Statistical Learning:

A compilation of insightful Q&As

Contents

[Chapter 1. Linear Regression 3](#_Toc75108558)

[Collinearity among features 3](#_Toc75108559)

[Chapter 2. Classification 4](#_Toc75108560)

[Encoding categorical variables 4](#_Toc75108561)

[Chapter 4: Resampling – Cross validation and Bootstrap 5](#_Toc75108562)

[Capability of bootstrap dataset to replicate real world variability 5](#_Toc75108563)

Chapter 1. Linear Regression

# Collinearity among features

How does collinearity among features affect the model fit, particularly in the context of linear regression?

**Answer:**

It does not seem to affect model fit score, at least in the context of Linear Regression. This is demonstrated in Collinearity and its effect on model fit in Linear Regression.ipynb. If you want to see the answer only without all the code, the [blog in medium](https://einchako.medium.com/does-feature-collinearity-affect-model-fit-4618c5dc79ea) might be the place to go.

# Collinearity in polynomial regression

Answer:

I posted [this question in stackoverflow and the answer](https://stackoverflow.com/questions/67914111/doesnt-introduction-of-polynomial-features-lead-to-increased-collinearity) suggests that although polynomial features do introduce multicollinearity they are not of concern if model fit score is not affected. Anyways, if we are applying polynomial regression, it means that we are more inclined towards predictive applications and not towards inferential studies. Then a reduction in model interpretability is a cost we should be willing to pay for achieving a better fit.

Chapter 2. Classification

# Encoding categorical variables

How ordinally encoding categorical variables with more than two classes affects model output? Ordinal encoding is encoding the classes as 1, 2, 3 etc. Is the effect similar for the case of response and feature variables?

**Answer:**

The demonstration of this effect on model output is shown in the file How ordinal encoding of categorical variable for more than 2 classes affects model output.ipynb. It also gives valuable insights that might go into deciding what form of encoding (ordinal or one-hot) to use in which situation. There is a [blog as well that explains the answer](https://einchako.medium.com/categorical-variables-how-to-encode-them-and-why-cf0e6f571821).

Chapter 4: Resampling – Cross validation and Bootstrap

# Capability of bootstrap dataset to replicate real world variability

The bootstrap methodology claims to estimate the variability of the entire training dataset- i.e., when various training datasets are chosen in multiple attempts? We will need to verify if the bootstrap methodology indeed has this capability.

Answer: