GENERAL ASSEMBLY

FEATURE SELECTION

Joseph Nelson, Data Science Immersive

AGENDA

- What is Feature Selection?
- Types of Feature Selection
- Regularization
- Coding Implementation
- Sklearn documentation

WHAT IS FEATURE SELECTION?

- And why is it useful?

WHAT IS FEATURE SELECTION?

- We have many potential features that may be used, but only some may have predictive power
- Envision text data or music data – many features produced, only some have predictive power
- Given an n x d pattern matrix (n patterns in d dimensional space), generate an n x m pattern matrix, where m << d
- Three types of feature selection: bottom up, top down, and random shuffling

- Selection vs extraction: Selection chooses current features, extraction produces interactions of current features (linear or non-linear combinations)

BOTTOM UP FEATURE SELECTION

- Iteratively adding additional features to our model based on performance
- We’ll discuss two types:
- 1. Sequential forward selection – Start with an empty set (X=0) and iteratively add best performing features
- Disadvantage:

BOTTOM UP FEATURE SELECTION

- Iteratively adding additional features to our model based on performance
- We’ll discuss two types:
- 1. Sequential forward selection – Start with an empty set (X=0) and iteratively add best performing features
- Disadvantage: once a feature is added, it cannot be discarded. Doesnot consider feature interaction

BOTTOM UP FEATURE SELECTION

- Iteratively adding additional features to our model based on performance
- We’ll discuss two types:
- 1. Sequential forward selection (SFS) – Start with an empty set (X=0) and iteratively add best performing features
- Disadvantage: once a feature is added, it cannot be discarded. Doesnot consider feature interaction
- 2. Sequential backward selection (SBS) – Start with X= D, iteratively delete least significant features
- Disadvantage:

BOTTOM UP FEATURE SELECTION

- Iteratively adding additional features to our model based on performance
- We’ll discuss two types:
- 1. Sequential forward selection (SFS) – Start with an empty set (X=0) and iteratively add best performing features
- Disadvantage: once a feature is added, it cannot be discarded. Doesnot consider feature interaction
- 2. Sequential backward selection (SBS) – Start with X= D, iteratively delete least significant features
- Disadvantage: all those above, plus computationally more expensive than SFS

- Generalized versions of these searches: do the same, but model entire subsets

BOTTOM UP FEATURE SELECTION

- Sequential floating forward search (SFFS):
- **Step 1: Inclusion**. Use the basic SFS method to select the most significant feature with respect to X and include it in X. Stop if d features have been selected, otherwise go to step 2.
- **Step 2: Conditional exclusion.** Find the least significant feature k in X. If it is the feature just added, then keep it and return to step 1. Otherwise, exclude the feature k. Note that X is now better than it was before step 1. Continue to step 3.
- **Step 3: Continuation of conditional exclusion.** Again find the least significant feature in X. If its removal will (a) leave X with at least 2 features, and (b) the value of J(X) is greater than the criterion value of the best feature subset of that size found so far, then remove it and repeat step 3. When these two conditions cease to be satisfied, return to step 1.

TOP DOWN FEATURE SELECTION

- Another way to select features is to impose a global constraint on the model. For example, in the case of text vectorization, we could impose that a feature needs to have a document frequency higher than a certain threshold to be considered relevant
- Fairly common in text data, visual/auditory data

RANDOM SHUFFLING FEATURE SELECTION

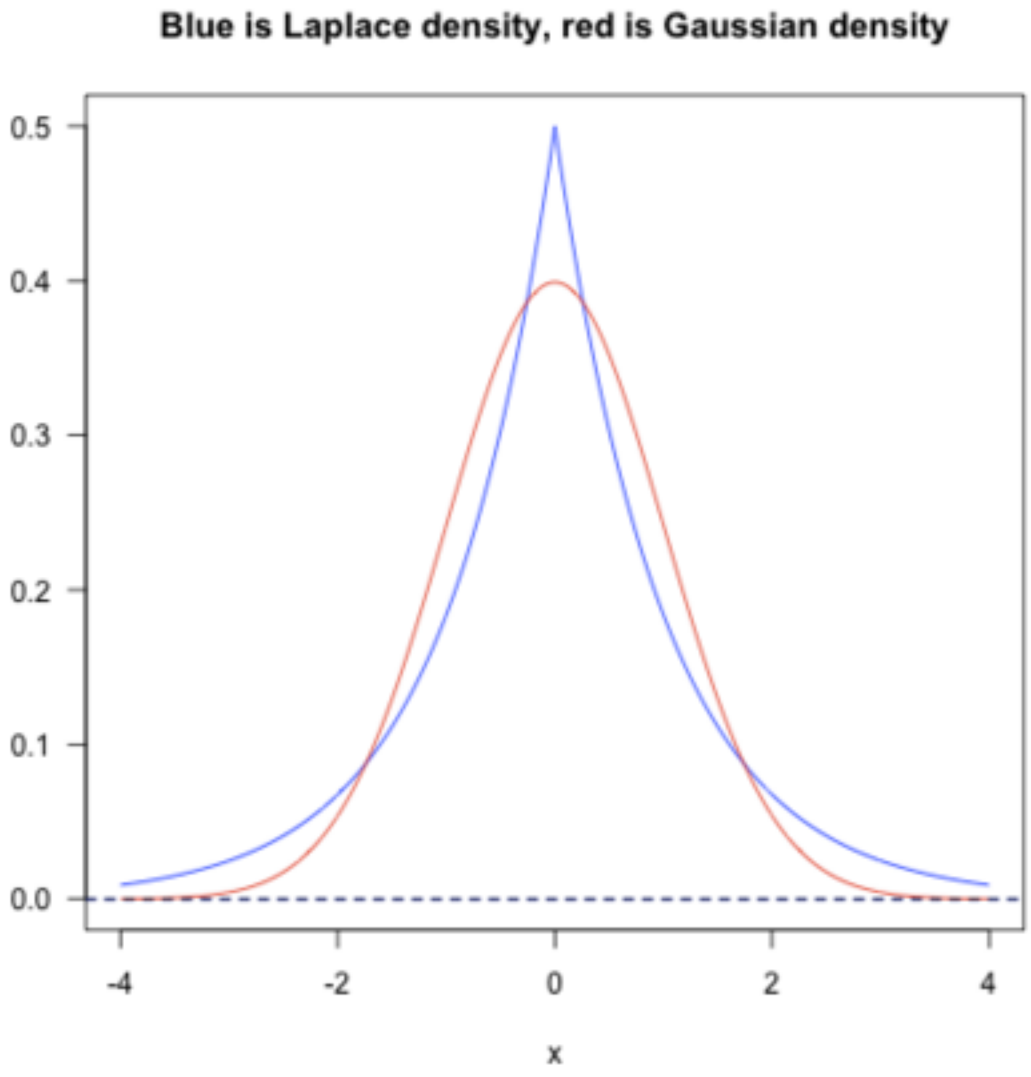
- There are other ways to check if a feature has any predictive power, such as random shuffling. First, we calculate the score of the model, then we randomize the values along that column. If the feature has any predictive power, this should yield a worse score. On the other hand, if the feature has no predictive power at all, this will result in no change in the score and thus we can toss that feature.

REGULARIZATION

- Regularization is an example of a top-down technique that works with parametric models (such as Logistic Regressions and Support Vector Machines). It imposes a global constraint on the values of the parameters that define the model. The regularized model is found solving a new minimization problem where two terms are present: the term defining the model and the term defining the regularization.

L1 AND L2 REGULARIZATION

- Regularization works by adding the penalty associated with the coefficient values to the error of the hypothesis. This way, an accurate hypothesis with unlikely coefficients would be penalized while a somewhat less accurate but more conservative hypothesis with low coefficients would not be penalized as much.



L1 AND L2 REGULARIZATION

- L1: Stricter constraints: more angular distribution for inclusion
- L2: Looser constraints, increases inclusion of features

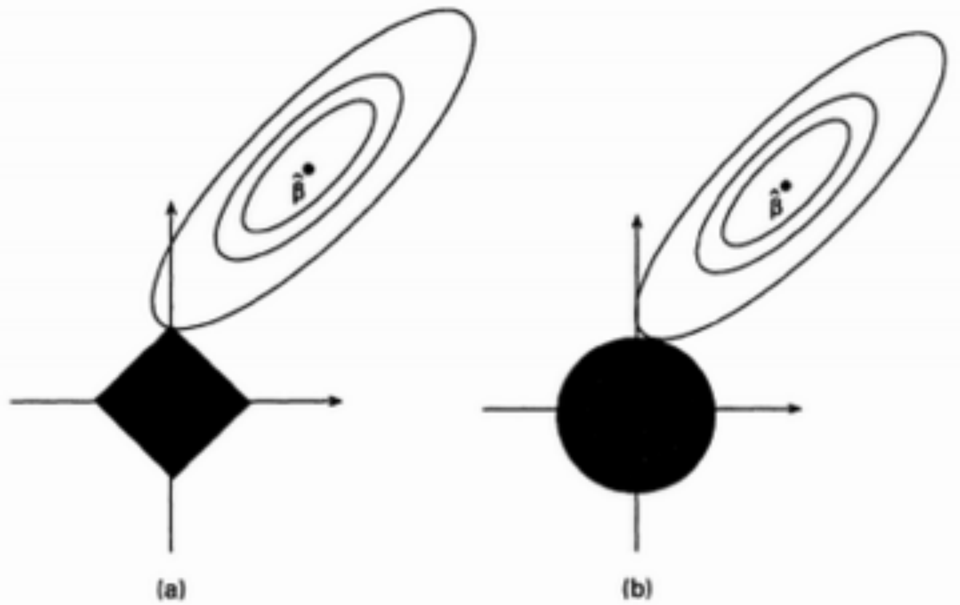


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

REGULARIZATION

- Bottom line: regularization allows us to reduce over fitting by generalizing the values we have available in our training set.