Statistical Learning:

A compilation of insightful Q&As

**Contents**

[Chapter 1. Linear Regression 3](#_Toc76314312)

[Collinearity among features 3](#_Toc76314313)

[Collinearity in polynomial regression 3](#_Toc76314314)

[Chapter 2. Classification 4](#_Toc76314315)

[Encoding categorical variables 4](#_Toc76314316)

[Class weights and threshold limit 4](#_Toc76314317)

[Instability of logistic regression for well-separated classes 4](#_Toc76314318)

[Chapter 4: Resampling – Cross validation and Bootstrap 5](#_Toc76314319)

[Capability of bootstrap dataset to replicate real world variability 5](#_Toc76314320)

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).

# Class weights and threshold limit

In Logistic Regression there is a concept of threshold, which is a probability limit *p* such that, for *Pr(y=1|x) > p, ypredicted = 1*. This threshold limit can be varied to control the proportion of true positives and false negatives in the total number of erroneous predictions. The usual choice of *p=0.5* sometimes leads to very large proportion of false negatives although the overall error rate is minimized. By keeping *p<0.5* we reduce the proportion of false negatives at the cost of a slightly increase in false positivity and error rate. The excessive false negativity rate as compared to false positivity – is this caused due to lower number of *ytrue=1* entries as compared to *ytrue=0* entries in the training dataset? Because if it is, then modifying class-weights might be a substitute of threshold manipulation.

Class-weights simply decide the weight of samples belonging to each unique response class in the overall loss function. It is a generic concept meant to undo imbalance in datasets. As explained above, we need to check if manipulating class-weights achieves the same result as manipulating threshold limit.

**Answer:**

The comparison of the methods can be found in [my blog](https://einchako.medium.com/manipulating-class-weights-and-decision-threshold-cb7d8d9433a4) and in [Class weights and threshold in Logistic Regression.ipynb](https://github.com/AnirbanChakraborty06/StatisticalLearning-ConfusionToClarity/blob/main/AnswersAndDemos/Class%20weights%20and%20threshold%20in%20Logistic%20Regression.ipynb).

# Instability of logistic regression for well-separated classes

During the study conducted for ‘[Class weights and threshold limit](#_Class_weights_and)’, it was observed that datasets generated in multiple attempts with the same simulation function were leading to different fit scores under linear regression. The variance in the fit scores was too massive to ignore – moving from 50% to 87%. Does it indicate towards some generic shortcoming of logistic regression?

**Answer:**

The phenomenon is demonstrated in [Instability of logistic regression for largely separated classes.ipynb](https://github.com/AnirbanChakraborty06/StatisticalLearning-ConfusionToClarity/blob/main/AnswersAndDemos/Instability%20of%20logistic%20regression%20for%20largely%20separated%20classes.ipynb). The blog related to this is [here](https://einchako.medium.com/the-instability-of-logistic-regression-1bb35f9054bf).

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:

The discussion regarding this show in [my blog](https://www.analyticsvidhya.com/blog/2021/06/bootstrap-the-source-of-its-power/) and in [Capability of bootstrap methodology to replicate real world variability.ipynb](https://github.com/AnirbanChakraborty06/StatisticalLearning-ConfusionToClarity/blob/main/AnswersAndDemos/Capability%20of%20bootstrap%20methodology%20to%20replicate%20real%20world%20variability.ipynb).