CSE 417T
Introduction to Machine Learning

Lecture 12

Instructor: Chien-Ju (CJ) Ho

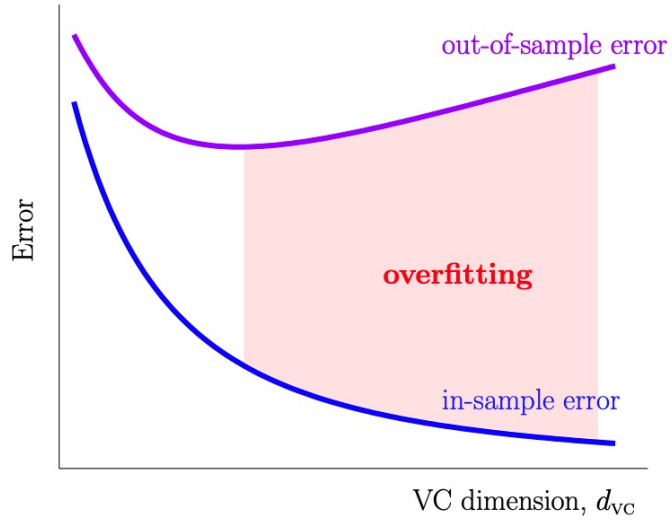
Logistics

- Homework 3: Due **Mar 19 (Friday)**
  - Keep track your own late days
- Homework 1 returned: Regrade requests till this Friday
  - Please be concise and polite
- Exam 1: **Mar 23 (Tuesday)**
  - By default, everyone is expected to take it during lecture time
    - Let me know **by this Friday (via private Piazza post)** if you can't do it during lecture time
  - Follow Piazza announcements
  - A dummy exam is on Gradescope
    - Try the feature of file uploading and test scenarios that you might encounter
  - I'll give some practice questions next week
  - Next Thursday lecture will be a review lecture

Recap

Overfitting and Its Cures

- Overfitting
  - Fitting the data more than is warranted
  - Fitting the noise instead of the pattern of the data
  - Decreasing E_{in} but getting larger E_{out}
  - When H is too strong, but N is not large enough



- Regularization
  - Intuition: Constraining H to make overfitting less likely to happen

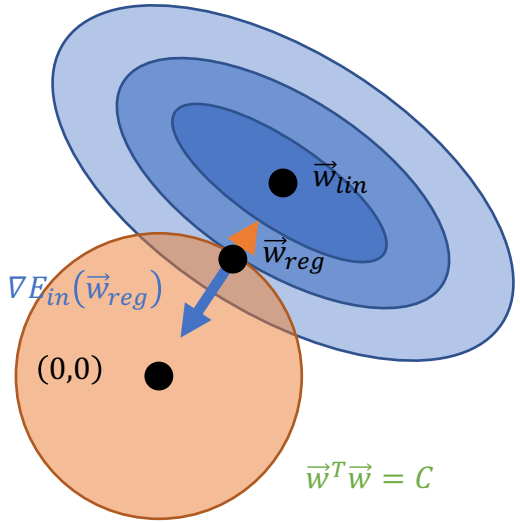
- Validation
  - Intuition: Reserve data to estimate E_{out}

Regularization (Constraining H)

- Weight decay

$$H(C) = \{h \in H_Q \text{ and } \vec{w}^T\vec{w} \leq C\}$$

- Algorithm: Find $g \in H(C)$ such that $g \approx f$



Constrained optimization

minimize $E_{in}(\vec{w})$

subject to $\vec{w}^T\vec{w} \leq C$

equivalent

Unconstrained optimization

minimize $E_{in}(\vec{w}) + \frac{\lambda_C}{N}\vec{w}^T\vec{w}$

Augmented error

Augmented Error

$$E_{aug}(h, \lambda, \Omega) = E_{in}(\vec{w}) + \frac{\lambda}{N}\Omega(h)$$

- Key components
  - Ω: Regularizer
  - λ: Amount of regularization

- Does the form look familiar? Recall in the VC Theory (treating δ as a constant)
  - $E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{d_{vc}\frac{\ln N}{N}}\right)$

- If we pick the right Ω, E_{aug} can be a good proxy for E_{out}

How to Pick the Right Ω

- Intuition: pick Ω that leads to "smoother" hypothesis
  - Overfitting is due to noise
  - Informally, noise is generally "high frequency"

- Computation: prefer Ω that makes the optimization easier (e.g., convex/differentiable)
  - Similar to picking the error measure

- We might have some other objective in mind
  - Ex: L-1 regularizer leads to weight vectors with more 0s

Summary of Regularization

- Regularization is everywhere in machine learning

- Two main ways of thinking about regularization
  - Constraining H to make overfitting less likely to happen
    - Will discuss more regularization methods in the 2nd half of the semester
    - Pruning for decision trees, early stopping / dropout for neural networks, etc
  - Define augmented error E_{aug} to better approximate E_{out}
    - $E_{aug}(h, \lambda, \Omega) = E_{in}(\vec{w}) + \frac{\lambda}{N}\Omega(h)$

- We show the equivalence of the two for weight decay
  - The conceptual equivalence is general with Lagrangian relaxation (will cover later in the semester)

Today’s Lecture

The notes are not intended to be comprehensive. They should be accompanied by lectures and/or textbook.

Let me know if you spot errors.

Prevent Overfitting

$$E_{out}(g) = E_{in}(g) + \text{overfit penalty}$$

- Regularization
  - Choose a regularizer Ω to approximate the penalty
- Validation
  - Directly estimate E_{out} (The goal of learning is to minimize E_{out})

Review of Test Set (Estimate E_{out})

- Out-of-sample error $E_{out}(g) = \mathbb{E}_{\vec{x}}[e(g(\vec{x}), y)]$
  - Key: $\vec{x}$ need to be out of sample

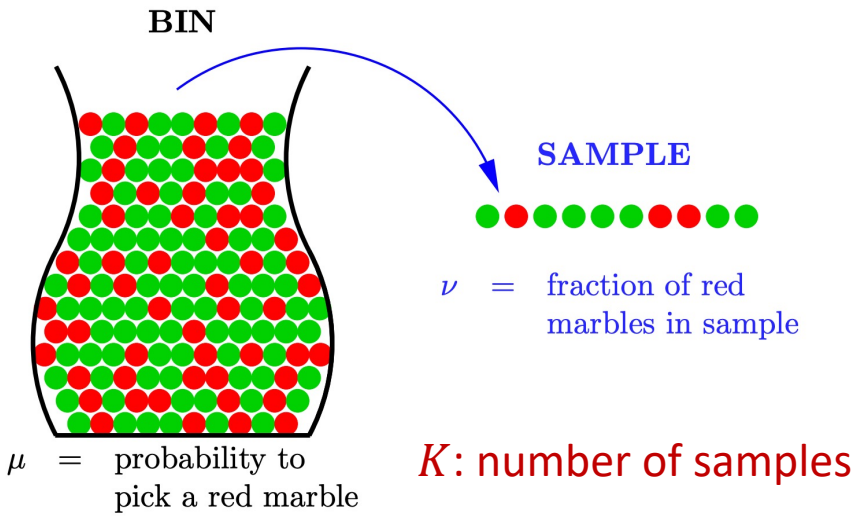
- Test set $D_{test} = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_K, y_K)\}$
  - Reserve K data points
  - None of the data points in test set can be involved in training

- Using the data in test set to estimate E_{out}
  - Since all data points in D_{test} are out of sample

Test Set

- Test set $D_{test} = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_K, y_K)\}$
- For a g learned using only training set

- Let $E_{test}(g) = \frac{1}{K}\sum_{k=1}^{K} e(g(\vec{x}_k), y_k)$
  - $E_{test}(g)$ is an unbiased estimate of $E_{out}(g)$
    - $\mathbb{E}[E_{test}(g)] = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[e(g(\vec{x}_k), y_k)] = E_{out}(g)$
  - Single-hypothesis Hoeffding bound applies
    - $E_{out}(g) \leq E_{test}(g) + O\left(\sqrt{\frac{1}{K}}\right)$

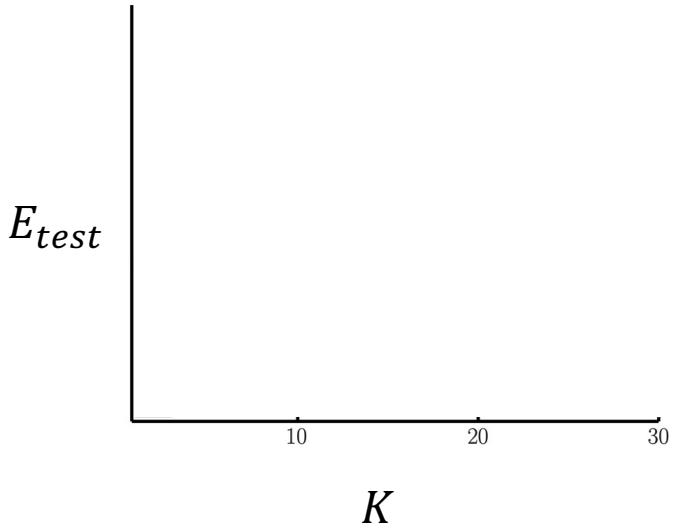


K: number of samples

Where are Test Set From?

- Given a data set D of N points
  - $D = D_{train} \cup D_{test}$
  - Reserving K points for test set means we only have $N - K$ points for training

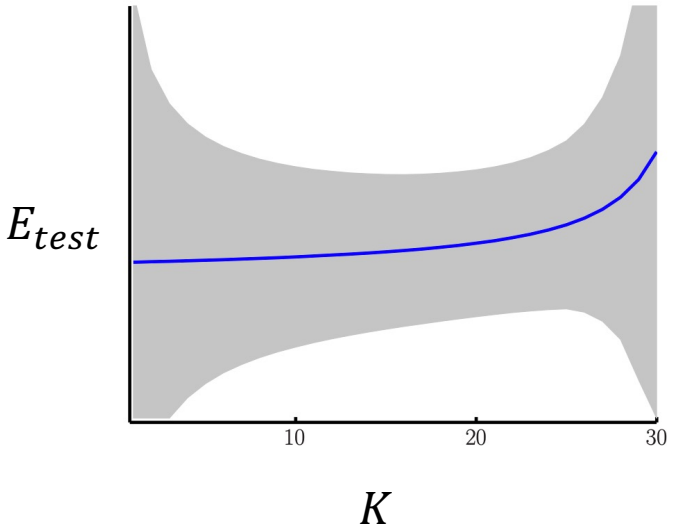
- Effect of the choice of K



Where are Test Set From?

- Given a data set D of N points
  - $D = D_{train} \cup D_{test}$
  - Reserving K points for test set means we only have $N - K$ points for training

- Effect of the choice of K



Rule of Thumb: $K^* = \frac{N}{5}$

Utilizing the Whole D

- Process:
  - $D = D_{train} \cup D_{test}$ where $|D_{test}| = K, |D_{train}| = N - K$
  - Learn some hypothesis g^- using only D_{train}
  - Estimate $E_{out}(g^-)$ using D_{test}
- Can we do better than g^- ?
  - Yes! Learn g using the entire D; return g and $E_{test}(g^-)$
- Generally (Informal, not theoretically proven)
  - Training on more data leads to better learned hypothesis
  - $E_{out}(g) \leq E_{out}(g^-)$

Validation: Beyond Test Set

What if we want to estimate E_{out} multiple times?

Validation: Beyond Test Set

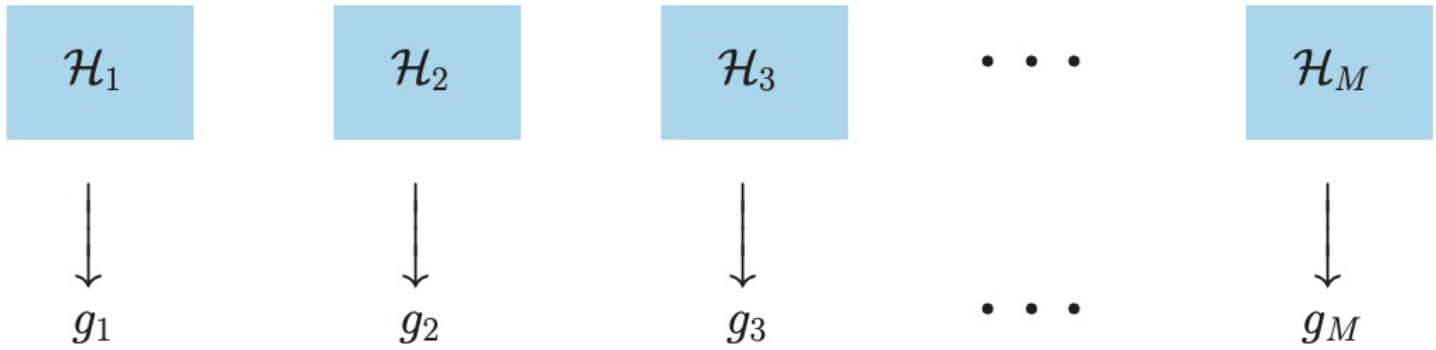
- Model selection:
  - Should I use linear models or decision trees?
  - Should I set the regularization parameter λ to 0.1, 0.01, or 0.001?
    - A model with different λ can be considered as different model

- Validation set
  - $D = D_{train} \cup D_{val}$

  - Key difference to the test set
    - D_{val} will be used multiple times
    - We need to account for the multiple usage of D_{val}

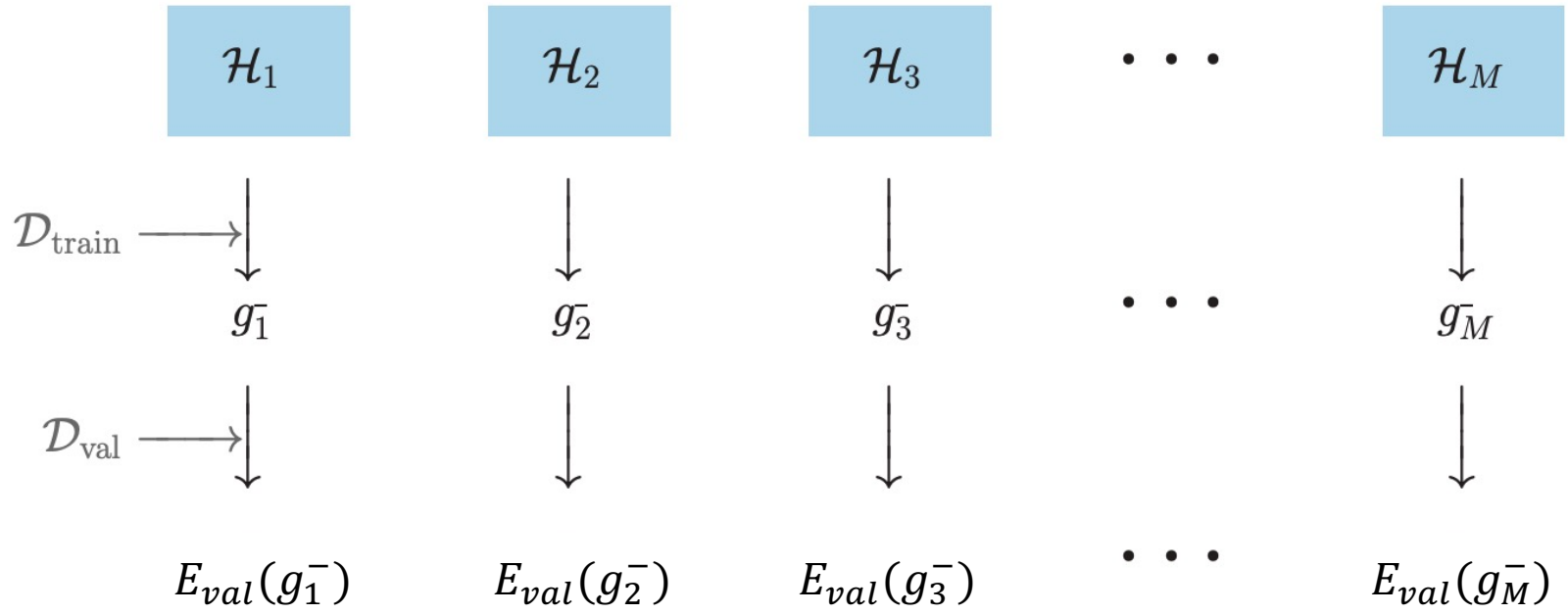
Model Selection

- Which model should we choose?



Model Selection using Validation

- Which model should we choose?



Key: D_{val} is used to choose from M hypothesis

Choose H_{m^*} such that $E_{val}(g_{m^*}^-) \leq E_{val}(g_m^-)$ for all m

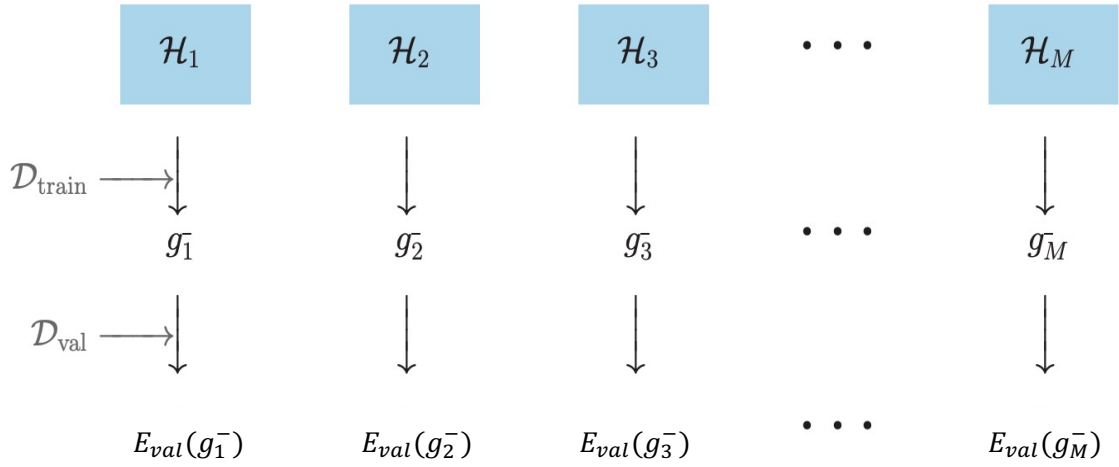
Question…

- Which of the following is true?

(a) $\mathbb{E}[E_{val}(g_{m^*}^-)] = E_{out}(g_{m^*}^-)$

(b) $\mathbb{E}[E_{val}(g_{m^*}^-)] \leq E_{out}(g_{m^*}^-)$

(c) $\mathbb{E}[E_{val}(g_{m^*}^-)] \geq E_{out}(g_{m^*}^-)$



Choose H_{m^*} such that $E_{val}(g_{m^*}^-) \leq E_{val}(g_m^-)$ for all m

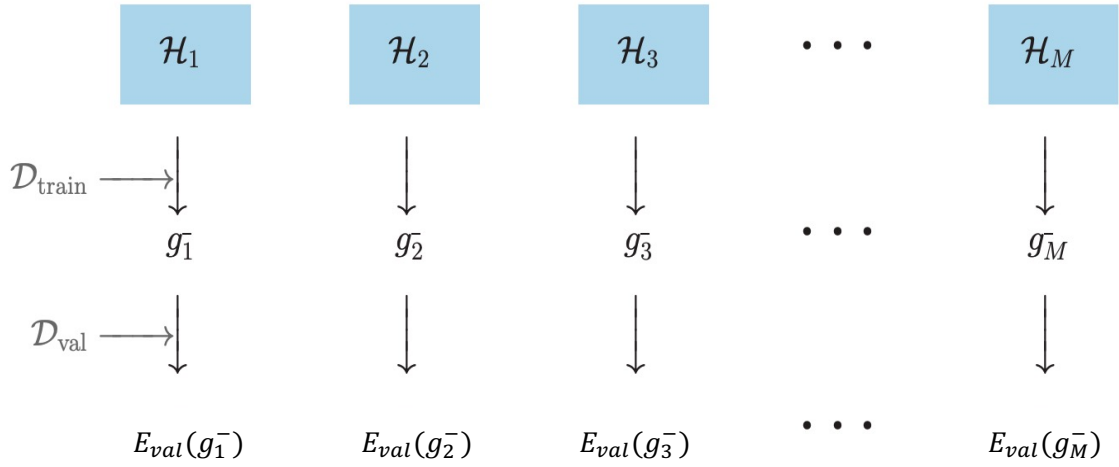
Question…

- Which of the following is true?

(a) $\mathbb{E}[E_{val}(g_{m^*}^-)] = E_{out}(g_{m^*}^-)$

(b) $\mathbb{E}[E_{val}(g_{m^*}^-)] \leq E_{out}(g_{m^*}^-)$

(c) $\mathbb{E}[E_{val}(g_{m^*}^-)] \geq E_{out}(g_{m^*}^-)$



Choose H_{m^*} such that $E_{val}(g_{m^*}^-) \leq E_{val}(g_m^-)$for all m

Equivalent to use D_{val} to choose from $H = \{g_1^-, \ldots, g_M^-\}$

$$E_{out}(g_{m^*}^-) \leq E_{val}(g_{m^*}^-) + O\left(\sqrt{\frac{\ln M}{K}}\right)$$ => Hoeffding Bound for Multiple Hypothesis

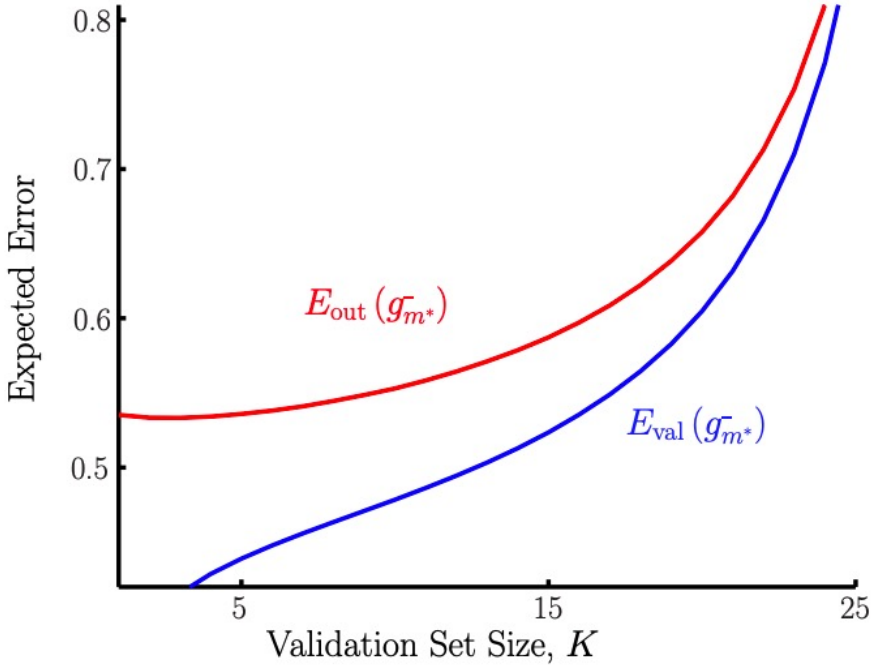
Question…

- Which of the following is true?

(a) $\mathbb{E}[E_{val}(g_{m^*}^-)] = E_{out}(g_{m^*}^-)$

(b) $\mathbb{E}[E_{val}(g_{m^*}^-)] \leq E_{out}(g_{m^*}^-)$

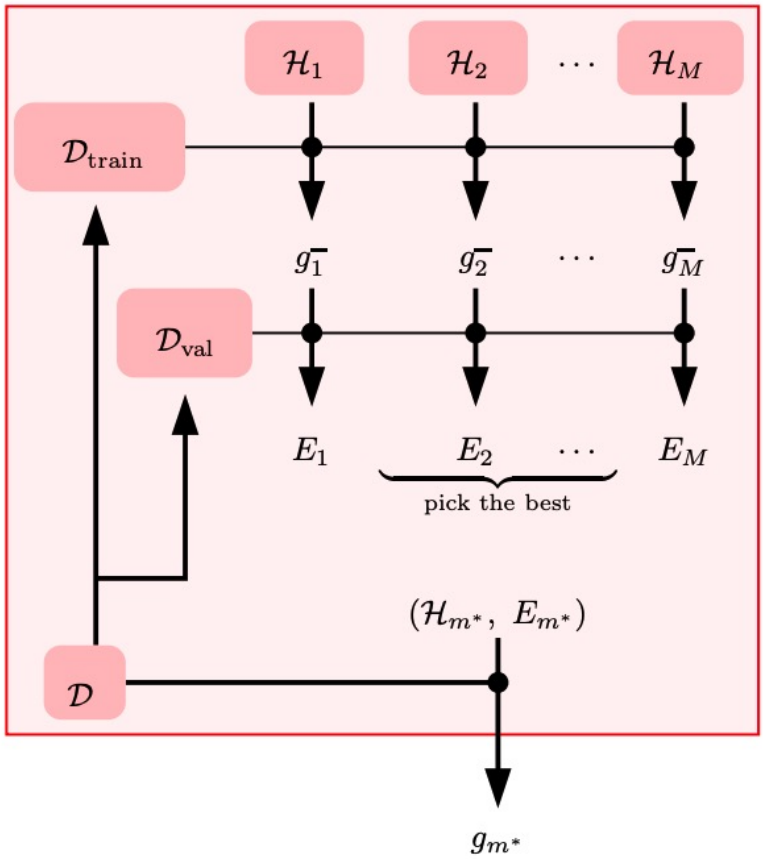
(c) $\mathbb{E}[E_{val}(g_{m^*}^-)] \geq E_{out}(g_{m^*}^-)$

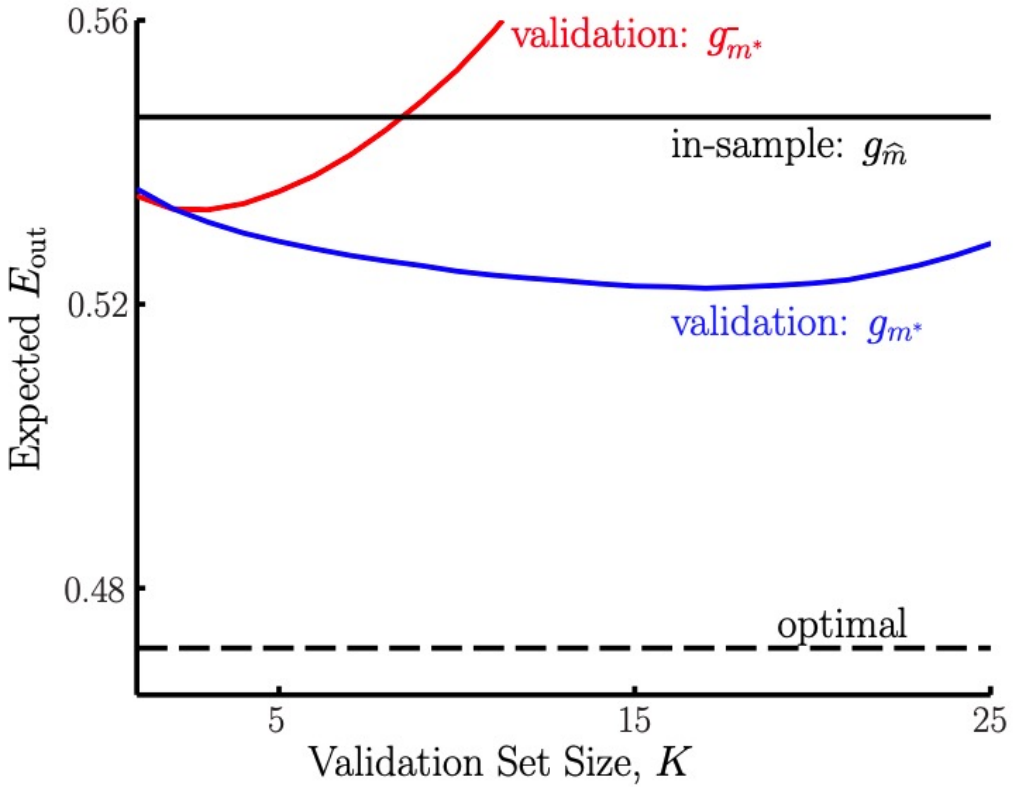


Equivalent to use D_{val} to choose from $H = \{g_1^-, \ldots, g_M^-\}$

$$E_{out}(g_{m^*}^-) \leq E_{val}(g_{m^*}^-) + O\left(\sqrt{\frac{\ln M}{K}}\right)$$ => Hoeffding Bound for Multiple Hypothesis

Utilizing the Whole D





$g_{\widehat{m}}$: the hypothesis minimizes in-sample error over $\{H_1, \ldots, H_M\}$

	Outlook	Relationship to E_{out}
E_{in}		
E_{val} (when not used for model selection)		
E_{test}		

	Outlook	Relationship to E_{out}
E_{in}	Incredibly optimistic	
E_{val} (when not used for model selection)	Slightly optimistic	
E_{test}	Unbiased	

	Outlook	Relationship to E_{out}
E_{in}	Incredibly optimistic	VC-bound
E_{val} (when not used for model selection)	Slightly optimistic	Hoeffding's bound (multiple hypotheses)
E_{test}	Unbiased	Hoeffding's bound (single hypothesis)

Note that the outlook comparisons are "in expectation"
If you only get one "draw" of $D_{train}, D_{val}, D_{test}$, you cannot say anything "for certain"

Remember that ML results are under the condition "with high probability"

The Dilemma When Choosing K

- The main ideas behind validation

Want large K
(E_{val} estimates E_{out} well)

$$E_{out}(g) \approx E_{out}(g^-) \approx E_{val}(g^-)$$

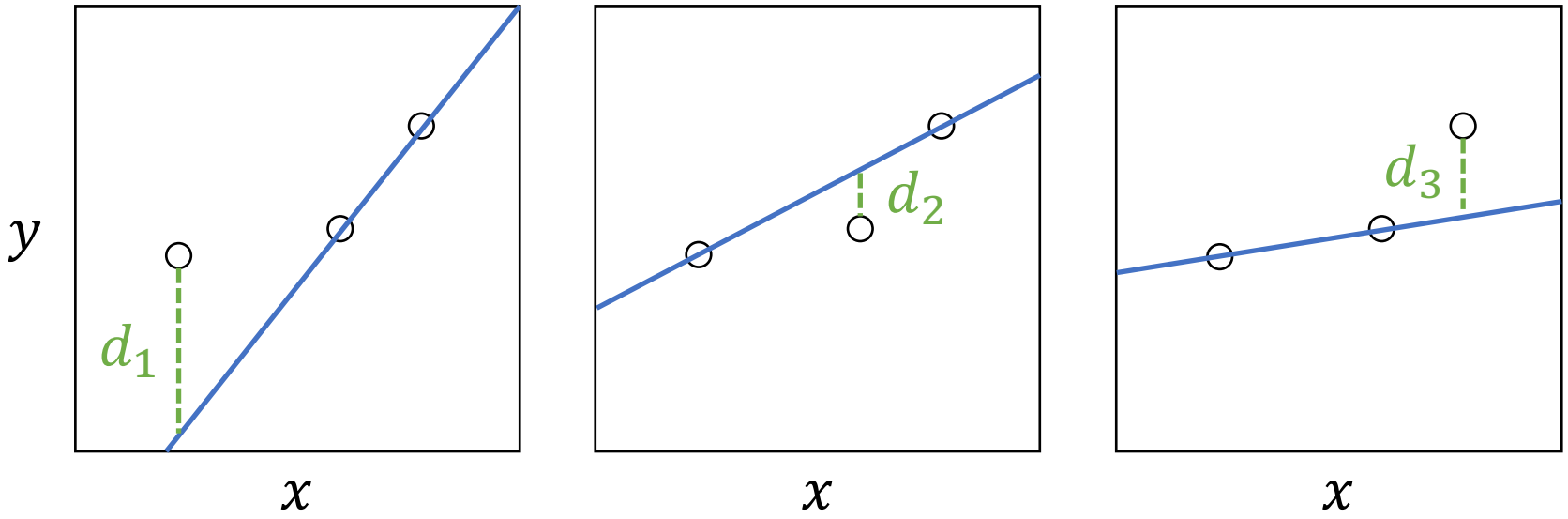
Want small K
(didn't sacrifice too much training data)

Leave-One-Out Cross Validation (LOOCV)

Getting the best of the both world

Intuition: Setting $K = 1$ but do it many times…

Illustrative Example



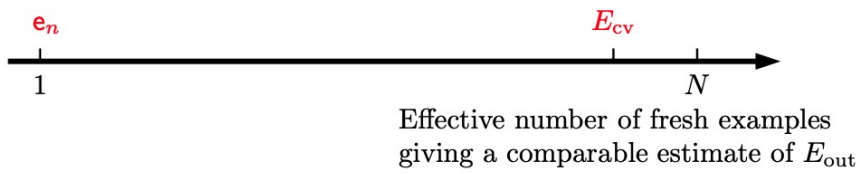
$$E_{cv} = \frac{1}{3}(d_1^2 + d_2^2 + d_3^2)$$

Properties of LOOCV

- LOOCV is unbiased (If *not* used for model selection)
  - E_{CV} is an unbiased estimator of $\bar{E}_{out}(N-1)$

(expected E_{out} when learning on $N-1$ points)

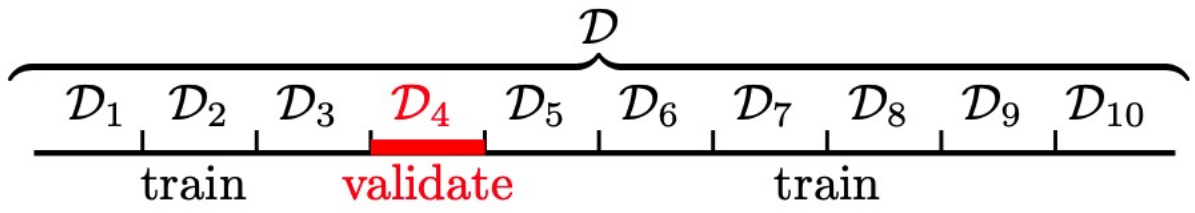
- The "effective number" of examples in E_{CV} estimation is high for LOOCV



- However, LOOCV is computationally expensive
  - Need to train N models, each on $N-1$ points

V-Fold Cross Validation

- Split D into V equally sized data sets: $D_1, D_2, \ldots, D_V$
  - Let g_i^- be the hypothesis learned using all data sets except D_i
  - Let $e_i = E_{val}(g_i^-)$ where the validation uses data set D_i
- The V-fold cross validation error is $\frac{1}{V}\sum_{i=1}^{V} e_i$



- Practical rule of thumb: $V = 10$

Three Learning Principles

Occam’s Razor

Sampling Bias

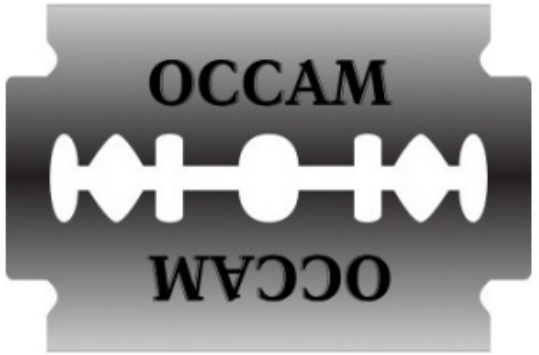
Data Snooping

Occam's Razor

"An explanation of the data should be made as simple as possible, but no simpler."
-- Einstein?

"entia non sunt multiplicanda praeter necessitatem"
(entities must not be multiplied beyond necessity)
-- William of Occam

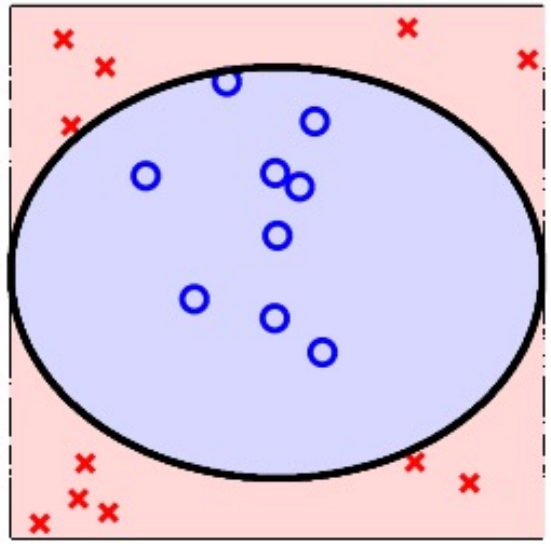
"trimming down"
unnecessary explanation

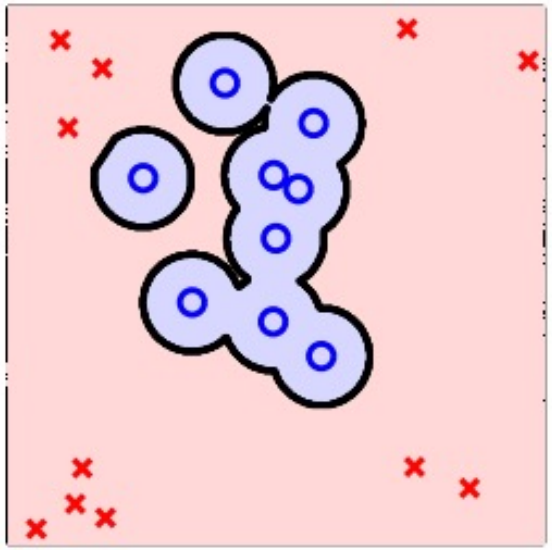


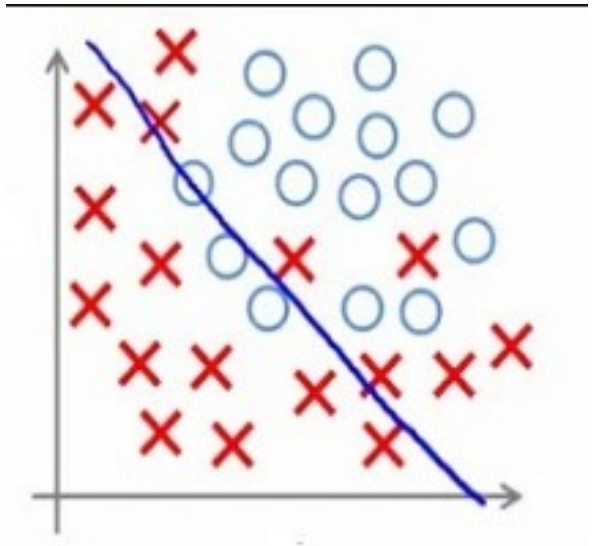
The simplest model that fits the data is also the most plausible

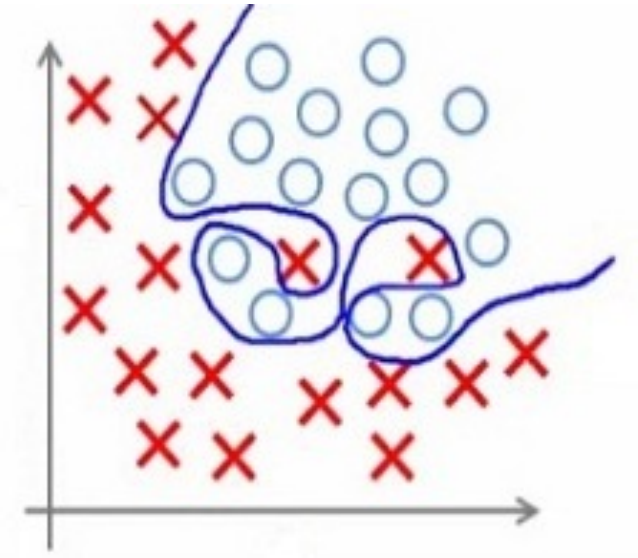
What does it mean to be simple?

Why is simple better?









Simple Model?

- For a hypothesis set H to be simple
  - # dichotomies it can generate is small
  - VC Dimension is small

- For a hypothesis h to be simple
  - lower order polynomial
  - smaller weights (think about the regularization)
  - easy to describe?
  - fewer number of parameters (fewer bits to describe)

Simple Model?

Connection:

A hypothesis set with *simple* hypotheses should be *simple*

Consider a hypothesis h can be specified by ℓ bits

$\Rightarrow H$ contains all such h

$\Rightarrow$ The size of H is 2^{ℓ}

Simple: small model complexity / VC dimension / size of hypothesis set

Why is Simple Better?

simple -> small VC dimension -> good generalization, less overfitting, …

Simple $\mathcal{H}$

$\Rightarrow$ small growth function $m_{\mathcal{H}}(N)$

$\Rightarrow$ if data labels are generated randomly, the probability of fitting perfectly is?

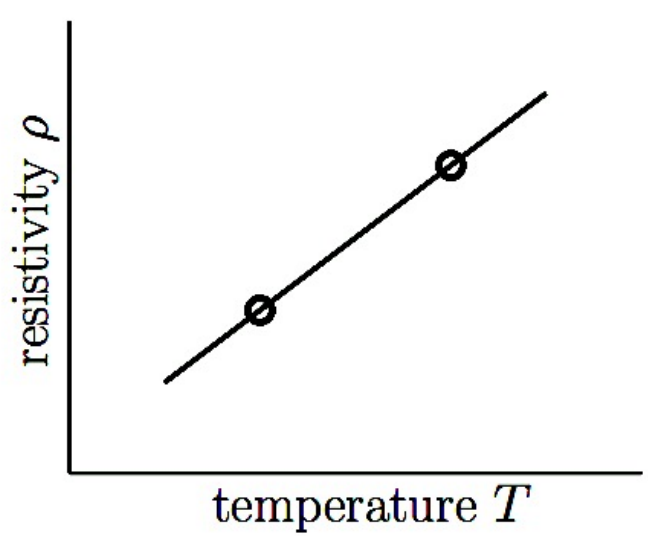
$$\frac{m_{\mathcal{H}}(N)}{2^N}$$

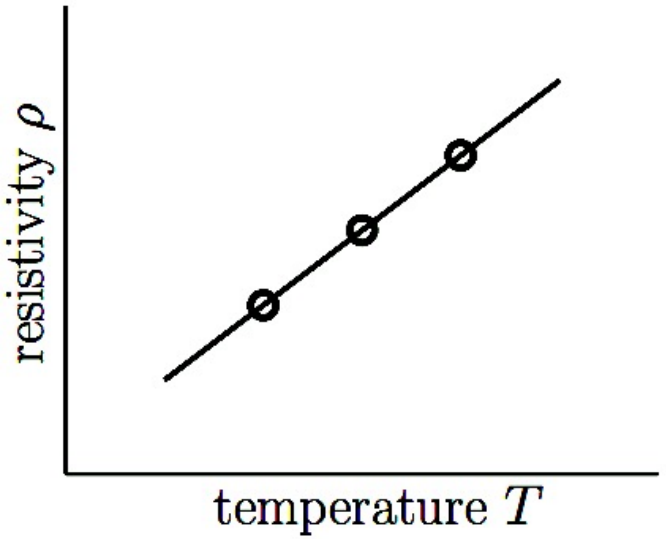
$\Rightarrow$ more significant when fit really happens

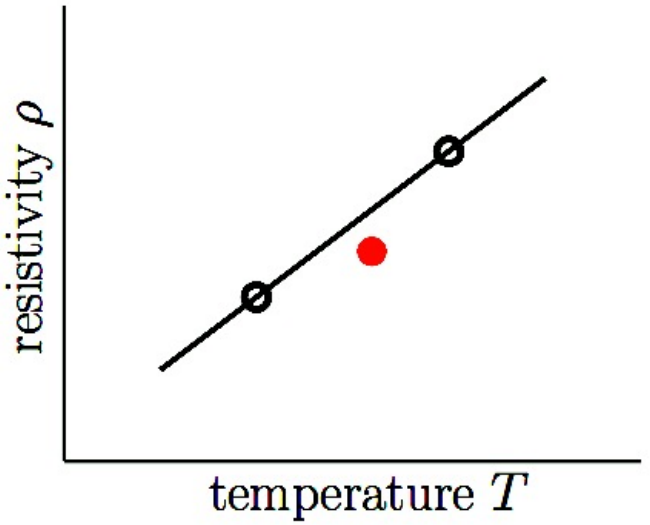
Falsifiability is important!

Falsifiability

Say you want to examine whether resistivity is linear in temperature (assume no measure error)







A Classical Puzzle

Imagine you got an email before each Cardinals game for the first 5 games.

Before Game 1: “Cardinals will win” -> Cardinals wins Game 1
Before Game 2: ”Cardinals will lose” -> Cardinals loses Game 2
….

Before Game 6:
If you pay me $50 dollars, I’ll tell you whether Cardinals will win or not

It’s not falsifiable:
Imagine if this person contacts 2^{10} persons, split them into two groups each game
2^5 persons will receive perfect prediction for the first 5 games

Occam’s Razor

Sampling Bias

Data Snooping

1948 US Presidential Election

- Truman vs. Dewey
- Chicago Daily Tribune decided to run a phone poll of how people <u>voted</u>

AN AMERICAN PAPER FOR AMERICANS

Chicago Daily Tribune

THE WORLD'S GREATEST NEWSPAPER

HOME

VOL. CVII - NO. 264 WEDNESDAY, NOVEMBER 3, 1948 FOUR CENTS - PAY NO MORE

DEWEY DEFEATS TRUMAN

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

REPUBLICAN TICKET AHEAD OF 1944 VOTE

Town Balloting Gives Trend

Probable Winners

SENATOR—Wayland Brooks (R.)*
GOVERNOR—Dwight H. Green (R.)*
LT. GOV.—Richard Yates Rowe (R.)
SECRETARY OF STATE—William G. Stratton (R.)
AUDITOR—Elmer A. Murray (R.)
TREASURER—Elmer H. Droste (R.)
ATTY. GEN.—George F. Barrett (R.)*
*Re-elected.

BY ARTHUR EVANS

Early returns last night indicated that the Dewey-Warren Republican ticket and the state ticket headed by Sen. Brooks and

Tops Coghlan in Hot Race for Attorney

Probable Winners

STATE'S ATTORNEY—John S. Boyle (D.) or Malachy J. Coghlan (R.)
RECORDER OF DEEDS—*Victor L. Schlaeger (D.)
CIRCUIT COURT CLERK—*John E. Conroy (D.) or Mrs. Mabel Reinecke (R.)
CORONER—*A. L. Brodie (D.) or Frank Kalieu (R.)
SUPERIOR COURT CLERK—*Henry Sonnenschein (D.)
SANITARY DISTRICT TRUSTEES (3)—Frank W. Chesrow (D.), John A. Culleton (D.), and Casimir Grigilk (D.) or Walter E. McCarron (R.)
MUNICIPAL COURT BAILIFF—*Albert J. Horns (D.)
MUNICIPAL COURT CLERK—*Joseph L. Gill (D.)
*Re-elected.

BY GEORGE TAGGE

Heavy Democratic leads in Chicago wards darkened the outlook last night for Republican candidates for state's attorney and other Cook county offices.

RECORD CITY VOTE SEEN IN LATE TALLIES

Suburban Ballot Near 375,000

A high record vote seemed certain for Chicago yesterday in the first post-war Presidential election.

A turnout of 1,900,000 was indicated in the final sampling of precincts in each of the 50 wards, according to Chief Clerk John S. Rusch of the Chicago election board. This prediction was based on a projection of 1,789,776 one hour before the polls closed.

TOPS 1944 RECORD

The old record was established in 1944.

BULLETINS ON ELECTIONS

COOK COUNTY

President--749 pcts. of 5023 in Cook county: Truman D 182,780, Dewey R 128,916, Wallace P 575.

Senator--481 pcts. of 5023 in Cook county: Douglas D 128,596, Brooks R 70,465.

Governor--481 pcts. of 5023 in Cook county Stevenson D 133,374, Green R 68,973.

Lieutenant Governor--14 pcts. of Cook county: Dixon D 3,985, Rowe R 1,495.

State's Attorney--524 pcts. of 5023 in Cook county: Boyle D 134,680, Coghlan R80,-100, Heller P 3,959.

Secretary of State--14 pcts. of 4143 in Chicago: Barret D 4,-003, Stratton R 1,494.

State Treasurer --1 pcts. of 4,143 in Chicago: Smith D 3,994, Droste R 1,475.

Secretary of State--4 precincts of 9,231 in Illinois, Barrett D 415, Stratton R 633.

NATION

Birmingham, Ala., Nov. 2 (AP)--Gov. Thurmond, the State Rights Democratic candidate, won Alabama's 11 electoral votes.

Alabama 2 boxes of 2,408, Thurmond 116, Dewey 37, Wallace 3.

Arkansas--2 districts of 2,217Truman 85; Dewey 2, Wallace 2.

Florida--60 precincts of 1,523, Dewey 8,983, Thurmond 3,173, Truman 11,657,Wallace 1,151; governor. 39 precincts, Acker R 2,369, Warren D 9,743.

Georgia-105 out of 1,730 precincts Truman 2,272, Dewey 385, Thurmond 690, Wallace

Ohio-1,975 polling places of 9,710, Truman 187,850, Dewey 221,233, Wallace 1,-389; governor. 32 polling places Herbert R 3,264, Lausche D 2,-978.

Oklahoma-senator 6 precincts of 3701 Kerr D 164 Risley R 89.

Rhode Island-4 districts of 266, Truman 545, Dewey 1,896, Wallace 9.

Rhode Island senator. 4 districts of 266; Green D 557, Hazard R 1,237.

Columbia, S.C., Nov. 2-(AP)-Gov. J. Strom Thurmond, States' rights presidential candidate, captured his home state's eight electoral college votes today.

Early Count Gives G.O.P. Senate Edge

BY WALTER TROHAN

Republicans took an early lead in one of the hottest campaigns in history for control of the senate. G.O.P. retention of control of the senate seemed assured when New Mexico appeared to be deserting solid Democratic rule.

Patrick J. Hurley, Republican, took an early lead over former Agriculture Secretary Anderson in New Mexico, one of eight pivotal states. This offset what was expected to be a certain Democratic gain in Oklahoma, where former Gov. Kerr, Democrat, leaped ahead of Rep. Risley(R.). The G.O.P. was also ahead in Kentucky where Sen. Cooper

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

Sizable Electoral Margin Seen

BY ARTHUR SEARS HENNING

Dewey and Warren won a sweeping victory in the Presidential election yesterday.

The early returns shoved the Republican ticket leading Truman so far that the complete returns would disclose that Dewey won the Presidency by an overwhelming majority of electoral votes.

Herbert Brownell, manager of the Dewey

Truman


Chicago Daily Tribune
DEWEY DEFEATS TRUMAN
G.O.P. Sweep Indicated in State; Boyle Leads in

What happened?

One explanation: we cannot claim anything for certain.

However, there are bigger issues here…

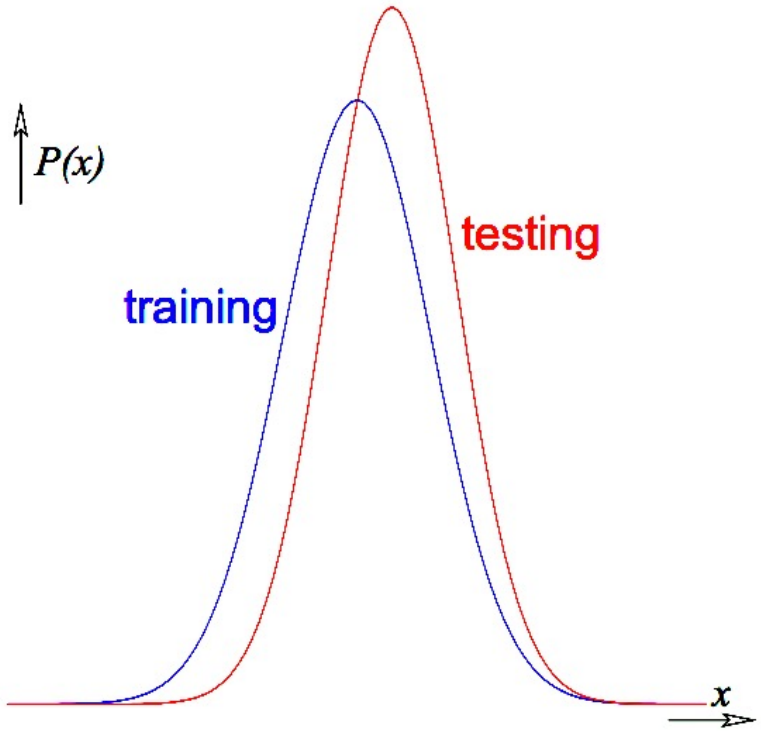
- Phones are expensive in 1948…
- Dewey was more favored in rich populations

- Imagine you are polling from people in DC/Texas/NY to predict who will win the presidential election…

Sampling Bias

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

What can we do….

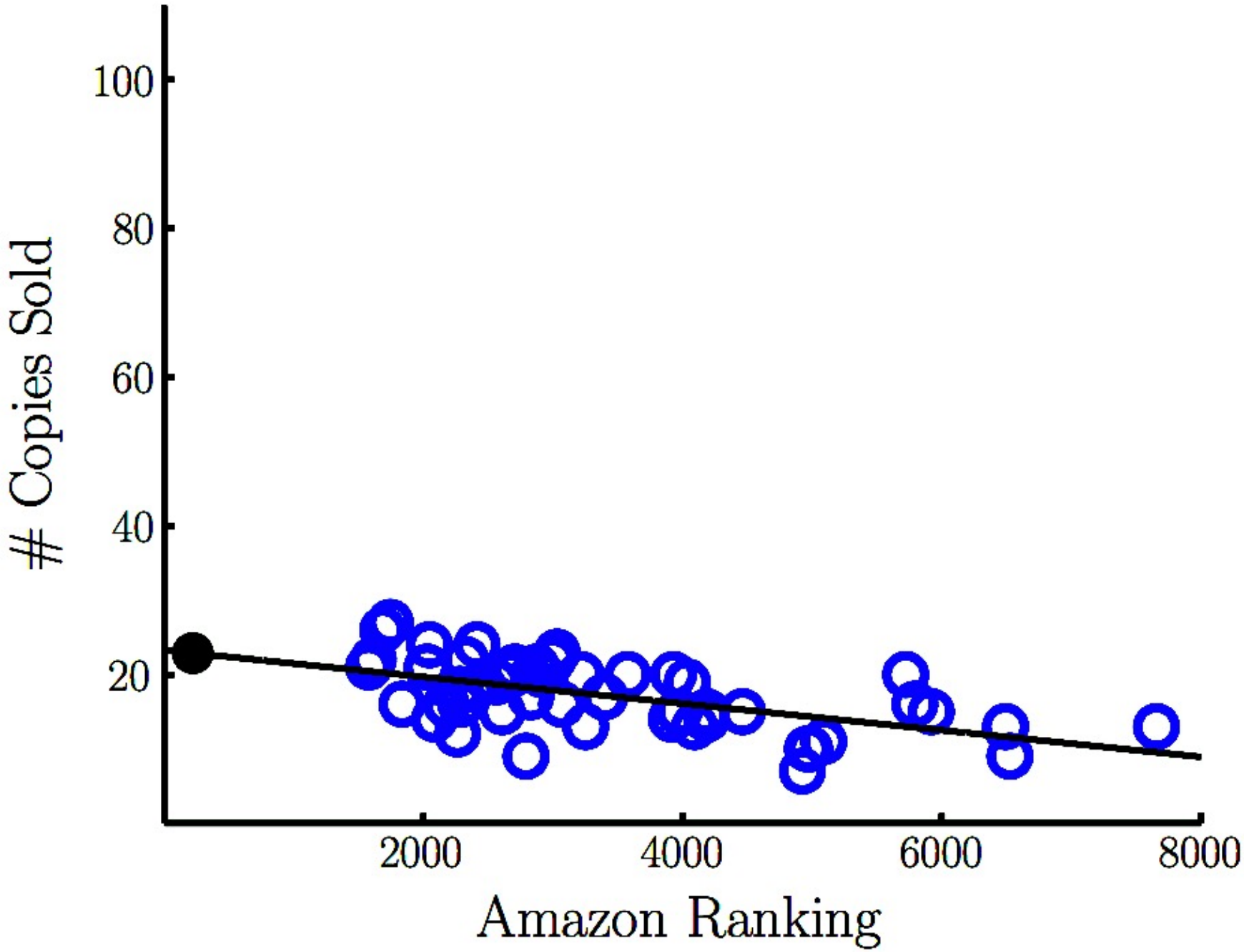


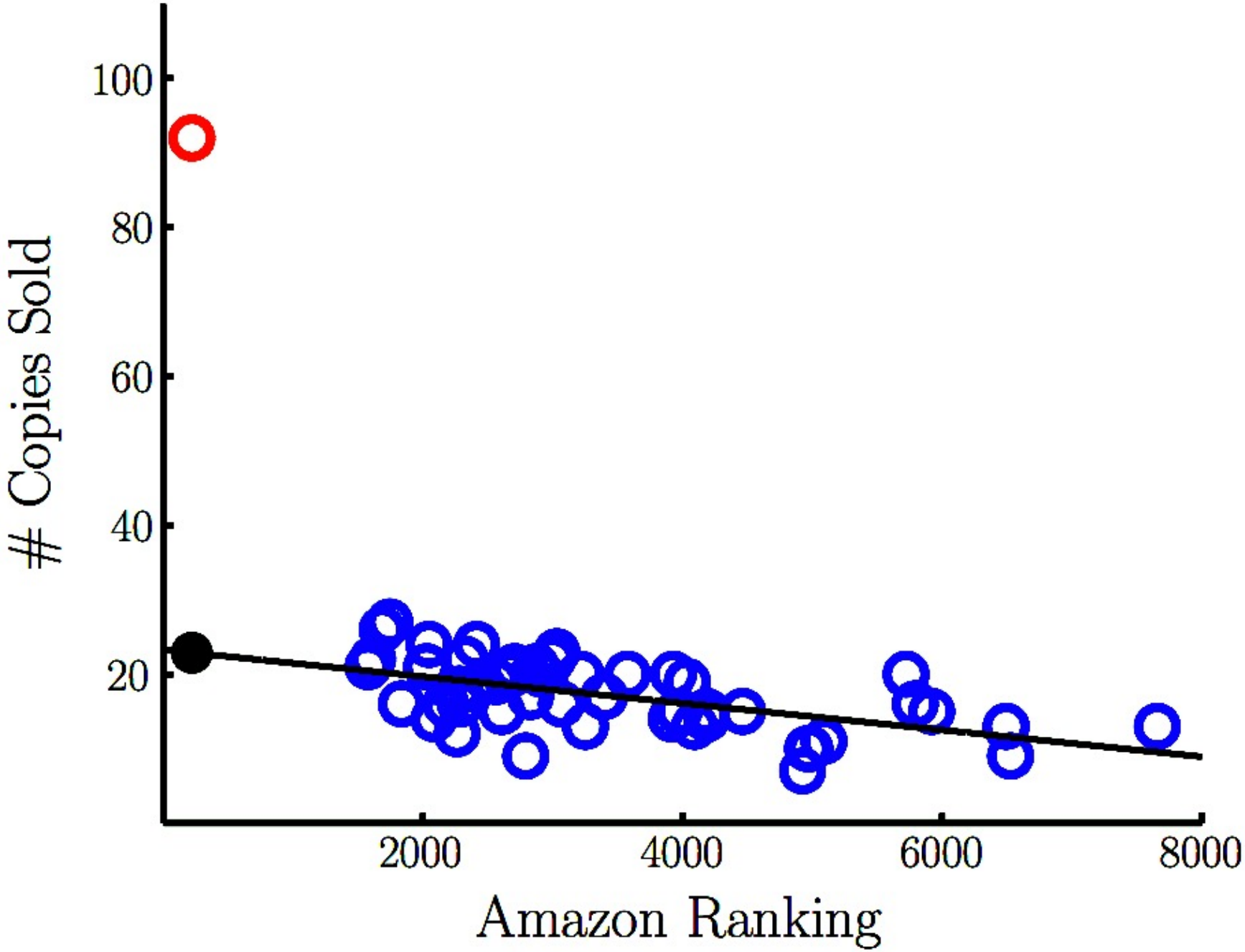
Make sure the training and test distributions are as close as possible…

- Example: importance weighting

Not always possible….

- If you don’t have access to some region of points in training, but they appear in the testing distribution


100
80
60
40
20
Copies Sold
2000
4000
6000
8000
Amazon Ranking


100
80
60
40
20
Copies Sold
2000
4000
6000
8000
Amazon Ranking

Credit card example

- Determine whether to approve credit cards given applicants' financial information
- Banks have lots of data:
  - Customer information
  - Whether they are good customers or not
- Are there any issues here?

age	32 years
gender	male
salary	40,000
debt	26,000
years in job	1 year
years at home	3 years
...	...

Approve for credit?