Bias-Variance Tradeoff

Cornell CS 3/5780 · Spring 2026

1. Setting

- Training data $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ drawn i.i.d. from $P(X, Y)$
- Regression: $y \in \mathbb{R}$ with squared loss
- Today's question is about generalization: what is my expected test error? (after training on D)
- **Definition**: The *expected label* given $\mathbf{x} \in \mathbb{R}^d$ (recall Bayes optimal prediction)

$$\bar{y}(\mathbf{x}) = E_{y|\mathbf{x}}[Y] = \int_y y \, \Pr(y|\mathbf{x}) \partial y$$

- Question: is $\bar{y}(\mathbf{x})$ a perfect prediction? When is it better/worse?

2. Expected Test Error (given h)

- Apply machine learning algorithm $\mathcal{A}$ to learn a hypothesis h
- **Notation**: $h_D = \mathcal{A}(D)$
- For a specific hypothesis h_D learned on dataset D:

$$E_{(\mathbf{x},y)\sim P}\left[(h_D(\mathbf{x}) - y)^2\right] = \int_x \int_y (h_D(\mathbf{x}) - y)^2 \Pr(\mathbf{x}, y) \partial y \partial \mathbf{x}$$

- This measures: How well does this particular hypothesis generalize?
- **Key observation**: h_D is a random variable!
- Different training sets D lead to different hypotheses
- The hypothesis depends on which data points were sampled

3. Expected Test Error (Given $\mathcal{A}$)

- Taking expectation over both test data **and** training data:

$$E_{\substack{(\mathbf{x},y)\sim P \\ D\sim P^n}}\left[(h_D(\mathbf{x}) - y)^2\right] = \int_D \int_{\mathbf{x}} \int_y (h_D(\mathbf{x}) - y)^2 P(\mathbf{x}, y) P(D) \partial \mathbf{x} \partial y \partial D$$

- Evaluates the quality of algorithm $\mathcal{A}$ given the distribution $P(X, Y)$
- Note: D = training points and $(\mathbf{x}, y)$ = test point
- It is also useful to compute the *average hypothesis* over all possible training sets:

$$\bar{h}(\mathbf{x}) = E_{D\sim P^n}[h_D(\mathbf{x})] = \int_D h_D(\mathbf{x}) \Pr(D) \, \partial D$$

- "Average predictor" across all possible training datasets (weighted average with weight = probability)

4. Decomposition part 1

Goal: $E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-y\right)^2\right] = \underbrace{E_{\mathbf{x},D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)^2\right]}_{\text{Variance}} + E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-y\right)^2\right]$

Fill in below steps: Add and subtract, then expand, then simplify cross term

$$E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-y\right)^2\right] = E_{\mathbf{x},y,D}\left[\left[\left(h_D(\mathbf{x})-\quad\right)+\left(\quad-y\right)\right]^2\right]$$

$$= E_{\mathbf{x},D}\left[\left(h_D(\mathbf{x})-\quad\right)^2\right] + 2E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-\quad\right)\left(\quad-y\right)\right] + E_{\mathbf{x},y}\left[\left(\quad-y\right)^2\right]$$

$$\begin{aligned} E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-\quad\right)\left(\quad-y\right)\right] &= E_{\_\_\_\_}\left[E_{\_\_\_\_}\left[h_D(\mathbf{x})-\quad\right]\left(\quad-y\right)\right] \\ &= E_{\_\_\_\_}\left[\left(E_{\_\_\_\_}\left[h_D(\mathbf{x})\right]-\quad\right)\left(\quad-y\right)\right] \\ &= E_{\_\_\_\_}\left[\left(\quad-\quad\right)\left(\quad-y\right)\right] \\ &= 0 \end{aligned}$$

5. Decomposition part 2

Goal: $E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-y\right)^2\right] = \underbrace{E_{\mathbf{x},y}\left[\left(\bar{y}(\mathbf{x})-y\right)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[\left(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x})\right)^2\right]}_{\text{Bias}^2}$

Fill in below steps: Add and subtract, then expand, then simplify cross term

$$\begin{aligned} E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-y\right)^2\right] &= E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-\quad\right)+\left(\quad-y\right)^2\right] \\ &= E_{\mathbf{x},y}\left[\left(\quad-y\right)^2\right] + E_{\mathbf{x}}\left[\left(\bar{h}(\mathbf{x})-\quad\right)^2\right] + 2E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-\quad\right)\left(\quad-y\right)\right] \end{aligned}$$

$$\begin{aligned} E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-\quad\right)\left(\quad-y\right)\right] &= \\ &= \\ &= 0 \end{aligned}$$

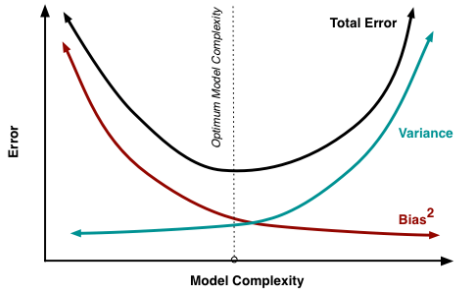
6. Bias-Variance Decomposition

$$\underbrace{E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-y\right)^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{\mathbf{x},D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)^2\right]}_{\text{Variance}} + \underbrace{E_{\mathbf{x},y}\left[\left(\bar{y}(\mathbf{x})-y\right)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[\left(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x})\right)^2\right]}_{\text{Bias}^2}$$

1. **Variance**: How much does h_D vary across different training sets?
   - Model "overfits" to particular training examples
   - Cause: Model is too complex relative to amount of training data
2. **Bias**: How far is the average hypothesis from the true expected label?
   - Model "underfits" the data, model class is not expressive enough
   - Cause: Model is too simple to capture the true pattern
3. **Noise**: Inherent unpredictability in the labels
   - Performance of the Bayes optimal hypothesis
   - Cause: inherent uncertainty and/or uninformative features

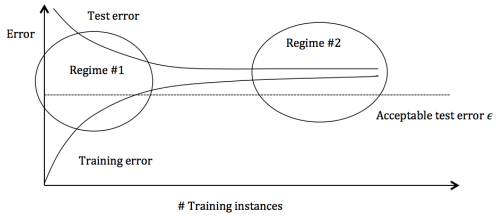
7. Bias-Variance Tradeoffs

Insight: tune the model complexity to trade off Variance and Bias



Diagnosis: where does poor performance come from?

1. High Variance (Regime 1)
   $\text{train err} < \epsilon < \text{test err}$
   indicates overfitting
2. High Bias (Regime 2)
   $\epsilon < \text{train err} \approx \text{test err}$
   indicates underfitting



8. Summary

Bias-Variance Decomposition:

$$\text{Expected Test Error} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

- **Bias**: Error from wrong model assumptions
- **Variance**: Error from sensitivity to training data
- **Noise**: Irreducible error from label randomness
- **Tradeoff**: Complex models have low bias but high variance
- **Goal**: Find the sweet spot that minimizes total error