Lecture 13:
Bias-variance tradeoff

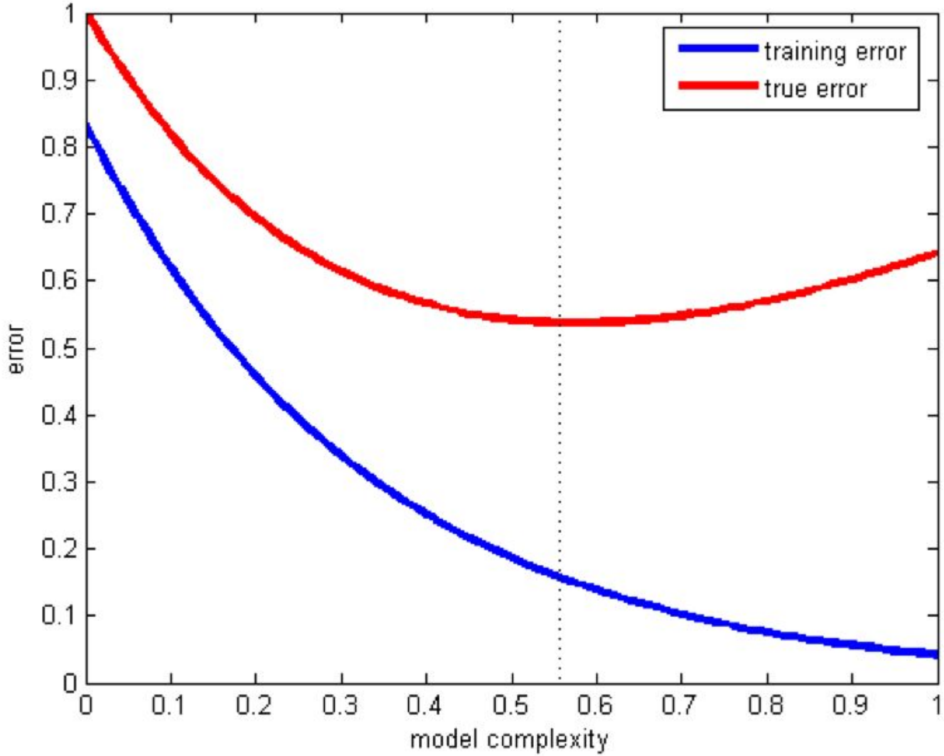
Madeleine Udell
Stanford University

Demo

https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/crime.ipynb

Training vs. test error

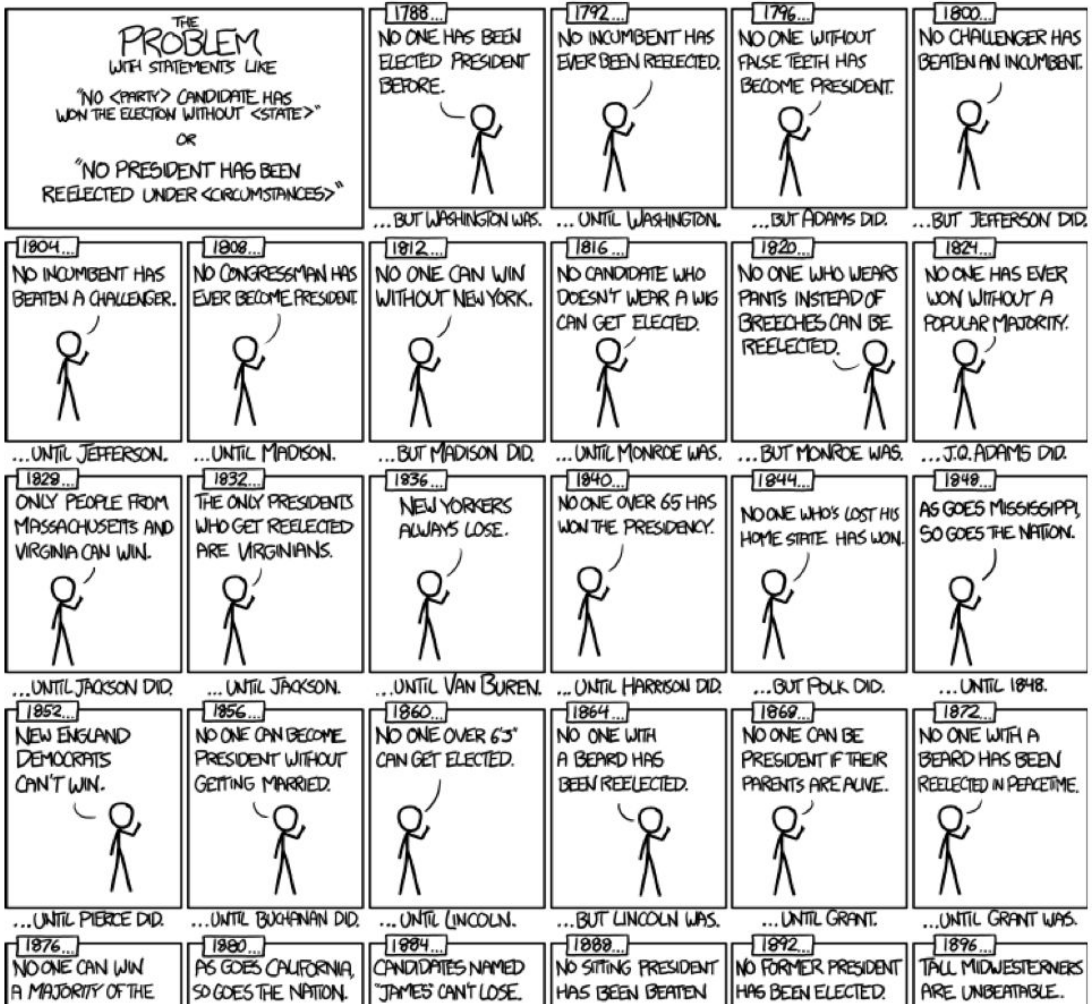
What about performance on new data?



http://stats.stackexchange.com/questions/23331/why-is-there-an-asymmetry-between-the-training-step-and-evaluation-step

Overfitting

The model fits the observed data well, but doesn't generalize well to new instances.


THE PROBLEM WITH STATEMENTS LIKE
"NO <PARTY> CANDIDATE HAS WON THE ELECTION WITHOUT <STATE>"
OR
"NO PRESIDENT HAS BEEN REELECTED UNDER <CIRCUMSTANCES>"
1788...
NO ONE HAS BEEN ELECTED PRESIDENT BEFORE.
...BUT WASHINGTON WAS.
1792...
NO INCUMBENT HAS EVER BEEN REELECTED.
... UNTIL WASHINGTON.
1796...
NO ONE WITHOUT FALSE TEETH HAS BECOME PRESIDENT.
...BUT ADAMS DID.
1800...
NO CHALLENGER HAS BEATEN AN INCUMBENT.
...BUT JEFFERSON DID.
1804...
NO INCUMBENT HAS BEATEN A CHALLENGER.
...UNTIL JEFFERSON.
1808...
NO CONGRESSMAN HAS EVER BECOME PRESIDENT.
...UNTIL MADISON.
1812...
NO ONE CAN WIN WITHOUT NEW YORK.
...BUT MADISON DID.
1816...
NO CANDIDATE WHO DOESN'T WEAR A WIG CAN GET ELECTED.
...UNTIL MONROE WAS.
1820...
NO ONE WHO WEARS PANTS INSTEAD OF BREECHES CAN BE REELECTED.
...BUT MONROE WAS.
1824...
NO ONE HAS EVER WON WITHOUT A POPULAR MAJORITY.
...J.Q. ADAMS DID.
1828...
ONLY PEOPLE FROM MASSACHUSETTS AND VIRGINIA CAN WIN.
...UNTIL JACKSON DID.
1832...
THE ONLY PRESIDENTS WHO GET REELECTED ARE VIRGINIANS.
... UNTIL JACKSON.
1836...
NEW YORKERS ALWAYS LOSE.
...UNTIL VAN BUREN.
1840...
NO ONE OVER 65 HAS WON THE PRESIDENCY.
...UNTIL HARRISON DID.
1844...
NO ONE WHO'S LOST HIS HOME STATE HAS WON.
...BUT POLK DID.
1848...
AS GOES MISSISSIPPI, SO GOES THE NATION.
...UNTIL 1848.
1852...
NEW ENGLAND DEMOCRATS CAN'T WIN.
...UNTIL PIERCE DID.
1856...
NO ONE CAN BECOME PRESIDENT WITHOUT GETTING MARRIED.
...UNTIL BUCHANAN DID.
1860...
NO ONE OVER 6'3" CAN GET ELECTED.
... UNTIL LINCOLN.
1864...
NO ONE WITH A BEARD HAS BEEN REELECTED.
...BUT LINCOLN WAS.
1868...
NO ONE CAN BE PRESIDENT IF THEIR PARENTS ARE ALIVE.
... UNTIL GRANT.
1872...
NO ONE WITH A BEARD HAS BEEN REELECTED IN PEACETIME.
...UNTIL GRANT WAS.
1876...
NO ONE CAN WIN A MAJORITY OF THE
1880...
AS GOES CALIFORNIA, SO GOES THE NATION.
1884...
CANDIDATES NAMED "JAMES" CAN'T LOSE.
1888...
NO SITTING PRESIDENT HAS BEEN BEATEN
1892...
NO FORMER PRESIDENT HAS BEEN ELECTED
1896...
TALL MIDWESTERNERS ARE UNBEATABLE.

Model complexity

If the model is too complex, you risk **overfitting** by “learning” noise.

If the model is not complex enough, you risk **underfitting** by ignoring signal.

Bias-variance tradeoff

Inherent tradeoff between capturing regularities in the training data and generalizing to unseen examples.

Bias: how closely does your model fit the observed data?

Variance: how much would your model fit vary from sample to sample?

Bias-variance tradeoff

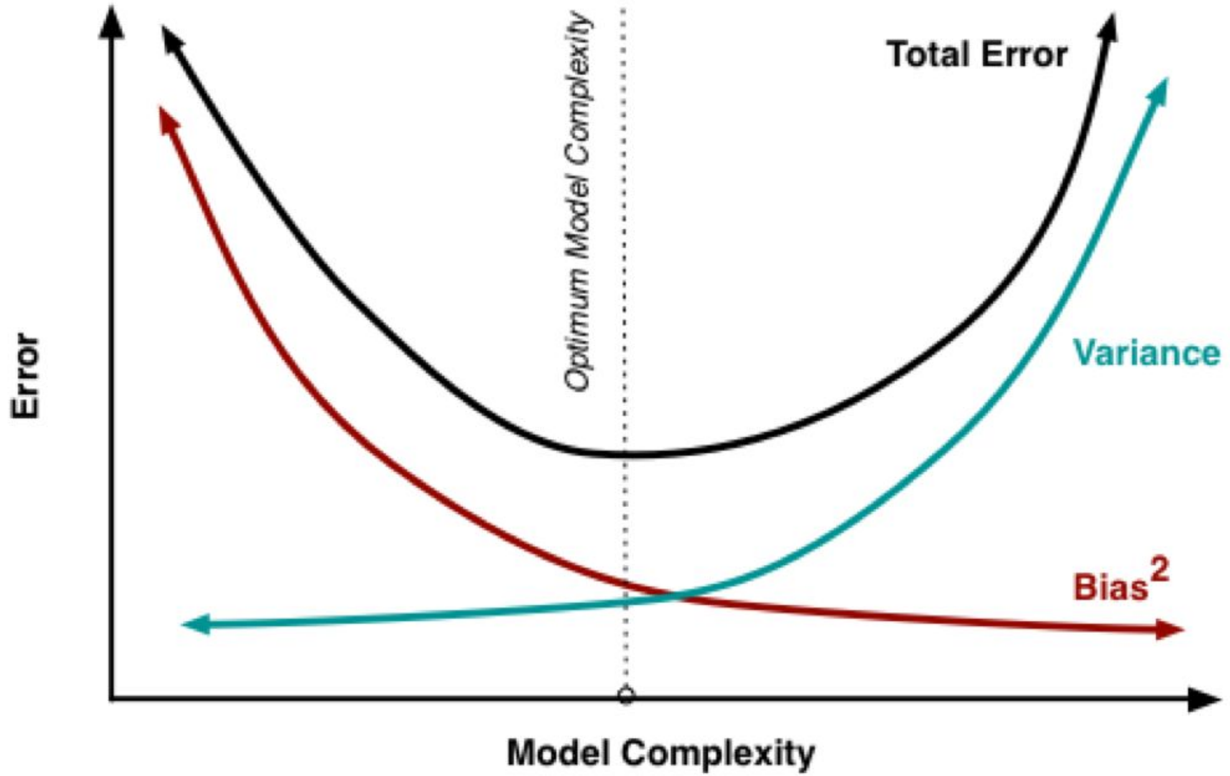
$$y_i = r(x_i) + \epsilon \qquad \mathbb{E}[\epsilon] = 0 \qquad \mathrm{Var}[\epsilon] = \sigma^2$$

Bias-variance tradeoff

$$y_i = r(x_i) + \epsilon \qquad \mathbb{E}[\epsilon] = 0 \qquad \mathrm{Var}[\epsilon] = \sigma^2$$

$$\mathbb{E}\left[(y - \hat{r}(x))^2\right] = \mathrm{bias}\left[\hat{r}(x)\right]^2 + \mathrm{Var}\left[\hat{r}(x)\right] + \sigma^2$$

Expectation is over random instances of the observed data. As model complexity increases, bias typically decreases and variance typically increases.



http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias-variance tradeoff

An example

$$y_i = 1 + x_i + x_i^2 + \epsilon$$ ← The true data generating process

Some bias and some variance

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$ ← The model you fit to the observed data

Bias-variance tradeoff

An example

$$y_i = 1 + x_i + x_i^2 + \epsilon$$ ← The true data generating process

Unbiased but more variance

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$$

Bias-variance tradeoff
An example

$$y_i = 1 + x_i + x_i^2 + \epsilon$$ ← The true data generating process

Unbiased but even more variance

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \hat{\beta}_3 x_i^3$$

Training vs. test error

Training error

Training error is computed on the observed data.

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{r}(x_i))^2$$

Training vs. test error

Test error

Test error is the expected error on new data.

$$\mathrm{Err} = \mathbb{E}\left[(Y - \hat{r}(X))^2\right]$$

Training vs. test error

Training error underestimates test error.

Model selection

Train, validate, test

Training set

Used to fit the models.

Validation set

Used to estimate generalization error for model selection.

Test set

Used to assess performance of the chosen model.

Validation/test set construction

Random subset
K-fold cross-validation
Leave-one-out cross-validation
Temporal partitioning

Validation/test set construction

Random subset

Validation/test set construction

K-fold cross-validation

1. Split the data into ***K*** parts.
2. For k^{th} part, train on other ***K-1*** pieces and validate on k^{th}.
3. Average error across the validation sets.

Validation/test set construction

Leave-one-out cross-validation [LOOCV, K = N]

1. Fit the data on all but the **k^{th}** point.
2. Use the fitted model to predict the **k^{th}** outcome.
3. Average error across the predicted outcomes.

There are computational tricks to avoid re-training the model every time.

Validation/test set construction

Temporal partitioning

How might you perform K-fold cross validation with time series data?

[Discuss with neighbors]