Understanding Generalization and Overfitting through bias & Variance

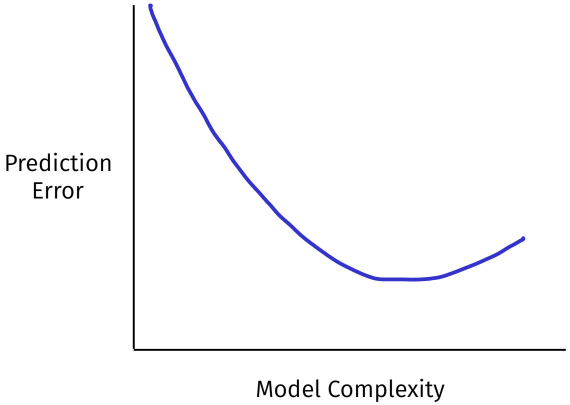
Evaluating model performance

We saw in the last classes how to estimate linear predictors by minimizing a squared loss objective function.

How do we evaluate whether or not our estimated predictor is good?

Measure: Validation error

Error vs. Model Complexity

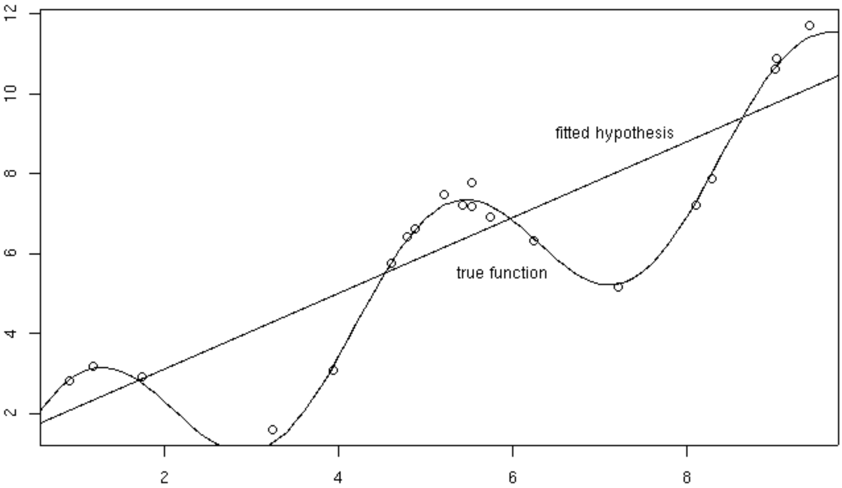


Sources of error

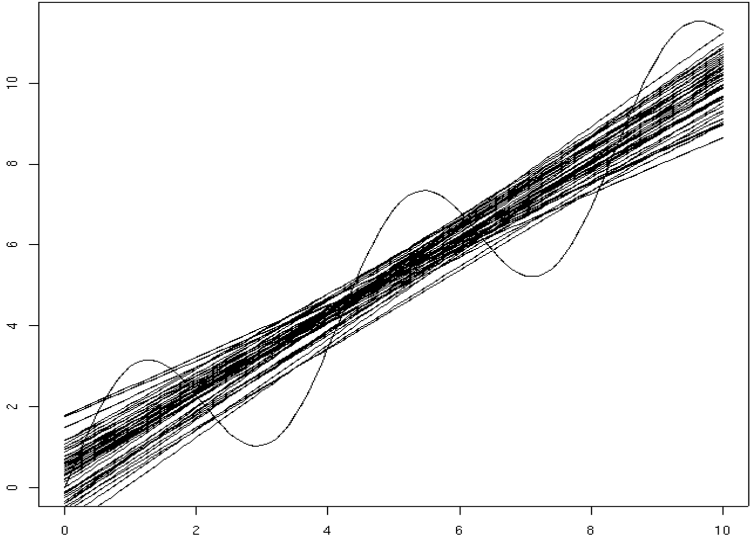
Three main sources of test error:

1. Bias
2. Variance
3. Noise

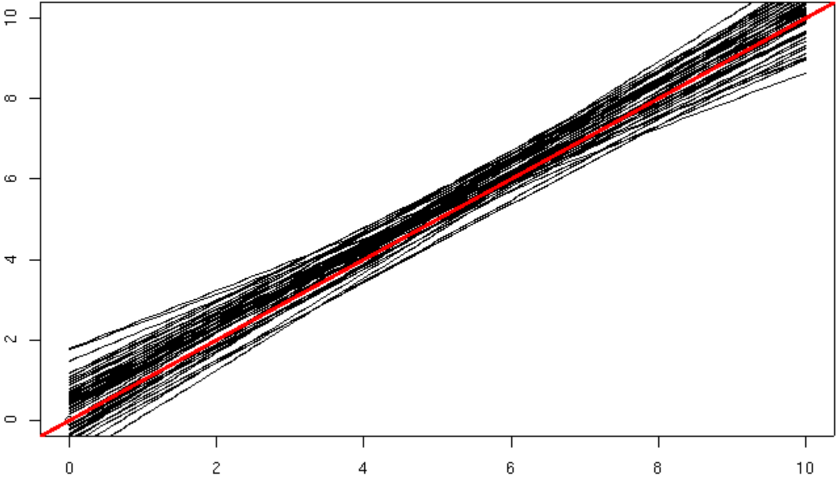
Example: function



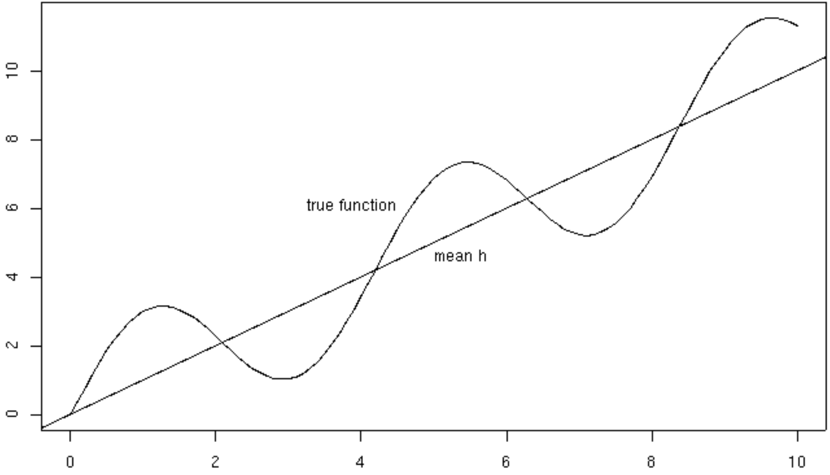
Fitting 50 lines after slight perturbation of points



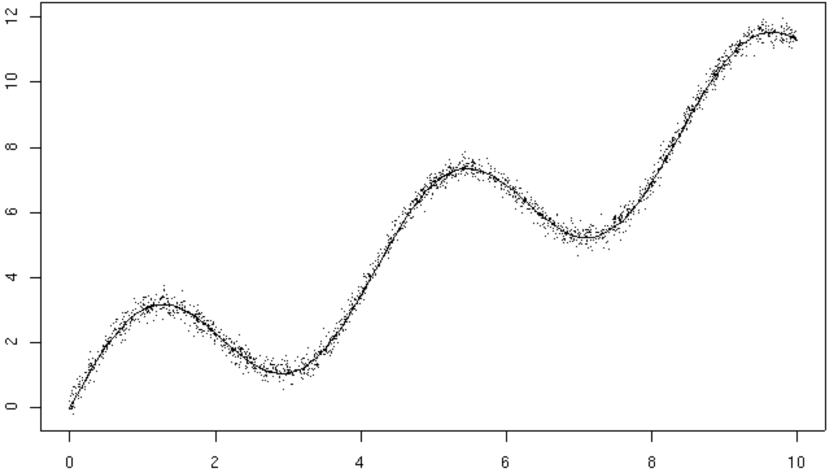
Variance after slight perturbation of points



Bias (with respect to non-linear fit)

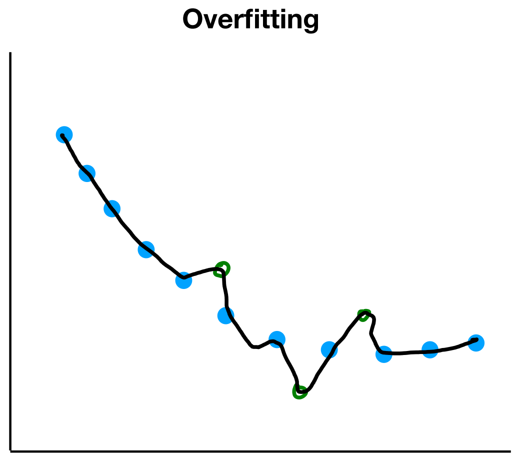


Noise



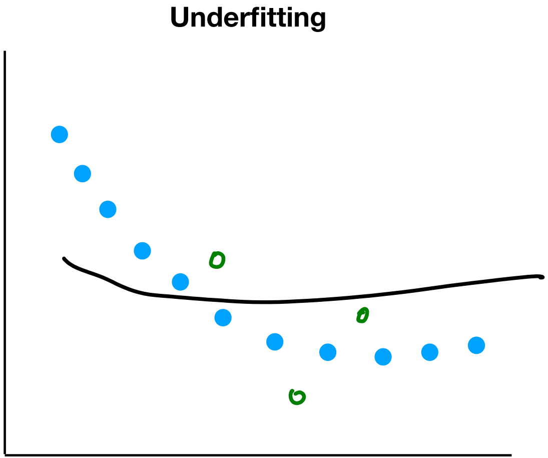
Overfitting

Overfitting: When the proposed hypothesis fits the training data too well



Underfitting

Underfitting: When the hypothesis is insufficient to fit the training data



Bias/Variance Decomposition for Regression

Bias-Variance Analysis in Regression

- Say the true underlying function is $y = g(\mathbf{x}) + \epsilon$ where ϵ is a r.v. with mean 0 and variance σ^2.
- Given a dataset of m samples, $\mathcal{D} = \{\mathbf{x}_i, y_i\}, i = 1 \dots m$, we fit a linear hypothesis parameterized by $\mathbf{w}$: $f_{\mathcal{D}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ to minimize the sum of squared errors $\sum_i (y_i - f_{\mathcal{D}}(\mathbf{x_i}))^2$
- Given a new test point $\hat{\mathbf{x}}$, whose corresponding $\hat{y} = g(\hat{\mathbf{x}}) + \hat{\epsilon}$, what is the expected test error for $\hat{\mathbf{x}}$, $\mathrm{Err}(\hat{\mathbf{x}}) = \mathbb{E}_{\mathcal{D},\hat{\epsilon}}[(f_{\mathcal{D}}(\hat{\mathbf{x}}) - \hat{y})^2]$?

Decomposing expected test error

$$
\begin{aligned}
\mathbb{E}[(f(\hat{\mathbf{x}}) - \hat{y})^2] &= \mathbb{E}[f(\hat{\mathbf{x}})^2 + \hat{y}^2 - 2f(\hat{\mathbf{x}})\hat{y}] \\
&= \mathbb{E}[f(\hat{\mathbf{x}})^2] + \mathbb{E}[\hat{y}^2] - 2\mathbb{E}[f(\hat{\mathbf{x}})]\mathbb{E}[\hat{y}] \\
&= \mathbb{E}[(f(\hat{\mathbf{x}}) - \overline{f}(\hat{\mathbf{x}}))^2] + \overline{f}(\hat{\mathbf{x}})^2 \\
&\quad + \mathbb{E}[\hat{y}^2] - 2\mathbb{E}[f(\hat{\mathbf{x}})]\mathbb{E}[\hat{y}] \\
&= \mathbb{E}[(f(\hat{\mathbf{x}}) - \overline{f}(\hat{\mathbf{x}}))^2] + \overline{f}(\hat{\mathbf{x}})^2 \\
&\quad + {\color{red}\mathbb{E}[\hat{y}^2]} - 2\overline{f}(\hat{\mathbf{x}})g(\hat{\mathbf{x}}) \qquad (1)
\end{aligned}
$$

where we have used the fact that $\mathbb{E}\left[(x - \mathbb{E}[x])^2\right] + (\mathbb{E}[x])^2 = \mathbb{E}[x^2]$

Decomposing expected test error

Applying the same trick used in Equation (1) to $\mathbb{E}[\hat{y}^2]$, we get

$$\begin{aligned}\mathbb{E}[(f(\hat{\mathbf{x}}) - \hat{y})^2] = &\ \mathbb{E}[(f(\hat{\mathbf{x}}) - \overline{f}(\hat{\mathbf{x}}))^2] + \overline{f}(\hat{\mathbf{x}})^2 \\ &+ \mathbb{E}[(\hat{y} - g(\hat{\mathbf{x}}))^2] + g(\hat{\mathbf{x}})^2 \\ &- 2\overline{f}(\hat{\mathbf{x}})g(\hat{\mathbf{x}})\end{aligned}$$

Bias-Variance decomposition

$$\begin{aligned}\mathbb{E}[(f(\hat{\mathbf{x}}) - \hat{y})^2] &= \mathbb{E}[(f(\hat{\mathbf{x}}) - \overline{f}(\hat{\mathbf{x}}))^2] \\ &+ (\overline{f}(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}}))^2 \\ &+ \mathbb{E}[(\hat{y} - g(\hat{\mathbf{x}}))^2]\end{aligned}$$

$$\mathbb{E}[(f(\hat{\mathbf{x}}) - \hat{y})^2] = \text{Variance}(f(\hat{\mathbf{x}})) + \text{Bias}(f(\hat{\mathbf{x}}))^2 + \sigma^2$$

Each error term

Bias: $\overline{f}(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}})$
Average error of $f(\hat{\mathbf{x}})$

Variance: $\mathbb{E}[(f(\hat{\mathbf{x}}) - \overline{f}(\hat{\mathbf{x}}))^2]$
Variance of $f(\hat{\mathbf{x}})$ across different training datasets

Noise: $\mathbb{E}[(\hat{y} - g(\hat{\mathbf{x}}))^2]$ $\mathbb{E}(\epsilon^2) = \sigma^2$
Irreducible noise

Illustrating bias and variance

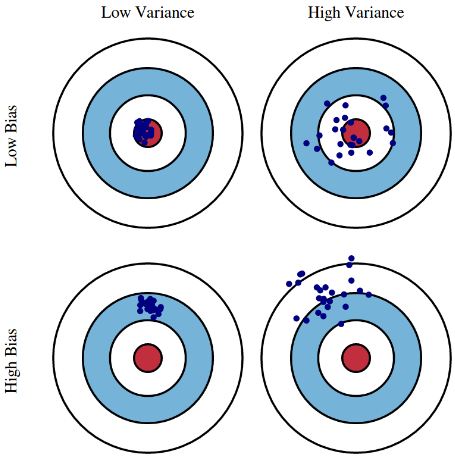


Image from `http://scott.fortmann-roe.com/docs/BiasVariance.html`

Model Selection

Given the bias-variance tradeoff, how do we choose the best predictor for the problem at hand? How do we set the model's parameters?

Measuring bias/variance

Bootstrap sampling: Repeatedly sample observations from a dataset with replacement

For each bootstrap dataset D_b, let V_b refer to the left-out samples which will be used for validation.

Train on D_b to estimate f_b and test on each sample in V_b

Measuring bias/variance

Bootstrap sampling: Repeatedly sample observations from a dataset with replacement

For each bootstrap dataset D_b, let V_b refer to the left-out samples which will be used for validation.

Train on D_b to estimate f_b and test on each sample in V_b

Compute bias and variance

Train-Validation-Test split

Divide the available samples into three sets:

1. Train set: Used to train the learning algorithm
2. Validation/Development set: Used for model selection and tuning hyperparameters
3. Test/Evaluation set: Used for final testing

Cross-Validation

k-fold Cross-Validation

Given: Training set $\mathcal{D}$ of m examples, set of parameters Θ
learner F, number of folds k
Split $\mathcal{D}$ into k folds, $\mathcal{D}_1, \ldots, \mathcal{D}_k$
For each $\theta \in \Theta$, do
for $i = 1 \ldots k$, do
Estimate $f_{i,\theta} = F_\theta(\mathcal{D} \setminus \mathcal{D}_i)$
$\mathrm{err}_\theta = \frac{1}{k} \sum_{i=1}^{k} \mathrm{Loss}(f_{i,\theta})$
Output: $\theta^* = \arg\min_\theta \mathrm{err}_\theta$
$f_{\theta^*} = F_\theta^*(\mathcal{D})$