ECE595 / STAT598: Machine Learning I
Lecture 38 Conclusion

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

Outline

- Welcome to the last lecture of this semester.
- It was a great experience discussing machine learning with you.

Today's Lecture:

- Debugging
- Three advices
  - Occam's Razor
  - Sampling Bias
  - Data Snooping
- Final remarks

Reference:

- Learning from Data, chapter 5
- http://cs229.stanford.edu/materials/ML-advice.pdf

Debugging ML

(Modified from Stanford CS 229 Lecture)

Debugging ML Algorithms

- You built a logistic regression model
- You solved this optimization problem

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}(h_{\boldsymbol{\theta}}(\boldsymbol{x}_n), y_n) + \lambda\|\boldsymbol{\theta}\|^2$$

- You test it using a testing test
- Very **big** testing error
- What do you do?

Some Common "Wisdom"

The common wisdom will tell you to try the followings:

- Try getting more training samples
- Try smaller sets of features
- Try large sets of features
- Try changing the features
- Run gradient descent for more iterations
- Try Newton's method
- Use a different λ
- Try SVM

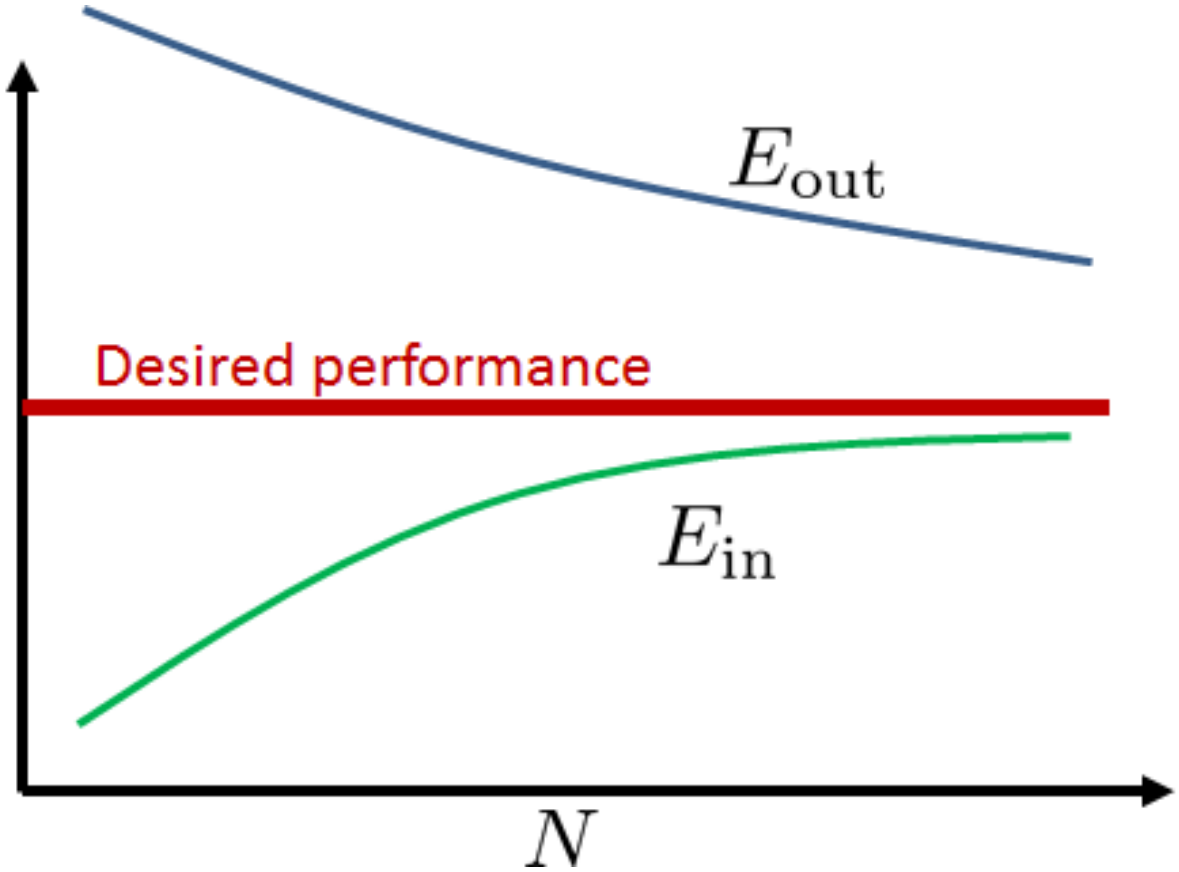
But which one should we try first?

Bias and Variance

Approach: You can try to inspect the bias and variance.

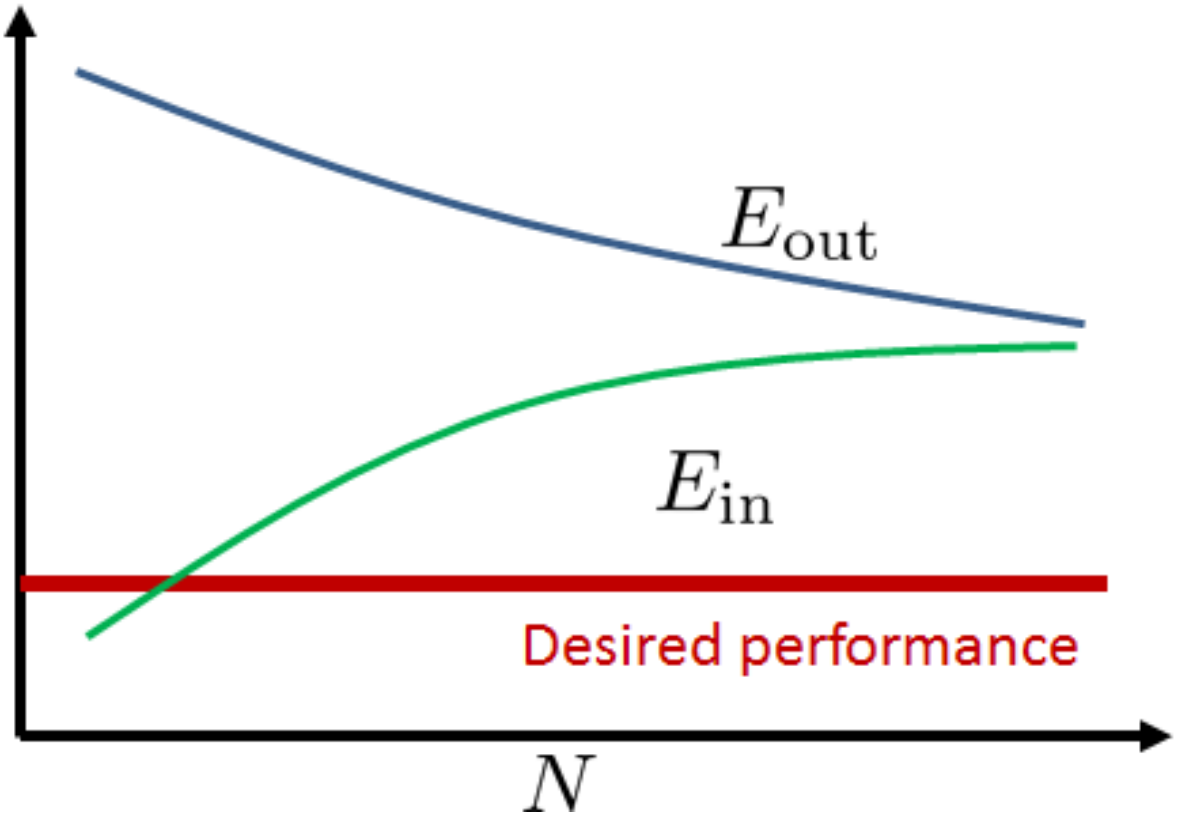
- Suspect 1: Overfitting (High variance)
- Suspect 2: Too few features (High bias)
- Run diagonstic
- Variance: Training error will be much lower than testing error
- Bias: Training error will be high

High Variance



- Test error stays high
- You need more training samples
- Reduce model complexity

High Bias



- Training error high
- Your model is not complicated enough
- Use more features

Some Common "Wisdom"

- Try getting more training samples Fixes high variance
- Try smaller sets of features Fixes high variance
- Try large sets of features Fixes high bias
- Try changing the features Fixes high bias
- Run gradient descent for more iterations
- Try Newton's method
- Use a different λ
- Try SVM

Objective Function or Optimization

Approach: You can try to inspect the optimization

- You tried logistic $\boldsymbol{\theta}_L$
- You tried SVM $\boldsymbol{\theta}_S$
- $\mathcal{L}$: **logistic** training loss function function
- E_{out}: Out-sample error
- Case A: $E_{\text{out}}(\boldsymbol{\theta}_S) < E_{\text{out}}(\boldsymbol{\theta}_L)$, $\mathcal{L}(\boldsymbol{\theta}_S) < \mathcal{L}(\boldsymbol{\theta}_L)$
- Problem: You did not optimize well
- Solution: Go back to your gradient descent
- Fix the optimization **algorithm**

Objective Function or Optimization

- You tried logistic $\boldsymbol{\theta}_L$
- You tried SVM $\boldsymbol{\theta}_S$
- $\mathcal{L}$: logistic training loss function function
- E_{out}: Out-sample error
- Case B: $E_{\text{out}}(\boldsymbol{\theta}_S) \ll E_{\text{out}}(\boldsymbol{\theta}_L)$, $\mathcal{L}(\boldsymbol{\theta}_S) > \mathcal{L}(\boldsymbol{\theta}_L)$
- Problem: Your regularization is too strong or too weak
- Solution: Adjust λ
- Fix your optimization **problem**

Some Common "Wisdom"

- Try getting more training samples Fixes high variance
- Try smaller sets of features Fixes high variance
- Try large sets of features Fixes high bias
- Try changing the features Fixes high bias
- Run gradient descent for more iterations Fixes convergence
- Try Newton's method Fixes convergence
- Use a different λ Fixes objective
- Try SVM Or compare to another method Fixes objective

Good Luck!

Two Approaches to Train a Model

Design:

- Build from scratch.
- Engineer your own sets of features.
- Implement it and hope it works.

Build-and-Fix:

- Download a code from GitHub.
- Implement something quick-and-dirty.
- Diagnostics.
- Fix and run.

Other Tricks

- **Ablation Study**:
- Remove one component and see the performance drop
- Identify the weak spot

- **Baseline Model**:
- Get a baseline model from the internet
- "Warm start" by feeding your own dataset
- It is hard to train ResNet50 using ImageNet from scratch

- **Make It Work First**:
- Danger of over-theorizing
- Often the problem is about your data set
- Often the problem is about the evaluation scheme

Three Advices

1. Occam's Razor

(AML Chapter 5)

Occam's Razor

An explanation of the data should be made
as simple as possible, but *no simpler*.

A. Einstein

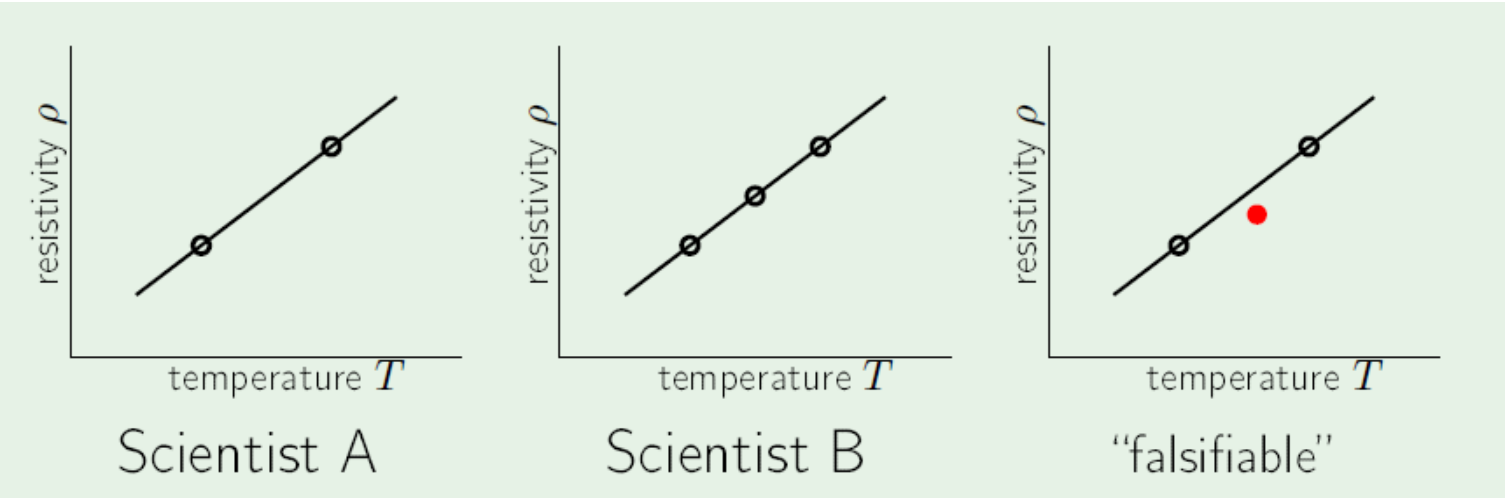
The simplest model that fits the data is also the most plausible.

AML Chapter 5

Why Simpler = Better?

- Two notions of complexity:
- Complexity of $\mathcal{H}$ – VC dimension
- Complexity of h – regularization
- Complex h implies complex $\mathcal{H}$
- E.g., Quadratic vs linear
- If a model is simple, then there will be few models in the family
- Why simple = better?
- If you have a complex hypothesis, then (of course) you can shatter
- This means nothing.
- But if you have a simple hypothesis that can also shatter
- This means a lot.

Who has more Scientific Evidence?

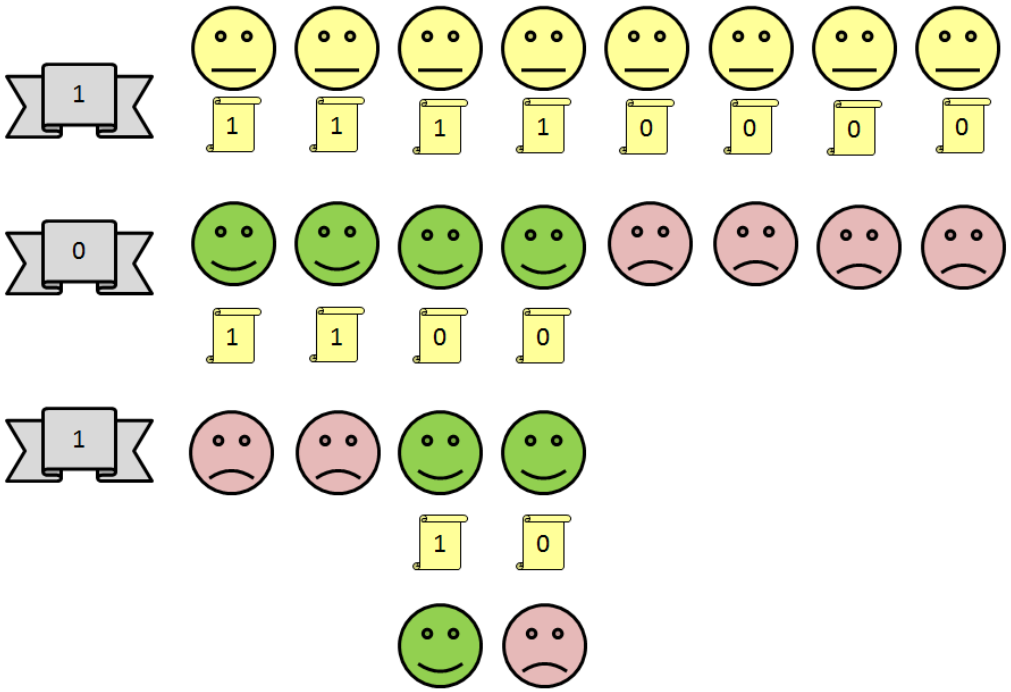


- Having a fit means nothing
- Can the data falsify your hypothesis?
- Hypothesis: resistivity is linear to temperature
- Scientist C has the most convincing evidence
- Scientist B is okay as long as the measurements are exact
- Scientist A has shown no evidence
- Any two point can give a line!
- There is no way for the data to say your hypothesis is wrong

Football Game

- Someone sent you a letter, predicting the winner of a game.
- You read the letter, and you watched the game. It was correct.
- The person sent you a letter one week later, predicting the next game.
- You watched the second game. It was correct.
- Repeat for 5 weeks.
- You received a letter. Pay $50 to get the next prediction.
- Should you pay?
- The prediction fits well to the data.

Football Game



Football ~~Game~~ Spam

- Should you pay?
- The prediction fits well to the data.
- No. Because he had sent 2^5 letters to 2^5 different people.
- You are just one of those.
- You happened to be the lucky guy who got all 5 games correct.
- A fit means nothing.

Three Advices

Sampling Bias
(AML Chapter 5)

Sampling Bias

- In 1948, Truman ran against Dewey for president election
- Chicago Daily Tribune ran a phone interview of how people voted
- Dewey won the poll
- On the actual election day, Truman won


Chicago Daily Tribune

HOME

DEWEY DEFEATS TRUMAN

G.O.P. Sweep Indicated in State; Boyle Leads in City

REPUBLICAN TICKET AHEAD OF 1944 VOTE

Tops Coghlan in Hot Race for Attorney

RECORD CITY VOTE SEEN IN LATE TALLIES

BULLETINS ON ELECTIONS

Early Count Gives G.O.P. Senate Edge

PUTS G.O.P. BACK IN THE WHITE HOUSE

What could have gone wrong?

- Nothing wrong with Hoeffding inequality
- Not the fault of δ

$$\mathbb{P}[|E_{\text{in}} - E_{\text{out}}| > \epsilon] \leq \delta$$

- In 1948, phone were expensive
- Training set $\neq$ population set
- Garbage in - garbage out

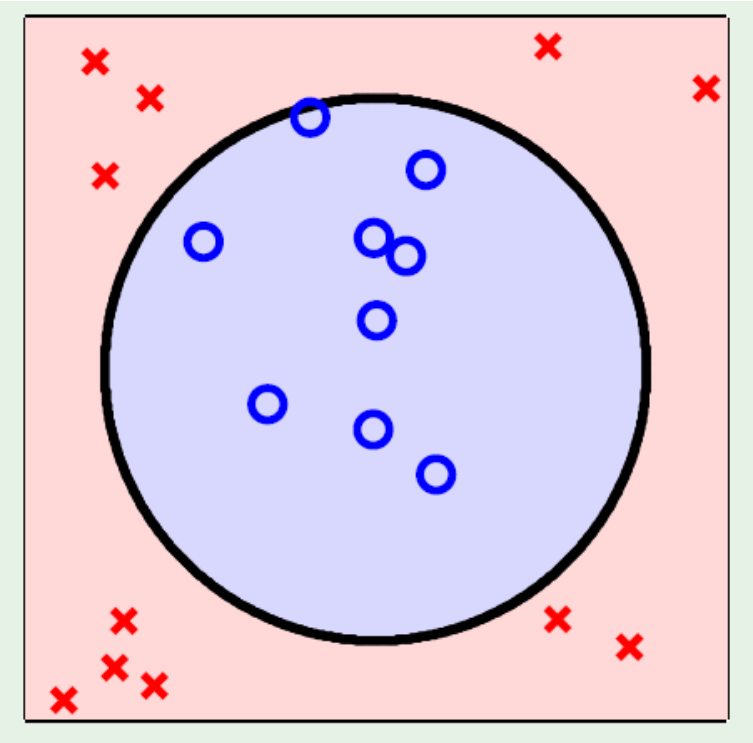
If the data is sampled in a biased way,
learning will produce a similarly biased solution.

AML Chapter 5

Three Advices

Data Snooping
(AML Chapter 5)

Data Snooping



Data Snooping

- Suppose you have a data set
- You choose a transform

$$\boldsymbol{z} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$$

- How about

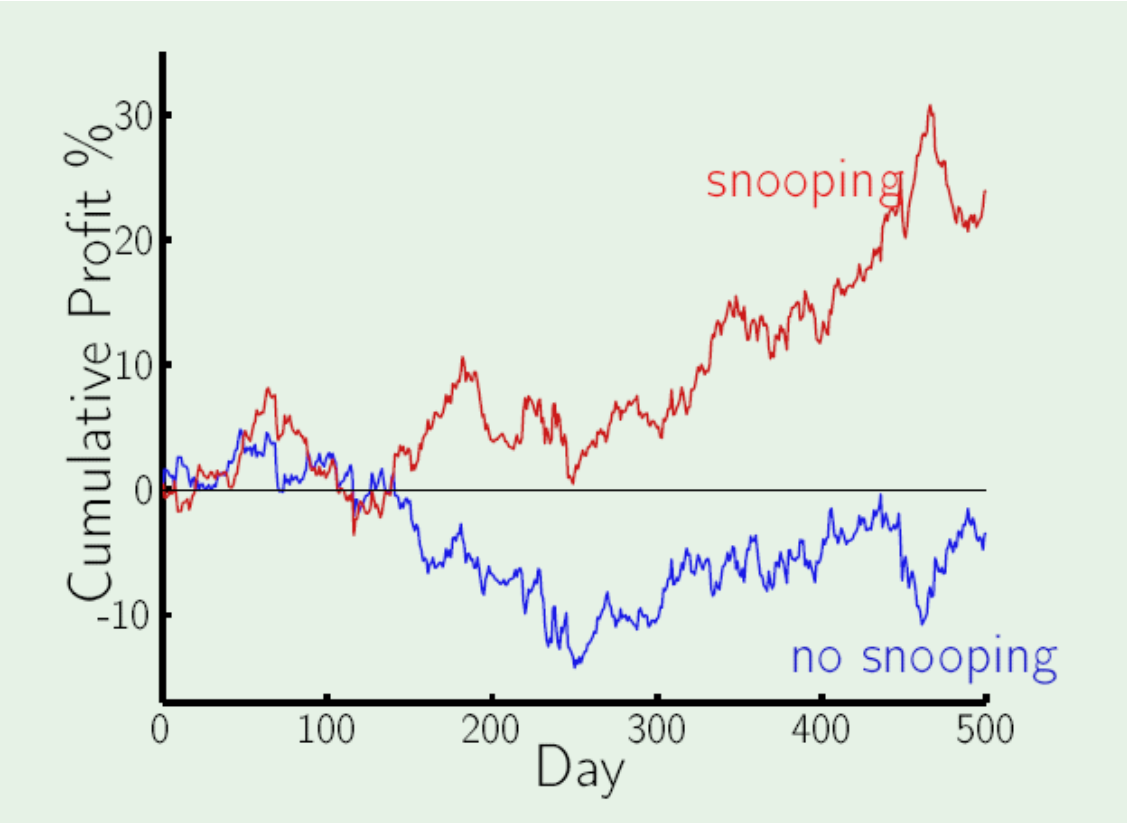
$$\boldsymbol{z} = (1, x_1^2, x_2^2)$$
$$\boldsymbol{z} = (1, x_1^2 + x_2^2).$$

- You did the model selection in your brain
- You saw the data
- Okay if you know a priori from the physics that it is circle.
- Not okay if you look at the data and decide

Forecasting

- Predict US dollar to British pound
- Collect data $\mathcal{D}$
- Normalize
- Split into $\mathcal{D}_{\text{traing}}$ and $\mathcal{D}_{\text{test}}$
- Lock $\mathcal{D}_{\text{test}}$ in a safe
- Train on $\mathcal{D}_{\text{traing}}$
- Open the safe, and test on $\mathcal{D}_{\text{test}}$
- You did a good prediction!
- You sell your algorithm to a bank

Forecasting



Data Snooping

If you torture the data long enough, it will confess.

AML Chapter 5

If a data set has affected any step n the learning process, its ability to assess the outcome has been compromised.

AML Chapter 5

Acknowledgement

Acknowledgement

- College of Engineering
- School of ECE
- Faculty
- TAs:
  - Guanzhe Hong
  - Tolunay Sefi
- Supporting Staff
- You! Thanks for bearing us during the COVID-19 outbreak.
- It is our first time recording videos in a studio. If it looks idiot, please give us a smile.
- Your continuous support is very important.
- Remember to do the course evaluation.