Interpretability in ML & Sparse Linear Models

Unchitta Kanjanasaratool

Mentor: Denali Molitor

UCLA Department of Mathematics unchitta@ucla.edu

Interpretability in ML: Why It's Important

Definition (non-mathematical): Interpretability is the degree to which a human can consistently explain or interpret why a model makes certain decisions.

Several components: model transparency, holistic model interpretability, modular-level interpretability, local interpretability for a single prediction or a group of predictions.

Proposed levels of evaluating interpretability (Doshi-Velez & Kim, 2017):

1. Application level
2. Human level
3. Function level

Interpretability in ML: Why It's Important

Interpretability helps explain *why* a model makes certain predictions.

We may not always need interpretability e.g. in low-risk or extensively studied problems, but knowing the *why* can help us learn more about **why a model might fail.**

Transparency & Ethics: The EU for example mandates that automated decisions must be explainable and should respect fundamental rights.

Interpretability also satisfies human curiosity and learning.

Interpretable Models: Linear Regression

Assume a linear relationship between the input and response variables, for example

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon \,,$$

we can predict

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_p X_p \,,$$

where the $\hat{\beta}$'s are the coefficients in the residual sum of squares (RSS) minimization problem, namely

$$\hat{\beta} = \arg\min_{\beta 0 \ldots \beta p} \Sigma_{i=0, i=p} (y_i - \hat{y}_i)^2 .$$

The linear coefficients (the β's) make this model easy to interpret on a modular level, and help us see the level of influence each variable has on the prediction.

Sparse Models & Lasso Regression

In practice however, interpreting linear models can still be hard if there are too many variables. There are certain variations of linear regression that can help with this problem, such as the *sparse models*. An example would be the "least absolute shrinkage and selection operator," or **lasso regression** which minimizes

$$\text{RSS subject to } \sum_{j=0,j=p} |\beta_j| \leq s.$$

This is equivalent to minimizing, with regularization, the quantity

$$\text{RSS} + \lambda \sum_{j=0,j=p} |\beta_j|,$$

where the second term in the quantity is some constant lambda times the L1 norm of the coefficients. (Lambda normally chosen using cross-validation to yield the minimal β's).

Sparse Models & Lasso Regression

Intuitively, lasso penalizes large models and selects small coefficients. Unintuitively (but we will see why), it also yields a model where a number of the coefficients are 0 or essentially close to 0. **This is an example of a *sparse regression model*, which penalizes large models (i.e. lots of features) and performs variable selection.**

The penalization is controlled via *lambda*: the larger it is, the bigger the sparsity. If lambda is small enough, lasso will yield the same results as the least square estimates (standard linear regression).

Sparsity & Interpretability

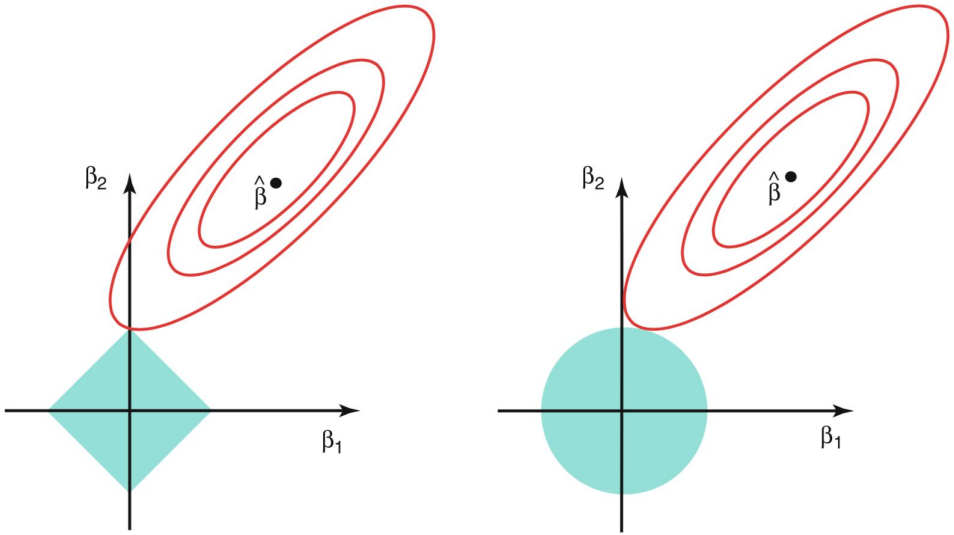
Having fewer variables often mean better interpretability.

Your model is explained by only a number of significant features, which reduces complexity and increases explainability. This is especially important when your data has hundreds or thousands of features; the complexity may be beyond human comprehension and there may not be enough observations.

Other methods for introducing sparsity to linear models include feature selection processes, subset selection and step-wise procedures e.g. forward and backward selection, sparse PCA.

Sparse Property of Lasso

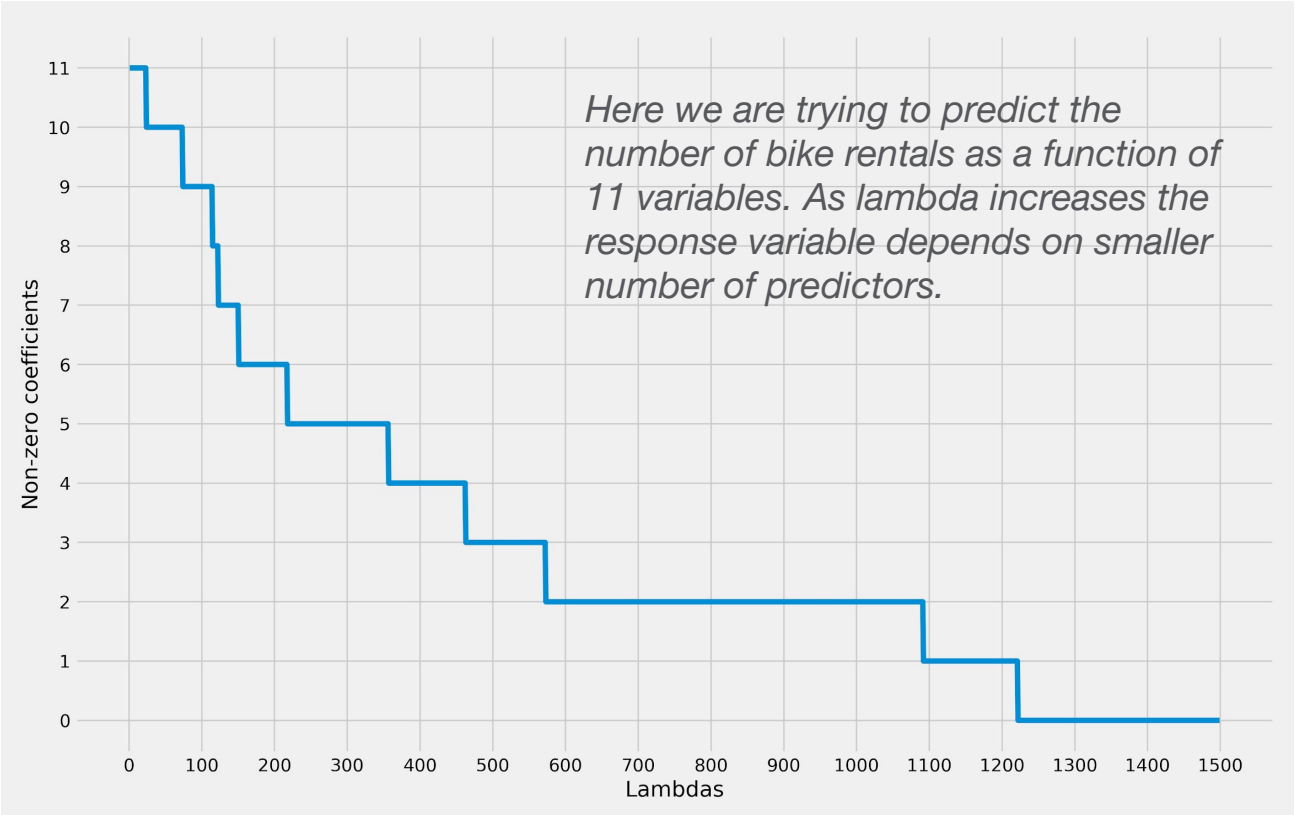
The variable selection property of lasso regression comes from the L1 regularizer. Consider the following figure in 2D problems. In higher dimensions the constraint region becomes polytopes (shapes with flat sides/sharp edges).



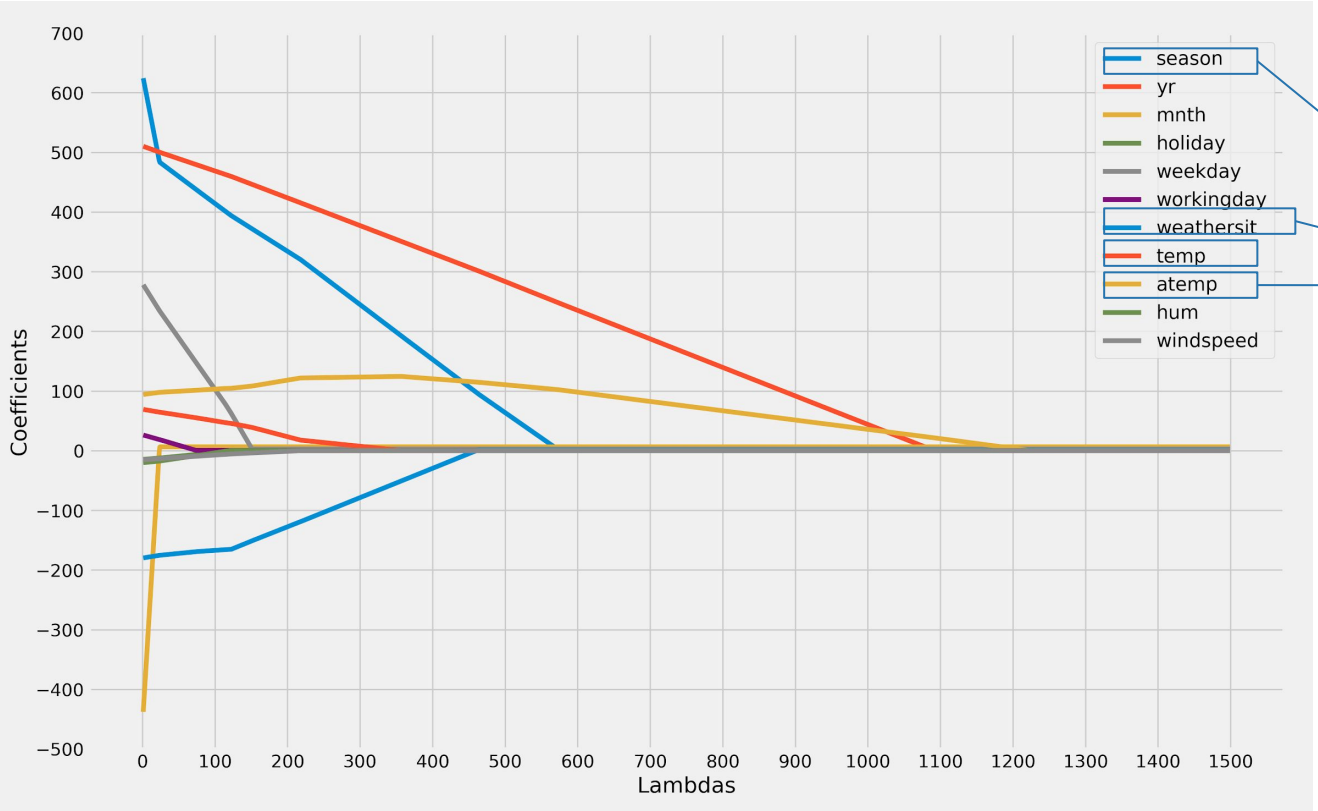
The red ellipses represent the contours of the RSS of lasso (left) and ridge regression (right). The solid green areas are the corresponding constraint functions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$.

Courtesy of Introduction to Statistical Learning, G. James, et al.

Example: UCI Bike Sharing Dataset

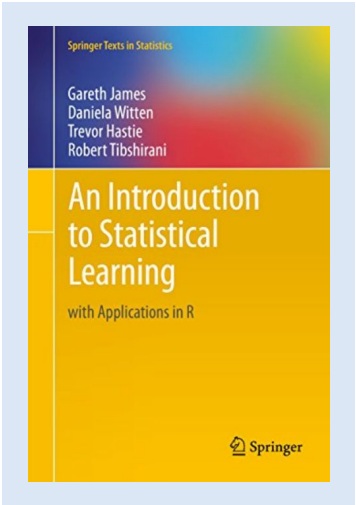


Example: UCI Bike Sharing Dataset

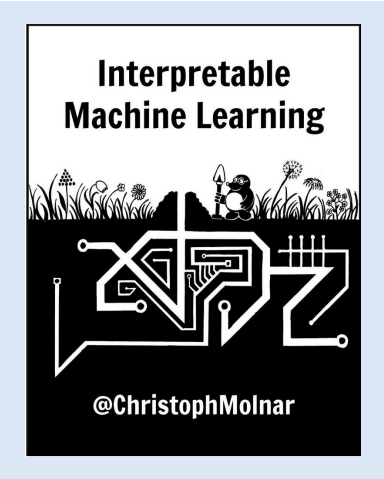


Likely the most significant features

Resources



Introduction to Statistical Learning with Applications in R, Gareth James, et al.



Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Christopher Molnar