GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

REGULARIZATION FOR REGRESSION

Recall our earlier discussion of **overfitting**.

Recall our earlier discussion of **overfitting**.

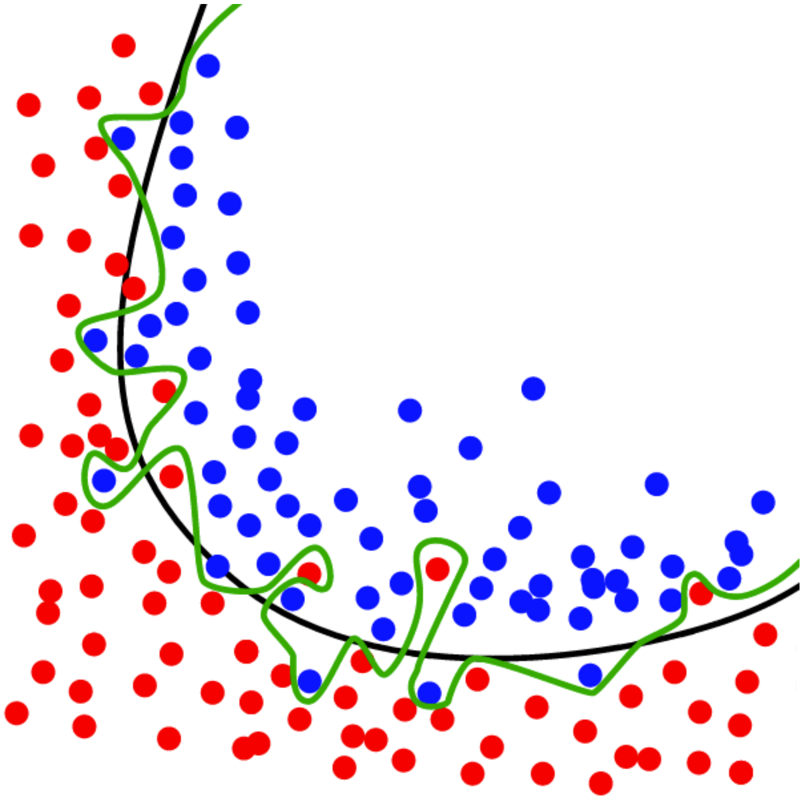
When we talked about this in the context of classification, we said that it was a result of matching the training set too closely.

Recall our earlier discussion of **overfitting**.

When we talked about this in the context of classification, we said that it was a result of matching the training set too closely.

In other words, a model that is overfit has learned from the **noise** in the dataset instead of just the **signal**.

OVERFITTING EXAMPLE (CLASSIFICATION)



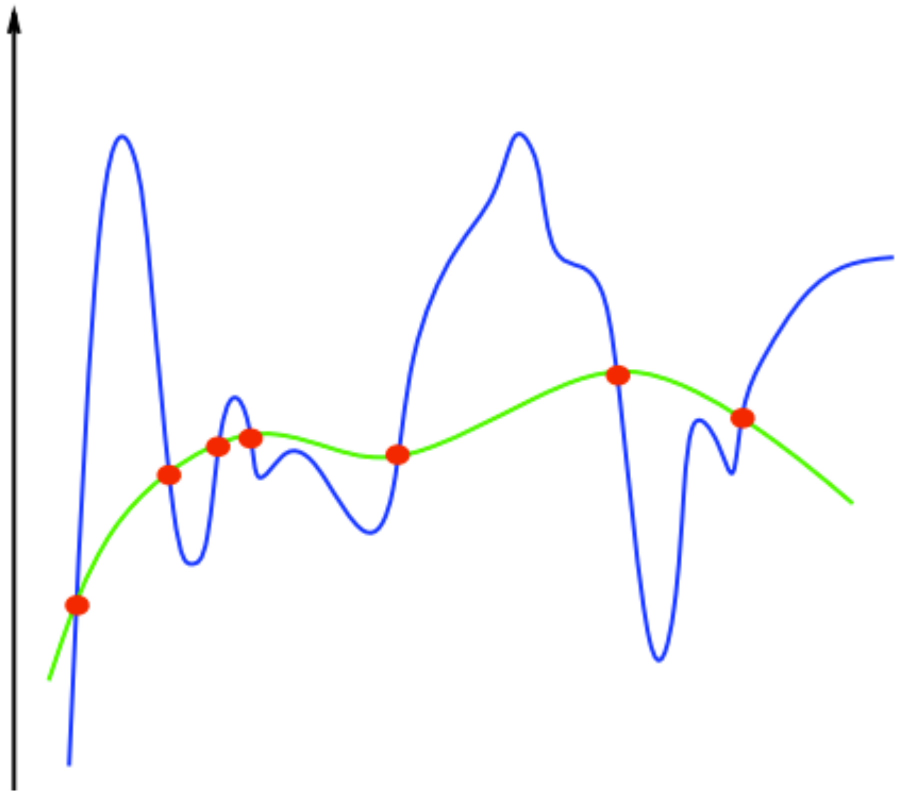
source: http://upload.wikimedia.org/wikipedia/commons/1/19/Overfitting.svg

The same thing can happen in regression.

It's possible to design a regression model that matches the noise in the data instead of the signal.

This happens when our model becomes *too complex* for the data to support.

OVERFITTING EXAMPLE (REGRESSION)



source: http://www.mit.edu/~9.520/spring12/slides/class02/class02.pdf

Q: How do we define the **complexity** of a regression model?

Q: How do we define the **complexity** of a regression model?

A: One method is to define complexity as a function of the size of the coefficients.

Q: How do we define the **complexity** of a regression model?

A: One method is to define complexity as a function of the size of the coefficients.

Ex 1: $\Sigma |\beta_i|$

Ex 2: $\Sigma \beta_i^2$

Q: How do we define the **complexity** of a regression model?

A: One method is to define complexity as a function of the size of the coefficients.

Ex 1: $\Sigma\,|\beta_i|$ this is called the **L1-norm**

Ex 2: $\Sigma\,\beta_i{}^2$ this is called the **L2-norm**

These measures of complexity lead to the following **regularization** techniques:

These measures of complexity lead to the following **regularization** techniques:

L1 regularization: $y = \Sigma\, \beta_i x_i + \varepsilon \quad st. \quad \Sigma\, |\beta_i| < s$

These measures of complexity lead to the following **regularization** techniques:

L1 regularization: $y = \Sigma\, \beta_i x_i + \varepsilon \quad st. \quad \Sigma\, |\beta_i| < s$

L2 regularization: $y = \Sigma\, \beta_i x_i + \varepsilon \quad st. \quad \Sigma\, \beta_i^2 < s$

These measures of complexity lead to the following **regularization** techniques:

L1 regularization: $y = \sum \beta_i x_i + \varepsilon \quad st. \quad \sum |\beta_i| < s$

L2 regularization: $y = \sum \beta_i x_i + \varepsilon \quad st. \quad \sum \beta_i^2 < s$

NOTE

L1 regularization is also known as *lasso* regularization. L2 regularization is also known as *ridge* regression.

These measures of complexity lead to the following **regularization** techniques:

L1 regularization: $y = \Sigma\, \beta_i x_i + \varepsilon \quad st. \quad \Sigma\, |\beta_i| < s$

L2 regularization: $y = \Sigma\, \beta_i x_i + \varepsilon \quad st. \quad \Sigma\, \beta_i^2 < s$

NOTE

L1 regularization is also known as *lasso* regularization. L2 regularization is also known as *ridge* regression.

Regularization refers to the method of preventing **overfitting** by explicitly controlling model **complexity**.

These regularization problems can also be expressed as:

L1 regularization: $min(\|y - x\beta\|^2 + \lambda\|x\|)$

L2 regularization: $min(\|y - x\beta\|^2 + \lambda\|x\|^2)$

These regularization problems can also be expressed as:

L1 regularization: $min(\|y - x\beta\|^2 + \lambda\|x\|)$

L2 regularization: $min(\|y - x\beta\|^2 + \lambda\|x\|^2)$

This (Lagrangian) formulation is what we'll use.

These regularization problems can also be expressed as:

L1 regularization: $min(\|y - x\beta\|^2 + \lambda\|x\|)$

L2 regularization: $min(\|y - x\beta\|^2 + \lambda\|x\|^2)$

This (Lagrangian) formulation is what we'll use.

NOTE

Lasso tends to eliminate coefficients, so it's useful for reducing the number of features. L2 tends to make coefficients small but not necessarily zero.

Q: What are bias and variance?

Q: What are bias and variance?

A: Bias refers to predictions that are *systematically* inaccurate.

Q: What are bias and variance?

A: Bias refers to predictions that are *systematically* inaccurate.
Variance refers to predictions that are *generally* inaccurate.

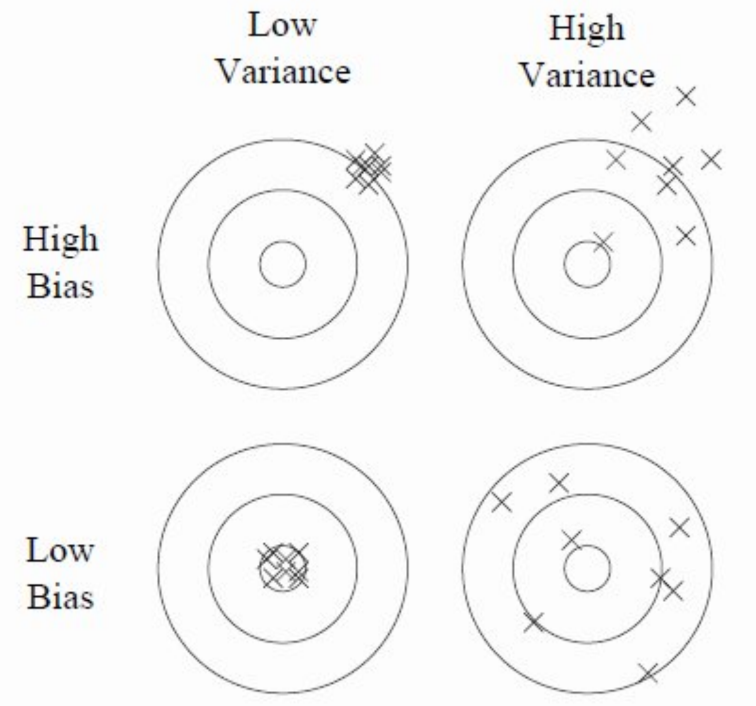


Figure 1: Bias and variance in dart-throwing.

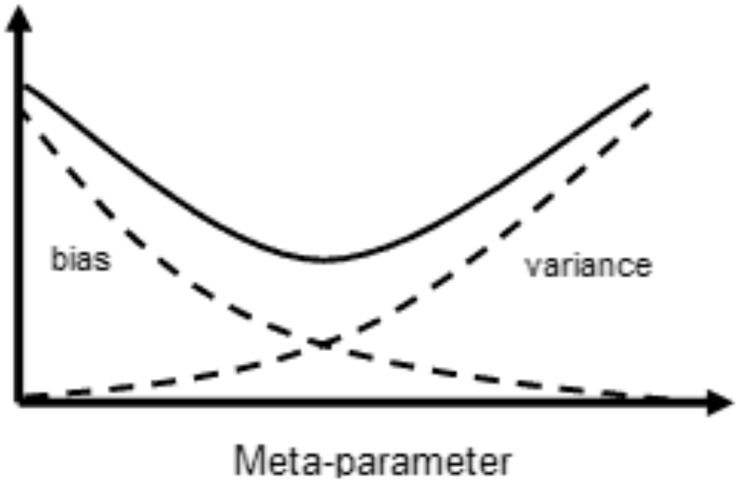
source: http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

Q: What are bias and variance?

A: Bias refers to predictions that are *systematically* inaccurate.
Variance refers to predictions that are *generally* inaccurate.

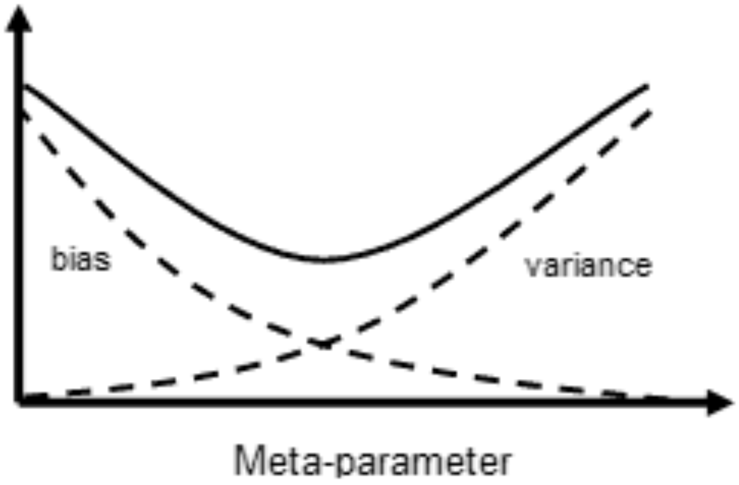
Generalization error can be decomposed into a bias component and variance component.

This is another example of the **bias-variance tradeoff**.



source: http://www.isu.edu/chem/images/kalivasmeta.gif

This is another example of the **bias-variance tradeoff**.



NOTE

The “meta-parameter” (or “hyperparameter”) here is the lambda we saw above.

source: *http://www.isu.edu/chem/images/kalivasmeta.gif*

This tradeoff is regulated by a **hyperparameter** λ, which we've already seen.

L1 regularization: $min(\|y - x\beta\|^2 + \lambda\|x\|)$
L2 regularization: $min(\|y - x\beta\|^2 + \lambda\|x\|^2)$

So regularization represents a method to trade away some variance for a little bias in our model, thus achieving a better overall fit.

This tradeoff is regulated by a **hyperparameter** λ, which we've already seen.

L1 regularization: $min(\|y - x\beta\|^2 + \lambda\|x\|)$

L2 regularization: $min(\|y - x\beta\|^2 + \lambda\|x\|^2)$

NOTE

Combining the regularization terms (with a balancing parameter) we have *elastic net* regularization.

So regularization represents a method to trade away some variance for a little bias in our model, thus achieving a better overall fit.