# Tools & Models for Data Science

## Over-Fitting and Regularization

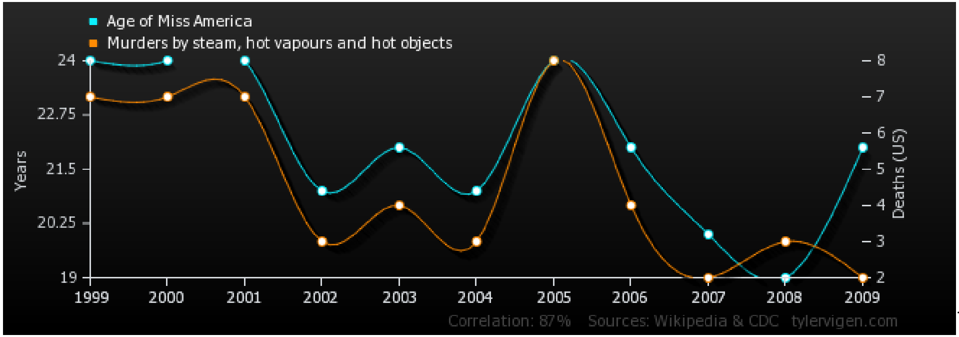
Chris Jermaine & Risa Myers

Rice University

## Over-Fitting

- Fundamental problem in data science
  - Given enough hypotheses to check...
  - One of them is bound to be true

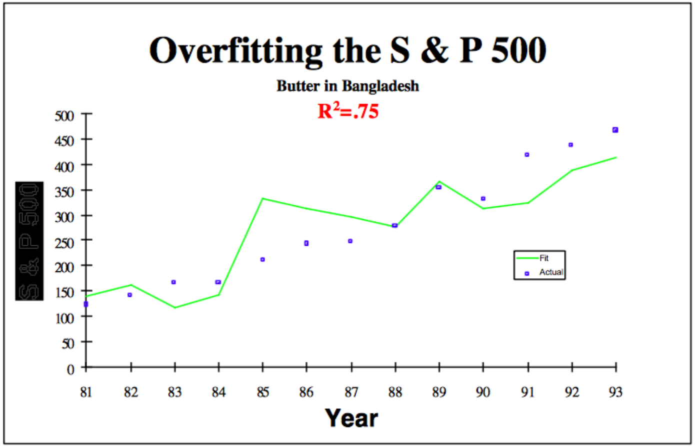
# Miss America and Murder-By-Steam



[1]

[1]`http://tylervigen.com/view_correlation?id=2948`

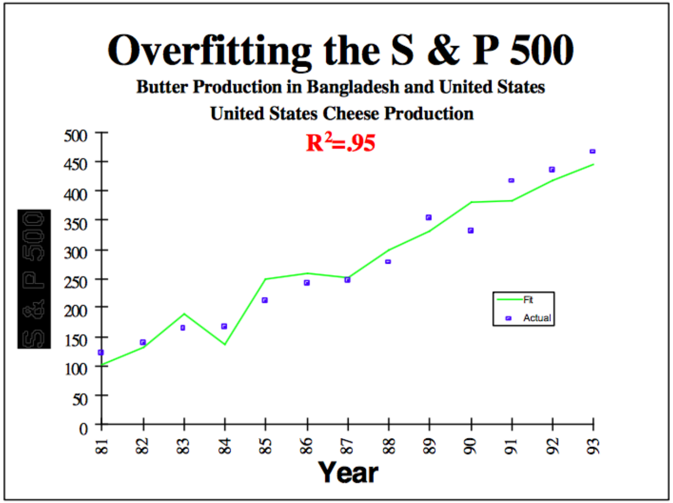
## Predicting the S&P 500



[2]

[2] LeinweberDJ.Stupiddataminertricks: overfittingtheS&P500.JournalofInvesting.2007;16(1):15.

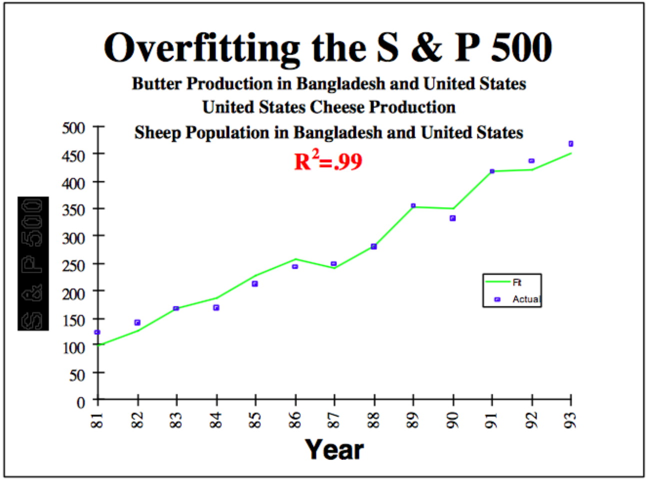
## Predicting the S&P 500



[3]

[3] LeinweberDJ.Stupiddataminertricks: overfittingtheS&P500.JournalofInvesting.2007;16(1):15.

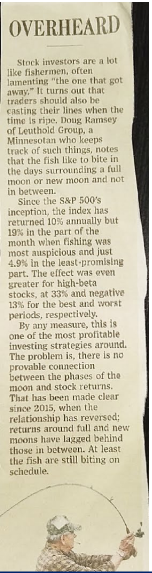
## Predicting the S&P 500



[4]

[4] LeinweberDJ.Stupiddataminertricks: overfittingtheS&P500.JournalofInvesting.2007;16(1):15.

# Fishing and the S&P 500



OVERHEARD

Stock investors are a lot like fishermen, often lamenting "the one that got away." It turns out that traders should also be casting their lines when the time is ripe. Doug Ramsey of Leuthold Group, a Minnesotan who keeps track of such things, notes that the fish like to bite in the days surrounding a full moon or new moon and not in between.

Since the S&P 500's inception, the index has returned 10% annually but 19% in the part of the month when fishing was most auspicious and just 4.9% in the least-promising part. The effect was even greater for high-beta stocks, at 33% and negative 13% for the best and worst periods, respectively.

By any measure, this is one of the most profitable investing strategies around. The problem is, there is no provable connection between the phases of the moon and stock returns. That has been made clear since 2015, when the relationship has reversed; returns around full and new moons have lagged behind those in between. At least the fish are still biting on schedule.

"Since the S&P 500's inception, the index has returned 10.3% annually but 19% in the part of the month when fishing was most auspicious and just 4.9% in the least-promising part. The effect was even greater for high-beta stocks at 33.4% and negative 13.4% for the best and worst periods, respectively."

"By any measure this is one of the most profitable investing strategies around. The problem is, there is no provable connection between the phases of the moon and stock returns. That has been made clear since 2015, when the relationship has reversed–returns around full and new moons have lagged those in between. At least the fish are still biting on schedule."[a]

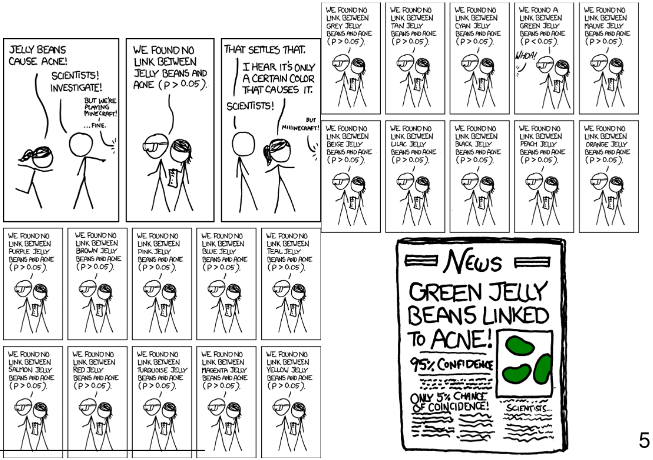
[a] https://blogs.wsj.com/moneybeat/2018/06/01/this-trading-strategy-is-the-reel-deal/

## None of these Models Likely to Generalize

- That means:
  - They’ve learned the input data
  - Not any underlying truth
  - When deployed in the field, likely to fail

# "Data Mining"

- Was originally a derogatory term used by stats
  - Meant that you could always find something if you look hard enough


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

[5]

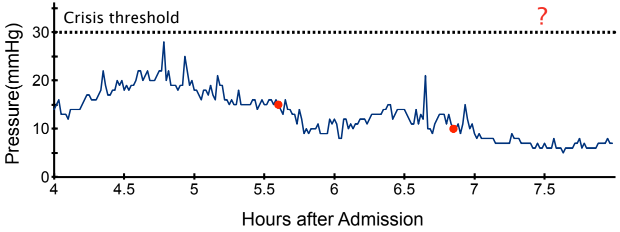
[5]`https://xkcd.com/882/`

## Detecting Over-Fitting

- Detection method number 1: sniff test
    - E.g.: do the regression coefficients make sense?
- Detection method number 2: independent validation and test sets
    - Avoid temptation!
- Detection is important...
    - Avoiding altogether is as well!

## Sniff Test Example

- Predict if patient will be in crisis
- 4-hour time series epochs
- 239 patients
- Researchers started with > 1,000 features
- Reduced to 50
  - including 68th and 143rd most recent observations



## Avoiding Over-Fitting: Occam’s Razor

- The Razor stated simply: When you have many hypotheses that match observed facts equally well, the simplest one is preferred.
  - Been around for a long time!
  - Credited to William of Ockham (died 1347)
  - First stated explicitly by John Punch, 1639: “Entities must not be multiplied beyond necessity”

## Avoiding Over-Fitting: Occam’s Razor

- Of course, it’s not that simple
  - You never have a large number of equally good hypotheses
  - Best you can do: have a preference towards simple models...
  - Under the assumption they will generalize well

## Occam’s Razor Example

- Example:
  - What’s the next number in this sequence?
  - 1, 3, 5, 7, ?

## Overly Complex Sequence

- 1, 3, 5, 7, ?
  - 9

    $f(x) = x + 2$

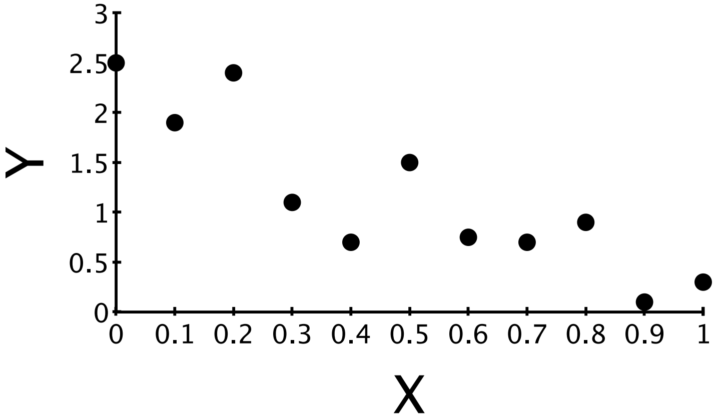
## Overly Complex Sequence

- 1, 3, 5, 7, ?
  - 217341[6]

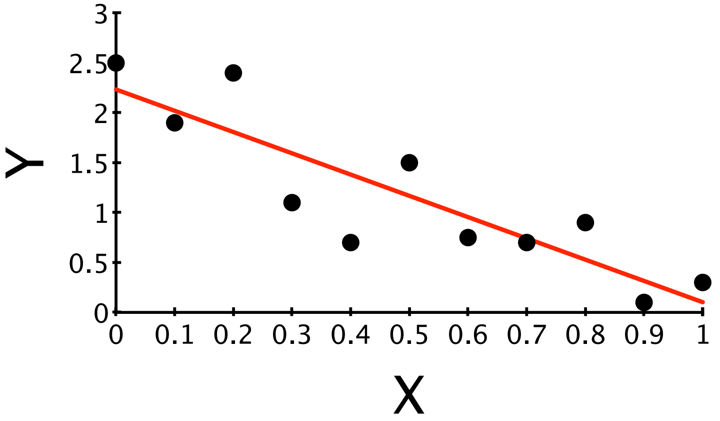
$$f(x) = \frac{18111}{2}x^4 - 90555x^3 + \frac{633885}{2}x^2 - 452773x + 217331$$

[6] https://ml.berkeley.edu/blog/2017/07/13/tutorial-4/

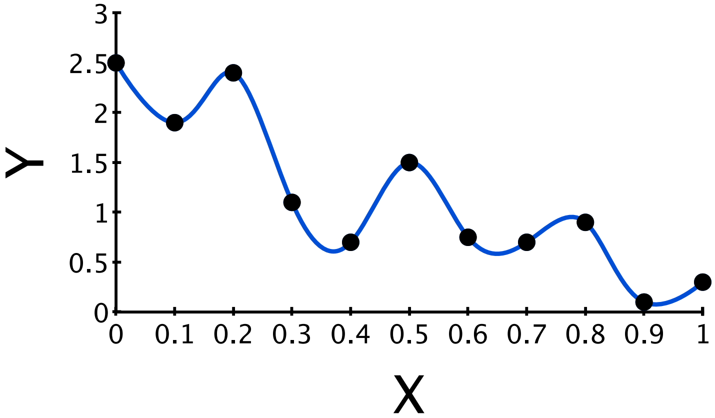
## Avoiding Over-Fitting: Example Data



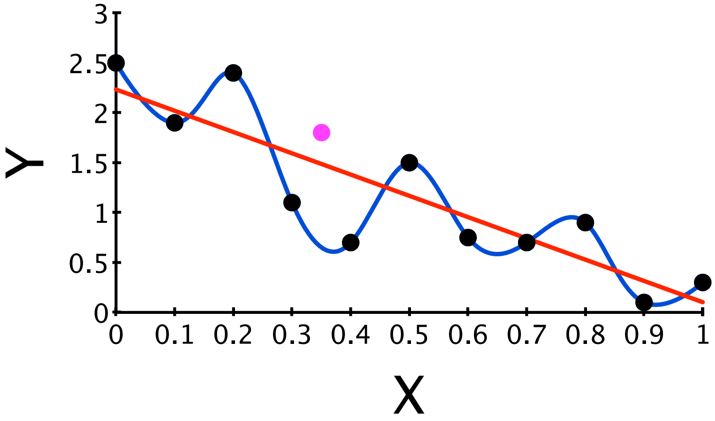
## Avoiding Over-Fitting: Example Linear Regression



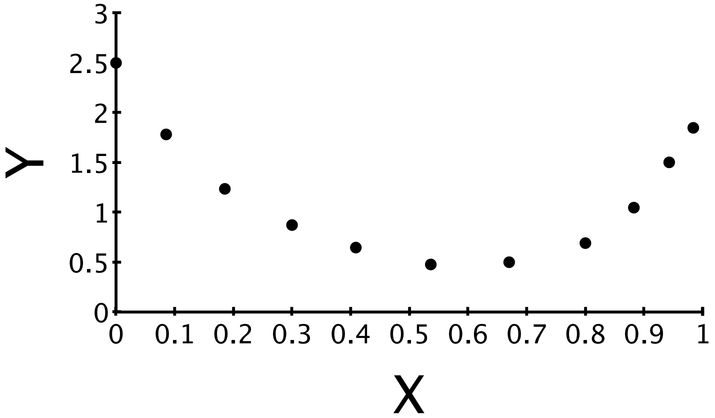
# Avoiding Over-Fitting: Example Over-fitting



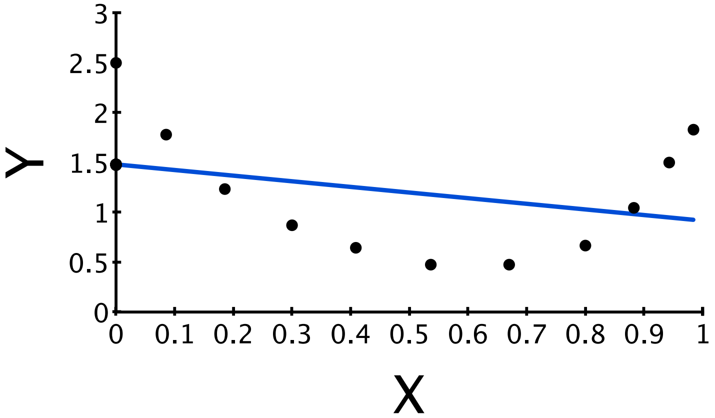
## Avoiding Over-Fitting: Example New Point



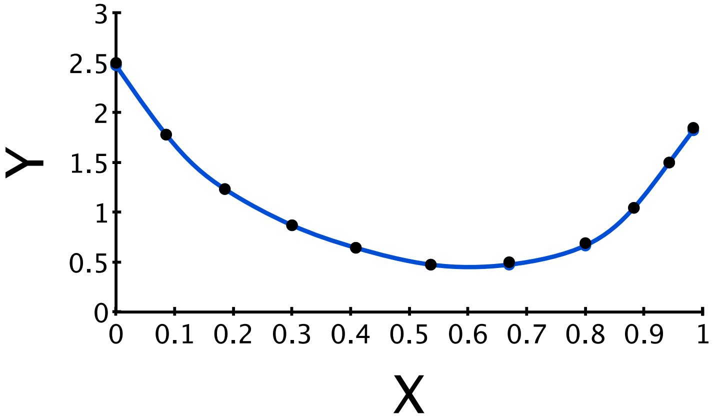
## Avoiding Under-Fitting: Example Data



## Avoiding Under-Fitting: Example Under-fitting

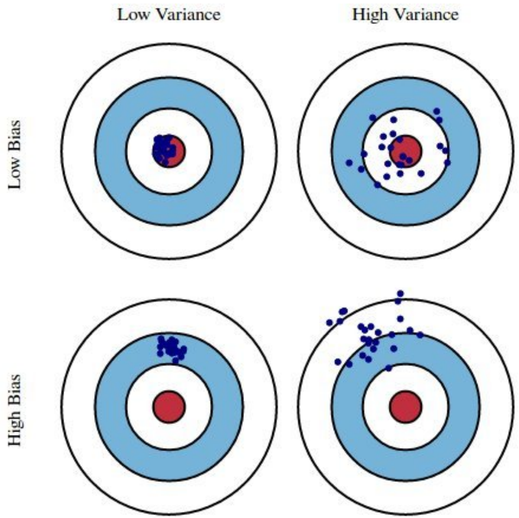


# Avoiding Under-Fitting: Example Fitting



## The Bias-Variance Trade-Off

- In a nutshell, we have two main sources of error in supervised learning
  - "Bias": error from incorrect model assumptions
  - "Variance": sensitivity of model to bad data

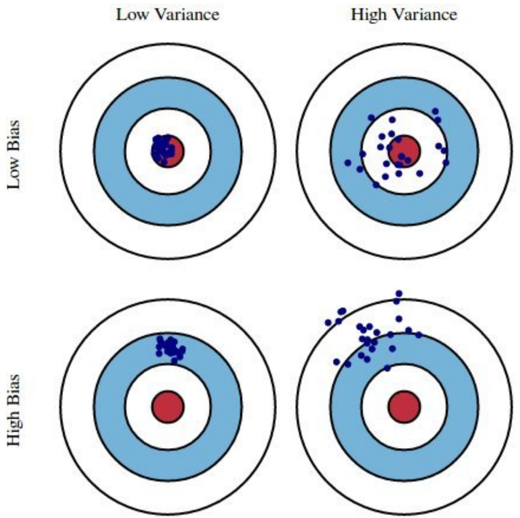


[a]

[a] http://scott.fortmann-roe.com/docs/BiasVaria

## What is Bias?

- Difference between what our model predicts and the truth
- Imperfect models
  - Simplifying assumptions to make the model easier to learn
  - What are common simplifying assumptions?

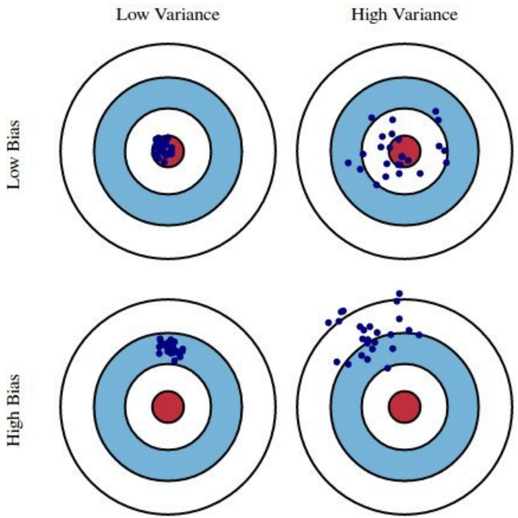


[a]

[a]http://scott.fortmann-roe.com/docs/BiasVariar

# What is Bias?

- Difference between what our model predicts and the truth
- Imperfect models
  - Simplifying assumptions to make the model easier to learn
  - What are common simplifying assumptions?
    - iid data
    - data distributed normally

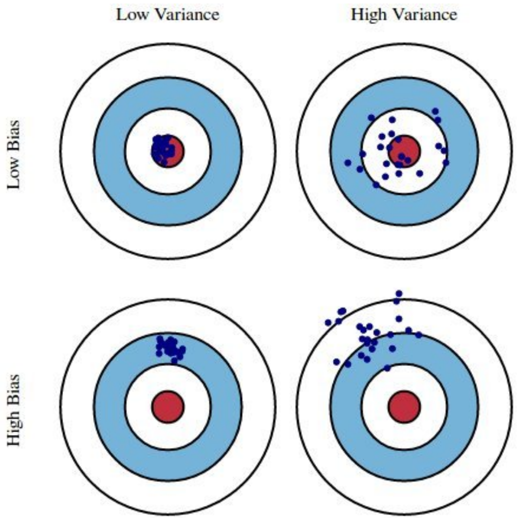


[a]

[a]http://scott.fortmann-roe.com/docs/BiasVaria

# What is Variance?

- Variability of a model's prediction for a data point
- If we build multiple models, how consistent are the predictions across these models?
- Sensitivity of the output to small fluctuations in the training set
- The model is highly dependent on our selection of training data
  - A good distribution of training data will enable us to predict new data well
  - If our training data has a lot of outliers or non-standard values, our predictions will be off

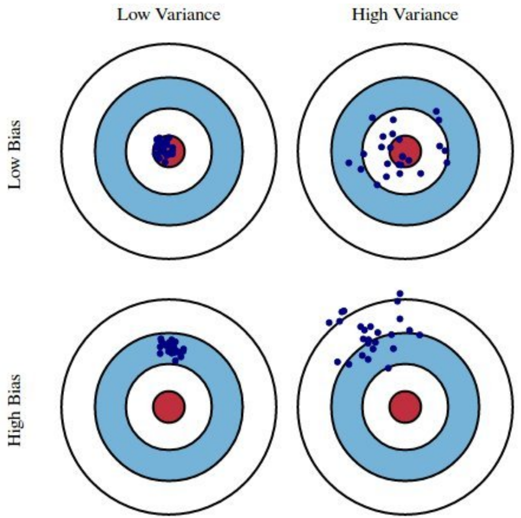


[a]

[a] http://scott.fortmann-roe.com/docs/BiasVariar

# The Bias-Variance Target

- Suppose we have a number of DIFFERENT training sets for the same problem
- Train the same algorithm on each of the training sets
- Each hit on the target is generated by a single model
- A perfect classifier would hit the bulls-eye



[a]

[a] http://scott.fortmann-roe.com/docs/BiasVaria

## Understanding the Trade-Off

- Expected squared error of any prediction is:

$$E[(Y - \hat{f}(X))^2]$$

  - $Y$ is the value we are trying to predict from X
  - $\hat{f}(.)$ is the model we are learning (is a random variable!)
  - $X$ is the observed data (ex: set of regressors)

## Some Definitions

- $Var(Y) = E[Y^2] - E^2[Y]$
- $Bias^2(\hat{f}(x)) = E^2[\hat{f}(x) - Y]$

## Understanding the Trade-Off

- Expected squared error of any prediction is $E[(Y-\hat{f}(X))^2]$
  - Expanding:

$$\begin{aligned} E[(Y-\hat{f}(x))^2] &= E[Y^2+\hat{f}^2(x)-2Y\hat{f}(x)] \\ &= E[Y^2]+E[\hat{f}^2(x)]-E[2Y\hat{f}(x)] \\ &= Var(Y)+E^2[Y]+E[\hat{f}^2(x)]-E[2Y\hat{f}(x)] \\ &= Var(Y)+E^2[Y]+Var(\hat{f}(x))+E^2[\hat{f}(x)]-E[2Y\hat{f}(x)] \\ &= Var(Y)+Var(\hat{f}(x))+(E^2[Y]-E[2Y\hat{f}(x)]+E^2[\hat{f}(x)]) \\ &= Var(Y)+Var(\hat{f}(x))+Bias^2(\hat{f}(x)) \end{aligned}$$

  - Means error is a sum of:
  - “Looseness” of relationship between $X$ and $Y$
  - Sensitivity of the learner to variability of the training data (variance)
  - Inability of the learner $\hat{f}$ to learn the relationship between $X$ and $Y$ (bias)
- It is the second one (variance) that leads to over-fitting

# Understanding the Trade-Off

- How did we derive the last step?
  - Note that $Bias(\hat{f}(x)) = E[\hat{f}(x) - Y]$. So:

$$\begin{aligned} Bias^2(\hat{f}(x)) &= E^2[\hat{f}(x) - Y] \\ &= (E[\hat{f}(x) - Y])(E[\hat{f}(x) - Y]) \\ &= (E[\hat{f}(x)] - E[Y])(E[\hat{f}(x)] - E[Y]) \\ &= E^2[Y] - 2E[\hat{f}(x)]E[Y] + E^2[\hat{f}(x)] \\ &= E^2[Y] - E[2Y\hat{f}(x)] + E^2[\hat{f}(x)] \end{aligned}$$

  - So

$$\begin{aligned} &Var(Y) + Var(\hat{f}(x)) + (E^2[Y] - E[2Y\hat{f}(x)] + E^2[\hat{f}(x)]) \\ &= Var(Y) + Var(\hat{f}(x)) + Bias^2(\hat{f}(x)) \end{aligned}$$

## Irreducible Error

- Since we have:

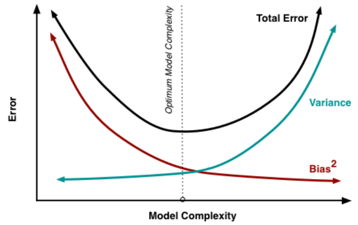
$$E[(Y-\hat{f}(x))^2] = Var(Y) + Var(\hat{f}(x)) + Bias^2(\hat{f}(x))$$

- Means error of supervised learner is a sum of:
  - “Looseness” of relationship between $x$ and $Y$ ($Var(Y)$)
  - Sensitivity of the learner to variability of the training data ($Var(\hat{f}(x))$)
  - Inability of the learner $\hat{f}$ to learn the relationship between $X$ and $Y$ ($Bias^2(\hat{f}(x))$)
- *It is the sensitivity of the learner to variability of the training data*
- *that leads to over-fitting*

## Ideally, Reduce Both Bias and Variance!

- Unfortunately, not possible
- “In real life”...
  - Exceedingly general models are a problem
  - They are difficult to train (lots of local mins, need tons of data)
  - So you design a model to solve your problem
  - Lowers $Var(\hat{f}(X))$ (damage due to over-fitting)
  - But increases so-called “inductive bias” (bias introduced in the model you chose)
  - Best we can do: choose sweet spot where error is minimized
  - To avoid necessity of high model bias, use methods to avoid over-fitting

# Finding the Sweet Spot



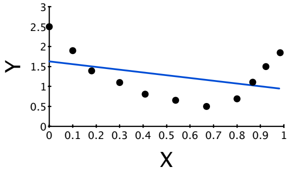
$$\frac{dBias}{dComplexity} = -\frac{dVariance}{dComplexity}$$

[7]

- In practice, there's no way to measure this
  - Try models of different complexity
  - Choose the one with the overall lowest error
  - Relies on choice of error measurement

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

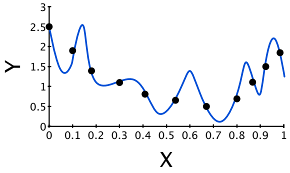
## Underfitting vs. Overfitting



Low Complexity
High Bias
Low Variance

Complexity

$\longrightarrow$



High Complexity
Low Bias
High Variance

## Regularization

- One way to deal with overfitting
- Massively important idea in data science
  - In a nutshell
    - Give learning algorithm ability to choose complexity of model
    - Simpler model: lowers $Var(\hat{f}(X))$, raises inductive bias
    - Complex model: raises $Var(\hat{f}(X))$, lowers inductive bias
- Automatically choose the correct model complexity
  - Balance trade-off between bias and variance
  - To get lowest error
- Done by adding a penalty term to objective function
  - Penalizes model for complexity

# Regularization

- Typically, penalty is based on $l_p$ norm
- Recall, $l_p$ norm of a vector $\langle x_1, x_2, \ldots \rangle$ computed as:

$$\left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

- Common penalties are $l_1$, $l_2$

## Example: Logistic Regression

- Standard objective function to minimize is:

$$\sum_i y_i \theta_i - \log(1 + e^{\theta_i})$$

where $\theta_i$ is $\sum_j x_{i,j} r_j$ or $X_i \cdot \boldsymbol{r}$

- Change objective function to:

$$\sum_i \left( y_i \theta_i - \log(1 + e^{\theta_i}) - \lambda (||r||_p)^p \right)$$

$(||r||_p)^p$ is the $l_p$ norm of the regression coefficients raised to power $p$

- We maximize the objective function, which is why we subtract the penalty term
- $\lambda$ is the regularization parameter

## Example: Logistic Regression

- If $p = 1$ have “the lasso”
- If $p = 2$ have “ridge regression”
- $\lambda$ controls the magnitude of the penalty

$$Loss = \sum_i \left(y_i\theta_i - \log(1 + e^{\theta_i})\right) + \lambda\,(||r||_p)^p$$

- As the $l_p$ norm of the regression coefficients gets bigger, the value of the Loss function increases
- Models with smaller norms will be preferred
  - Typically, try different values of $\lambda$ during validation
  - Usually scale $\lambda$ with $n$
  - Keeps loss balanced between prediction and regularization parts as data size changes

## Ridge Regression

- As known as [Andrey] Tikhonov regularization
- Used when we want an solution to a problem that does not have a unique solution
- “Ridge" local optimums
- Reduces coefficient values
- Has higher bias, but lower variance than least squares estimator
- $l_2$ norm (sum of squares)

## Lasso Regression

- Least Absolute Shrinkage and Selection Operator
- Includes
  - Feature selection
  - Regularization
- Credited to Robert Tibshirani 1996
- Designed for least squares models
- Primary goal to improve accuracy by:
  - Reducing the number of features used in the final model
  - This means it actually eliminates features by letting their regression coefficients be 0
  - Forces the sum of the regression coefficients to be less than a fixed value
- $l_1$ norm (sum of absolute values)

# Choosing between $l_1$ and $l_2$

- $l_1$
  - Can eliminate features
  - ...at the cost of accuracy
  - Good for
    - high dimensional data
    - sparse datasets (lots of features with no values)
    - Classification
- $l_2$
  - Distributes errors among the features
  - ...often more accurate
  - More computationally intensive
  - Good for
    - Dense features
    - Regression

## Closing Remarks

- When regularizing, important to normalize data
  - That is, transform so mean = 0 and variance = 1
  - ? Why?

## Closing Remarks

- When regularizing, important to normalize data
  - That is, transform so mean = 0 and variance = 1
  - Why?
    - In a linear model, input dims with larger values generally lead to larger outputs
    - For a small input value to lead to a large output, it needs a large multiplier
    - But we penalize large multipliers when regularizing
    - To avoid bias towards dims with larger values, need to normalize

## Questions?

- What do we know now that we didn’t know before?

- How can we use what we learned today?