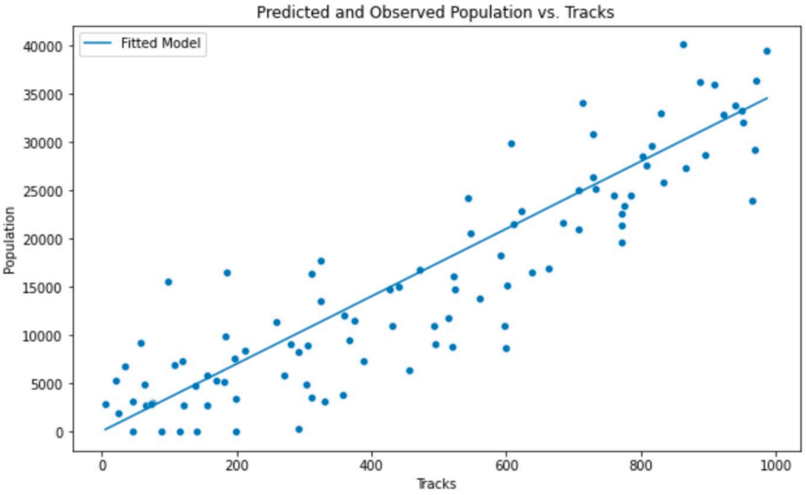
Data Analysis – BootCamp B4: Classification, Regularization and Sampling

Collaborators:

1.Inspect the plot below. It shows a scatterplot between two variables (# of tracks on x axis and population of animals on y axis). The line is the fitted regression model. We know that the model was fit using Ridge regression without an intercept term (i.e. $\theta_0=0$). Which of the following statements are true? (Exam April 2022 question)



A The fit is inaccurate due to the large amounts of scatter around the line, which suggests that linear regression is inappropriate.

B There is a curvature in the relationship, which suggests that linear regression is inappropriate.

C The training loss would decrease if the regularization hyperparameter λ was 0.

D The plot displays high variance, which suggests that λ should be increased to reduce model complexity.

C is correct.

While there is scatter around the line, the general trend suggests that linear regression is appropriate since there is no curvature in the relationship. This rules out the first two options. Since λ restricts the magnitude of our parameters, removing regularization by setting λ = 0 would improve our fit and therefore reduce training loss. The plot does display high variance in the observations themselves, not the predictions necessarily. The variance in the observations is not linked to model complexity or the regularization hyperparameter. Further, λ being increased would lead to a poorer fit.

2.You are given the following optimization problem for solving a learning task.

$$argmin_\theta \left[\sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda \sum_{p=1}^{d} \theta_p^2 \right]$$

Which of the following statements are true? (May 2021 exam)

i)There are *d* data points
ii) There are *n* data points
iii) The data is *d* dimensional

iv) This is a classification problem

v) This is a linear model

vi) This problem uses LASSO ($L1$) regularization

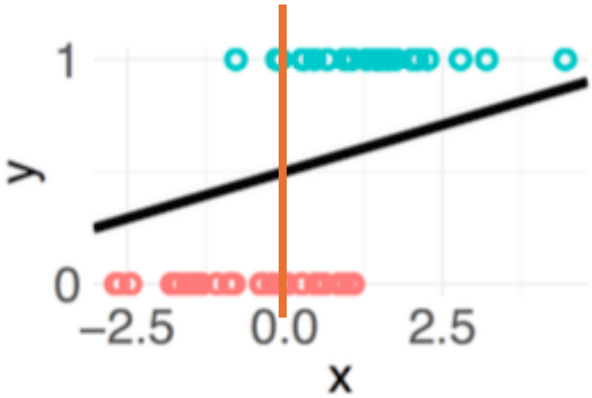
vii) Larger values of λ imply increased regularization

viii) Larger values of λ will likely increase variance

ix) Larger values of λ will likely increase bias

x) None of the above are true

3. Suppose a friend of yours obtains a dataset with a single feature x and an output y (binary). We would like to build a binary classification predicting y from x. Your friend argues that the data are linearly separable by drawing the line on the following plot of the data. Discuss with each other and argue whether this statement is correct.



The scatter plot of x against y isn't the graph you should be looking at. We see e.g. that for data points around x=0, there are multiple points that are either in class y=0 or y=1. Therefore, there is not one line (i.e. a hyperplane) that can separate the data.

4. You want to build a model to classify whether a tweet spreads misinformation—because clearly, **X** isn't doing that anymore. 10M tweets from 10K users over the last 24 months # tweets/user follows a long-tail distribution You estimate that 1% of tweets are misinformation.

a. How would you sample 100K tweets to label?

b. You get 100K labels from 20 annotators and want to look at some labels to estimate the quality. How many labels would you look at? How would you sample?

c. Imagine an endless stream of tweets (like X, but with even less oversight), they can't fit all in memory. How to sample 10K tweets such that each tweet has an equal chance of being selected?

a.Use stratified sampling to ensure enough misinformation tweets. Perhaps also sample proportionally across the long-tail distribution of users

b. No perfect number exists. The more the better. Look at e.g. 500 labels for a quick assessment. To account for different annotators and different categories (misinfo vs. non-misinfo) use stratified random sampling.

c. Use reservoir sampling (as we discussed in class) to maintain a uniform random sample of 10K tweets.