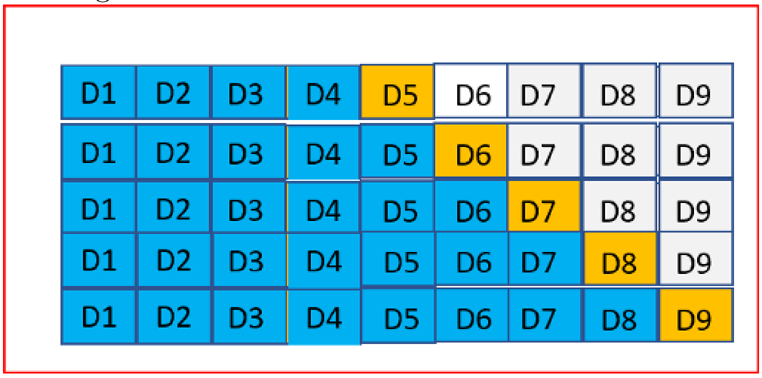
Springboard - Blog

https://www.springboard.com/blog/machine-learning-interview-questions/

1. Bias vs Variance
   - **Bias** is due to erroneus or overly simplistic assumptions in learning algorithm
   - Usually underfitting your data
   - **Variance** typically due to too much complexity in learning algorithm
   - Makes the model sensitive to high degrees of variation in training data.
   - Too much noise from training data
   - If you make data more complex and add more variables, you'll lose bias but gain variance.
2. Supervised vs Unsupervised learning
   - Supervised requires labeled data. Unsupervised does not.
3. How is KNN different from k-means clustering?
   - KNN is a supervised classification algorithm.
   - K-means clustering is unsupervised.
   - Works very similarly
   - KNN required labelled data
   - K means clustering requires only a set of unlabeled point and a threshold
   - The algorithm will gradually *learn* how to cluster them by computing mean of the distance between different points.
4. How does a ROC curve work
   - graphical representation of constrast between true and false positive rate at various threholds.
   - Used as a proxy for trade-off between sensitivity of model (true positive) vs the fall-out or probability it will trigger a false alarm (false positives)
   - Think about recall and precision in this case.
     - *ex.* You'd have perfect recall (there are actually 10 apples, and you predicted there would be 10) but 66.7% precision because out of the 15 events you precited, only 10 (the apples) are correct.
5. Baye's Theorem?
   - $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
   - Leads to a branch of ML called Naive Bayes classifier
6. Why is "Naive" Bayes naive?
   - Used a lot in text mining
   - It's naive because it makes an assumption that is virtually impossible in real-life data.
     - conditional probability is calculated as the pure product of the individial probabilities of components.
     - This implies absolute independence of features - condition probably never met in real life.
   - Anohter way put, if a Naive Bayes classifier figured that you liked pickles and ice-cream would probably naively recommend you a pickle ice-cream.

7. Difference between L1 and L2 regularization
    - Regularization helps solve over-fitting problems in ML
    - Simple model will be very poor generalization of data.
    - Complex model may not perform well in test due to over-fitting.
    - Regulatization refers to adding a penalty term to objective function and control model complexity using that penalty term.
    - Ridge regression used L_2 norm for regularization.
        - Ridge regression is used to analyze techniques that suffer from multicollinearity.
        - When multicollinearity occurs, least squares estimates are unbiased, but variances are large so they might be far from true value.
    - L2 forces weights to be small but does not make them zero and does non sparse solution
    - Lasso regression uses L1 regression
        - Type of linear regression that uses shrinkage.
        - Shrinkage is where data values are *shrunk* towards a central point, like the mean.
        - Creates simple sparse models (models with fewer parameters)
    - Main difference between L1 and L2 is the difference between the constraint region.
8. Favorite algorithm
    - Optical flow is pattern of apparent motion of objects, surfaces, and edges caused by the relative motion between an observer and a scene.
    - Optical flow methods try to calculate motion between between two images which are taken at times t and $t + \Delta t$ at every voxel position.
    - This sort of method are called differential since they are based on local Taylor Series.
    - They use partial derivatives with respect to spatial and temporal coordinates.
9. Difference between Type I and Type II error?
    - Type I error is rejection of true null hypothesis ("false" positive)
    - Type II error is not rejective the false null hypothesis
10. Fourier Transform
    - Converts signal from time to frequency domain
    - Decomposes generic function into a superposition of symmetric function.
11. Probability vs Likelihood [Need to expand]
    - Probability attaches to possible results
    - Likelihood attaches to hypothesis
12. Difference between generative and discriminative model
    - Generative model will learn categories of data
    - Discriminative will simply learn the different categories of data.
    - Discriminative will generally outperform generative models on classification tasks.
13. What CV technique would you use on time-series dataset
    - Time series is not randomly distributed but but chronologically ordered.
    - Forward chaining is one way of doing this

    - In time series CV each day is a test data and the previous day's data is the training set.



    - We start training the model with a minimum number of observations and use the next day's data to test the model and we keep moving through the data set.

14. Decision tree pruned

    - Pruning is what happens in decision trees when branches that have weak predictive power are remove.
    - This reduces model complexity and increases predictive accuracy of a decision tree model.
    - Reduced error pruning is simplest version
        - replace each
        - if it doesn't decrease predictive accuracy, keep it pruned.
    - What is more important to you - model accuracy or model performance?
        - Accuracy paradox
        - A simple model may have a high level of accuracy but be too crude to be useful
        - If you have a large dataset to detect fraud and majority of it are not-fraud data points, the probability of the model detecting fraud is a lot lower.
        - model designed to find fraud that asserted there was no fraud at all.

15. F1 score

    - F1 score is measure of model's performance
    - Weighted average of precision and recall of a model, with results tending to 1 being the best and those tending to 0 being the worst.
    - Used in classification tests where true negatives don't matter much.

16. How to handle an imbalanced dataset?

    - Imbalanced dataset is when 90% of the data is in one class
    - this accuracy can lead to a skewed of one prediction class
    - These are ways to get over the hump
        - Collect more data to even the imbalances in the dataset
        - Resample dataset to correct for imbalances
        - Try a different algorithm on your dataset

17. When should you use classification over regression?

    - Classification produces discrete values.
    - Regression is continuous
    - Seeing if a name is male or female vs how correlated they were with male and female names.

18. Example where ensemble techniques might be useful

- Ensemble technique uses a combination of learning algorithms to optimize better predictive performance.
- Typically reduces overfitting in models and and makes model more robust (unlikely to be influence by small changes in the training data).
- *ex.* bagging

Building a recommendation system

1. Simply put, the engine computes the co-occurrence matrix from a history matrix of events and actions.
2. Then, we have to apply statistics to filter out the sufficiently anomalous signals to be interesting as a recommendation.
3. We need to be able to extract **relevant indicators** from the co-occurrence matrix. This is what makes a good recommendation system.
4. Creating an item-to-item indicator matrix is called an **item-item model**. Another one is a **user-item model**.

n-item and user-item

5. To create user-item model we could apply a simple matrix factorisation.
   - Class of collaborative filtering
   - Works by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices.
6. Another way is to train a neural network to predict the scores of a user-item input.
7. Usually item-item models are more robust and produce better results when we don't invest into feature engineering and model tuning
8. Recommending same things over and over is bad and produces **content fatigue**
9. There are two ways to improve the values:
   - Anti-Flood: Penalize the second and third recommendations if they have the same similarity scores to the top recommendation.
   - Dithering: Add a wildcard recommendation to create interesting new data points for the recommendation to keep learning about other content.