Data Science

Q1. What is Data Science? List the differences between supervised and unsupervised learning.

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. How is this different from what statisticians have been doing for years? The answer lies in the difference between explaining and predicting: statisticians work a posteriori, explaining the results and designing a plan; data scientists use historical data to make predictions.

The differences between supervised and unsupervised learning are:

Supervised	Unsupervised
Input data is labelled	Input data is unlabeled
Split in training/validation/test	No split
Used for prediction	Used for analysis
Classification and Regression	Clustering, dimension reduction, and density estimation

Q2. What is Selection Bias?

Selection bias is a kind of error that occurs when the researcher decides what has to be studied. It is associated with research where the selection of participants is not random. Therefore, some conclusions of the study may not be accurate.

The types of selection bias include:

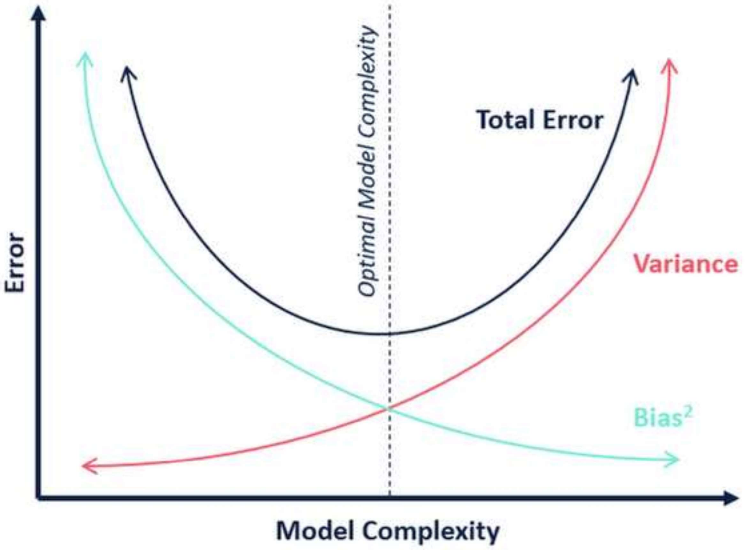
- Sampling bias: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
- Time interval: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.
- Data: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
- Attrition: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

Q3. What is bias-variance trade-off?

Bias: Bias is an error introduced in the model due to the oversimplification of the algorithm used (does not fit the data properly). It can lead to under-fitting.
Low bias machine learning algorithms — Decision Trees, k-NN and SVM
High bias machine learning algorithms — Linear Regression, Logistic Regression

Variance: Variance is error introduced in the model due to a too complex algorithm, it performs very well in the training set but poorly in the test set. It can lead to high sensitivity and overfitting.
Possible high variance – polynomial regression

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



Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.

1. The k-nearest neighbor algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbors that contribute to the prediction and in turn increases the bias of the model.
2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.
3. The decision tree has low bias and high variance, you can decrease the depth of the tree or use fewer attributes.
4. The linear regression has low variance and high bias, you can increase the number of features or use another regression that better fits the data.

There is no escaping the relationship between bias and variance in machine learning. Increasing the bias will decrease the variance. Increasing the variance will decrease bias.

Q4. What is a confusion matrix?

The confusion matrix is a 2X2 table that contains 4 outputs provided by the binary classifier.

	Predict +	Predict -
Actual +	TP	FN (II error)
Actual -	FP (I error)	TN

A data set used for performance evaluation is called a test data set. It should contain the correct labels and predicted labels. The predicted labels will exactly the same if the performance of a binary classifier is perfect. The predicted labels usually match with part of the observed labels in real-world scenarios.
A binary classifier predicts all data instances of a test data set as either positive or negative. This produces four outcomes: TP, FP, TN, FN. Basic measures derived from the confusion matrix:

1. $Error\ Rate\ = \frac{FP+FN}{P+N}$

2. $Accuracy\ = \frac{TP+T}{P+N}$

3. $Sensitivity\ (Recall\ or\ True\ positive\ rate) = \frac{TP}{TP+FN} = \frac{TP}{P}$

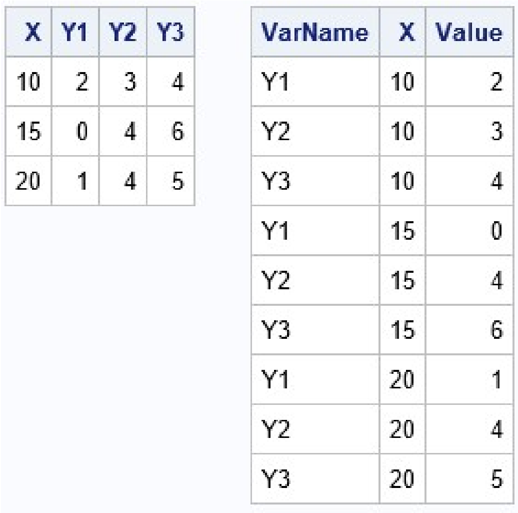
4. $Specificity\ (True\ negative\ rate) = \frac{TN}{TN+FP} = \frac{TN}{N}$

5. $Precision\ (Positive\ predicted\ value) = \frac{TP}{TP+FP}$

6. $F - Score\ (Harmonic\ mean\ of\ precision\ and\ recall) = \frac{2\,TP}{(2\,TP + FP + FN)}$

Q5. What is the difference between “long” and “wide” format data?

In the wide-format, a subject’s repeated responses will be in a single row, and each response is in a separate column. In the long-format, each row is a one-time point per subject. You can recognize data in wide format by the fact that columns generally represent groups (variables).

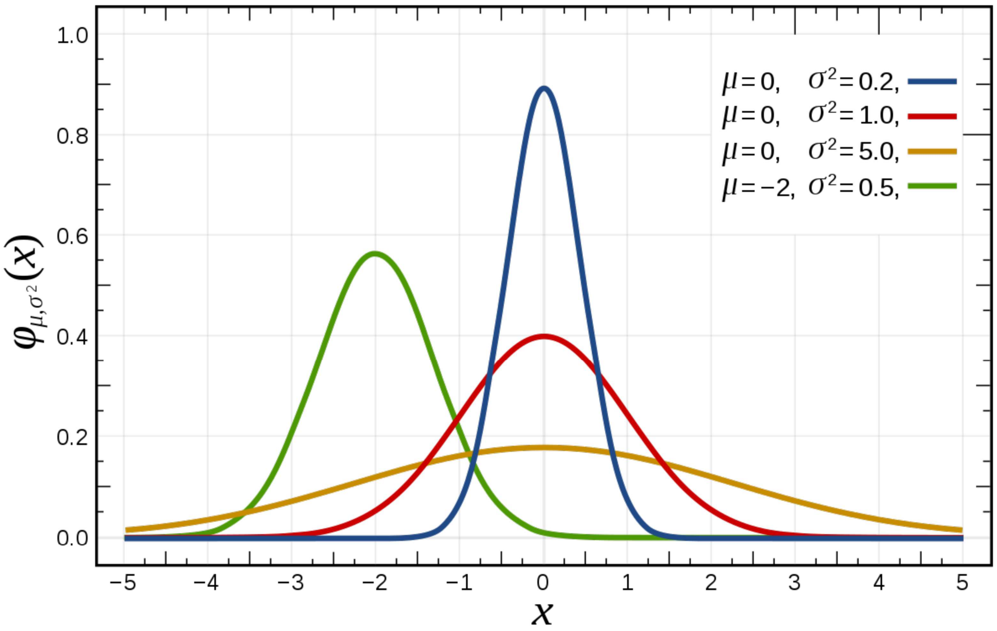


X	Y1	Y2	Y3
10	2	3	4
15	0	4	6
20	1	4	5

VarName	X	Value
Y1	10	2
Y2	10	3
Y3	10	4
Y1	15	0
Y2	15	4
Y3	15	6
Y1	20	1
Y2	20	4
Y3	20	5

Q6. What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.



The random variables are distributed in the form of a symmetrical, bell-shaped curve. Properties of Normal Distribution are as follows:

1. Unimodal (Only one mode)
2. Symmetrical (left and right halves are mirror images)
3. Bell-shaped (maximum height (mode) at the mean)
4. Mean, Mode, and Median are all located in the center
5. Asymptotic

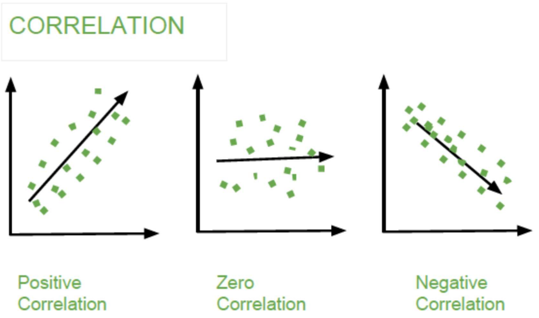
Q7. What is correlation and covariance in statistics?

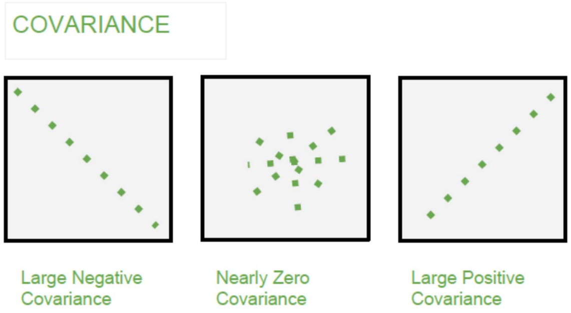
Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related. Given two random variables, it is the covariance between both divided by the product of the two standard deviations of the single variables, hence always between -1 and 1.

$$\rho = \frac{Cov(X,Y)}{\sigma(X)\,\sigma(Y)} \in [-1,1]$$

Covariance is a measure that indicates the extent to which two random variables change in cycle. It explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$





Q8. What is the difference between Point Estimates and Confidence Interval?

Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.

A confidence interval gives us a range of values which is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented by $1 - \alpha$, where α is the level of significance.

Q9. What is the goal of A/B Testing?

It is a hypothesis testing for a randomized experiment with two variables A and B.

The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business. *It can be used to test everything from website copy to sales emails to search ads. An example of this could be identifying the click-through rate for a banner ad.*

Q10. What is p-value?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is the minimum significance level at which you can reject the null hypothesis. The lower the p-value, the more likely you reject the null hypothesis.

Q11. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?

- $Probability\ of\ not\ seeing\ any\ shooting\ star\ in\ 15\ minutes\ is\ = 1 - P(Seeing\ one\ shooting\ star) = 1 - 0.2 = 0.8$
- $Probability\ of\ not\ seeing\ any\ shooting\ star\ in\ the\ period\ of\ one\ hour = (0.8)^4 = 0.4096$

- $Probability\ of\ seeing\ at\ least\ one\ shooting\ star\ in\ the\ one\ hour\ = 1 - P(Not\ seeing\ any\ star) = 1 - 0.4096 = 0.5904$

Q12. How can you generate a random number between 1 – 7 with only a die?

Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes. To get our 7 equal outcomes we have to reduce this 36 to a number divisible by 7. We can thus consider only 35 outcomes and exclude the other one. A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice. All the remaining combinations from (1,1) till (6,5) can be divided into 7 parts of 5 each. This way all the seven sets of outcomes are equally likely.

Q13. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

$$P(Having\ two\ girls\ given\ one\ girl) = \frac{1}{2}$$

Q14. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?

There are two ways of choosing the coin. One is to pick a fair coin and the other is to pick the one with two heads.

$$Probability\ of\ selecting\ fair\ coin\ = \frac{999}{1000} = 0.999$$

$$Probability\ of\ selecting\ unfair\ coin\ = \frac{1}{1000} = 0.001$$

$$\begin{aligned} &Selecting\ 10\ heads\ in\ a\ row \\ &\quad = Selecting\ fair\ coin\ * Getting\ 10\ heads\ + Selecting\ unfair\ coin \\ &\quad = P(A) + P(B) \end{aligned}$$

$$P(A) = 0.999 * \left(\frac{1}{2}\right)^{10} = 0.999 * \left(\frac{1}{1024}\right) = 0.000976$$

$$P(B) = 0.001 * 1 = 0.001$$

$$\frac{P(A)}{P(A) + P(B)} = \frac{0.000976}{0.000976 + 0.001} = 0.4939$$

$$\frac{P(B)}{P(A) + P(B)} = \frac{0.001}{0.001976} = 0.5061$$

$$Probability\ of\ selecting\ another\ head\ = \frac{P(A)}{P(A)+P(B)} * 0.5\ +\ \frac{P(B)}{P(A)+P(B)} * 1 =$$
$$= 0.4939\ *\ 0.5\ +\ 0.5061\ =\ 0.7531$$

Q15. What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.).

$$Sensitivity\ = \frac{TP}{TP+FN}$$

Q16. Why is Re-sampling done?

https://machinelearningmastery.com/statistical-sampling-and-resampling/

- Sampling is an active process of gathering observations with the intent of estimating a population variable.
- Resampling is a methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter. Resampling methods, in fact, make use of a nested resampling method.

Once we have a data sample, it can be used to estimate the population parameter. The problem is that we only have a single estimate of the population parameter, with little idea of the variability or uncertainty in the estimate. One way to address this is by estimating the population parameter multiple times from our data sample. This is called resampling. Statistical resampling methods are procedures that describe how to economically use available data to estimate a population parameter. The result can be both a more accurate estimate of the parameter (such as taking the mean of the estimates) and a quantification of the uncertainty of the estimate (such as adding a confidence interval).

Resampling methods are very easy to use, requiring little mathematical knowledge. A downside of the methods is that they can be computationally very expensive, requiring tens, hundreds, or even thousands of resamples in order to develop a robust estimate of the population parameter.

The key idea is to resample form the original data — either directly or via a fitted model — to create replicate datasets, from which the variability of the quantiles of interest can be assessed without long-winded and error-prone analytical calculation. Because this approach involves repeating the original data analysis procedure with many replicate sets of data, these are sometimes called computer-intensive methods. Each new subsample from the original data sample is used to estimate the population parameter. The sample of estimated population parameters can then be considered with statistical tools in order to quantify the expected value and variance, providing measures of the uncertainty of the estimate. Statistical sampling methods can be used in the selection of a subsample from the original sample.

A key difference is that process must be repeated multiple times. The problem with this is that there will be some relationship between the samples as observations that will be shared across multiple subsamples. This means that the subsamples and the estimated population parameters are not strictly identical and independently distributed. This has implications for statistical tests performed on the sample of estimated population parameters downstream, i.e. paired statistical tests may be required.

Two commonly used resampling methods that you may encounter are k-fold cross-validation and the bootstrap.

- Bootstrap. Samples are drawn from the dataset with replacement (allowing the same sample to appear more than once in the sample), where those instances not drawn into the data sample may be used for the test set.
- k-fold Cross-Validation. A dataset is partitioned into k groups, where each group is given the opportunity of being used as a held out test set leaving the remaining groups as the training set. The k-fold cross-validation method specifically lends itself to use in the evaluation of predictive models that are repeatedly trained on one subset of the data and evaluated on a second held-out subset of the data.

Resampling is done in any of these cases:

- Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points
- Substituting labels on data points when performing significance tests
- Validating models by using random subsets (bootstrapping, cross-validation)

Q17. What are the differences between over-fitting and under-fitting?

In statistics and machine learning, one of the most common tasks is to fit a model to a set of training data, so as to be able to make reliable predictions on general untrained data.

In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfitted, has poor predictive performance, as it overreacts to minor fluctuations in the training data.

Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

Q18. How to combat Overfitting and Underfitting?

To combat overfitting:

1. Add noise
2. Feature selection
3. Increase training set

4. L2 (ridge) or L1 (lasso) regularization; L1 drops weights, L2 no
5. Use cross-validation techniques, such as k folds cross-validation
6. Boosting and bagging

7. Dropout technique

8. Perform early stopping

9. Remove inner layers

To combat underfitting:

1. Add features
2. Increase time of training

Q19. What is regularization? Why is it useful?

Regularization is the process of adding tuning parameter (penalty term) to a model to induce smoothness in order to prevent overfitting. This is most often done by adding a constant multiple to an existing weight vector. This constant is often the L1 (Lasso - $|\alpha|$) or L2 (Ridge - α^2). The model predictions should then minimize the loss function calculated on the regularized training set.

Q20. What Is the Law of Large Numbers?

It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of frequency-style thinking. It says that the sample means, the sample variance and the sample standard deviation converge to what they are trying to estimate. According to the law, the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed.

Q21. What Are Confounding Variables?

In statistics, a confounder is a variable that influences both the dependent variable and independent variable.

If you are researching whether a lack of exercise leads to weight gain:
lack of exercise = independent variable
weight gain = dependent variable
A confounding variable here would be any other variable that affects both of these variables, such as the age of the subject.

Q22. What Are the Types of Biases That Can Occur During Sampling?

a. Selection bias
b. Under coverage bias
c. Survivorship bias

Q23. What is Survivorship Bias?

It is the logical error of focusing aspects that support surviving some process and casually overlooking those that did not work because of their lack of prominence. This can lead to wrong conclusions in numerous different means. *For example, during a recession you look just at the survived businesses, noting* *that they are performing poorly. However, they perform better than the rest, which is failed, thus being removed from the time series.*

Q24. What is Selection Bias? What is under coverage bias?

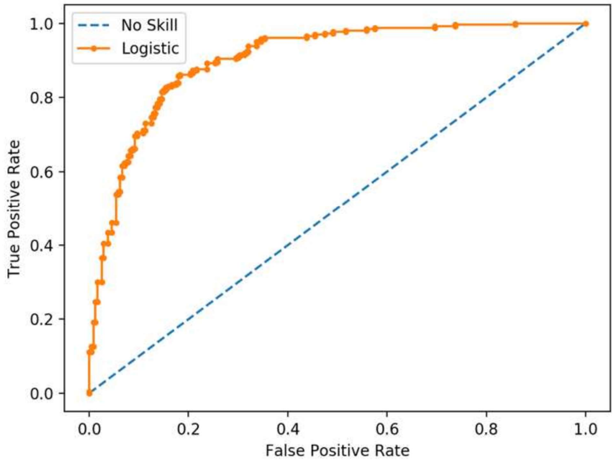
https://stattrek.com/survey-research/survey-bias.aspx

Selection bias occurs when the sample obtained is not representative of the population intended to be analyzed. *For instance, you select only Asians to perform a study on the world population height.*
Under coverage bias occurs when some members of the population are inadequately represented in the sample. *A classic example of under coverage is the Literary Digest voter survey, which predicted that Alfred Landon would beat Franklin Roosevelt in the 1936 presidential election. The survey sample suffered from under coverage of low-income voters, who tended to be Democrats.*
How did this happen? The survey relied on a convenience sample, drawn from telephone directories and car registration lists. In 1936, people who owned cars and telephones tended to be more affluent. Under coverage is often a problem with convenience samples.

Q25. Explain how a ROC curve works?

The ROC curve is a graphical representation of the contrast between true positive rates and false positive rates at various thresholds. It is often used as a proxy for the trade-off between the sensitivity (true positive rate) and false positive rate.

- $TPR\ = \frac{TP}{P} = \frac{TP}{TP+FN}$
- $TNR\ = \frac{TN}{TN+FP} = \frac{TN}{N}$
- $FPR\ = \frac{FP}{TN+FP}$
- $FNR = \frac{FN}{FN+T}$



Q26. What is TF/IDF vectorization?

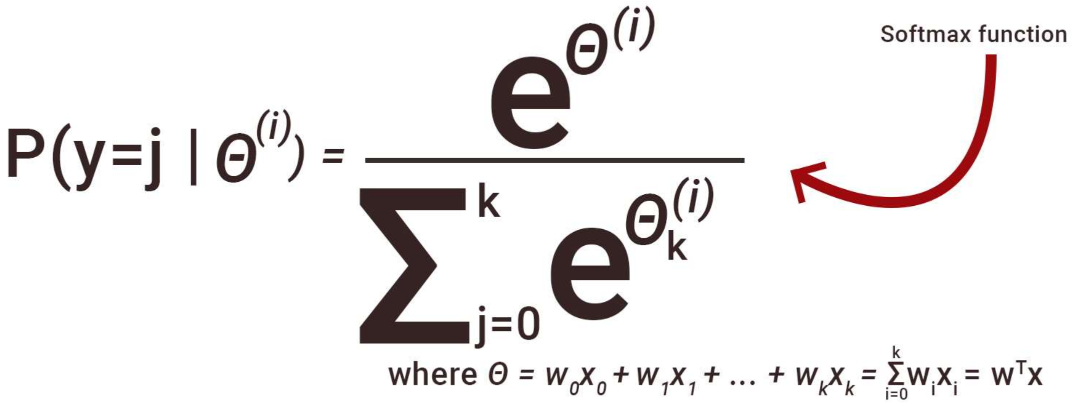
TF-IDF is short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

- $TF = \frac{\#\,\text{'word' in doc}}{\text{tot}\ \#\ \text{words in doc}}$
- $IDF = \log\left(\frac{\#\ \text{docs with 'word' in it}}{\text{tot docs in collection}}\right)$

The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Q27. Why we generally use Soft-max (or sigmoid) non-linearity function as last operation in-network? Why RELU in an inner layer?

It is because it takes in a vector of real numbers and returns a probability distribution. Its definition is as follows. Let x be a vector of real numbers (positive, negative, whatever, there are no constraints). Then the i-eth component of soft-max(x) is:


$P(y=j \mid \theta^{(i)}) = \frac{e^{\theta^{(i)}}}{\sum_{j=0}^{k} e^{\theta_k^{(i)}}}$
Softmax function
where $\theta = w_0x_0 + w_1x_1 + \ldots + w_kx_k = \sum_{i=0}^{k} w_ix_i = w^Tx$

It should be clear that the output is a probability distribution: each element is non-negative and the sum over all components is 1.

RELU because it avoids the vanishing gradient descent issue.