Model Selection

Polynomial Regression

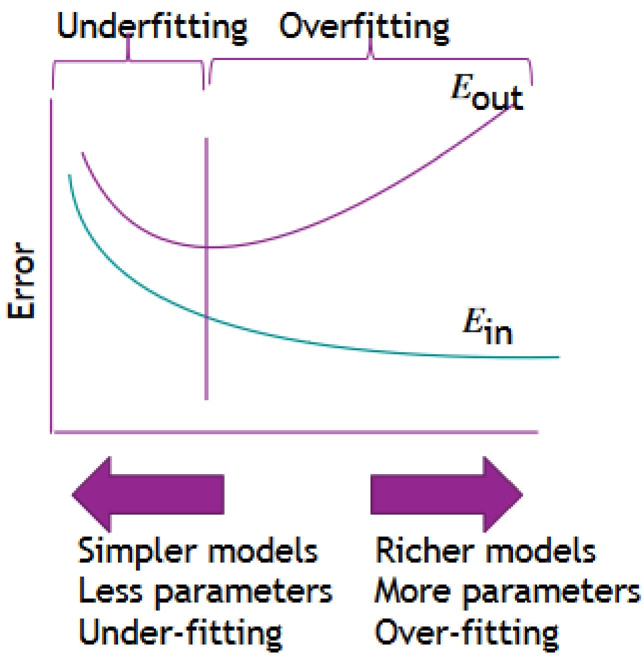
polynomial transform $\Phi_2(x) = [1, x_1, x_2, x_1^2, x_1x_2, x_2^2] = z$

$\hat{y} = \hat{w}^T z = \hat{w}^T \Phi(x)$

Underfitting and Overfitting

What can go wrong with choosing the hypothesis with the smallest cost?

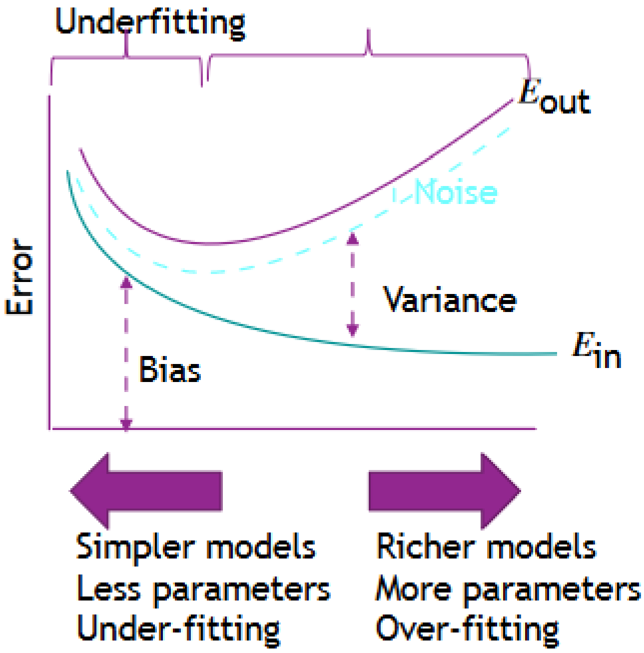
1. Limited Hypothesis class. No function in our hypothesis class can model the data well - **biased solution**
2. Limited Data. We might model the noise and not the true pattern. Small changes to the data causes the hypothesis to change - **high variance solution**



Understanding error: bias and variance

$E_{out}(g) = bias + variance + noise$

- noise: irreducible error
- bias: error of average hypothesis (estimated from N examples) from the true function $f(x) + \epsilon$
    - too simple model (low degree) → add some features, create more complex hypothesis
- variance: how much would the prediction for an example change if the hypothesis was fit on a different set of N points
    - too complex model (high degree) → remove some features, go back to simpler hypothesis



Given: Dataset $D = \{(x^{(1)}, y^{(1)}), ..., (x^{(N)}, y^{(N)})\}$

Learn: If I had a different set of N training examples, I would get a different hypothesis $g^{(D)}(x)$

Expected prediction: $\bar{g}(x) = E_D\left[g^{(D)}(x)\right]$

Intuitive approximation: $\bar{g}(x) = \frac{1}{k}\Sigma_{i=1}^{k} g_i^{(D_i)}(x)$ for $D_1, ..., D_k$

For a hypothesis (e.g. $y = w_0$), cannot fit data because the limitation of the hypothesis itself

$bias(x) = (f(x) - \bar{g}(x))^2$

$$bias = E_x[(f(x) - \bar{g}(x))^2] \approx \frac{1}{N}\Sigma_{i=1}^{N}(f(x^{(i)})) - \bar{g}(x^{(i)}))^2$$

For a hypothesis (e.g. y = w0 + w1x), the difference in hypothesis space (e.g. y = 1+2x; y = -1-2x)

$$var(x) = E_D[(g^D(x) - \bar{g}(x))^2] \approx \frac{1}{L}\Sigma_{l=1}^{L}(g_l^{(D_l)}(x) - \bar{g}(x))^2$$

$$var = E_x[E_D[(g^{(D)}(x) - \bar{g}(x))^2]] \approx \frac{1}{N}\Sigma_{i=1}^{N}\frac{1}{L}\Sigma_{l=1}^{L}(g_l^{(D_l)}(x^{(i)}) - \bar{g}(x^{(i)}))^2$$

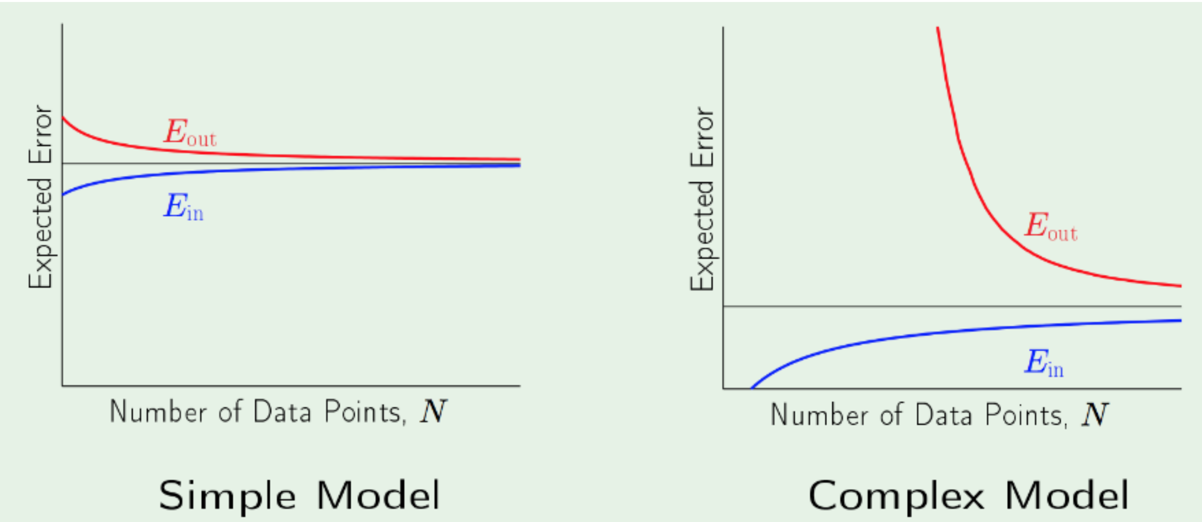
Generalization error (for average D): bias, variance, noise decomposition

$$E_{out}(g^{(D)}) = E_x[(g^{(D)}(x) - y)^2]$$

$$E_D[E_{out}(g^{(D)})] = E_D[E_x[(g^{(D)}(x) - y)^2]] = E_x[E_D[(g^{(D)}(x) - y)^2]] + \sigma^2$$

→ noise

Learning Curves



Confidence

Hoeffding inequality for sample size K, random variables bounded in [a,b], that probability that the average v of random variables deviate from its average μ by more than ϵ:

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 K/(b-a)^2} = \delta$$

With probability $1 - \delta$ the true error is within ϵ of the average error on the test set.

K-fold cross validation

Dividing data into K sets $D_1, D_2, ..., D_k$

for i = 1 to K

train on $D - D_i$

let g_i^- be the fitted model, validation error $e_i = E_{val}(g_i^-)$

return $E_{cv} = \frac{1}{K}\Sigma_{i=1}^{K} e^i$

Regularization—Preventing overfitting

bias $\uparrow$ variance $\downarrow$ large λ: high bias, low var small λ: low bias, high var

$E_{lasso}(w) = E_{in}(w) + \lambda(|w_1| + ... + |w_d|)$ Least Absolute Selection and Shrinkage Operator

$E_{ridge}(w) = E_{in}(w) + \lambda(w_1^2 + ... + w_d^2)$ Note: drop w_0^2

$\nabla E_{ridge}(w) = \frac{2}{N}(X^TXw - X^Ty) + 2\lambda I'w = 0$

$w_{ridge} = (X^TX + N\lambda I')^{-1}X^Ty$