Bias Variance Tradeoff

(Arguably) the most important concept in machine learning!

* Recall from the Regularization & Ridge Regression lecture that, especially in the presence of multicollinearity or lower sample sizes, OLS can overfit to noise and is sensitive to idiosyncracies in the training set, but is unbiased.
Constants like zero or the overall mean are wrong (biased) but extremely stable relative to idiosyncracies in the training set.
Ridge Regression interpolated b/t these two extremes by shrinking OLS estimates towards zero, making OLS stabler and less prone to overfitting.

Q How can we quantify the sensitivity of an estimator to the random idiosyncrasies of a training dataset?

Model Suppose $y_i = f(x_i) + \varepsilon_i$

for some "true" underlying function f and noise ε_i with $\mathbb{E}\varepsilon_i = 0$, so $\mathbb{E}(y_i | x_i) = f(x_i)$.

Goal of ML: estimate f (obtain $\hat{f}$) as "best" as possible from a training dataset $D = \{(x_i, y_i)\}_{i=1}^n$

$\hat{f} = \hat{f}(x; D)$

e.g.
- OLS
- Ridge
- overall mean
- etc.

Want our estimator $\hat{f}$ to be as "close" to the true f as possible;
on average we want $\hat{f}$ to be as close to f as possible,

$$MSE(f, \hat{f}) := \mathbb{E}_{\mathcal{D}}\left[\left(f(x) - \hat{f}(x; \mathcal{D})\right)^2\right]$$

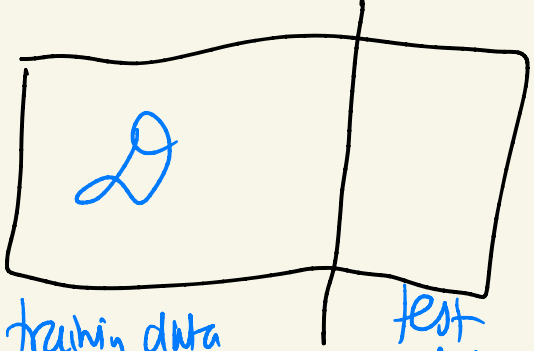
↓ Averaging over the randomness in the training set $\mathcal{D}$

We don't actually observe f, so

$$MSE = \mathbb{E}\left[\left(Y(x) - \hat{f}(x; \mathcal{D})\right)^2\right]$$

↓ Out-of-sample testing data

full dataset



used to create $\hat{f}$

used to evaluate $\hat{f}$

$$MSE(x;\mathcal{D}) = \mathbb{E}\left[\left(Y - \hat{f}(x;\mathcal{D})\right)^2\right]$$

$$= \mathbb{E}\,(Y-\hat{f})^2 \quad \text{where } \hat{f}=\hat{f}(x;\mathcal{D})$$

$$x \mapsto \hat{f}(x)$$

$$= \mathbb{E}\left(Y^2 - 2Y\hat{f} + \hat{f}^2\right)$$

$$= \mathbb{E}\,Y^2 - 2\,\mathbb{E}(Y\hat{f}) + \mathbb{E}(\hat{f}^2)$$

$Y = f + \varepsilon$

$$= \mathbb{E}(f+\varepsilon)^2 - 2\,\mathbb{E}\left[(f+\varepsilon)\hat{f}\right] + \mathbb{E}\hat{f}^2$$

$$= \mathbb{E}\left[f^2 + 2f\varepsilon + \varepsilon^2\right]$$

$$-2\,\mathbb{E}\left[f\hat{f} + \varepsilon\hat{f}\right] + \mathbb{E}\hat{f}^2$$

$x \mapsto f(x)$ is an (unknown) fixed / constant

$$= f^2 + 2f\,\mathbb{E}\varepsilon + \mathbb{E}\varepsilon^2$$

$$-2f\,\mathbb{E}\hat{f} - 2\,\mathbb{E}(\hat{f})\,\mathbb{E}(\varepsilon) + \mathbb{E}\hat{f}^2$$

$\mathbb{E}\varepsilon = 0$

$\mathbb{E}\varepsilon = 0$

$\mathbb{E}\varepsilon^2 = \sigma^2$

$$= f^2 - 2f\,\mathbb{E}\hat{f} + \mathbb{E}\hat{f}^2 + \mathbb{E}\varepsilon^2$$

$$= f^2 - 2f\,\mathbb{E}\hat{f} + \left(\mathbb{E}\hat{f}\right)^2 + \mathbb{E}\hat{f}^2 - \left(\mathbb{E}\hat{f}\right)^2 + \mathbb{E}\varepsilon^2$$

$$= \left(f - \mathbb{E}\hat{f}\right)^2 + \left[\mathbb{E}\hat{f}^2 - \left(\mathbb{E}\hat{f}\right)^2\right] + \mathbb{E}\varepsilon^2$$

$Y = f + \varepsilon$

out-of-sample

$$MSE(x; \mathcal{D}) = \mathbb{E}_{\mathcal{D}}\left[\left(Y - \hat{f}(x; \mathcal{D})\right)^2\right]$$

$$= \left(f - \mathbb{E}_{\mathcal{D}}\hat{f}\right)^2 + \left[\mathbb{E}_{\mathcal{D}}\left(\hat{f}^2\right) - \left(\mathbb{E}_{\mathcal{D}}\hat{f}\right)^2\right] + \mathbb{E}\varepsilon^2$$

$$= Bias\left(\hat{f}\right)^2 + VAR\left(\hat{f}\right) + \text{IRREDUCIBLE ERROR}$$

out-of-sample

$$MSE\left(\hat{f}\right) = Bias\left(\hat{f}\right)^2 + VAR\left(\hat{f}\right) + \text{IRREDUCIBLE ERROR}$$

* $\text{Bias}(\hat{f})^2 = (f - \mathbb{E}_{\mathcal{D}}\hat{f})^2$

Bias is: averaged over $N \to \infty$ draws of the training dataset $\mathcal{D}$, how close is $\hat{f}$ (which is fit from $\mathcal{D}$) to the "true" underlying function f?

* $\text{VaR}(\hat{f}) = \mathbb{E}_{\mathcal{D}}(\hat{f} - \mathbb{E}_{\mathcal{D}}\hat{f})^2$

Variance is: averaged over $N \to \infty$ draws of the training dataset $\mathcal{D}$, how close is $\hat{f}$ to the average estimator across N draws of the training set?

In other words, how variable is $\hat{f}$ across different draws of the training dataset? How sensitive is $\hat{f}$ to the random idiosyncracies of the training set?

overfitting = memorizing the noise of the training set rather than the true underlying trend.
High variance estimators are more prone to overfitting.

* $\mathbb{E}\varepsilon^2$ is irreducible error,
the noise inherent to the problem
(e.g. for pure offers, σ_ε^2 is the inherent noise of scoring a certain number of runs in a half inning)

* Bias-Variance Tradeoff for Fare Effects:

1. Overall mean — very low variance
   very high bias
2. OLS — low bias
   high variance
3. Ridge — introduces bias with the penalty term $+\lambda \sum_j \beta_j^2$ in order to lower variance!

Takeaway to make better predictions, sometimes it helps to introduce some bias to lower the variance!

* We will be returning to the bias-variance tradeoff throughout the ML unit.