S2: Occam's razor, Bayes Method, concept of regularization and possible conceptual contradiction

Scribe: 2 – Lecture 19

Student: Pranjal Umesh Kalekar

CS 7840: Foundations and Applications of Information Theory (fa24)

https://northeastern-datalab.github.io/cs7840/fa24/

Lecturers: Wolfgang Gatterbauer, Javeed Aslam

Version Jan 1, 2001 <replace with the date of submission>

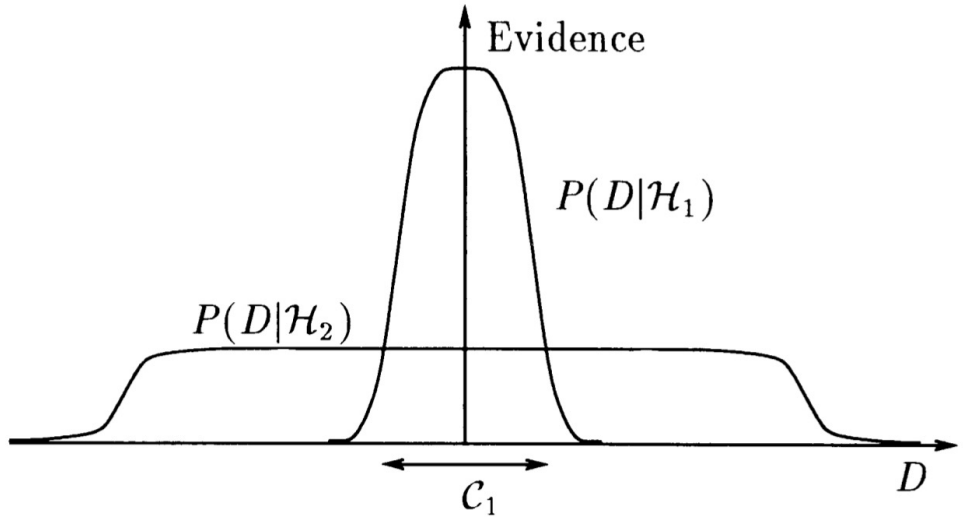
Understanding of Occam's razor

If there are two methods covering same data ‘ Good Enough’ then the simple model is preferred over the complex one.

Simpler models are also better because they make a strong predictions. Such models are therefore falsifiable (one can easily find something they don't predict, and see if it is true) and, in probabilistic terms, put a lot of the "probability mass" or "likelihood" on a few specific phenomena. Thus, when such a specific phenomenon *does* occur, simpler models explain it better than a more complex theory, which spread the probability mass over more possibilities

Bayes Theorem and its Embodiment of Occam's Razor

As we have seen during the class, Bayes embodies Occam's razor by rewarding models in proportion to how much much they predicted the data that occurred.



A simple model H1 makes only a limited range of predictions, shown by P(DlH1); a more powerful model H2 , that has, for example, more free parameters than H1, is able to predict a *greater* variety of data sets. This means however that H2 does not predict the data sets in region C1 as strongly as H1. Assume that equal prior probabilities have been assgined to the two models. Then if the data set falls in region C1 , the less powerful modelH1 will be the more probable model.

Occam's Razor and the concept of regularization

By my understanding, in simplest words regularization is adding penalties to the features that are not contributing to the data of interest.

Two types of regression:
Lasso(L1): forcing large negative coefficients to zero excluding them from impacting prediction.
Ridge(L2): instead of forcing large coefficients to zero, replacing it with a comparatively small negative coefficients

Clearly this indicates decrease in the impact of less impacting parameters towards prediction – leading to less free features – and eventually to a simpler model – AHA! Occam's Razor!!

These three concepts have same ideology right, But I have a question:

When I regularize the mode, it penalizes the features affecting less and its purpose is to avoid overfitting so basically isn't it generalizing the model? Kind of like what H2 was doing On P2. and As we saw – by Bayesian and Occam's razor's terminology – H1 is preferred over H2…. So is this a conflict?

- So solve this I dug a little deeper to find out about *Bias and variance trade off*,

The way I visualize this concept is
bias ----> overfitting
variance------> generalization

There is often a tradeoff between the bias and variance contributions to the estimation error, which makes for a kind of "uncertainty principle" (Grenander 1951). Typically, variance is reduced through "smoothing," via a combining, for example, of the influences of samples that are nearby in the input (x) space. This, however, will introduce bias, as details of the regression function will be lost; for example, sharp peaks and valleys will be blurred.

Geman, S., Bienenstock, E., & Doursat, R. (1992). *Neural networks and the bias-variance dilemma.* Neural Computation, 4(1), 1-58.

This paper derives error estimation equation via Bias/Variance and tried smoothing over using both trades to obtain the optimal fit figure.2, example 3

Even the talk by Ilya Sutskever (OpenAI)(referenced multiple times by prof. Wolfgang) speaks about generalization and it's power in unsupervised learning.

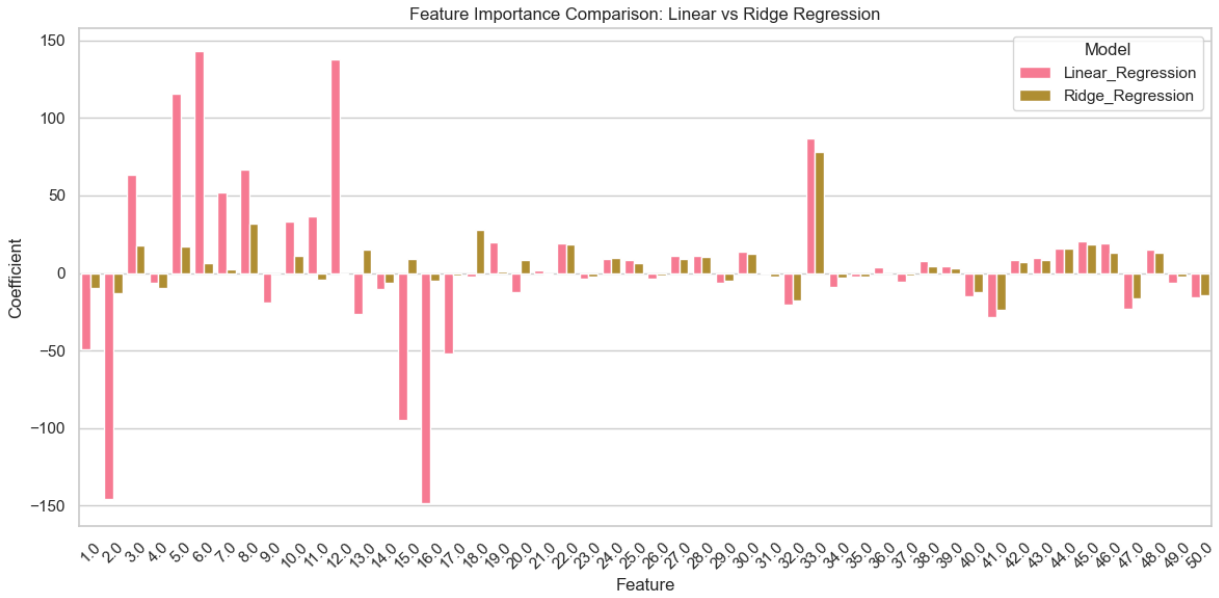
I was curious about how regularization affect feature importance during model training

Ideally the regularization should penalize least affecting features and somehow smooth out importance across the effective features

- Taking inspiration from Neural Computation(1992) I trained linear regression and ridge regression on an synthetic generated data with high multicollinearity and noise – I did try training the models on well constructed diabetes dataset but obviously the results were not significantly noticeable
- Results: even though the performance difference between two models is not much, one can clearly observe the smoothing in feature importance in regularized model

Model Performance Metrics:

	Model	Dataset	R^2	RMSE	Mean Residual	Std Residual
0	Linear Regression	train	0.774	73.521	0.000	73.521
1	Linear Regression	val	0.154	108.532	19.479	106.770
2	Linear Regression	test	-0.031	115.795	29.102	112.078
3	Ridge Regression	train	0.753	76.955	-0.000	76.955
4	Ridge Regression	val	0.319	97.361	21.302	95.002
5	Ridge Regression	test	0.136	105.988	20.055	104.073



References:

- https://blogs.iq.harvard.edu/occams_razor_an, Occam's Razor And Thinking about Evolution, Amy Perfors, Nov, 2005.
- https://link.springer.com/chapter/10.1007/978-94-017-2219-3_3 , Bayesian Interpolation, David J.C. MacKay, Maximum Entropy and Bayesian Methods, 1992
- https://gatterbauer.name/download/2008_Interpreting_versus_Teaching_Facts_Classroom.pdf , Wolfgang Gatterbauer
- https://en.wikipedia.org/wiki/Regularization_(mathematics) , relating regularization and Occam's razor
- https://www.dam.brown.edu/people/documents/bias-variance.pdf
- Coding file included in the submission.