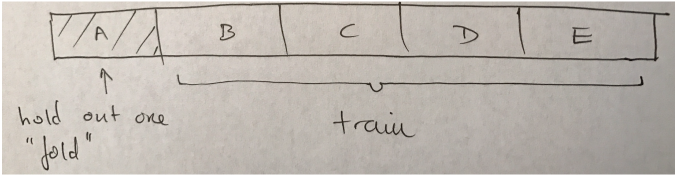
# Bias-Variance Decomposition

February 11, 2021

# Model Selection via Cross Validation




- For each model (e.g., linear, or neural network, or random forest), measure its performance by "holding out" one fold at a time (e.g., with 5 'experiments' as per here)
- For example, train BCDE / validate A; train ACDE / validate on B; etc. Measure average validation error.
- Choose model with best, average validation error. Then train on all data.

# The Bias-Variance Decomposition

$$\textit{generalization error} = \underbrace{\textit{systematic error}}_{\text{bias}} + \underbrace{\textit{sensitivity of prediction}}_{\text{variance}}$$

- Simple models under-fit: will deviate from data (high bias) but will not be influenced by peculiarities of data (low variance).
- Complex models over-fit: will not deviate systematically from data (low bias) but will be very sensitive to data (high variance).

Note: the right tradeoff between bias and variance depends on the amount of data. More data, can use more complex models.

# Bias-Variance: Analysis (1 of 4)

- Define the <u>trained model</u> $f_{\mathcal{D}}(\mathbf{x}) \in \mathbb{R}$.
  - Data $\mathcal{D}$ is a random variable, sampled $\mathcal{D} \sim P^N$ (for distr. $P$).
- Consider some new input $\mathbf{x}$. Conditioned on $\mathbf{x}$, true target $y$ is a random variable (may be noise.)

We're interested in the <u>generalization error</u> at $\mathbf{x}$:

$$\mathbb{E}_{\mathcal{D},\, y|\mathbf{x}}[(y - f_{\mathcal{D}}(\mathbf{x}))^2],$$

where the expectation is taken wrt $\mathcal{D}$ and $y$.

# Bias-Variance: Analysis (2 of 4)

- Define the <u>true conditional mean</u>, $\overline{y} = \mathbb{E}_{y|\mathbf{x}}[y]$.

The generalization error at $\mathbf{x}$ is:

$$\mathbb{E}_{\mathcal{D},\, y|\mathbf{x}}[(y - f_{\mathcal{D}}(\mathbf{x}))^2] = \mathbb{E}_{\mathcal{D},\, y|\mathbf{x}}[(y - \overline{y} + \overline{y} - f_{\mathcal{D}}(\mathbf{x}))^2]$$
$$= \underbrace{\mathbb{E}_{y|\mathbf{x}}[(y - \overline{y})^2]}_{\text{noise}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\overline{y} - f_{\mathcal{D}}(\mathbf{x}))^2]}_{\text{bias+var}} + \underbrace{2\mathbb{E}_{\mathcal{D},\, y|\mathbf{x}}[(y - \overline{y})(\overline{y} - f_{\mathcal{D}}(\mathbf{x}))]}_{0} \tag{1}$$

The last term can be written as

$$2\mathbb{E}_{\mathcal{D}}[\overline{y} - f_{\mathcal{D}}(\mathbf{x})] \cdot \mathbb{E}_{y|\mathbf{x}}[y - \overline{y}] = 2\mathbb{E}_{\mathcal{D}}[\overline{y} - f_{\mathcal{D}}(\mathbf{x})] \cdot 0 = 0.$$

# Bias-Variance: Analysis (3 of 4)

- Define the prediction mean $\overline{f}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]$.

Expanding the second term in (1), we have

$$\mathbb{E}_{\mathcal{D}}[(\overline{y} - f_{\mathcal{D}}(\mathbf{x}))^2] = \mathbb{E}_{\mathcal{D}}[(\overline{y} - \overline{f}(\mathbf{x}) + \overline{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2] =$$

$$\underbrace{(\overline{y} - \overline{f}(\mathbf{x}))^2}_{\text{bias squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\overline{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2]}_{\text{variance}} + \underbrace{2\mathbb{E}_{\mathcal{D}}[(\overline{y} - \overline{f}(\mathbf{x}))(\overline{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))]}_{0} \tag{2}$$

The last term can be written as

$$2(\overline{y} - \overline{f}(\mathbf{x}))\mathbb{E}_{\mathcal{D}}[\overline{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x})] = 2(\overline{y} - \overline{f}(\mathbf{x}))(0) = 0.$$

# Bias-Variance: Analysis (4 of 4)

Substituting (2) back into (1), we have:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D},\, y|\mathbf{x}}[(y - f_{\mathcal{D}}(\mathbf{x}))^2] = \\
&\mathbb{E}_{y|\mathbf{x}}[(y - \overline{y})^2] + (\overline{y} - \overline{f}(\mathbf{x}))^2 + \mathbb{E}_{\mathcal{D}}[(\overline{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2] \\
&= \mathsf{noise}(\mathbf{x}) + (\mathrm{bias}(f(\mathbf{x})))^2 + \mathrm{var}_{\mathcal{D}}(f_{\mathcal{D}}(\mathbf{x})).
\end{aligned}
$$

Depends on noise, and (i) systematic error (or bias), and (ii) sensitivity of the predictor to data (or variance.)

Considering the expectation over $\mathbf{x}$, the generalization error is:

$$
\mathbb{E}_{\mathbf{x}}\left[\mathsf{noise}(\mathbf{x}) + (\mathrm{bias}(f(\mathbf{x})))^2 + \mathrm{var}_{\mathcal{D}}(f_{\mathcal{D}}(\mathbf{x}))\right]
$$

# The Bias-Variance Tradeoff

- If model fits the training data perfectly and there is a small amount of data then the variance will be high (overfits!)
- If model is very simple, then the variance will be low but the bias high (underfits!)
- As $N \to \infty$ the variance $\mathbb{E}_{\mathcal{D}}[(\overline{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2]$ falls, can use a more complex model.

# The Bias-Variance Tradeoff

- If model fits the training data perfectly and there is a small amount of data then the variance will be high (overfits!)
- If model is very simple, then the variance will be low but the bias high (underfits!)
- As $N \to \infty$ the variance $\mathbb{E}_{\mathcal{D}}[(\overline{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2]$ falls, can use a more complex model.