CS 189 Introduction to Machine Learning
Spring 2019 Jonathan Shewchuk

DIS7

1 Bias and Variance

Oftentimes, such as in linear regression, we model the data-generating process as a noisy measurement of a true underlying response.

$$y_i = f(x_i) + \epsilon_i$$

Where ϵ_i is a zero-mean random noise variable.

We use machine learning techniques to build a hypothesis model $h(x)$ which is fit to the data as an approximation of $f(x)$. We usually don't know $f(x)$, but in the experiment which generated the plots on the next pages, we know $f(x)$ is a straight line.

$$f(x) = wx + b$$

The figures on the next pages show attempts to fit 0-degree, 1-degree, and 2-degree polynomials to f using different subsets of training data.

(a) The third figure is an attempt to fit a quadratic $h(x) = ax^2 + bx + c$ when the underlying f is a line. Why does the quadratic model learn a non-zero a? Why didn't it learn straight lines?

(b) When evaluating models, what do we mean by "bias" of a model-estimation method? Explain the differences we see in the bias for polynomials of degree 0, 1, and 2.

(c) When evaluating models, what do we mean by "variance" of a model-estimation method? Explain the differences we see in the variance for polynomials of degrees 0, 1, and 2.

(d) We can decompose the least squares risk function into bias and variance as done in Lecture Note 12.

$$\begin{aligned}
\mathbb{E}[(h(x) - y)^2] &= \mathbb{E}[h(x)^2] + \mathbb{E}[y^2] - 2\mathbb{E}[y(h(x)] \\
&= \mathrm{Var}(h(x)) + \mathbb{E}[h(x)]^2 + \mathrm{Var}(y) + \mathbb{E}[y]^2 - 2\mathbb{E}[y]\mathbb{E}[h(x)] \\
&= (\mathbb{E}[h(x)] - \mathbb{E}[y])^2 + \mathrm{Var}(h(x)) + \mathrm{Var}(y) \\
&= \underbrace{(\mathbb{E}[h(x) - f(x)])^2}_{\text{bias squared of method}} + \underbrace{\mathrm{Var}(h(x))}_{\text{variance of method}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible error}}
\end{aligned}$$

We can decompose the error this way over the entire dataset, or we can decompose an individual point's error into these three components.

Now, observe the last figure. Why is the variance larger for points near the left and right extremes, and smaller for points in the middle?

(e) Why is our estimate of the bias not zero for the 1- and 2-degree models? Would it be zero if we generated an infinite number of datasets?

(f) How are bias and variance related to overfitting and underfitting?

(g) Does training error provide a measure of bias, variance, or both? How about validation and test error?

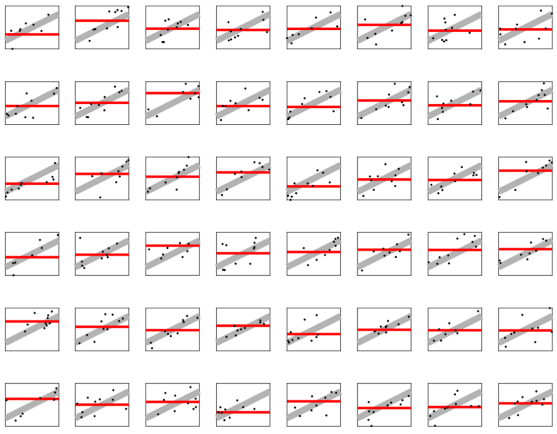
(h) How can we interpret the bias-variance trade-off in hard- and soft- margin SVM? Recall that the soft margin SVM objective is

$$\min |w|_2^2 + C \sum_i \xi_i \quad \text{subject to} \quad y_i(x_i^\top w + \alpha) \geq 1 - \xi_i; \quad \xi_i \geq 0$$

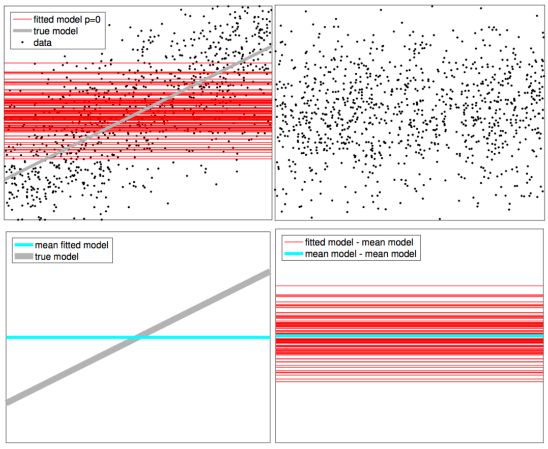
(i) How can we interpret the bias-variance trade-off in LDA and QDA?

The figures on the left show many different models fit on subsets of training data for degrees $p = 0, 1, 2$. The figures on the right, the top left shows all learned models on top of the true model and data. The top right shows the noise of each data point, or the residual after subtracting $y - f(x)$. The bottom left shows the average learned model on top of the true model, and the figure on the bottom right shows all learned models on top of the average learned model.

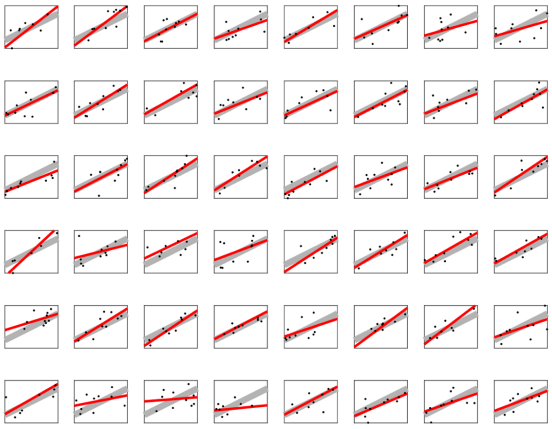
Fitting A Model over Multiple Datasets: $p = 0$



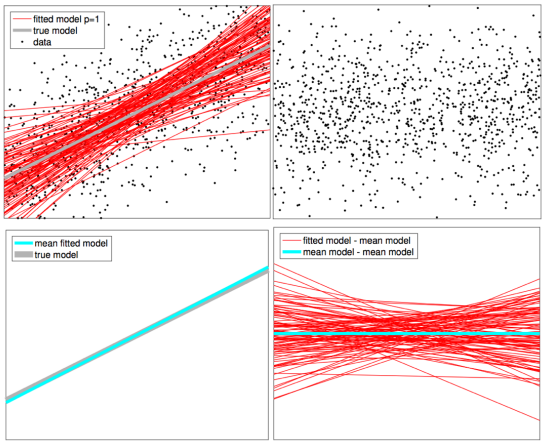
Bias and Variance in Model Selection: $p = 0$



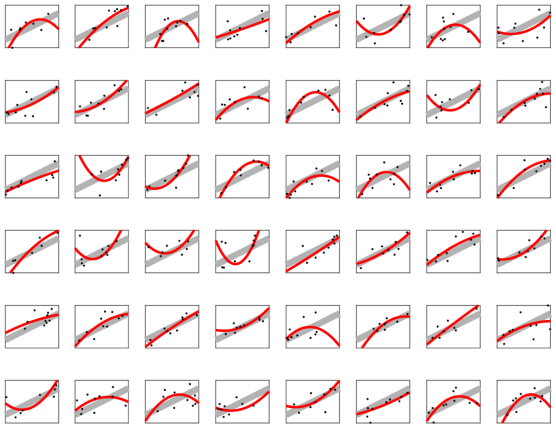
Fitting A Model over Multiple Datasets: $p = 1$



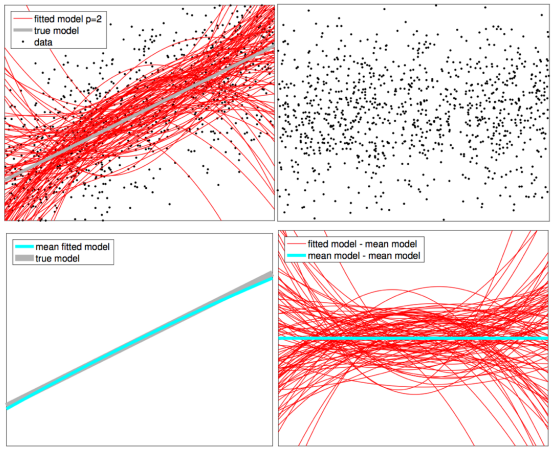
Bias and Variance in Model Selection: $p = 1$



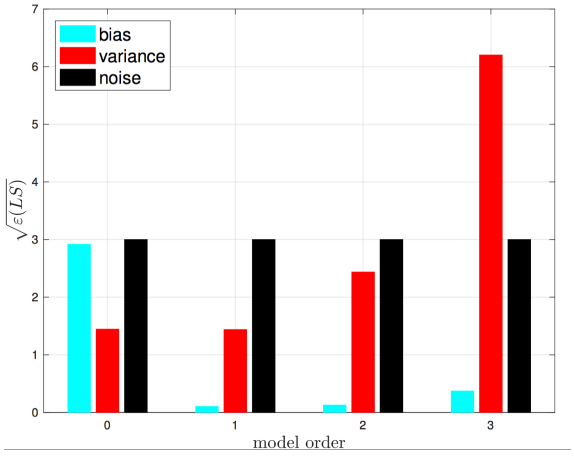
Fitting A Model over Multiple Datasets: $p = 2$



Bias and Variance in Model Selection: $p = 2$



Bias and Variance: Underfitting vs. Overfitting



Variation of Prediction Error with Model Order