CS 189 Introduction to Machine Learning
Summer 2019 Marc Khoury & Brijen Thananjeyan 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?
Solution: The second-order approximation is curving to fit the noise better, because the data includes noise. In the absence of noise, all $\{x_i, y_i\}$ would lie on the same line. In that case, the second-order approximations would learn straight lines, because the data would never suggest curvature.

(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.
Solution: Bias measures how close the average hypothesis (over all possible training sets) can come to the true underlying value $f(x)$, for a fixed value of x. Low bias means that, on average (that is, on average over infinite possible datasets), the regressor $h(x)$ accurately estimates $f(x)$. The degree-0 polynomial has the most bias, because a constant is not expressive enough to learn a sloped line. The assumptions of the degree-0 polynomial are too restrictive to allow us to learn the model accurately. The degree-1 model has the lowest bias because the assumptions of the model estimator are exactly correct. The degree-2 model has low bias because it is expressive enough to capture the first model.

(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.
Solution: Variance measures the variance of the hypothesis (over all possible training sets), for a fixed value of x. A low variance means that the prediction does not change much as the

training set varies. An unbiased method (bias = 0) could have large variance. The degree-0 polynomial has the least variance. It doesn't change as much from data subset to data subset because it always ignores x and guesses the average value of y in the training data. The degree-2 polynomial had the most variance because it will heavily curve to fit noise if it trains on a noisier data subset.

(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?
Solution: Interpolation (estimating within the region of your data) is more stable and reliable than extrapolation (making predictions outside of the main distribution of your training data). Interpolation is more robust because a large deviation from ground truth in the middle of the data range will cause lots of misclassifications, but a large deviation at one of the extremes of the distribution would cause relatively few misclassifications.

(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?
Solution: Bias describes an expectation over all possible datasets, but we are estimating the bias using a finite number of datasets. Our measure of the bias would be zero for the 1-degree model if we used an infinite number of datasets. In the case of the degree-2 model, the curvature would also approach zero as the number of datasets tends to infinity, but more slowly.

(f) How are bias and variance related to overfitting and underfitting?
Solution: High variance models are prone to overfitting. They are more prone to fit the noise of the data rather than the underlying distribution because their assumptions are too weak. High bias models are prone to underfitting, because their assumptions are too restrictive and inexpressive to fit the underlying distribution well.

(g) Does training error provide a measure of bias, variance, or both? How about validation and test error?
Solution: Training error only provides a measure of bias. For example, training error will fool you if you use a 100-degree polynomial to fit a line. The model will heavily overfit and the training error will be zero, but the model will fail to generalize because of the variance

of your model-estimation method. Validation and test error measure generalization ability, so they measure both bias and variance.

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

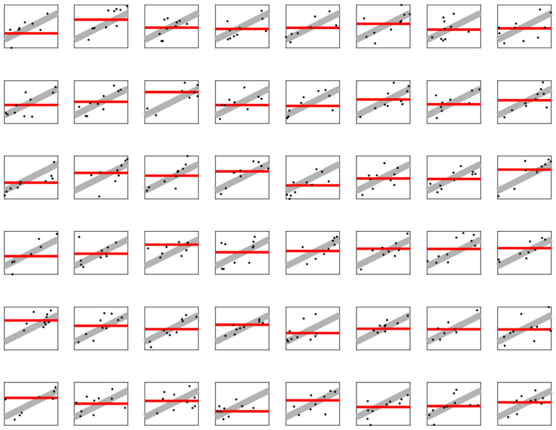
$$\min |w|_2^2 + C \sum_i \xi_i \qquad \text{subject to} \quad y_i(x_i^\top w + \alpha) \geq 1 - \xi_i; \quad \xi_i \geq 0$$

Solution: Hard-margin SVM is essentially soft-margin SVM with the C parameter tending to infinity. Increasing the C parameter shrinks the margin, weighs misclassified points more heavily, and results in more support vectors. One misclassified data point is allowed to have a larger impact on the learned model. Thus, increasing C increases the variance of the model. However, reducing C close to 0 results in underfitting - the margin grows larger because we are allowed to misclassify everything and we have little gravity towards learning the true model - which corresponds to high bias.

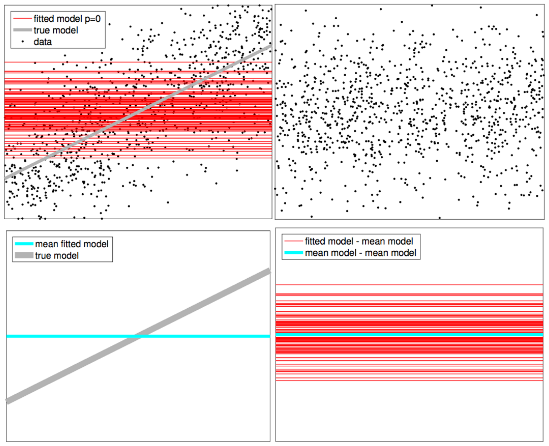
(i) How can we interpret the bias-variance trade-off in LDA and QDA?
Solution: LDA makes a strong assumption which provides regularity: that all class covariance matrices are equal. The model learned by LDA varies less over possible datasets, but it risks underfitting data where class covariance matrices really are unequal. Thus, QDA has more variance and usually (but not always) less bias. For example, if the decision boundary is linear and LDA has zero bias, then QDA can't have less bias than that.

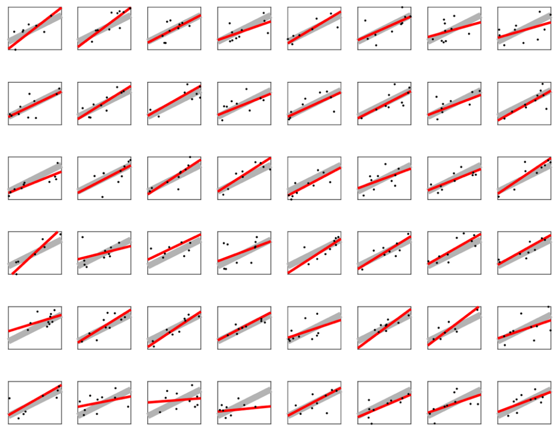
The figures on the left show many different models fit on subsets of training data for degrees $p = 0, 1, 2$. The figures on the right, the top left shows all learned models on top of the true model and data. The top right shows the noise of each data point, or the residual after subtracting $y - f(x)$.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

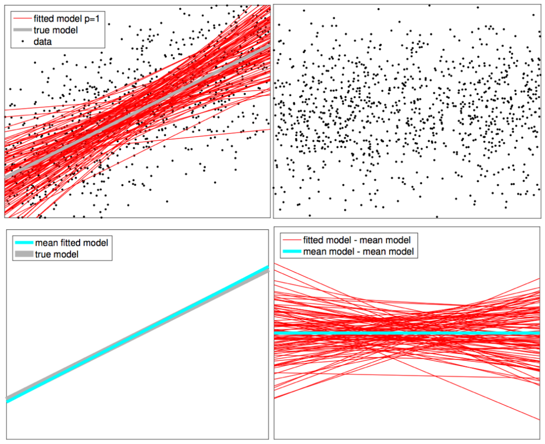


Bias and Variance in Model Selection: $p = 0$

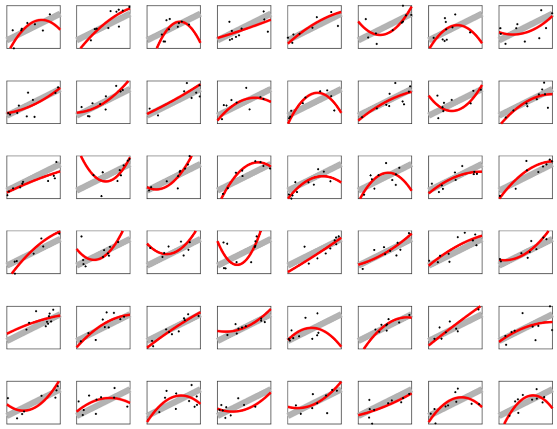




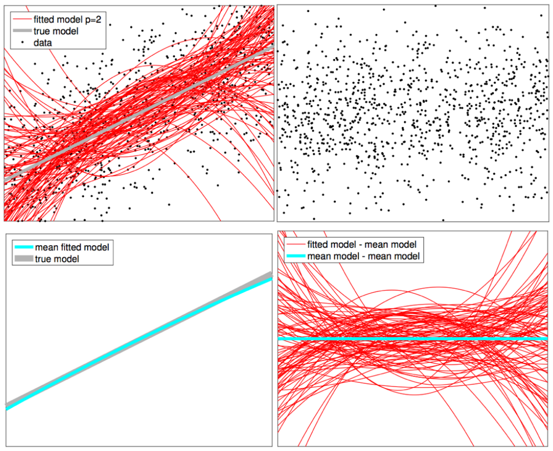
Fitting A Model over Multiple Datasets: $p = 1$



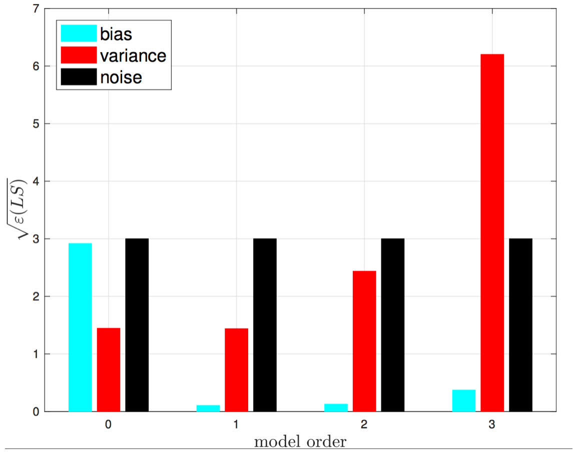
Bias and Variance in Model Selection: $p = 1$



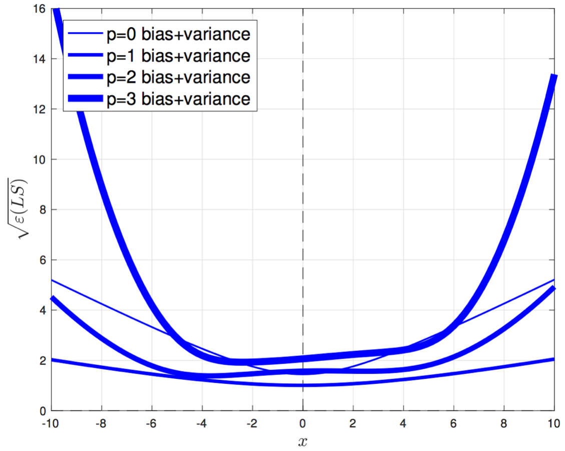
Fitting A Model over Multiple Datasets: $p = 2$



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



Bias and Variance: Underfitting vs. Overfitting



Variation of Prediction Error with Model Order