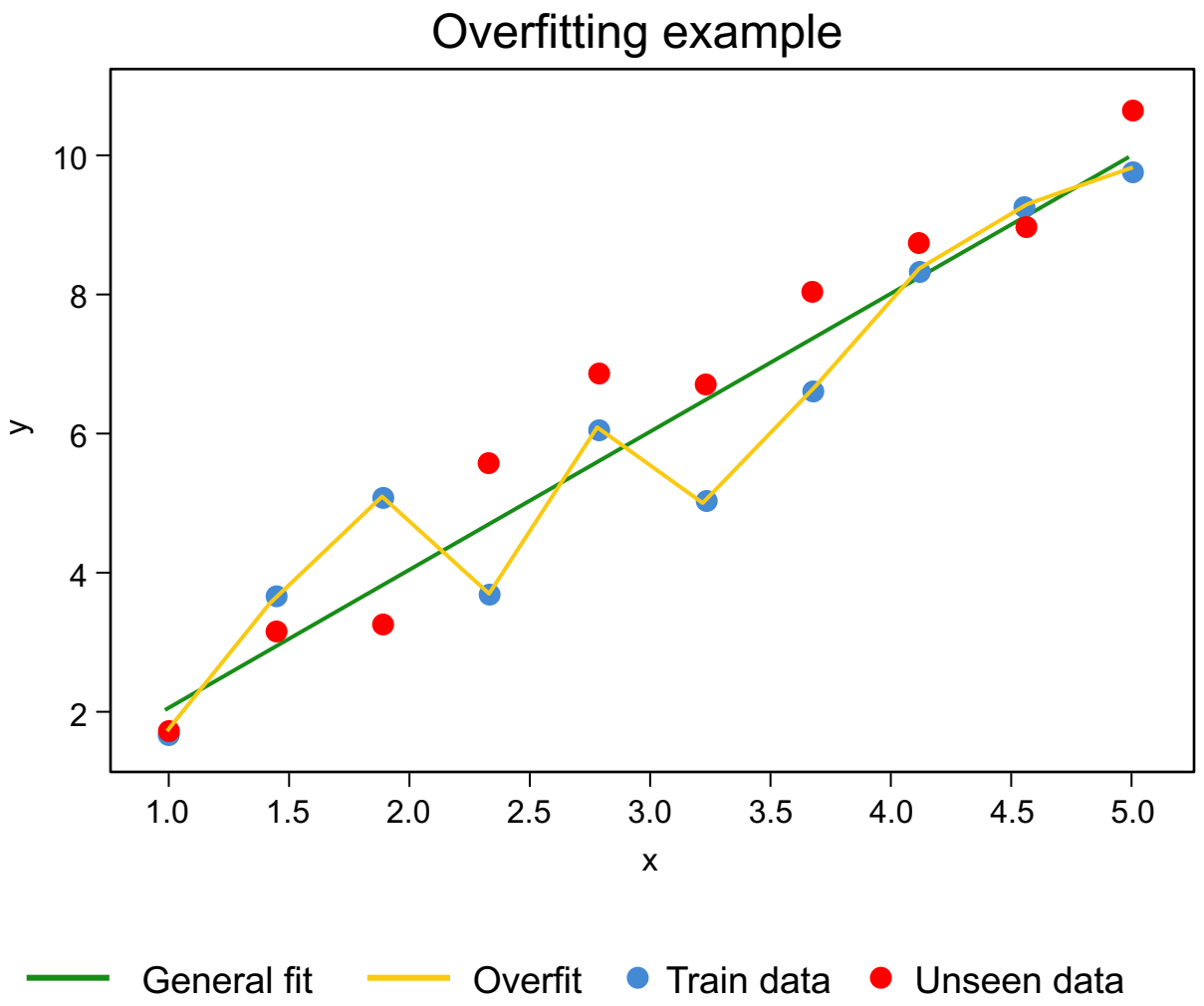
Overfitting

One of the Biggest Dangers in Modeling Is Overfitting (aka Bias)

- Overfitting means the model is fit more to the dataset than the problem.
- An overfit model will not perform as well on new (unknown) data.
- We need some tools to combat overfitting.

Illustration of Overfitting

It is possible to do a perfect fit (orange line) of the data (blue) you are modeling. However, if new data (red) is introduced, the model trained perfectly on the original data is actually a worse predicter than a general fit (green).



Overfitting

- Overfitting is the bane of most advanced models.
- While using a validation set can monitor for overfitting, the goal is to prevent overfitting in the first place.

DataScience@SMU

Regularization

Regularization

Regularization is the concept of a “penalty” for the coefficients or slopes.

- Our coefficients, m, in linear models play the role of slopes.
- We need to penalize for the size of the coefficients, not their value.
  - In other words, do not differentiate between negative and positive.
  - For normalized data, the coefficients/slopes are also the variable importance.
  - So, unless there is a very strong interaction that lowers the loss significantly, we want to keep the absolute value of the slopes small.

Look at a Loss Function

J is our loss, P is our prediction.

$$J = \frac{1}{n}\sum_{i=1}^{n}(Y_i - P_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=0}^{k} m_j x_{ij})^2$$ Replace P, with the actual formula

$$J = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=0}^{k} m_j x_{ij})^2 + \lambda \sum_{j=0}^{k} f(m_j)$$ Introduce the penalty

The Regularization Penalty

Our new penalty is:

$$\lambda \sum_{j=1}^{k} f(m_j)$$

- λ is the “strength” of the penalty, λ=0 returns to the original regression (linear or logistic).
- λ must be found experimentally by varying the value of λ and fitting the model and looking at the results.
  - It’s a “hyperparameter” that must be tuned.
  - Every problem has a different λ.

The Form of the Penalty

- Notice nothing has been said about f(m).
- We need to find a good function of the coefficients that penalizes weak correlations and promotes, or at least does not hinder, strong correlations.
- We will introduce the two most popular forms in the next sections.

DataScience@SMU

L1 and Lasso

L1 Regularization

- L1 regularization is also called Lasso.
  - The terms mean exactly the same thing.
- L1 uses the penalty term: $\lambda \sum_{j=0}^{k} |m_j|$
- The penalty is thus the absolute value of the coefficients.
- To remember, L1 can be used to stand for "1st order."

Loss Function in Depth

$$J = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \underbrace{\sum_{j=0}^{k} m_j x_{ij}}_{\text{Assuming all m's are zero, this portion contributes the sum of all targets to the loss—the predictions are all zero! As the magnitude of m increases (positive or negative), the predictions will get closer to the target and this portion contributes } \textbf{less} \text{ to the loss.}})^2 + \underbrace{\lambda \sum_{j=0}^{k} |m_j|}_{\text{If all m's are zero—this contributes nothing to the loss. That's good, since we want zero loss. However, as the predictions improve by increasing the magnitude of the m's, this term will grow in size.}}$$

If λ is large, the penalty will quickly overwhelm the prediction term. If λ is small, then it will not prevent overfitting. This happens for all types of regularization, not just L1!

Remember Y, x are **fixed**. That is your data. Your data will not change! The only thing that can change is your coefficients m.

What Makes L1 Special

- Lasso is actually an acronym.
  - Least Absolute Shrinkage and Selection Operator
- This implies that somehow L1 is involved with selection.
  - It is!
- L1 introduces sparsity—it reduces many of the coefficients to zero and thus can be used as feature selection.
- The primary use of L1 is for **feature selection**.

DataScience@SMU

L2/Ridge Regularization

L2 Regularization

- L2 regularization is also called ridge.
  - The terms mean exactly the same thing.
- L2 uses the penalty term: $\lambda \sum_{j=0}^{k} {m_j}^2$
- The penalty is thus the absolute value of the coefficients.
- To remember, L2 can be used to stand for "2nd order."

What Makes L2 Special

- L2/ridge regression provide a good general method to prevent overfitting.
- L2 does **not** induce sparsity, but it allows all the features to contribute.
- Your primary weapon against overfitting should be L2 regularization.
  - Many algorithms, such as XGBoost, implement it by **default**.

DataScience@SMU

L2 and No Sparsity

Why Do L1 and L2 Have Different Behaviors?

- For such a simple difference, L1 and L2 behave very differently
  - L1: feature selection/sparsity
  - L2: no sparsity
- How does this happen?

Geometric Diagram

Here, β is used to denote the coefficient, rather than m. What this diagram is saying is that as the prediction gets larger and larger (the concentric ovals), they will eventually come into contact with the constrained region of the coefficients (where the penalty term contributes more than the prediction term). Because the forbidden region of L1 is diamond shaped, when the regions intersect, it is invariable when one of the coefficients is zero. However, L2, the region is circular, and thus the intersection is unlikely to be when one of the coefficients is zero.

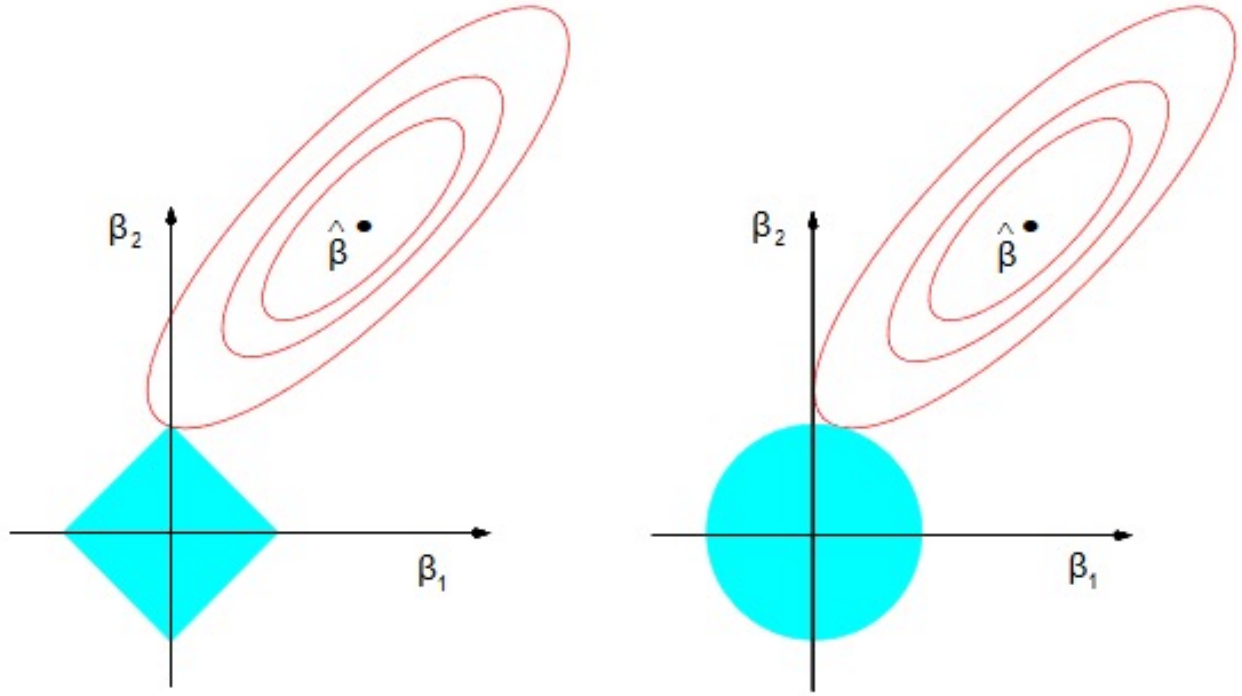


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Diagram courtesy: "The Elements of Statistical Learning," Second Ed. Hastie, Tibshirani, and Friedman, pg. 71

DataScience@SMU