \frac{x_1 + x_2 + \ldots + x_{n_b}}{n_b}$$

$$\sigma_b^2 = \frac{(x_1 - \mu_1)^2 + \ldots + (x_n - \mu_n)^2}{n_b}$$



- What if we don't know the source?
- If we knew parameters of the Gaussians (μ, σ^2)
  - can guess whether point is more likely to be a or b



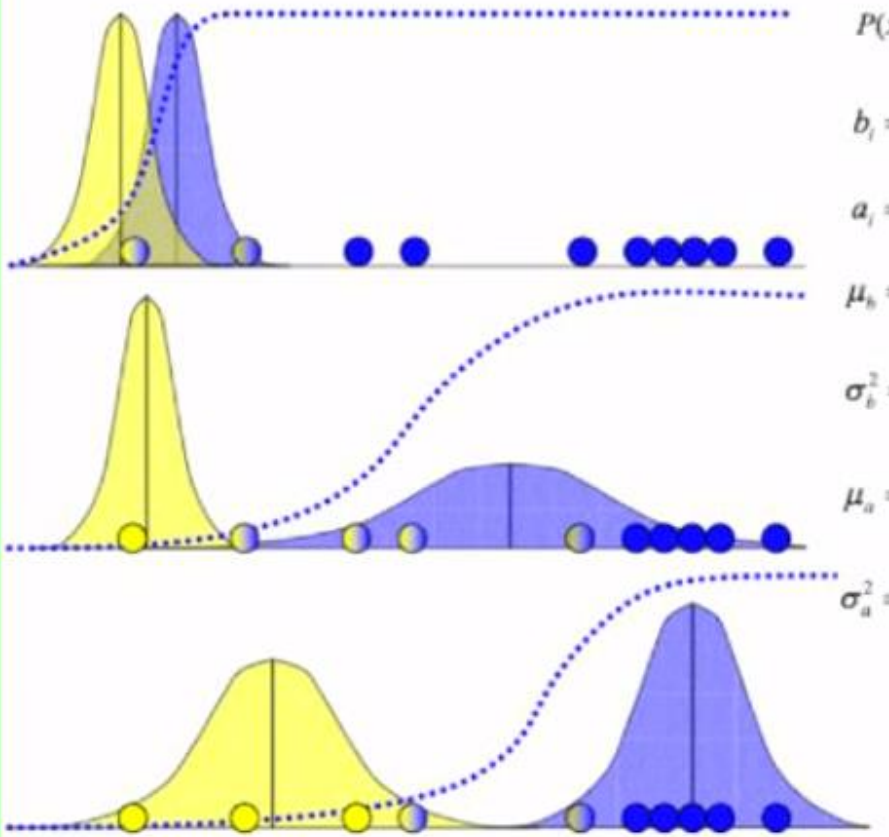
Expectation Maximization (EM)

- Chicken and egg problem
  - need (μ_a, σ_a^2) and (μ_b, σ_b^2) to guess source of points
  - need to know source to estimate (μ_a, σ_a^2) and (μ_b, σ_b^2)
- EM algorithm
  - start with two randomly placed Gaussians (μ_a, σ_a^2), (μ_b, σ_b^2)
  - for each point: $P(b|x_i)$ = does it look like it came from b?
  - adjust (μ_a, σ_a^2) and (μ_b, σ_b^2) to fit points assigned to them

The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step.

EM pick K

EM: 1-d example



$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right)$$

$$b_i = P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$a_i = P(a \mid x_i) = 1 - b_i$$

$$\mu_b = \frac{b_1 x_1 + b_2 x_2 + \ldots + b_n x_n}{b_1 + b_2 + \ldots + b_n}$$

$$\sigma_b^2 = \frac{b_1 (x_1 - \mu_1)^2 + \ldots + b_n (x_n - \mu_n)^2}{b_1 + b_2 + \ldots + b_n}$$

$$\mu_a = \frac{a_1 x_1 + a_2 x_2 + \ldots + a_n x_n}{a_1 + a_2 + \ldots + a_n}$$

$$\sigma_a^2 = \frac{a_1 (x_1 - \mu_1)^2 + \ldots + a_n (x_n - \mu_n)^2}{a_1 + a_2 + \ldots + a_n}$$

could also estimate priors:

$P(b) = (b_1 + b_2 + \ldots b_n) / n$

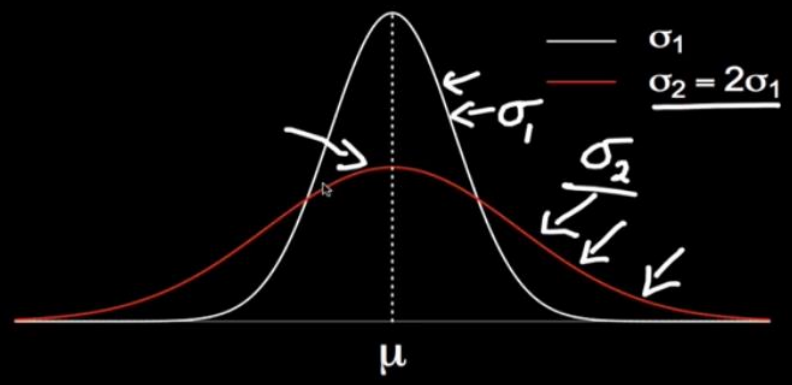
$P(a) = 1 - P(b)$

How to pick K?

- Probabilistic model $L = \log P(x_1 \ldots x_n) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} P(x_i \mid k) P(k)$
  - tries to "fit" the data (maximize likelihood)
- Pick K that makes *L* as large as possible?
  - *K* = *n*: each data point has its own "source"
  - may not work well for new data points
- Split points into training set T and validation set V
  - for each *K*: fit parameters of T, measure likelihood of V
  - sometimes still best when *K* = *n*
- Occam's razor: pick "simplest" of all models that fit
  - Bayes Inf. Criterion (BIC): $\max_p \{ L - \frac{1}{2} p \log n \}$
  - Akaike Inf. Criterion (AIC): $\min_p \{ 2p - L \}$

L … **likelihood**, how well our model fits the data
p … **number of parameters** how "simple" is the model

Two normal distributions:



The probability density function (pdf) is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

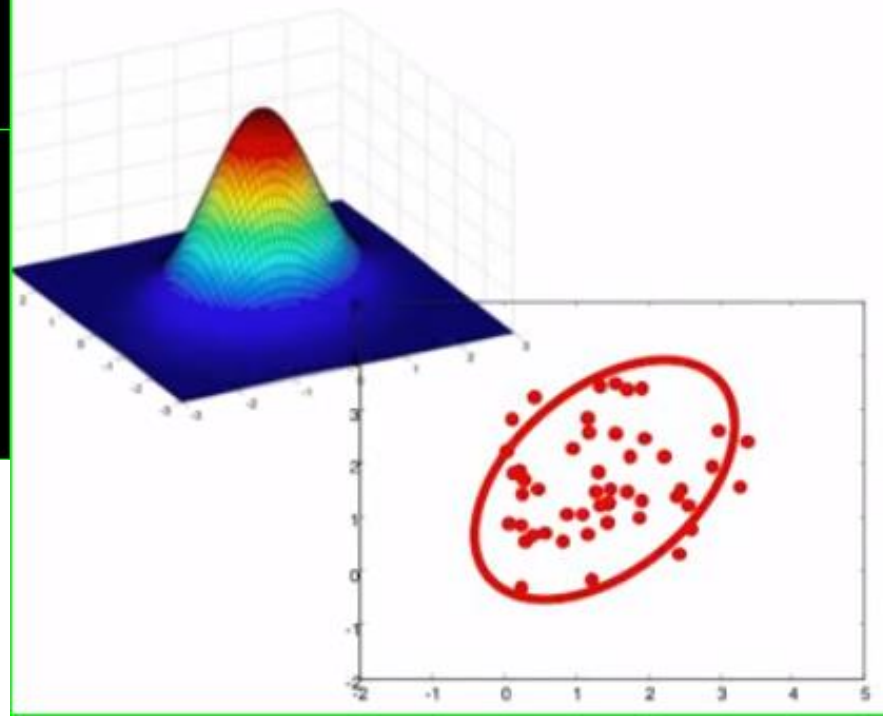
Multivariate Gaussian models

- Similar to univariate case

$$\mathcal{N}(\underline{x}\,;\,\underline{\mu},\Sigma) = \frac{1}{(2\pi)^{d/2}}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu})\Sigma^{-1}(\underline{x}-\underline{\mu})^T\right\}$$

μ = length-d row vector
Σ = d x d matrix

$|\Sigma|$ = matrix determinant



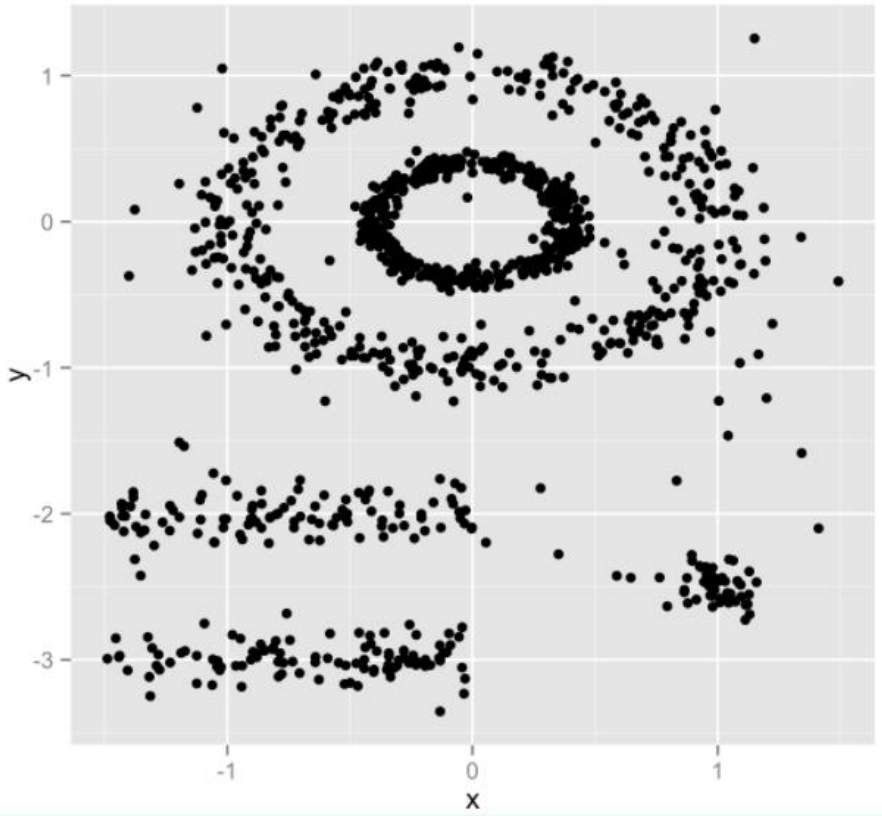
Maximum likelihood estimate:

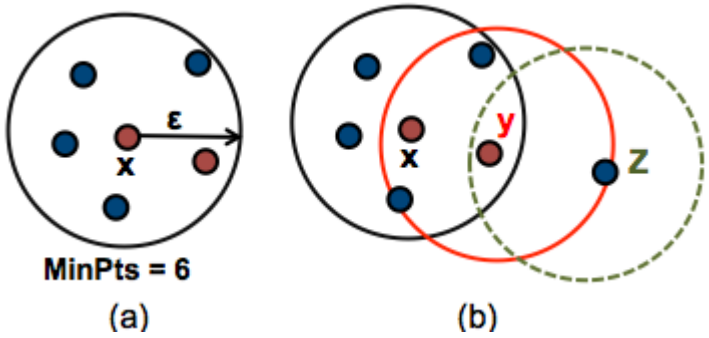
$$\hat{\mu} = \frac{1}{m}\sum_j \underline{x}^{(j)}$$

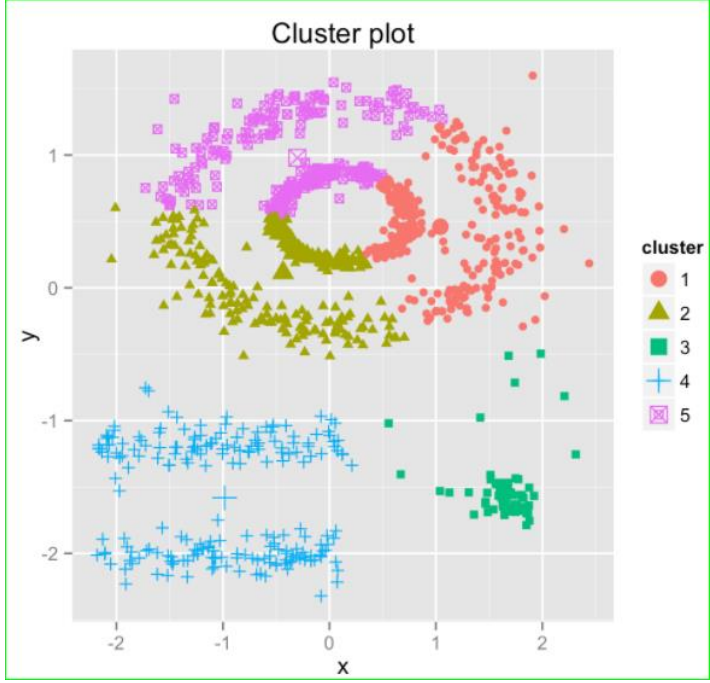
$$\hat{\Sigma} = \frac{1}{m}\sum_j (\underline{x}^{(j)} - \underline{\hat{\mu}})^T(\underline{x}^{(j)} - \underline{\hat{\mu}})$$

(average of dxd matrices)

DBSCAN [Density based]





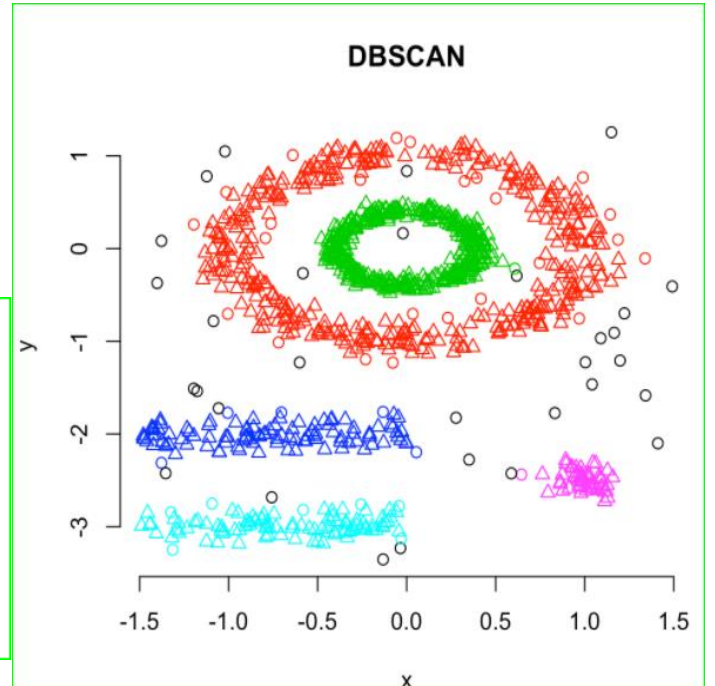


K-MEANS

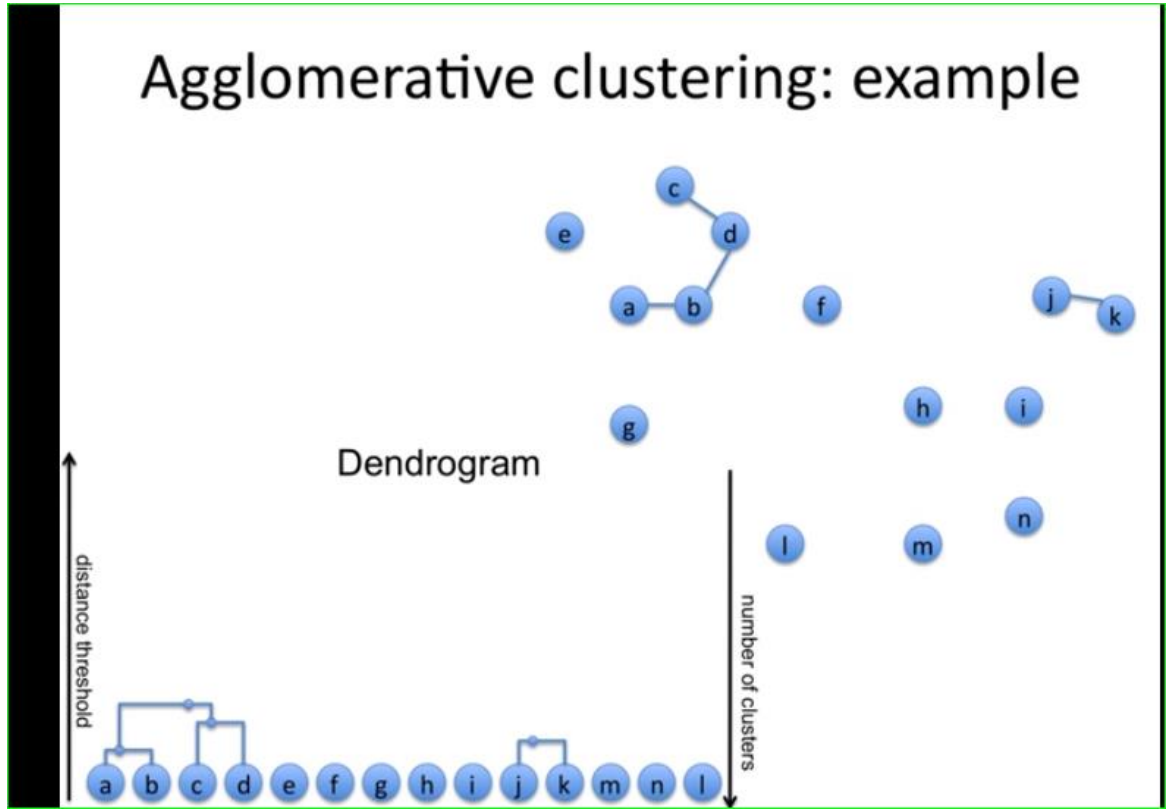
We define 3 terms, required for understanding the **DBSCAN algorithm**:

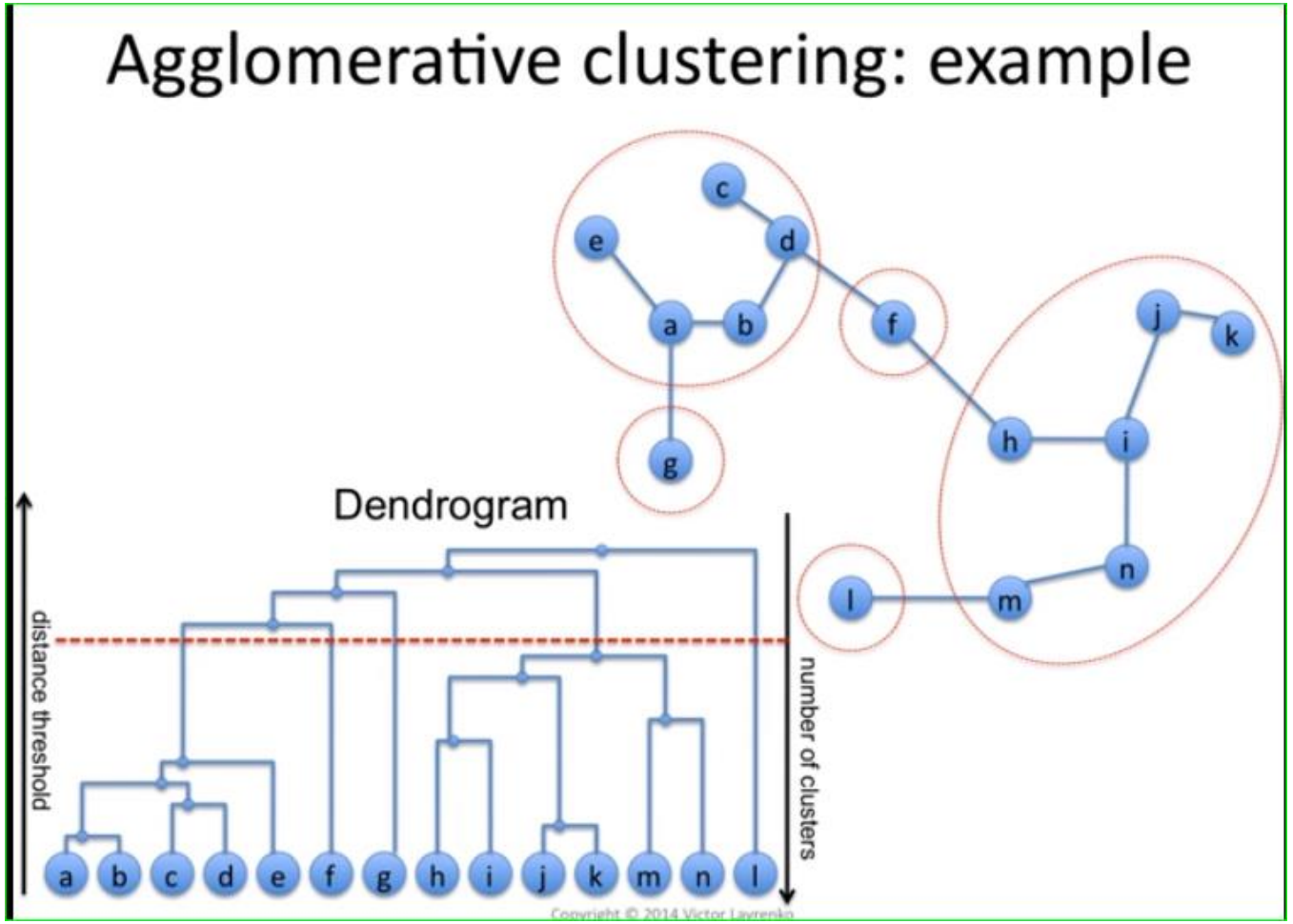
- **Direct density reachable**: A point "A" is **directly density reachable** from another point "B" if: i) "A" is in the ϵ-neighborhood of "B" and ii) "B" is a core point.
- **Density reachable**: A point "A" is **density reachable** from "B" if there are a set of core points leading from "B" to "A.
- **Density connected**: Two points "A" and "B" are **density connected** if there are a core point "C", such that both "A" and "B" are **density reachable** from "C".

A **density-based cluster** is defined as a group of density connected points. The algorithm of density-based clustering (DBSCAN)



Hierarchical clustering

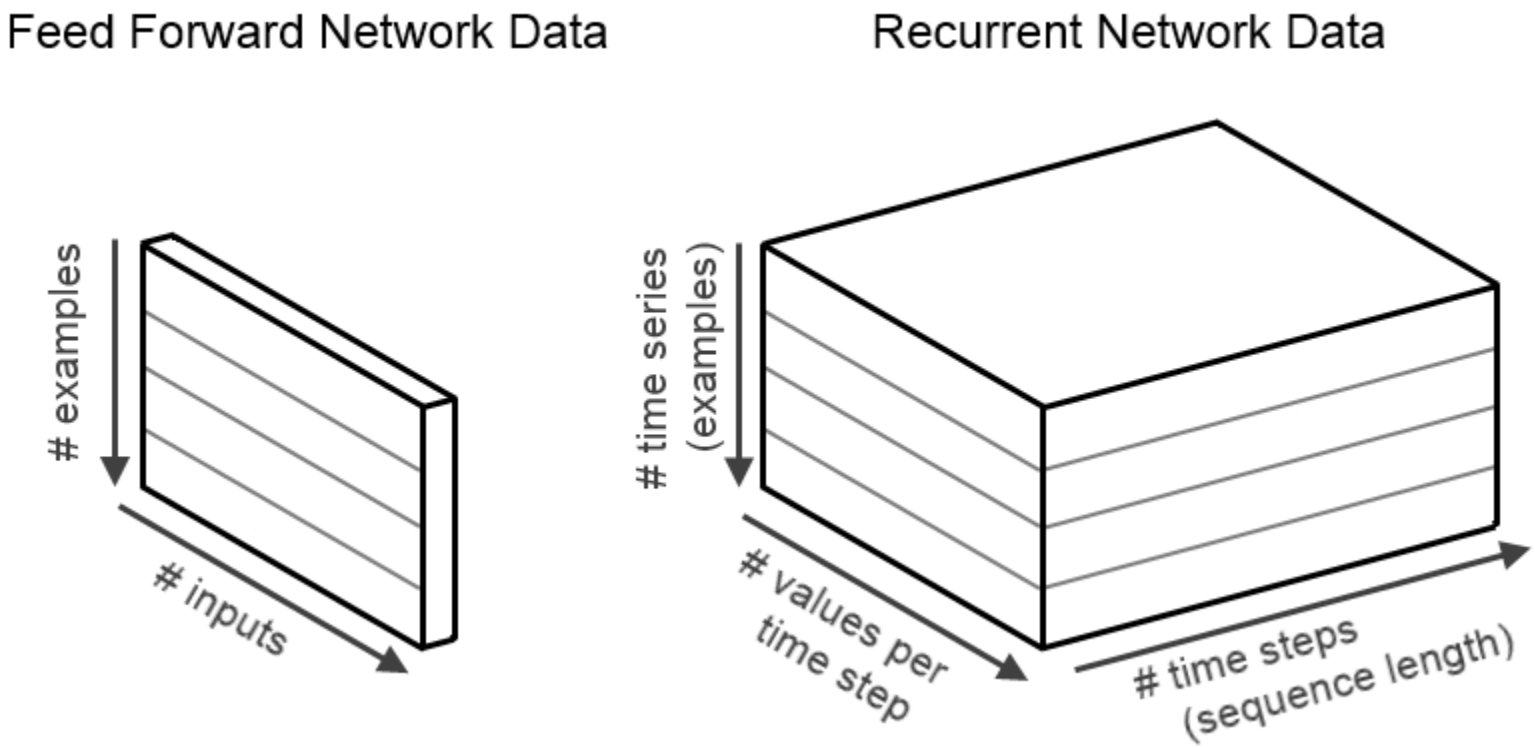




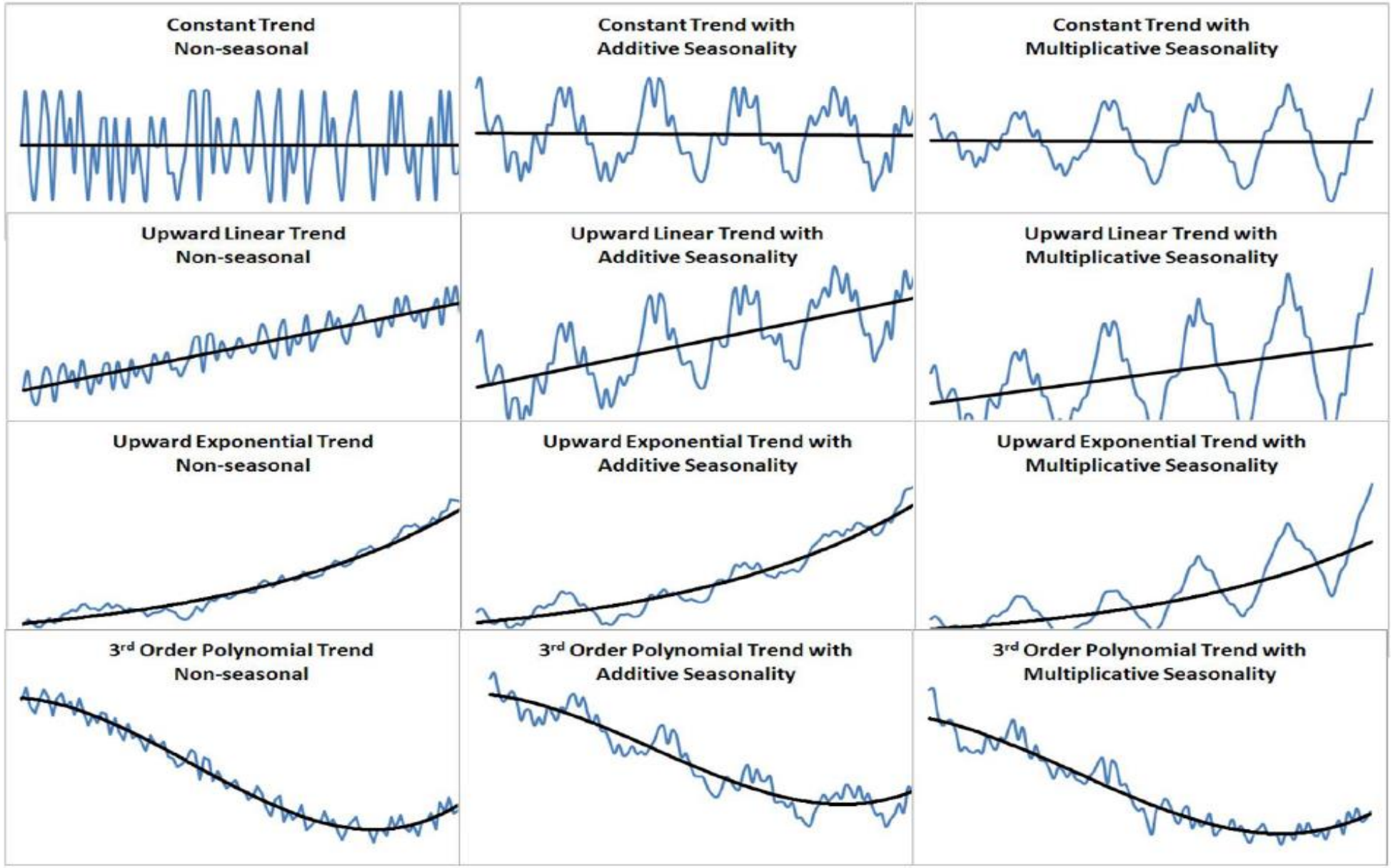
Why forecasting is important

- Departments throughout the organization depend on forecasts to formulate and execute their plans.
- Finance needs forecasts to project cash flows and capital requirements.
- Human resources need forecasts to anticipate hiring needs.
- Production needs forecasts to plan production levels, workforce, material requirements, inventories, etc.
- Demand is not the only variable of interest to forecasters.
- Manufacturers also forecast worker absenteeism, machine availability, material costs, transportation and production lead times, etc.
- Besides demand, service providers are also interested in forecasts of population, of other demographic variables, of weather, etc.

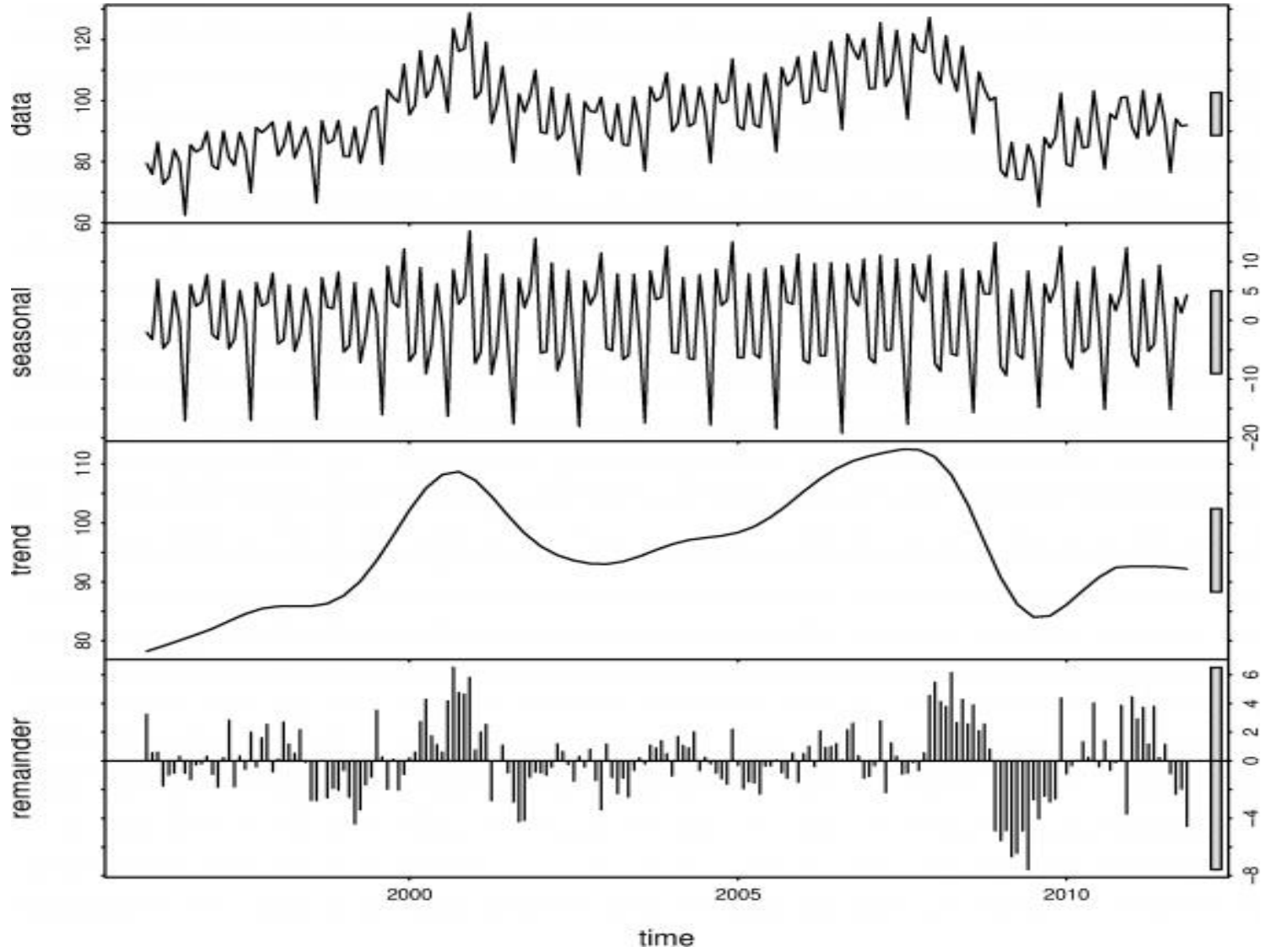
When to use time series analysis



Trend, Seasonality [Additive, Multiplicative]



Decomposing time series data



Time series metrics

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \hat{X}_i)^2}{n}},$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| X_i - \hat{X}_i \right|$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{X_i - \hat{X}_i}{X_i} \right| * 100$$

ARIMA [autoregressive integrated moving average]

ARIMA terminology

- A non-seasonal ARIMA model can be (almost) completely summarized by three numbers:

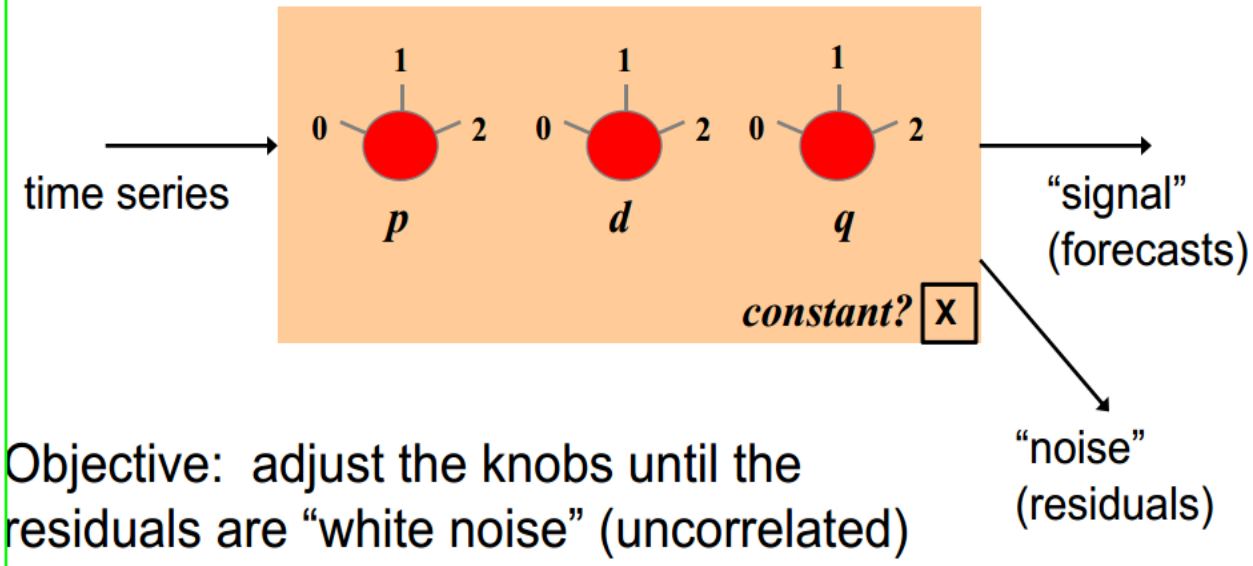
  ***p* = the number of *autoregressive* terms**

  d* = the number of *nonseasonal differences

  ***q* = the number of *moving-average* terms**

- This is called an "ARIMA(*p*,*d*,*q*)" model
- The model may also include a *constant* term (or not)

The ARIMA "filtering box"



Objective: adjust the knobs until the residuals are "white noise" (uncorrelated)

Forecasting equation of Y in ARIMA

Forecasting equation for *y*

constant — AR terms (lagged values of *y*)

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} \ldots - \theta_q e_{t-q}$$

MA terms (lagged errors)

By convention, the AR terms are + and the MA terms are –

Not as bad as it looks! Usually $p+q \leq 2$ and either $p=0$ or $q=0$ (pure AR or pure MA model)

I of ARIMA [Use Dickey fuller test]

ARIMA forecasting equation

- Let Y denote the *original* series
- Let y denote the *differenced* (stationarized) series

No difference (d=0): $y_t = Y_t$

First difference (d=1): $y_t = Y_t - Y_{t-1}$

Second difference (d=2): $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$

$= Y_t - 2Y_{t-1} + Y_{t-2}$

Note that the second difference is not just the change relative to two periods ago, i.e., it is *not* $Y_t - Y_{t-2}$. Rather, it is the change-in-the-change, which is a measure of local "acceleration" rather than trend.

Interpretation of AR & MA terms

Interpretation of AR terms

A series displays autoregressive (AR) behavior if it apparently feels a "restoring force" that tends to pull it back toward its mean.

- In an AR(1) model, the AR(1) coefficient determines how fast the series tends to return to its mean. If the coefficient is near zero, the series returns to its mean quickly; if the coefficient is near 1, the series returns to its mean slowly.
- In a model with 2 or more AR coefficients, the sum of the coefficients determines the speed of mean reversion, and the series may also show an oscillatory pattern.

Interpretation of MA terms

A series displays moving-average (MA) behavior if it apparently undergoes random "shocks" whose effects are felt in two or more consecutive periods.

- The MA(1) coefficient is (minus) the fraction of last period's shock that is still felt in the current period.
- The MA(2) coefficient, if any, is (minus) the fraction of the shock two periods ago that is still felt in the current period, and so on.

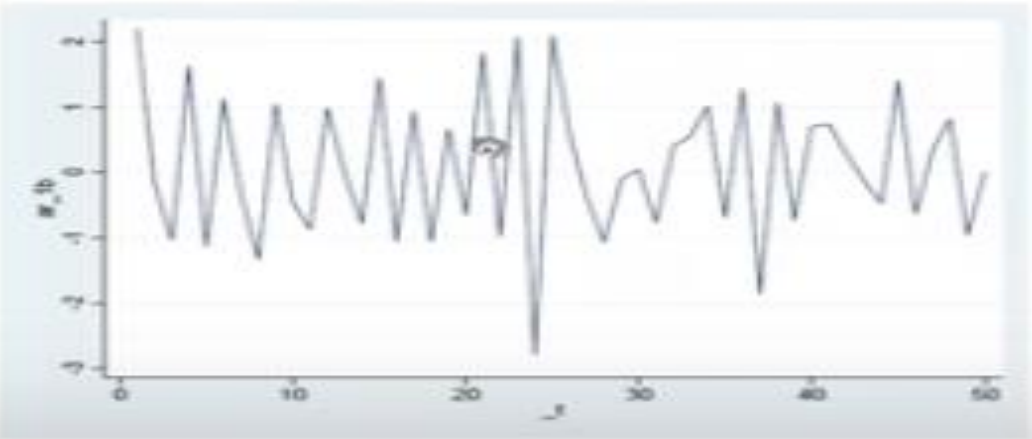
AR terms

Autoregressive (AR) models

- Autoregressive (AR) models are models in which the value of a variable in one period is related to its values in previous periods.
- AR(p) is an autoregressive model with p lags: $y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$
  where μ is a constant and γ_p is the coefficient for the lagged variable in time t-p.
- AR(1) is expressed as: $y_t = \mu + \gamma y_{t-1} + \epsilon_t = \mu + \gamma(Ly_t) + \epsilon_t$ or $(1 - \gamma L)y_t = \mu + \epsilon_t$

AR(1) with $\gamma = 0.8$

AR(1) with $\gamma = -0.8$



MA terms

Definition [edit]

The notation MA(q) refers to the moving average model of order q:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

where μ is the mean of the series, the θ_1, ..., θ_q are the parameters of the model and the ε_t, ε_{t-1},..., ε_{t-q} are white noise error terms. The value of q is called the order of the MA model. This can be equivalently written in terms of the backshift operator B as

$$X_t = \mu + (1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t.$$

Thus, a moving-average model is conceptually a linear regression of the current value of the series against current and previous (unobserved) white noise error terms or random shocks. The random shocks at each point are assumed to be mutually independent and to come from the same distribution, typically a normal distribution, with location at zero and constant scale.

Interpretation [edit]

The moving-average model is essentially a finite impulse response filter applied to white noise, with some additional interpretation placed on it. The role of the random

Box Jenkins method

The following table summarizes how one can use the sample autocorrelation function for model identification.

Shape	Indicated Model
Exponential, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to identify the order of the autoregressive model.
Alternating positive and negative, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to help identify the order.
One or more spikes, rest are essentially zero	Moving average model, order identified by where plot becomes zero.
Decay, starting after a few lags	Mixed autoregressive and moving average (ARMA) model.
All zero or close to zero	Data are essentially random.
High values at fixed intervals	Include seasonal autoregressive term.
No decay to zero	Series is not stationary.

Lagrange multiplier

In mathematical optimization, the **method of Lagrange multipliers** (named after Joseph Louis Lagrange[1]) is a strategy for finding the local maxima and minima of a function subject to equality constraints.

For the case of only one constraint and only two choice variables (as exemplified in Figure 1), consider the optimization problem

maximize $f(x, y)$
subject to $g(x, y) = c$.

We assume that both f and g have continuous first partial derivatives. We introduce a new variable (λ) called a **Lagrange multiplier** and study the **Lagrange function** (or **Lagrangian** or **Lagrangian expression**) defined by

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda \cdot (g(x, y) - c),$$

where the λ term may be either added or subtracted. If $f(x_0, y_0)$ is a maximum of $f(x, y)$ for the original constrained problem, then there exists λ_0 such that (x_0, y_0, λ_0) is a stationary point for the Lagrange function (stationary points are those points where the partial derivatives of $\mathcal{L}$ are zero). However, not all stationary points yield a solution of the original problem. Thus, the method of Lagrange multipliers yields a necessary condition for optimality in constrained problems.[2][3][4][5][6] Sufficient conditions for a minimum or maximum also exist.

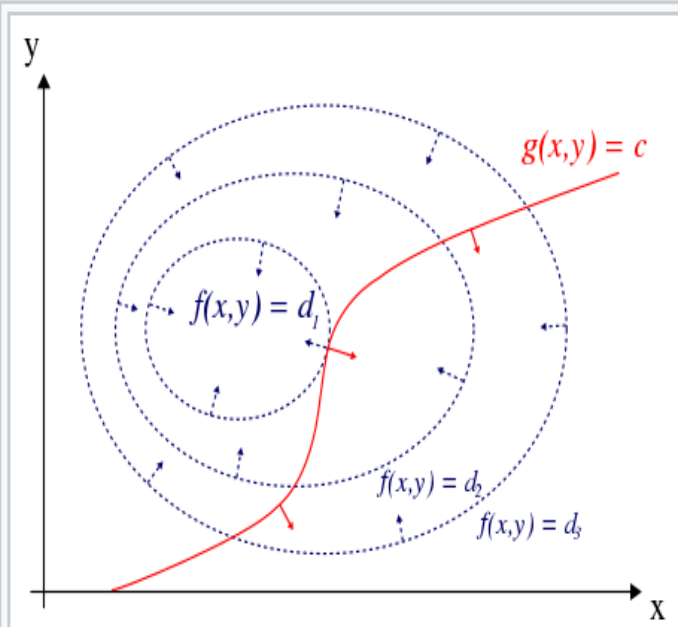
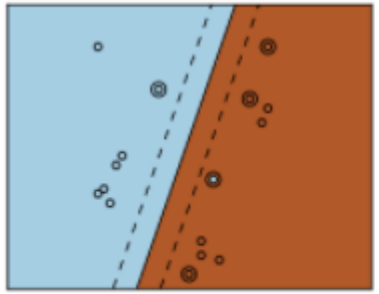
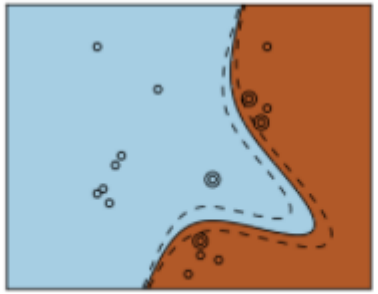


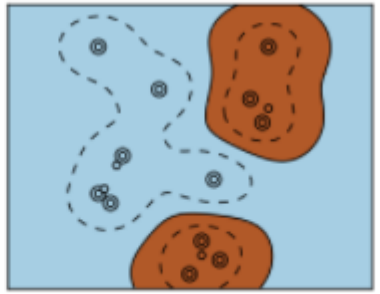
Figure 1: The red line shows the constraint $g(x, y) = c$. The blue lines are contours of $f(x, y)$. The point where the red line tangentially touches a blue contour is the maximum of $f(x, y)$, since $d_1 > d_2$.

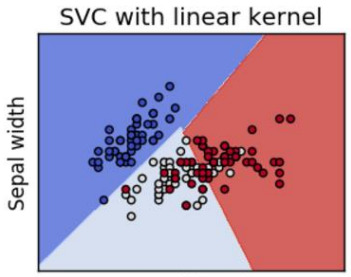
SVM kernel

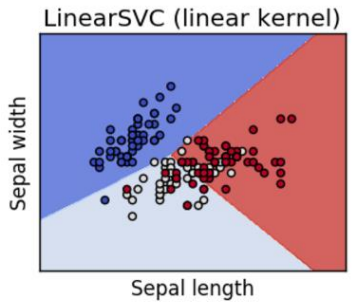
Three different types of SVM-Kernels are displayed below. The polynomial and RBF are especially useful when the data-points are not linearly separable.

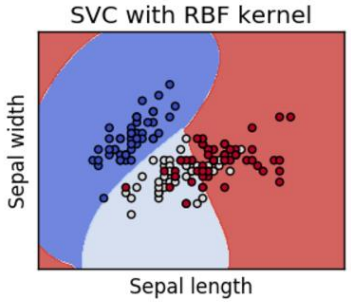


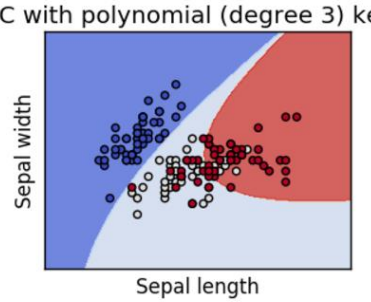










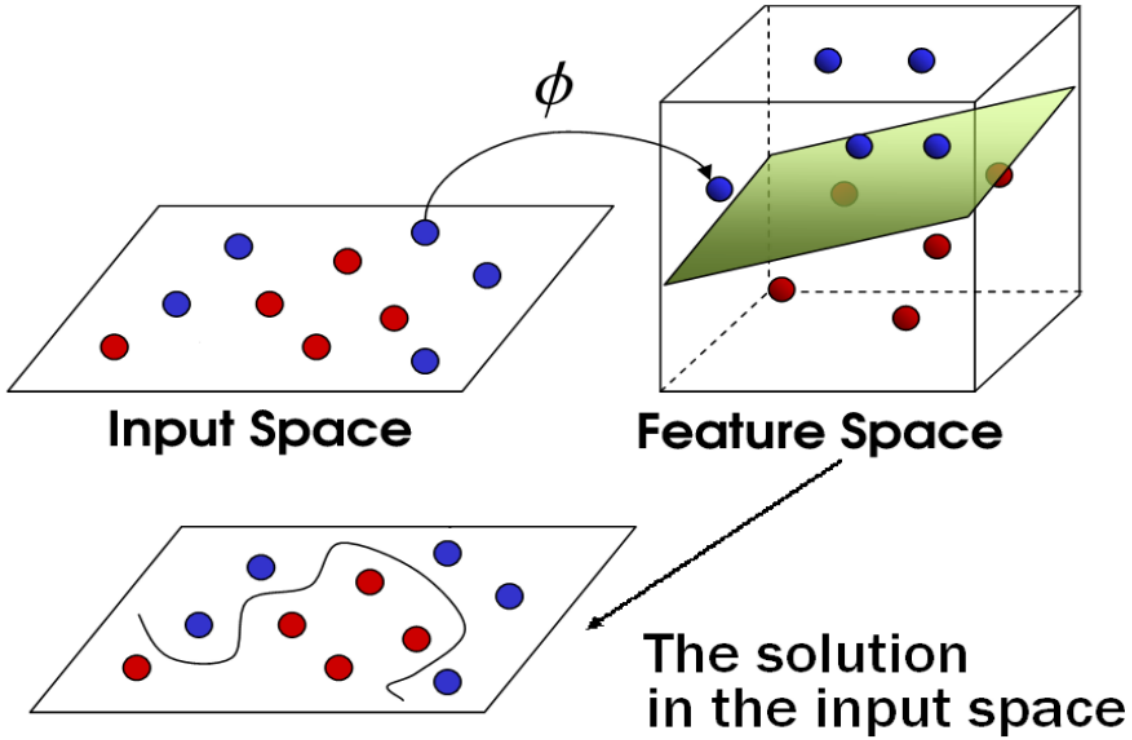


Higher dimensions VC (Vapnik-Chervonenkis) Dimension page 20

Non linear SVM

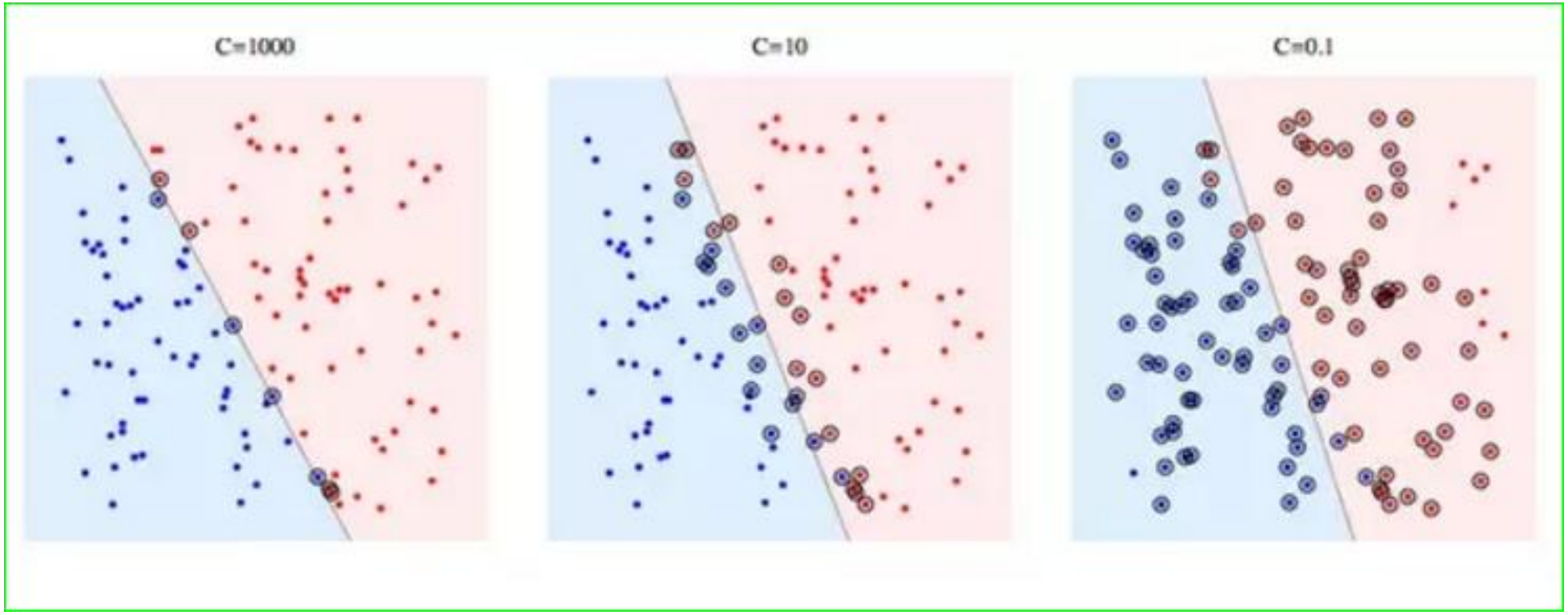
Non Linear SVM :

An illustration of the algorithm:



SVM – cost parameter (C)

C is the parameter for the soft margin cost function, which controls the influence of each individual support vector; this process involves trading error penalty for stability.



A large C gives you low bias and high variance. Low bias because you penalize the cost of misclassification a lot.
A small C gives you higher bias and lower variance.

C problem

MARSIAN Technologies LLP | EXPERIENCE PERFECTION

www.marsiantech.com

Soft Margin:
The “C” Problem

-“C” plays a major role in controlling overfitting.

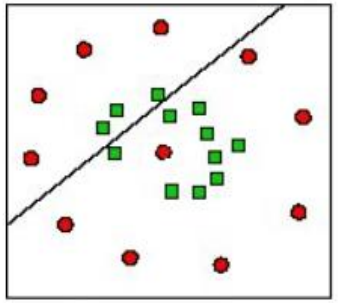
-Finding the “Right” value for “C” is one of the major problems of SVM:

-Larger C → less training samples that are not in ideal position (which means less training error that affects positively the Classification Performance (CP)) But smaller margin (affects negatively the (CP)).C large enough may lead us to overfitting (too much complicated classifier that fits only the training set)

-Smaller C → more training samples that are not in ideal position (which means more training error that affects negatively the Classification Performance (CP)) But larger Margin (good for (CP)). C small enough may lead to underfitting (naïve classifier)

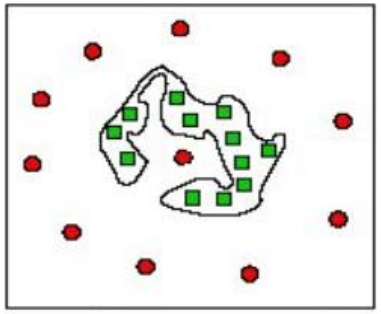
Soft Margin:
The “C” Problem: Overfitting and Underfitting

Under-Fitting

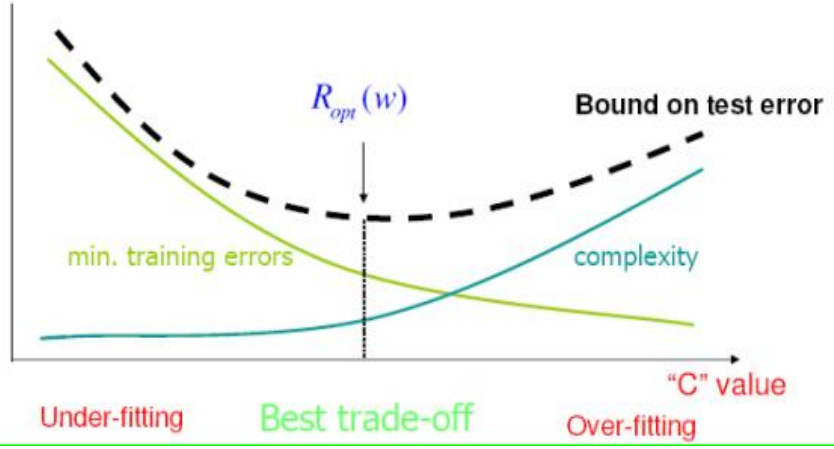


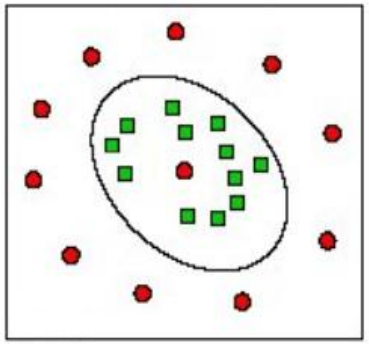
Too much simple!

Over-Fitting



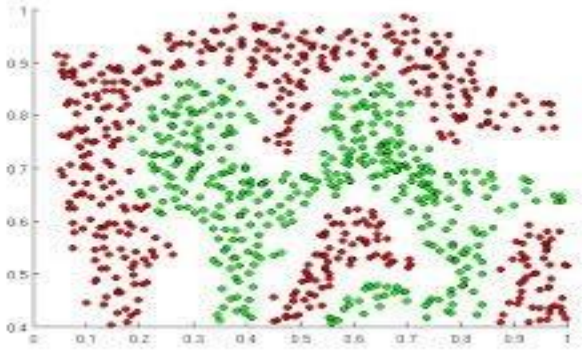
Too much complicated!





Trade-Off

Gamma



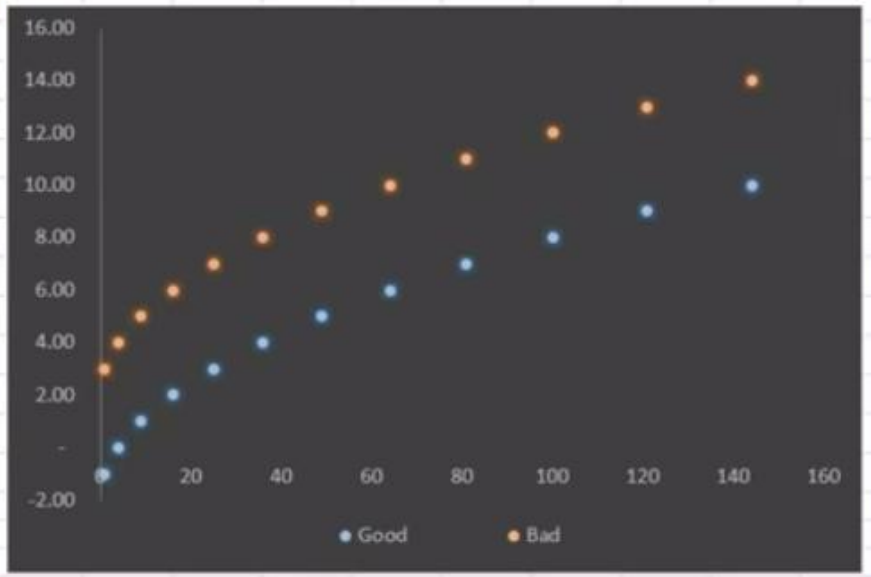
- They are not linearly separable in 2D so you want to transform them to a higher dimension where they will be linearly separable. Imagine "raising" the green points, then you can separate them from the red points with a plane (hyperplane)
- To "raise" the points you use the RBF kernel, gamma controls the shape of the "peaks" where you raise the points. A small gamma gives you a pointed bump in the higher dimensions, a large gamma gives you a softer, broader bump.
- So a small gamma will give you low bias and high variance while a large gamma will give you higher bias and low variance.
- You usually find the best C and Gamma hyper-parameters using Grid-Search.

What is kernel trick

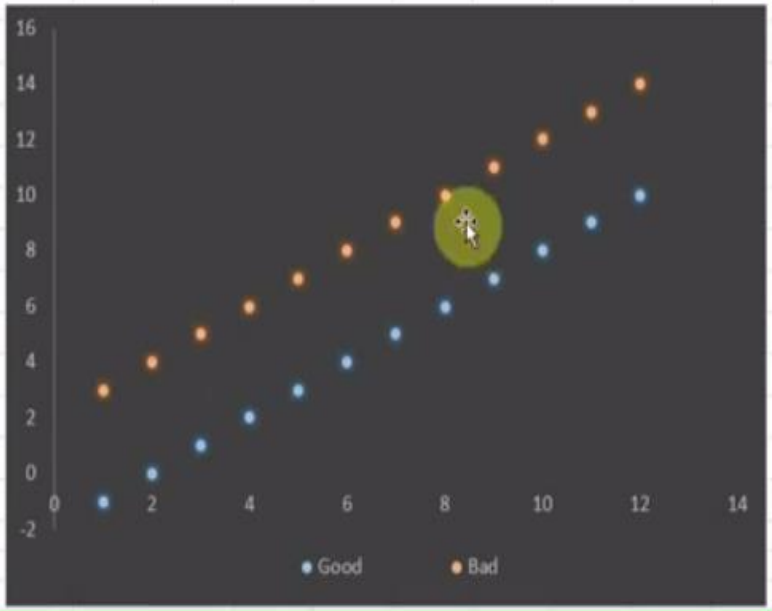
- For linear case, we tend to come up a function like
  - a1 + b1*x1 + b2*x2 + ... > c → class 1
  - a1 + b1*x1 + b2*x2 + ... <= c → class 2
- In case of non linear separation, one can have possibilities of
  - a1 + b1*x1^2 + b2*sqrt(x2)....
- One can introduce new variables -- like f1=x1^2, f2=sqrt(x2)
- This can provide linearly separable space

KERNEL TRICK: USING TRANSFORMATION OF VARIABLES TO CONVERT NON LINEAR SPACE TO LINEAR SPACE.

X	Good	Bad
1	-1.00	3.00
4	-	4.00
9	1.00	5.00
16	2.00	6.00
25	3.00	7.00
36	4.00	8.00
49	5.00	9.00
64	6.00	10.00
81	7.00	11.00
100	8.00	12.00
121	9.00	13.00
144	10.00	14.00



X1 (=sqrt(X))	Good	Bad
1	-1	3
2	0	4
3	1	5
4	2	6
5	3	7
6	4	8
7	5	9
8	6	10
9	7	11
10	8	12
11	9	13
12	10	14



SVM advantages and disadvantages

MARSIAN Technologies LLP | EXPERIENCE PERFECTION

www.marsiantech

The Advantages of SVM:

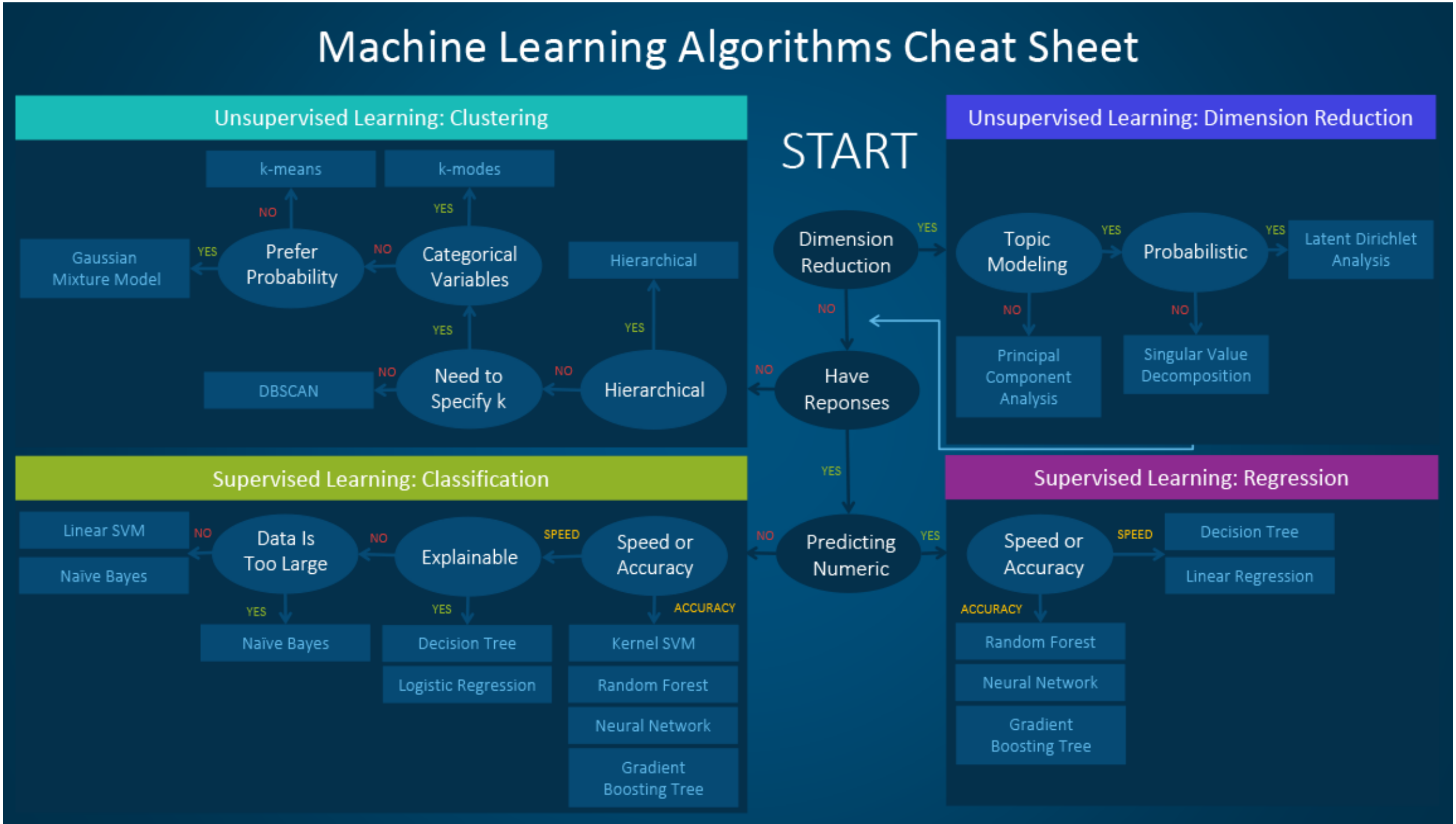
- ▶Based on a strong and nice Theory:
  - -In contrast to previous “black box” learning approaches, SVMs allow for some intuition and human understanding.
- ▶Training is relatively easy:
  - -No local optimal, unlike in neural network
  - -Training time does not depend on dimensionality of feature space, only on fixed input space thanks to the kernel trick.
- ▶Generally avoids over-fitting:
  - - Tradeoff between classifier complexity and error can be controlled explicitly.
- ▶SVMs have been demonstrated superior classification Accuracies to neural networks and other methods in many Apllications:

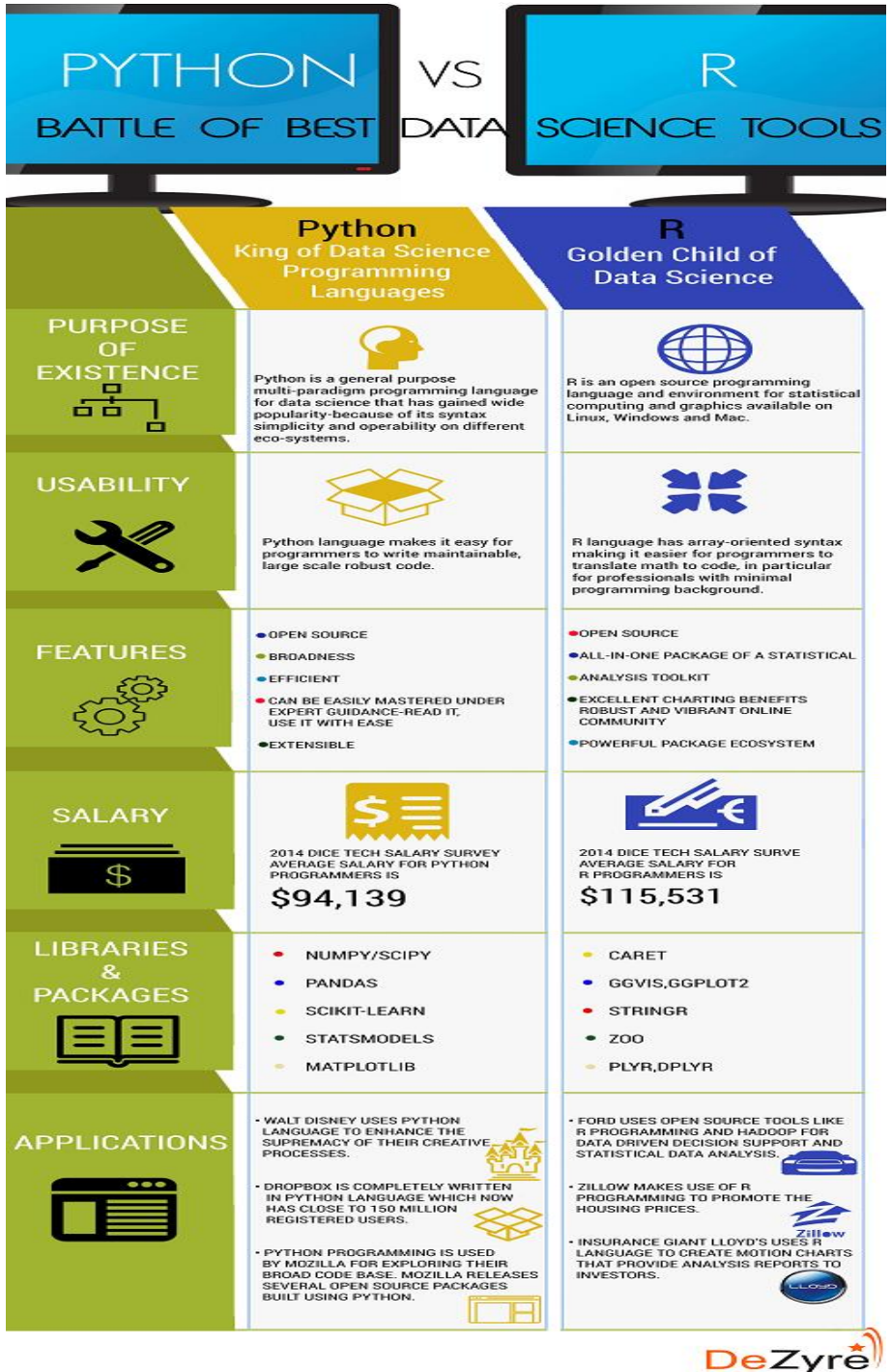
-generalize well even in high dimensional spaces under small training set conditions. Also it is robust to noise

The Drawbacks of SVM:

- ▶It is not clear how to select a kernel function in a principled manner.
- ▶What is the right value for the “Trade-off” parameter “C”:
  - We have to search manually for this value, Since we don’t have a principled way for that.
- ▶Tends to be expensive in both memory and computational time, especially for multiclass problems:
  - This is why some applications use SVMs for verification rather than classification . This strategy is computationally cheaper once SVMs are called just to solve difficult cases.

Machine Learning Algorithms Cheat Sheet



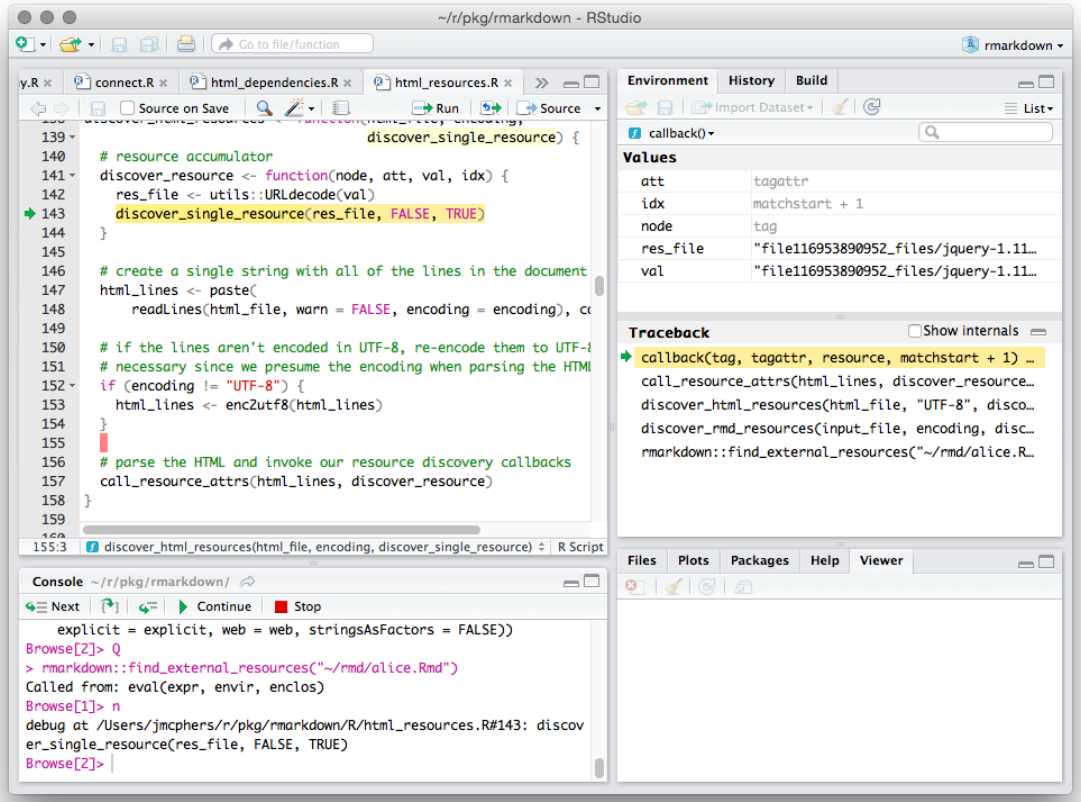


Feature	Python is Better	R Language is Better
Model Building	Both are Similar	Both are Similar
Model Interpretability	Not better than R.	R is better
Production	Python is Better	Not better than Python
Community Support	Not better than R.	R has good community support over Python.
Data Science Libraries	Both are similar.	Both are similar
Data Visualizations	Not better than R	R has good data visualizations libraries and tools.
Learning Curve	Learning Python is easier than learning R.	R has a steep learning curve.

R distribution dilemma [GPL-2, LGPL]

If R is going to be a significant part of your product (not only “enhances it”, but is a foundation of it, which means that **your application relies on R**), if you link to the R.DLL (or r.so, or r.dylib) in any way (both static or dynamic), if you share complex (this is undefined!) data between R and your application, if you convey the application to the end-users (especially merged with R), then you have to deliver sources of your application to all end-users of your product. This is their right to the product as the whole.

R debugging & logging



Use print statements

Log4r: logging library

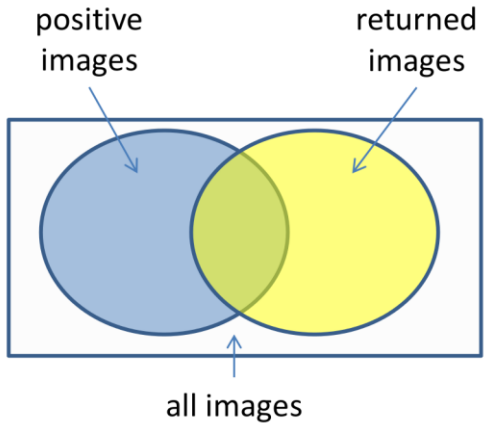
Precision and Recall

Precision:

Proportion of returned images that are positive

Recall:

Proportion of positive images that are returned



The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Confusion Matrix: Accuracy, Precision and Recall

Given a confusion matrix, it's easy to compute accuracy, precision and recall:

	Predicted class A	Predicted class B	Predicted class C	
Actual class A	50	80	70	200
Actual class B	40	140	120	300
Actual class C	120	220	160	500
	210	440	350	1000

$$Accuracy = \frac{50 + 140 + 160}{1000}$$

$$Precision_A = \frac{50}{50 + 40 + 120}$$

$$Recall_A = \frac{50}{50 + 80 + 70}$$

Confusion matrices can, themselves, be confusing sometimes ☹

Imbalance classification

http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

- An automated inspection machine which detect products coming off manufacturing assembly line may find number of defective products significantly lower than non defective products.
- A test done to detect cancer in residents of a chosen area may find the number of cancer affected people significantly less than unaffected people.
- In credit card fraud detection, fraudulent transactions will be much lower than legitimate transactions.
- A manufacturing operating under six sigma principle may encounter 10 in a million defected products

Ways to deal with imbalanced class problem

- Survey more people / collect more data
- Under-sample
- Over-sample
- SMOTE (Generate synthetic sample)
- Ensemble techniques
- Penalize models (cost sensitive classifiers)
- Try more measures

```
library(DMwR)
trainSplit$target <- as.factor(trainSplit$target)
trainSplit <- SMOTE(target ~ ., trainSplit, perc.over = 100, perc.under=200)
trainSplit$target <- as.numeric(trainSplit$target)
```

https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/

SVM vs Random forest

- **Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm and vice-versa.**
- SVM and Random Forest are both used in classification problems.
- a) If you are sure that your data is outlier free and clean then go for SVM. It is the opposite - if your data might contain outliers then Random forest would be the best choice
- b) Generally, SVM consumes more computational power than Random Forest, so if you are constrained with memory go for Random Forest machine learning algorithm.
- **c)** Random Forest gives you a very good idea of variable importance in your data, so if you want to have variable importance then choose Random Forest machine learning algorithm.
- d) Random Forest machine learning algorithms are preferred for multiclass problems.
- e) SVM is preferred in multi-dimensional problem set - like text classification
- but as a good data scientist, you should experiment with both of them and test for accuracy or rather you can use ensemble of many Machine Learning techniques.

Normality test

Assuming your dataset is called `words` and has a `counts` column, you can plot the histogram to have a visualization of the distribution:

```
hist(words$counts, 100, col="black")
```

where 100 is the number of bins

You can also do a normal Q-Q plot using

```
qqnorm(words$counts)
```

Finally, you can also use the Shapiro-Wilk test for normality

```
shapiro.test(word$counts)
```

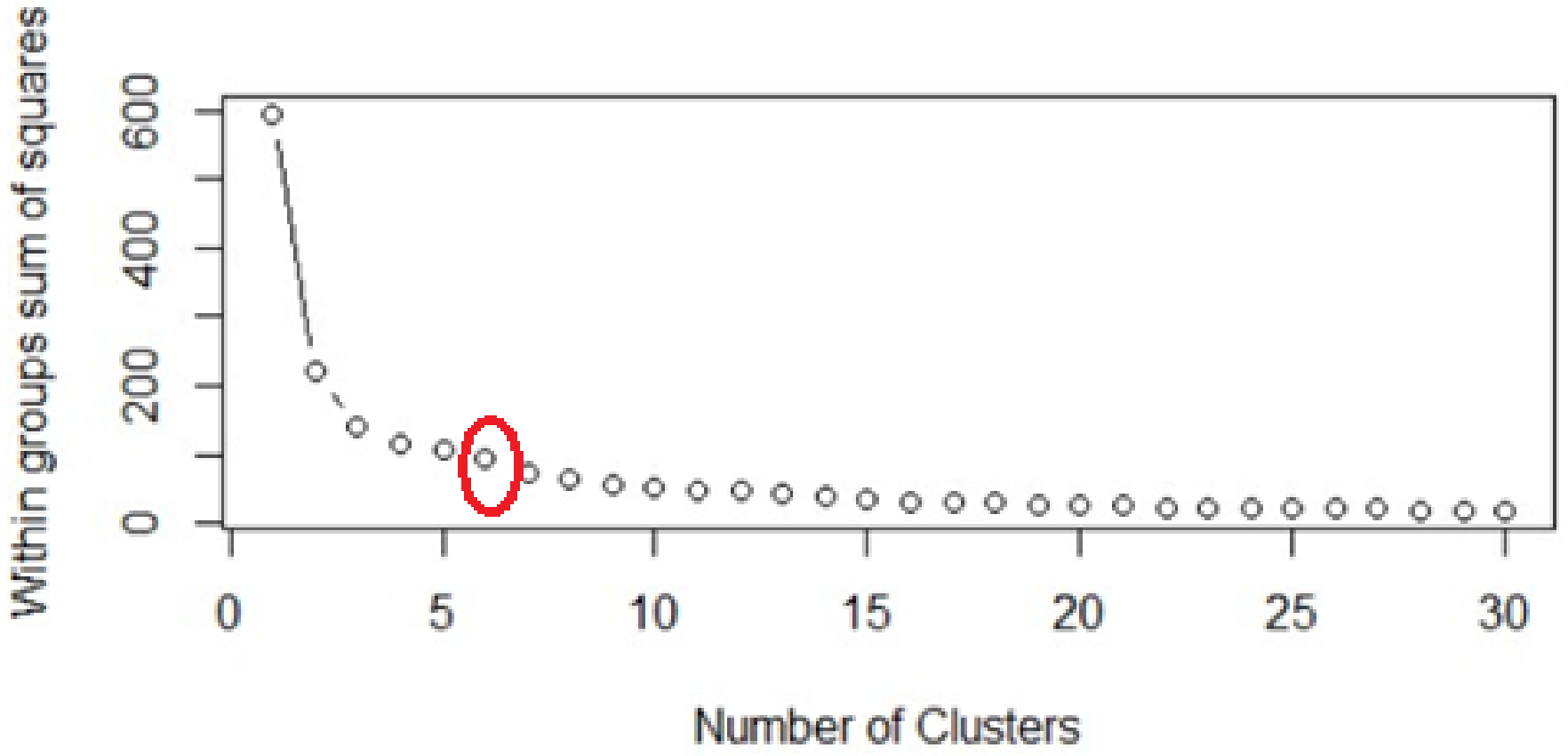
Tests of univariate normality include D'Agostino's K-squared test, the Jarque–Bera test, the Anderson–Darling test, the Cramér–von Mises criterion, the Lilliefors test for normality (itself an adaptation of the Kolmogorov–Smirnov test), the Shapiro–Wilk test, the Pearson's chi-squared test, and the Shapiro–Francia test.

How many variables how much data

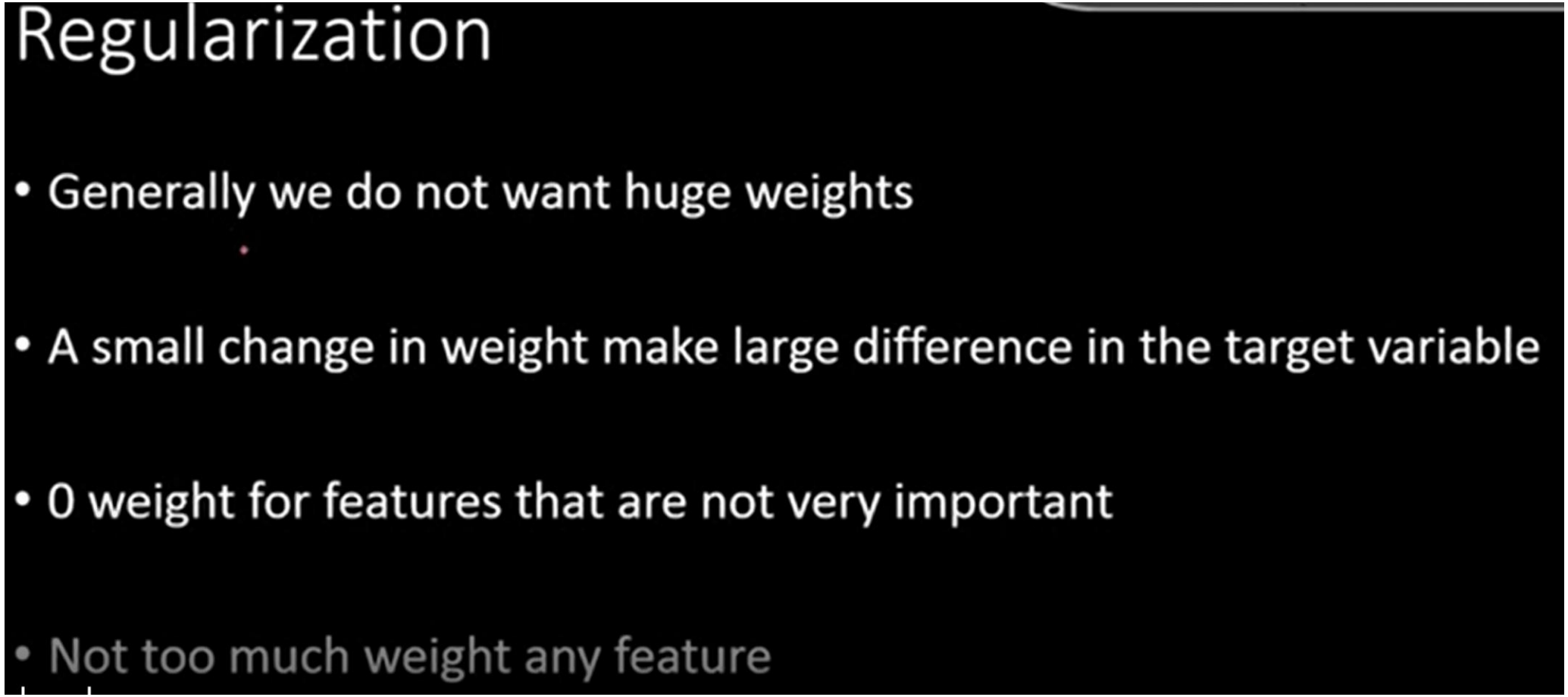
How Many Variables and How Much Data? Statisticians give us procedures to learn with some precision how many records we would need to achieve a given degree of reliability with a given dataset and a given model. Data miners' needs are usually not so precise, so we can often get by with rough rules of thumb. A good rule of thumb is to have 10 records for every predictor variable. Another, used by Delmaster and Hancock (2001, p. 68) for classification procedures, is to have at least $6 \times m \times p$ records, where m is the number of outcome classes and p is the number of variables.

Elbow plot

Red circled point in above graph i.e. Number of Cluster =6 is the point after which you don't see any decrement in WSS. This point is known as bending point and taken as K in K – Means.



Regularization



L1 & L2

L1 regularization on least squares:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2 + \lambda \sum_{i=1}^{k} |w_i|$$

L2 regularization on least squares:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2 + \lambda \sum_{i=1}^{k} w_i^2$$

L1 v/s L2

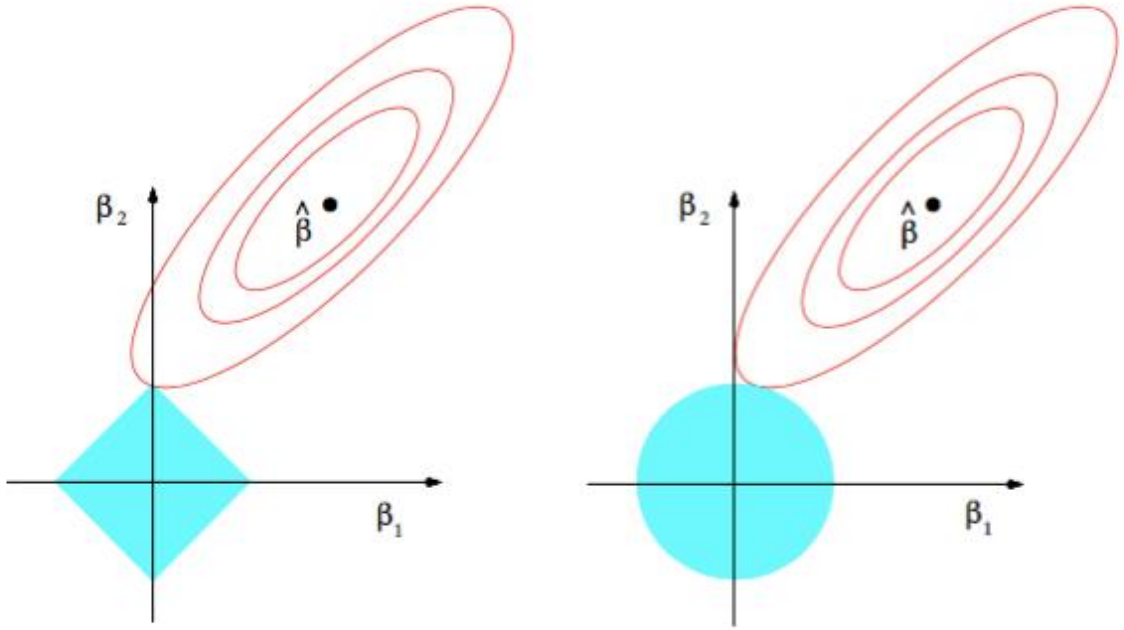
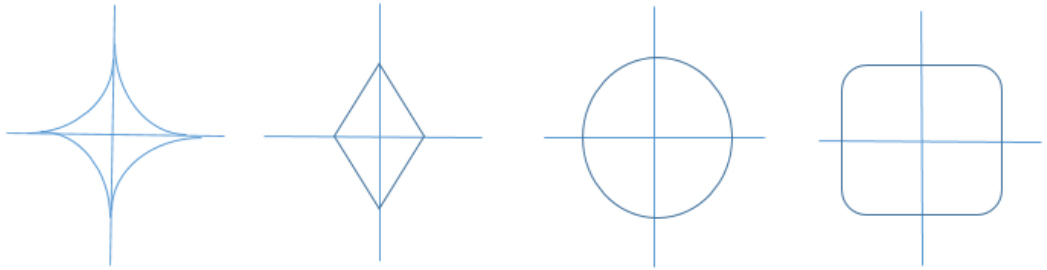


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

L2 regularization	L1 regularization
Computational efficient due to having analytical solutions	Computational inefficient on non-sparse cases
Non-sparse outputs	Sparse outputs
No feature selection	Built-in feature selection



p = 0.5 p =1 p = 2 p = 4

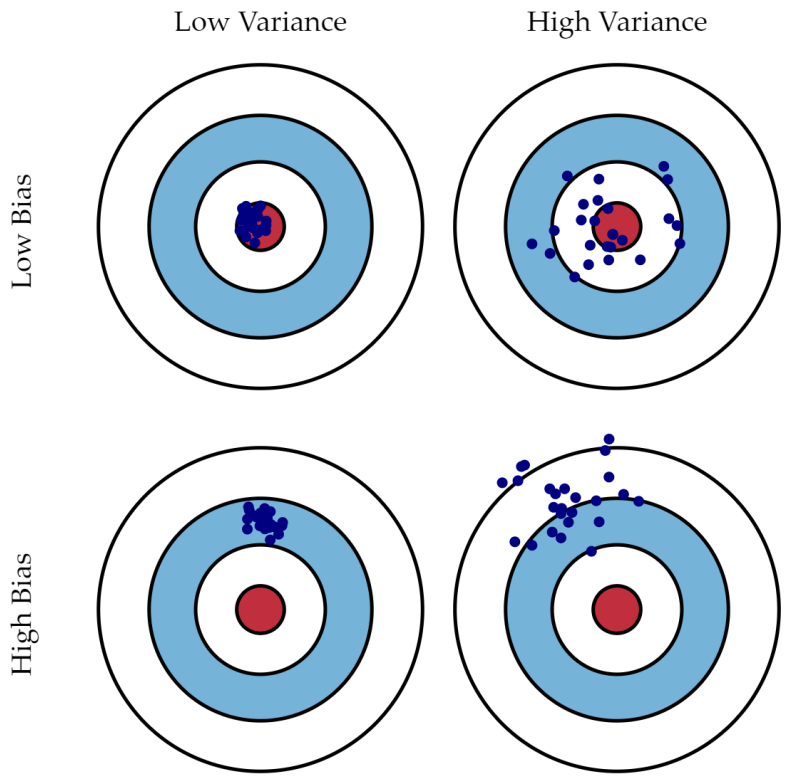
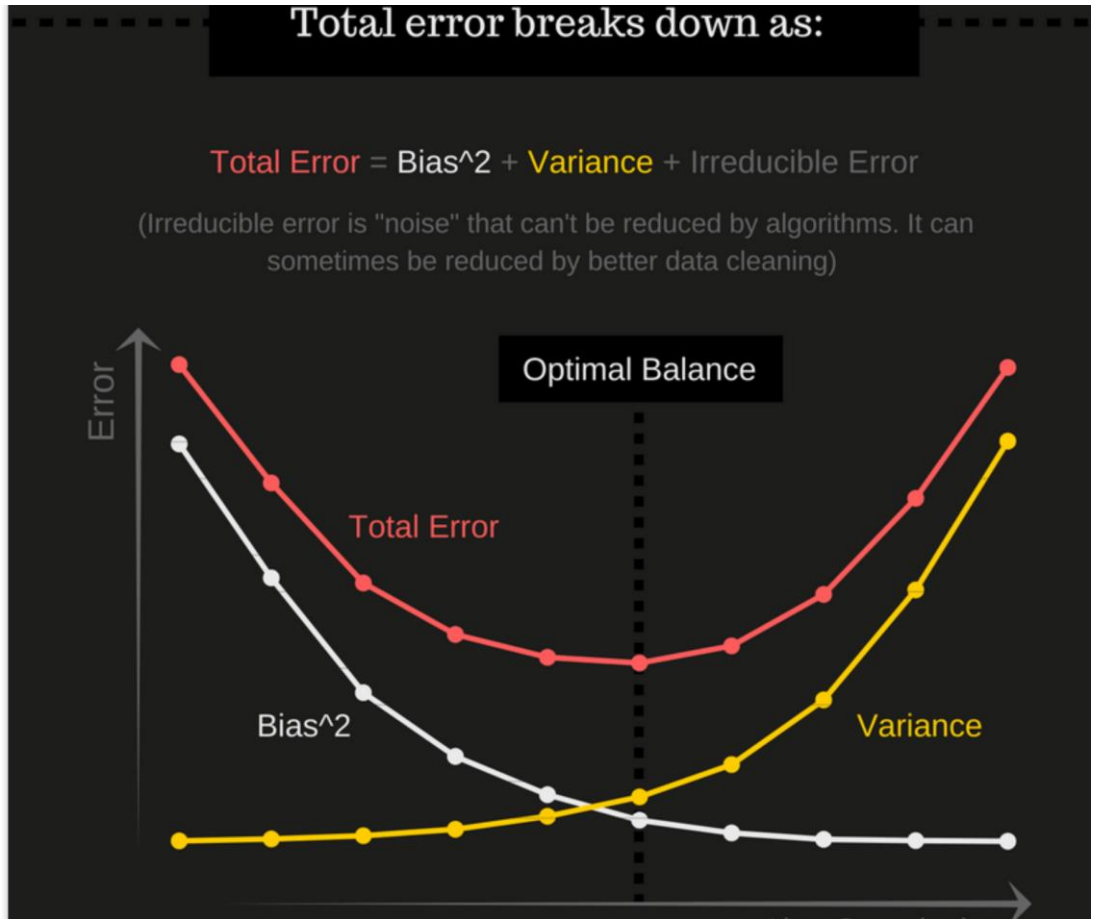


Fig. 1 Graphical illustration of bias and variance.

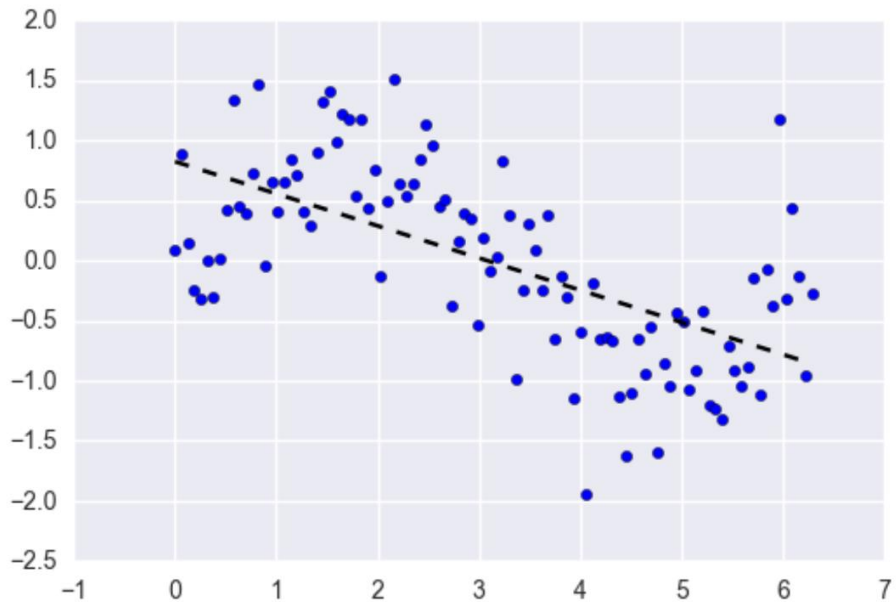

Total error breaks down as:
Total Error = Bias^2 + Variance + Irreducible Error
(Irreducible error is "noise" that can't be reduced by algorithms. It can sometimes be reduced by better data cleaning)
Optimal Balance
Error
Total Error
Bias^2
Variance

Error from Bias

Bias is the difference between your model's expected predictions and the true values.

That might sound strange because shouldn't you "expect" your predictions to be close to the true values? Well, it's not always that easy because some algorithms are simply too rigid to learn complex signals from the dataset.

Imagine fitting a linear regression to a dataset that has a non-linear pattern:



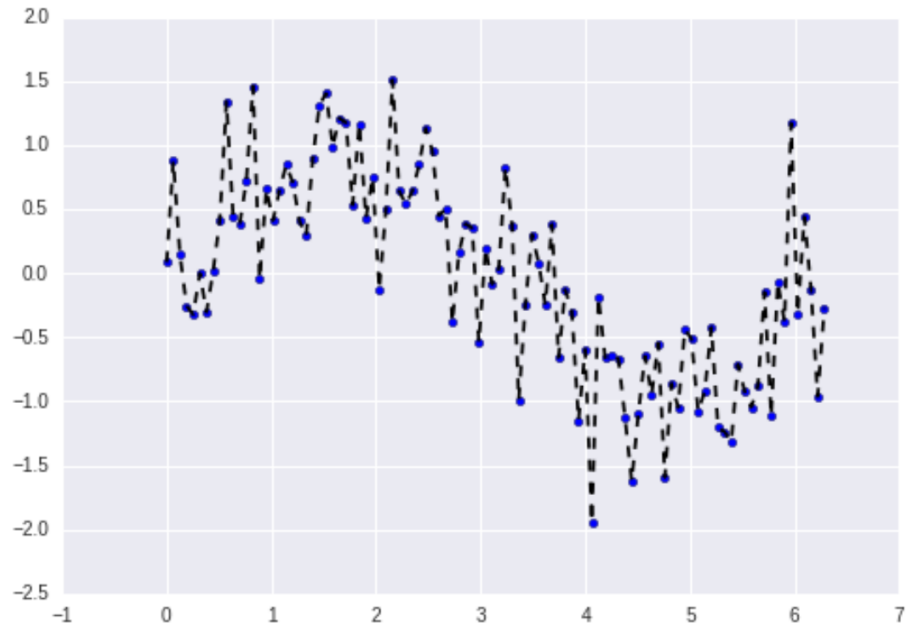
No matter how many more observations you collect, a linear regression won't be able to model the curves in that data! This is known as **under-fitting**.

Error from Variance

Variance refers to your algorithm's sensitivity to specific sets of training data.

High variance algorithms will produce drastically different models depending on the training set.

For example, imagine an algorithm that fits a completely unconstrained, super-flexible model to the same dataset from above:

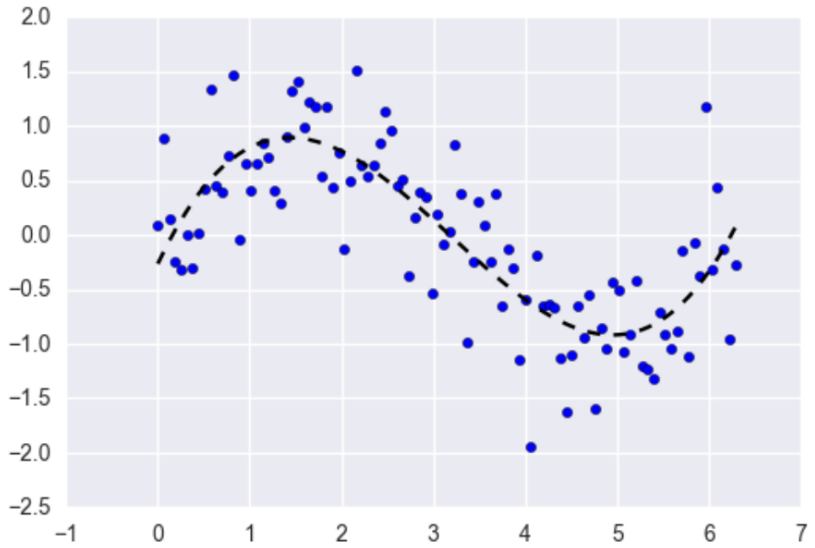


As you can see, this unconstrained model has basically memorized the training set, including all of the noise. This is known as **over-fitting**.

A proper machine learning workflow includes:

- Separate training and test sets
- Trying appropriate algorithms (No Free Lunch)
- Fitting model parameters
- Tuning impactful hyperparameters
- Proper performance metrics
- Systematic cross-validation

Finally, as you might have already concluded, an optimal balance of bias and variance leads to a model that is neither overfit nor underfit:



This is the ultimate goal of supervised machine learning - to isolate the **signal** from the dataset while ignoring the noise!