Exam Professional Machine Learning Engineer All Questions

View all questions & answers for the Professional Machine Learning Engineer exam

Go to Exam

EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 115 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 115
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You are building an ML model to predict trends in the stock market based on a wide range of factors. While exploring the data, you notice that some features have a large range. You want to ensure that the features with the largest magnitude don't overfit the model. What should you do?

A. Standardize the data by transforming it with a logarithmic function.

B. Apply a principal component analysis (PCA) to minimize the effect of any particular feature.

C. Use a binning strategy to replace the magnitude of each feature with the appropriate bin number.

D. Normalize the data by scaling it to have values between 0 and 1.

Show Suggested Answer

by mymy9418 at *Dec. 18, 2022, 2:35 a.m.*

Comments

Type your comment...

Submit

fitri001 6 months, 1 week ago

Selected Answer: D

D. Normalize the data by scaling it to have values between 0 and 1 (Min-Max scaling): This technique ensures all features contribute proportionally to the model's learning process.

pen_spark
expand_more It prevents features with a larger magnitude from dominating the model and reduces the risk of overfitting.expand_more

upvoted 4 times

fitri001 6 months, 1 week ago

A. Standardize the data by transforming it with a logarithmic function: While logarithmic transformation can help compress the range of skewed features, it might not be suitable for all features, and it can introduce non-linear relationships that might not be ideal for all machine learning algorithms.

B. Apply a principal component analysis (PCA) to minimize the effect of any particular feature: PCA is a dimensionality reduction technique that can be useful, but its primary function is to reduce the number of features, not specifically address differences in feature scales.

C. Use a binning strategy to replace the magnitude of each feature with the appropriate bin number: Binning can introduce information loss and might not capture the nuances within each bin, potentially affecting the model's accuracy.

upvoted 1 times

gscharly 6 months, 2 weeks ago

Selected Answer: D

agree with pico

upvoted 1 times

pico 11 months, 3 weeks ago

Selected Answer: D

Not A because a logarithmic transformation may be appropriate for data with a skewed distribution, but it doesn't necessarily address the issue of features having different scales.

upvoted 4 times

Krish6488 12 months ago

Selected Answer: D

Features with a larger magnitude might still dominate after a log transformation if the range of values is significantly different from other features. Scaling is better, will go with Option D

upvoted 1 times

envest 1 year, 3 months ago

by abylead: Min-Max scaling is a popular technique for normalizing stock price data. Logs are commonly used in finance to normalize relative data, such as returns.https://itadviser.dev/stock-market-data-normalization-for-time-series/

upvoted 1 times

[Removed] 1 year, 3 months ago

Selected Answer: D

The correct answer is D. Min-max scaling will render all variables comparable by bringing them to a common ground.

A is wrong for the following reasons:
1. It is never mentioned that all variables are positive. If some columns have negative values, log transformation is not applicable.
2. Log transformation of variables having small positive values (close to 0) will increase their magnitude. For example, ln(0.0001) = -9.2, which will increase this variable's effect considerably.

upvoted 2 times

djo06 1 year, 3 months ago

Selected Answer: D

D is the right answer

upvoted 1 times

NickHapton 1 year, 4 months ago

go for D, z-score. This question doesn't mention outlier, just large range.
reason why not log transformation:
log transformation is more suitable for addressing skewed distributions and reducing the impact of outliers. It compresses the range of values, especially for features with a large dynamic range. While it can help normalize the distribution, it doesn't directly address the issue of feature magnitude overpowering the model.

upvoted 1 times

SamuelTsch 1 year, 4 months ago

Selected Answer: A

Selected Answer: A

From my point of view, log transformation is more tolerant to outliers. Thus, went to A.

upvoted 1 times

tavva_prudhvi 1 year, 3 months ago

n cases where the data has significant skewness or a large number of outliers, option A (log transformation) might be more suitable. However, if the primary concern is to equalize the influence of features with different magnitudes and the data is not heavily skewed or has few outliers, option D (normalizing the data) would be more appropriate.

upvoted 1 times

coolmenthol 1 year, 4 months ago

Selected Answer: A

See https://developers.google.com/machine-learning/data-prep/transform/normalization

upvoted 2 times

Antmal 1 year, 5 months ago

Selected Answer: A

A is a better option because Log transform data used when we want a heavily skewed feature to be transformed into a normal distribution as close as possible, because when you normalize data using Minimum Maximum scaler, It doesn't work well with many outliers and its prone to unexpected behaviours if values go out of the given range in the test set. It is a less popular alternative to scaling.

upvoted 1 times

tavva_prudhvi 1 year, 3 months ago

If your data is heavily skewed and has a significant number of outliers, log transformation (option A) might be a better choice. However, if your primary concern is to ensure that the features with the largest magnitudes don't overfit the model and the data does not have a significant skew or too many outliers, normalizing the data (option D) would be more appropriate.

upvoted 1 times

M25 1 year, 6 months ago

Selected Answer: D

The challenge is the “scale” (significant variations in magnitude and spread):
https://stats.stackexchange.com/questions/462380/does-data-normalization-reduce-over-fitting-when-training-a-model,
apparently largely used anyhow: https://itadviser.dev/stock-market-data-normalization-for-time-series/.

upvoted 1 times

M25 1 year, 6 months ago

Even if binning “prevents overfitting and increases the robustness of the model”:
https://www.analyticsvidhya.com/blog/2020/10/getting-started-with-feature-engineering,
the disadvantage is that information is lost, particularly on features sharper than the binning:
https://www.kaggle.com/questions-and-answers/171942,
and then you need to reasonably re-adjust the binning to spot the moving target “trends” [excluding C]:
https://stats.stackexchange.com/questions/230750/when-should-we-discretize-bin-continuous-independent-variables-features-and-when.

upvoted 1 times

M25 1 year, 6 months ago

“(…) some features have a large range”, possible presence of outliers exclude standardization [excluding A]:
https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/.
“(…) a wide range of factors”, PCA transform the data so that it can be described with fewer dimensions / features:
https://en.wikipedia.org/wiki/Principal_component_analysis, but [excluding B]: it asks to “ensure that the features with largest magnitude don’t overfit the model”.

upvoted 1 times

niketd 1 year, 7 months ago

Selected Answer: D

The question doesn't talk about the skewness within each feature. It talks about normalizing the effect of features with large range. So scaling each feature within (0,1) range will solve the problem

upvoted 1 times

JamesDoe 1 year, 7 months ago

Really need more info to answer this: what does "large range" mean? Distribution follows a power law --> use log(). Or are they more evenly/linearly distributed --> use (0,1) scaling.

upvoted 1 times

guilhermebutzke 1 year, 7 months ago

Selected Answer: C

I think C could be a better choice. Bucketizing the data we can fix the distribution problem by bins.

in letter A, standardization by log could not be effective if the range of the data has negative and positive values.

In letter D, definitely normalization does not resolve the skew problem. Data normalization assumes that data has some normal distribution.

https://medium.com/analytics-vidhya/data-transformation-for-numeric-features-fb16757382c0

upvoted 3 times

TNT87 1 year, 8 months ago

Selected Answer: D

D. Normalize the data by scaling it to have values between 0 and 1.

Standardization and normalization are common techniques to preprocess the data to be more suitable for machine learning models. Normalization scales the data to be within a specific range (commonly between 0 and 1 or -1 and 1), which can help prevent features with large magnitudes from dominating the model. This approach is especially useful when using models that are sensitive to the magnitude of features, such as distance-based models or neural networks.

upvoted 1 times

FherRO 1 year, 8 months ago

Selected Answer: A

https://developers.google.com/machine-learning/data-prep/transform/normalization#log-scaling

upvoted 1 times

Load full discussion...