TOOL

Interview Prep Guide

In your weekly written assignments, you tackled individual questions that required you to recall or define machine learning concepts. Now, it's time to expand on this knowledge. The *Big Picture Questions* in this *Interview Prep Guide*, are designed to push your thinking further, asking you to explore these concepts in greater depth and a broader context. These questions aren't just a review—they're an opportunity to enhance your initial work by connecting ideas, weighing tradeoffs, and thinking like a practitioner.

Each Big Picture Question is modeled after common machine learning interview prompts. In a typical ML interview, you'll be asked to demonstrate your understanding of core concepts, explain tradeoffs, and apply your knowledge to real-world projects you've worked on. Completing this guide will help you solidify your mastery of key content areas and prepare you to confidently answer questions you're likely to encounter in interviews.

During lab sessions, you'll collaborate with peers to refine your understanding and co-create clear, well-reasoned responses to complex, real-world problems. This collaborative process mirrors the kind of teamwork and problem-solving you'll encounter in professional settings—and gives you the chance to learn from a range of perspectives. Our goal is for you to walk away with accurate, concise responses that double as a personalized "cheat sheet" for future interviews.

You may not get to every question in this guide—and that's OK. Think of the remaining questions as invitations for continued learning and exploration. Use them to deepen your understanding outside of lab and revisit them as part of your long-term interview prep.

This guide is downloadable and fillable, so you can access and update it as needed—whether you're in lab, preparing for an interview, or simply revisiting your learning.

Unit 1

1. Explain the distinction between supervised and unsupervised learning. Provide an example of a real-world problem for each type, highlighting why each approach is suitable.

2. Walk us through the key steps in the machine learning lifecycle, starting from problem formulation and ending with model deployment. Emphasize the significance of each stage in the process.

3. When selecting packages for a machine learning project, what factors would you consider, and how do you assess whether a particular package aligns with your project's specific requirements? Provide practical examples or experiences if possible.

4. Illustrate the difference between classification and regression problems by providing examples for each. Discuss the key distinctions in their objectives and how the choice between them depends on the nature of the problem.

5. Provide an example of a real-world problem where machine learning could be applied to support decision-making in an industry? Discuss the key characteristics of the problem and how machine learning might be leveraged to address it.

1. Why is data preparation so important to the machine learning development process?

2. Discuss why data visualization is a critical aspect of the data preparation process. Provide an example of how visualizing data can reveal insights that impact the subsequent steps of a machine learning project.

3. Define what an outlier is and explain how outliers can affect the outcomes of a machine learning model. What strategies can be employed to handle outliers during the data preparation phase?

4. Name a few libraries used for data analysis and visualization and explain when you would use each library.

1. Describe one advantage and one disadvantage of using the k-nearest neighbors algorithm for a movie recommendation system.

2. Explain the difference between k-nearest neighbors and decision trees. When would you decide to use one over the other?

3. Define hyperparameters in the context of machine learning models. Provide one example of a hyperparameter for both k-nearest neighbors and decision trees, explaining how adjusting these could affect the model's predictions.

4. Explain overfitting using decision trees as an example. What is one strategy you could use to prevent overfitting in decision trees?

5. Why do machine learning practitioners split data into training and testing sets? Discuss how this practice might impact the performance of a model.

Unit 4

1	Why is it good or useful to estimate the probability of something occurring?	It helps us to make more nuanced predictions. Probabilities allow the model to choose the action that has the highest expected benefit. We can also rank outcomes by how likely they are and allow us to weigh the risks of something occurring.
2	What factors or characteristics of the dataset influence the selection between logistic regression, decision trees and KNN?	If your data is more so simple and has a straight-line pattern, logistic regression works well. decision trees are better if things are more messy or non-linear. KNN just looks at what ' s close, so it needs clean, scaled data and isn ' t great with big datasets. trees are more forgiving, logistic needs prep, and KNN can get slow fast
3	In which scenarios would logistic regression be more suitable or preferable compared to a decision tree or KNN?	Depends on the data type you have. Linear datatype is more well adapted to logistic regression to handle a linear relationship. DT and KNNs are more in terms of nonlinear relationships. Logistic regression is more simple and requires less time to compute. Also useful if you care about probabilities of an event occurring.
4	Identify one key reason why logistic regression is suitable for predicting customer churn. Provide a brief example to illustrate your point.	
5	Explain one key role of the loss function in logistic regression.	
6	Describe how regularization helps prevent overfitting in a logistic regression model	

Unit 5

1. Why do we place significant importance on addressing overfitting and achieving generalization in machine learning?

2. Can you identify specific applications where minimizing false negatives is more crucial than minimizing false positives? Additionally, what evaluation metric would be appropriate in such situations?

3. What are some reasons why we might consider performing feature selection in machine learning? Imagine you have a dataset with information about various movies, such as genre, director, budget, and box office revenue. Explain how you would approach feature selection and why it might be useful in predicting a movie's success.

4. What is cross-validation and what is the benefit of performing cross-validation?

5. Explain model selection in the context of building a simple app that predicts whether a given day's weather will be rainy or sunny.

6. Describe out-of-sample validation as if you were explaining it to someone who has never studied machine learning. How does this method help in ensuring that a model performs well in real-life situations beyond just the classroom?

1. What are some real-world applications of ensemble learning?

2. When would it not be appropriate or optimal to use an ensemble model like Random Forest or GBDT?

3. Why is model stacking effective in improving performance?

4. Provide a brief explanation of how ensemble modeling could improve prediction accuracy in online advertising.

5. Using a simple example of predicting exam scores from study hours, explain the concept of bias and variance. What does the bias-variance tradeoff imply for model performance?

6. Choose one of the ensemble methods—bagging, boosting, or stacking—and explain how it helps reduce prediction error in weather forecasting.

1. What are some limitations/common challenges faced when training neural networks for computer vision tasks?

2. Briefly define deep learning. Mention one real-world application of deep learning that has significantly impacted the tech industry. Why is deep learning effective in this application?

3. How does a neural network differ from logistic regression in handling complex data patterns? Provide one advantage of using neural networks over logistic regression in voice recognition technology.

4. Outline the key steps involved in training a traditional neural network. How does the training process ensure that the network learns to predict accurately?

5. Explain why traditional neural networks are less effective for image data and why a specific architecture, like convolutional neural networks, is necessary. What does this architecture uniquely offer for processing images?

6. What are the key differences between traditional neural networks and convolutional neural networks in terms of structure and functionality? How does this difference make convolutional networks more suited for tasks like image classification?

Unit 8

1. Define Natural Language Processing (NLP) in simple terms. Identify one impactful real-world application of NLP and briefly explain how it utilizes NLP technologies.

2. Why is preprocessing of text data important in NLP? Give two examples of preprocessing techniques used in the NLP pipeline and explain their purpose.

3. Explain the concept of TF-IDF and its role in NLP. How is TF-IDF calculated for a document in a corpus? Illustrate with a simple example involving a few short texts.

4. How do vectorizers differ from word embeddings in representing text data? Mention one advantage of using word embeddings over vectorizers.

5. Compare a traditional neural network to a sequence-to-sequence model. Why are sequence-to-sequence models particularly suited for translation tasks in NLP? Briefly describe the key components of a sequence-to-sequence model.

6. Describe what a large language model (LLM) is and why it is needed.

1. Why is it crucial to have a well-defined problem before diving into the model development phase?

2. How would you approach a situation where a client presents you with a vague problem statement? What steps would you take to refine and clarify the problem?

3. How do you handle situations where the problem statement provided by stakeholders conflicts with available data or feasibility constraints? Provide examples of how you would navigate such scenarios.

4. Why is it essential to evaluate model performance using appropriate metrics? Discuss the importance of selecting evaluation metrics based on the problem domain and the potential pitfalls of relying solely on accuracy.

5. Why is ethical AI important? Provide a few real-world examples of algorithmic fairness failures.

6. How would you incorporate ethical considerations into the design and development of AI systems to maintain algorithmic fairness? Can you discuss any frameworks or guidelines you would follow?