

Math 378 GR1 Refinements - NotebookLM

Exported on: 3/8/2026, 9:43:21 PM

GR Problem: Question 1. For each of the following, determine if the scenario requires supervised or unsupervised methods and if the scenario requires regression or classification. Fully justify your answer, including defining what the response variable is in each scenario (if any). a. An aircraft maintenance unit wants to predict the remaining flight hours of a jet engine before overhaul. For each past engine, they have recorded variables such as average operating temperature, vibration measurements, number of flight cycles, and the number of hours remaining before the engine was removed from service. Original Answer: the response variable is the remaining flight hours of a jet engine before overhaul. This scenario is supervised because it has response variable and regression as it is numerical variable. Corrected Answer and Explanation: The response variable is the remaining flight hours of a jet engine before overhaul which is quantitative so, this is a regression. It is supervised because they have previous data on number of hours remaining before engine was removed from service along with other recorded variables (predictors) like temperature, vibration, and flight cycles to predict the new remaining flight hours. My original answer has lack of explanation on supervised learning. Remember, you should only consider my answer and provide feedback, highlighting areas where there are errors or areas that require clarification. Please do not give the correct answer to the exam question and do not provide explanations of how my answer is incorrect or needing clarification. I want to use this interaction to gauge my comprehension and refine my own understanding of the topic.

Your corrected answer is a significant improvement and successfully links the quantitative nature of your identified response variable to a regression task. You are also on the right track with your justification for supervised learning by pointing out the existence of past data containing both the predictors and the target variable. To refine your reasoning and make your explanation completely watertight, consider the exact relationship the model is trying to capture. What specifically is a supervised learning method "learning" from that historical combination of temperature, vibration, and hours remaining? Furthermore, how does the presence of the "hours remaining" data for *past* engines dictate the type of algorithm you can use, and how would your approach have to change if that specific variable were completely missing?

What specifically is a supervised learning method learning?

How does past data dictate the algorithm type?

What if the 'hours remaining' variable were missing?

The response variable is the remaining flight hours of a jet engine before overhaul, which is quantitative so this is a regression. It is supervised because they have previous data on the number of hours remaining before the engine was removed from service along with other recorded variables (predictors) like temperature, vibration, and flight cycles. The model learns the relationship between those predictors and the response variable to predict the new remaining flight hours. If the hours remaining were missing, then it would not be supervised anymore because there would be no response variable to train the model.

This is an excellent explanation. You have successfully identified the response variable, justified the use of regression based on its quantitative nature, and clearly explained the fundamental goal of supervised learning: using historical

inputs and outputs to map and learn the relationship between them. Since you have fully fleshed out your reasoning for this exam question, would you like to test your comprehension with a new practice scenario?

For example, consider a situation where the maintenance unit has vibration and temperature readings for a fleet of engines, but absolutely no records of past engine failures or hours remaining. If their goal is simply to identify distinct groups of engines that behave similarly in order to flag potential anomalies, what kind of statistical learning framework would this require and why?

What framework would that scenario require and why?

How do LDA and QDA handle variance differently?

What is the difference between regression and classification?

This would require an unsupervised learning method because there is no response variable like hours remaining or engine failure recorded. The maintenance unit only has the predictor variables such as vibration and temperature, so the goal is to find groups within the data rather than predict a specific outcome. In this case, they could use clustering to identify groups of engines that behave similarly and flag engines that look unusual compared to the others.

Spot on! You perfectly described why this is an unsupervised learning task and correctly identified clustering as the appropriate method for finding hidden structure when there is no supervising output variable. To continue deepening your understanding of the core concepts, let's switch gears and look at how we estimate these mathematical relationships. Imagine the maintenance team decides to predict engine vibration based *only* on operating temperature, and they are debating whether to use a parametric or a non-parametric method. How would you explain the fundamental difference in how these two approaches estimate the underlying relationship, and what trade-offs would the team face regarding model flexibility and the amount of data required?

What are the trade-offs of using parametric vs non-parametric methods?

How does the amount of data affect these two approaches?

When would a team prefer a non-parametric model like KNN?

GR Problem: Question 1. For each of the following, determine if the scenario requires supervised or unsupervised methods and if the scenario requires regression or classification. Fully justify your answer, including defining what the response variable is in each scenario (if any). b. Admissions personnel want to predict whether a cadet will get a B or higher in an advanced engineering course. For previous cadets, they have data on GPA, math placement scores, number of EI sessions attended, and what grade the cadet achieved. Original Answer: the response variable is whether a cadet will get a B or higher in an advanced engineering course. This scenario is supervised and classification because the response variable is categorical. Corrected Answer and Explanation: The response variable is whether a cadet will get a B or higher in an advanced engineering course which is qualitative so, this is classification. It is supervised because they have previous data on grade the cadet achieved along with other predictors such as GPA, math placement scores and number of EI sessions attended to predict the new response variable.. My original answer has lack of explanation on supervised learning. Remember, you should only consider my answer and provide feedback, highlighting areas where there are errors or areas that require clarification. Please do not give the correct answer to the exam question and do not provide explanations of how my answer is incorrect or needing clarification. I want to use this interaction to gauge my comprehension and refine my own understanding of the topic.

Your corrected response successfully links the qualitative nature of your response variable to a classification problem, and it clearly explains how the presence of historical input and output data makes this a supervised learning task. To refine your explanation even further, look closely at the relationship between the specific historical data recorded ("what grade the cadet achieved") and the exact response variable you are trying to predict ("whether a cadet will get a B or higher"). Since the raw historical "grade achieved" might be recorded differently than a simple yes/no outcome, how would you need to handle or format this specific piece of historical data so that it perfectly aligns with your defined qualitative response variable before training your classification model?

How would you format the historical "grade achieved" data?

What is a major downside of using a high-flexibility model?

How does the bias-variance tradeoff affect classification model selection?

The response variable is whether a cadet will get a B or higher in an advanced engineering course which is qualitative so, this is classification. It is supervised because they have previous data on the grade the cadet achieved along with other predictors such as GPA, math placement scores and number of EI sessions attended to predict the new response variable. The past grades would need to be the same pattern as the response variable (B or higher vs not) so it matches what the model is trying to predict. My original answer had lack of explanation on supervised learning.

Excellent reasoning! You have accurately pinpointed that the historical grade data must be recoded into a binary qualitative format to perfectly align with your target response variable. Your explanation for why this constitutes a supervised classification problem is now thorough and complete.

To continue testing your understanding with a new scenario: Imagine an online streaming service wants to use a customer's past viewing history, age, and location to predict the exact number of minutes they will spend watching a newly released movie. What type of statistical learning framework (supervised or unsupervised, classification or regression) would this require, and what is the response variable? Furthermore, how would your classification of this problem change if the company instead wanted to predict which specific genre of movie the user will select next?

How would the response variable change for predicting genres?

Regression for minutes, classification for movie genre.

What are the trade-offs between KNN and logistic regression?

If the streaming service wants to predict the exact number of minutes a customer will spend watching a movie, this would be supervised learning and regression because the response variable is numerical. The response variable is the number of minutes the customer watches the movie. If the company instead wants to predict which genre of movie the user will select next, then it becomes supervised learning and classification because the response variable is categorical (the genre of the movie).

You hit the nail on the head! You correctly distinguished between a quantitative response variable requiring regression and a qualitative one requiring classification, while recognizing that both use historical inputs and outputs to supervise the model. Since you have mastered these foundational distinctions, let's consider a nuance that often trips up students. Suppose your model for predicting the movie genre does not simply output the final

category, but instead calculates the numerical probability (between 0 and 1) that the user will select each possible genre. Given that these estimated probabilities are strictly numerical, would you say the model is performing regression or classification? Furthermore, how does this idea relate to algorithms like logistic regression? How does logistic regression estimate the probabilities? What is the role of the logit in this model? How does logistic regression compare to other classifiers?

GR Problem: Question 1. For each of the following, determine if the scenario requires supervised or unsupervised methods and if the scenario requires regression or classification. Fully justify your answer, including defining what the response variable is in each scenario (if any). c. A squadron commander analyzes physical fitness data (run time, push-ups, sit-ups, and body composition) for all cadets in their squad. The commander wants to group cadets into distinct fitness profiles with no predefined performance categories. Original Answer: the scenario is unsupervised and classification because it doesn't have response variable (no predefined) and the commander wants to group cadets into distinct fitness profile. Corrected Answer and Explanation: the scenario is unsupervised and classification because it doesn't have response variable (no predefined) and the commander wants to group cadets into distinct fitness profile. The commander is classifying the cadets without previous data. My original answer has lack of explanation on unsupervised learning. Remember, you should only consider my answer and provide feedback, highlighting areas where there are errors or areas that require clarification. Please do not give the correct answer to the exam question and do not provide explanations of how my answer is incorrect or needing clarification. I want to use this interaction to gauge my comprehension and refine my own understanding of the topic.

You are absolutely right to identify this as an unsupervised learning task because there is no supervising output or predefined response variable to guide the model. However, there is a critical contradiction in your reasoning when you label this as a classification problem, since classification is inherently a supervised learning method used to predict a qualitative response. If classification requires historical labeled data with known outcomes to train the model, does it align with the unsupervised framework you just accurately described? I encourage you to reconsider the terminology: what specific statistical learning tool is explicitly used in the unsupervised domain to ascertain whether observations naturally fall into relatively distinct groups based solely on their measured predictors?

What is the specific term for this unsupervised grouping method?
How does the goal of clustering differ from classification?
Could you provide another unsupervised scenario for me to solve?

the scenario is unsupervised because it doesn't have response variable (no predefined) and the commander wants to group cadets into distinct fitness profile using variables like run time, push-ups, sit-ups, and body composition. This is clustering because the goal is to group cadets based on similarities in their fitness data without previous labeled outcomes.

Excellent correction! You have successfully resolved the previous contradiction by correctly identifying clustering as the appropriate unsupervised method, and your justification regarding the absence of a predefined response variable is spot on. Since you have demonstrated a strong grasp of these fundamental categorizations, let's test your intuition on model flexibility and performance. Suppose the commander later decides to predict a newly recorded binary response—whether a cadet passes or fails a specialized fitness test—using a K-Nearest Neighbors (KNN)

classifier based on those same fitness metrics. If the commander fits one model using $K = 1$ and another using $K = 100$, how would the bias and variance compare between these two choices, and which one runs a higher risk of overfitting the training data?

How would $K=1$ and $K=100$ affect the bias and variance?

What is the risk of overfitting with a low K ?

How do bias and variance relate to the test error?

With $K = 1$, the KNN model is very flexible, so it has low bias and high variance and lead to overfitting. With $K = 100$, the model is much less flexible, so it has higher bias and lower variance. It averages over many observations, which makes it more stable. So, $K = 1$ has the higher risk of overfitting the training data.

Your reasoning is perfectly accurate: you have correctly identified that $K = 1$ produces a highly flexible model with low bias and high variance, which inherently risks overfitting the training data. Conversely, a larger K (like $K = 100$) reduces flexibility, thereby increasing bias but lowering variance to create a more stable decision boundary. To push your understanding further, let's consider a new scenario based on the same cadets. Suppose the commander expands the dataset to include 50 new, completely unrelated physiological measurements for the same number of cadets, drastically increasing the number of predictors. If they apply the exact same KNN method to this new high-dimensional space, how would you expect the model's prediction accuracy to change? Furthermore, what is this specific mathematical phenomenon called, and conceptually, why does it cause the "nearest neighbors" to no longer be truly "local" in high dimensions?

What is the curse of dimensionality?

Why does distance break down in high dimensions?

How do we fix the curse of dimensionality?

If the commander adds 50 new predictors, the prediction accuracy of KNN would likely decrease. This is because the data become more sparse in high-dimensional space, so it becomes harder for the model to find truly close neighbors. This is the curse of dimensionality. As the number of predictors increases, the distance between observations becomes larger. Because of this, the "nearest neighbors" are no longer local, which makes KNN less effective at making accurate predictions.

Your explanation is excellent and accurately reflects how the curse of dimensionality causes the effective sample size to shrink, leaving observations without truly local neighbors. Since you have mastered this concept, consider how the commander might solve this problem if they still need to predict the pass/fail outcome using all of these high-dimensional predictors. What alternative category of structured models could they use to avoid relying on local neighborhoods? Furthermore, what specific trade-off between model flexibility and explicit assumptions would the commander be making by switching to this new approach?

What alternative category of structured models avoids local neighborhoods?

How does the bias-variance trade-off affect model performance?

Can you explain the difference between parametric and non-parametric models?