Course Logistics

Instructor: Chen Zeng
Office: Science and Engineering Hall 4670
Office Phone: 202-994-6481
Office Hours: Thursday 2-3PM
E-Mail: chenz@gwu.edu

Textbook
"Bayesian Reasoning and Machine Learning" by David Barber, Cambridge University Press, 2012, ISBN13: 978-0521518147

Recommended Reading
"Pattern Recognition and Machine Learning" by Christopher M. Bishop, Springer 2007, ISBN-13: 978-0387310732
"Learning From Data" by Y. Abu-Mostafa, M. Magdon-Ismail, and H.T. Lin, AMLBook 2012, ISBN-13: 978-1600490064
Web Site Resources:
(Additional resources will be available in class and on Blackboard.)
https://blackboard.gwu.edu

WK	Date	Lec	HW	Lec	Sec
1	08/29 [T]	1/2		1	1.1
1				2	1.2
2	09/05 [T]	3/4		3	2.1
2				4	2.2
3	09/12 [T]	5/6		5	2.3
3			#1	6	2.4
4	09/19 [T]	7/8		7	3.1
4				8	3.2
5	09/26 [T]	9/10		9	3.3
5			#2	10	3.4
6	10/03 [T]	11/12		11	3.5
6				12	3.6
7	10/10 [T]	Break		13	3.7
7			#3	14	4.1
8	10/17 [T]	13/14		15	4.2
8				16	4.3
9	10/24 [T]	15/16		17	4.4
9			#4	18	4.5
10	10/31 [T]	17/18		19	4.6
10				20	5.1
11	11/07 [T]	19/20		21	5.2
11			#5	22	6.1
12	11/14 [T]	21/22		23	6.2
12				24	6.3
13	11/21 [T]	23/24		25	6.4
13			#6	26	6.5
14	11/28 [T]	25/26		27	6.6
14					
15	12/05 [T]	27/28			
15					
16	12/12 [T]	Makeup			
16	12/13:21	Finals			

Lec1 Part-1 Introduction
Lec2 Part-1 ML Types (examples)
Lec3 Part-2 Linear Model - Regression
Lec4 Part-2 Bias-Var
Lec5 Part-2 Regularization and Kernel
Lec6 Part-2 Functional Space
Lec7 Part-3 Classification
Lec8 Part-3 Decision Tree
Lec9 Part-3 Entropy vs ML
Lec10 Part-3 Random Forrest
Lec11 Part-3 SVM Kernel
Lec12 Part-3 SVM Kernel
Lec13 Part-3 Project Enhancer
Lec14 Part-4 Graph Models
Lec15 Part-4 Exact Inference
Lec15 Part-4 Inference as Traditional Statistics (vs Bayesian)
Lec16 Part-4 Deterministic Approx KL to LBP
Lec17 Part-4 Deterministic Approx KL to MF
Lec18 Part-4 Stochastic Approx MCMC
Lec19 Part-4 Bayesian Model Selection
Lec19 Part-5 PCA
Lec20 Part-5 Latent Variable
Lec21 Part-6 Neural Network (Universal)
Lec22 Part-6 Neural Network (BP)
Lec23 Part-6 Hopfield
Lec24 Part-6 Boltzmann Machine
Lec25 Part-6 BM and RBM Training
Lec26 Part-6 DL DCA RNA

HW1 (Due 09/17)
Enabled: Statistics Tracking
Attached Files: HW1.pdf (141.347 KB)

HW2 (Due 10/01)
Enabled: Statistics Tracking
Attached Files: HW2.pdf (206.948 KB)

HW3 (Due 10/15)
Enabled: Statistics Tracking
Attached Files: HW3.pdf (323.536 KB)

HW4 (Due 10/29)
Enabled: Statistics Tracking
Attached Files: HW4.pdf (251.023 KB)

HW5 (Due 11/12)
Enabled: Statistics Tracking
Attached Files: HW5.pdf (149.293 KB)

HW6 (Due 11/26)
Enabled: Statistics Tracking
Attached Files: HW6.pdf (798.562 KB)

Only one submission!!!

100 – 93	A
93 – 87	A-
87 – 81	B+
81 – 75	B
75 – 69	B-
69 – 63	C+
63 – 57	C
57 – 51	C-
51 – 0	F

Instructor: Chen Zeng
SEH 4750
Office Hr: Sun 2-5PM
chenz@gwu.edu
(301)-412-2910 (Cell)

The breakdown of the components of the grade is summarized below:

Full class attendance (2% bonus)

Homework Assignments (30%)

Term Project 1 and 2 (each worth 35%)

Report (25%)
- Clarity (5%)
- Innovation (5%)
- Results (5%)
- Data visualization (5%)
- Well-documented and user-friendly codes (5%)

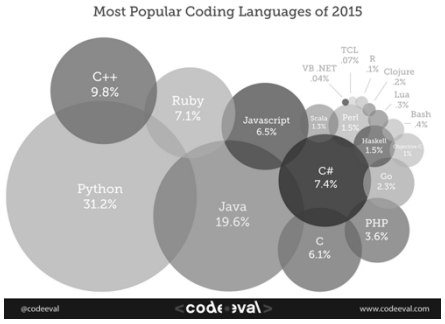
Presentation (10%)
- Concise (5%)
- Visualization (5%)

Mathematical Language

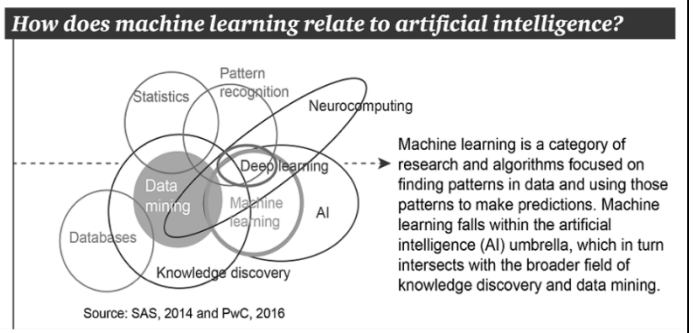
Algorithm Theory (complexity, branch and bound, combinatorial optimization)
Basic graph theory and algorithms
Convex Duality

1. Calculus (derivatives)
2. Optimization Gradient descent
3. Lagrange multipliers
4. Matrix (linear algebra) SVD/Eigenvalue

Programming Language

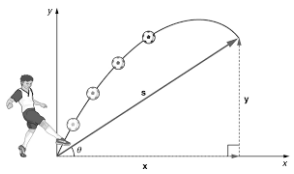


Statistical Language



http://usblogs.pwc.com/emerging-technology/a-look-at-machine-learning-infographic/

Learning?



Time of flight, $t = \frac{2v_0 \sin\theta}{g}$

Maximum height reached, $H = \frac{v_0^2 \sin^2\theta}{2g}$

Horizontal range, $R = \frac{v_0^2 \sin 2\theta}{g}$

F(X) = Y

Input (image/audio/video)

Output (Conclusion)

F(X) = Y

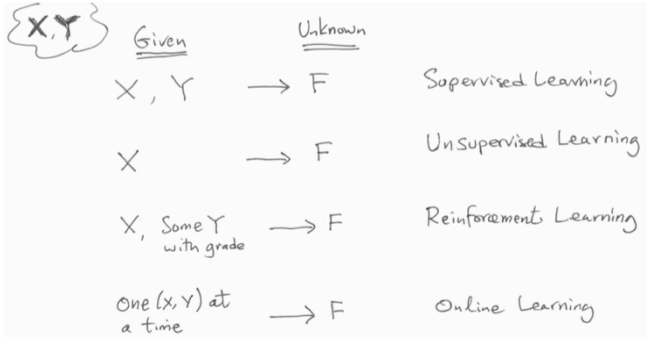
Knowledge (intuition/… = Math/Brain Neural Network/…)

F(X) = Y Scientist = Thinker

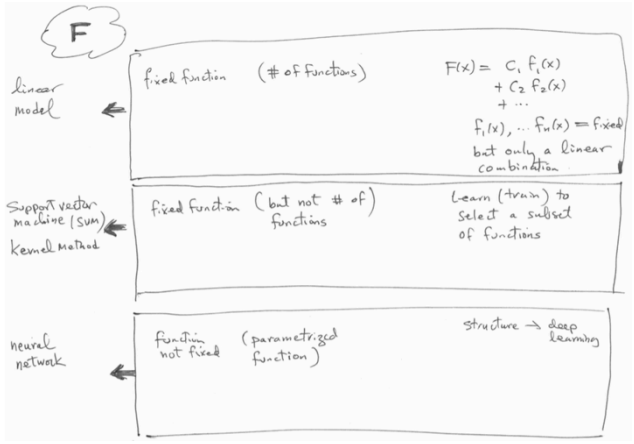
F(X) = Y Doctor = Doer

F(X) = Y Artist = Painter

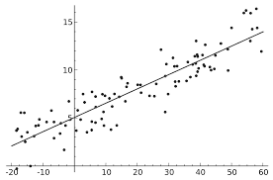
F(X) = Y



F(X) = Y



HS Science Class: linear fitting



Lec1 Part-1 Introduction
Lec2 Part-1 ML Types (examples)
Lec3 Part-2 Linear Model - Regression
Lec4 Part-2 Bias-Var
Lec5 Part-2 Regularization and Kernel
Lec6 Part-2 Functional Space
Lec7 Part-3 Classification
Lec8 Part-3 Decision Tree
Lec9 Part-3 Entropy vs ML
Lec10 Part-3 Random Forrest
Lec11 Part-3 SVM Kernel
Lec12 Part-3 SVM Kernel
Lec13 Part-3 Project Enhancer
Lec14 Part-4 Graph Models
Lec15 Part-4 Exact Inference
Lec15 Part-4 Inference as Traditional Statistics (vs Bayesian)
Lec16 Part-4 Deterministic Approx KL to LBP
Lec17 Part-4 Deterministic Approx KL to MF
Lec18 Part-4 Stochastic Approx MCMC
Lec19 Part-4 Bayesian Model Selection
Lec19 Part-5 PCA
Lec20 Part-5 Latent Variable
Lec21 Part-6 Neural Network (Universal)
Lec22 Part-6 Neural Network (BP)
Lec23 Part-6 Hopfield
Lec24 Part-6 Boltzmann Machine
Lec25 Part-6 BM and RBM Training
Lec26 Part-6 DL DCA RNA

Memory?

X=height
Y=Stable(Good/Everlasting …)
F(X)=Y

F(X) = Y

Input-Observation (image/audio/video)

Output (Conclusion) – extreme (min/max)

Knowledge (intuition/… = Math/Brain Neural Network/…)

F(X=2m) = Y (=mgX=9.8x2)
F(X=1m) = Y (=mgX=9.8x1)
F(X=0m) = Y (=mgX=0 最小)

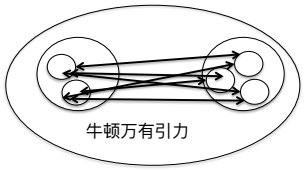
X=height Y=Stable(Good/Everlasting …)
F(X)=Y

F(X) = Y

Input-Observation (image/audio/video)

Output (Conclusion) – extreme (min/max)

Knowledge (intuition/… = Math/Brain Neural Network/…)

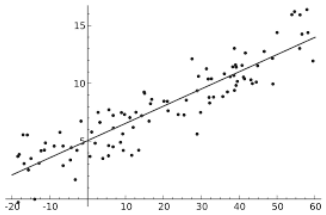


Everything you see: atoms!

Everything you see: pixels!

Learning Possible?

HS Science Class: linear fitting



Lec3 Part-2 Linear Model - Regression
Lec4 Part-2 Bias-Var
Lec5 Part-2 Regularization and Kernel
Lec6 Part-2 Functional Space
Lec7 Part-3 Classification
Lec8 Part-3 Decision Tree
Lec9 Part-3 Entropy vs ML
Lec10 Part-3 Random Forrest
Lec11 Part-3 SVM Kernel
Lec12 Part-3 SVM Kernel

Slides taken from

Learning From Data

A SHORT COURSE

Yaser S. Abu-Mostafa
California Institute of Technology

Malik Magdon-Ismail
Rensselaer Polytechnic Institute

Hsuan-Tien Lin
National Taiwan University

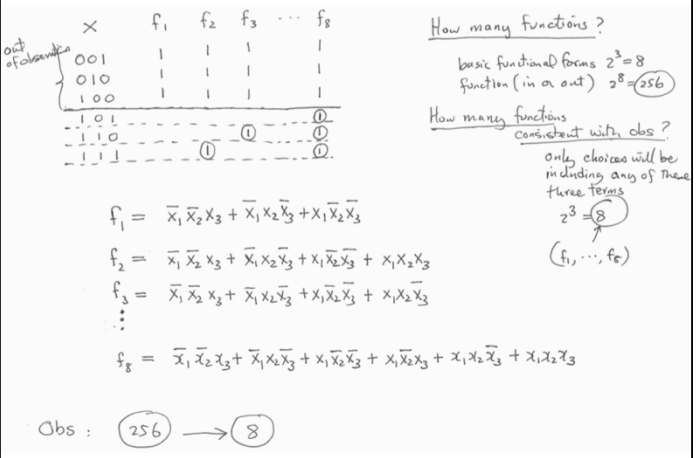
Learning Possible?

$\mathbf{x}_n$	y_n
0 0 0	○
0 0 1	●
0 1 0	●
0 1 1	○
1 0 0	●

$y_n = f(\mathbf{x}_n)$ for $n = 1, 2, 3, 4, 5$.

x	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	○	○	○	○	○	○	○	○	○	○
0 0 1	●	●	●	●	●	●	●	●	●	●
0 1 0	●	●	●	●	●	●	●	●	●	●
0 1 1	○	○	○	○	○	○	○	○	○	○
1 0 0	●	●	●	●	●	●	●	●	●	●
1 0 1		?	○	○	○	○	●	●	●	●
1 1 0		?	○	○	●	●	○	○	●	●
1 1 1		?	○	●	○	●	○	●	○	●

101 = x1 \bar x2 x3 (unique) how many such terms 2^3=8, how many function 2^8 (whether to include such term or not in the "sum" (OR)



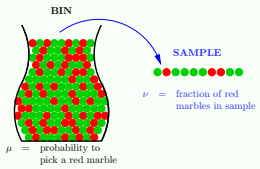
Taken from:

Learning From Data
Lecture 3
Is Learning Feasible?

Outside the Data
Probability to the Rescue
Learning vs. Verification
Selection Bias - A Cartoon

M. Magdon-Ismail
CSCI 4100/6100

Population Mean from Sample Mean



The BIN Model

- Bin with red and green marbles.
- Pick a sample of N marbles *independently*.
- μ: probability to pick a red marble.
  ν: fraction of red marbles in the sample.

  Sample $\longrightarrow$ the data set $\longrightarrow$ ν
  BIN $\longrightarrow$ outside the data $\longrightarrow$ μ

Can we say anything about $\boldsymbol{\mu}$ **(outside the data)** after observing $\boldsymbol{\nu}$ **(the data)**?

ANSWER: No. It is *possible* for the sample to be all green marbles and the bin to be mostly red.

Then, why do we trust polling (e.g. to predict the outcome of the presidential election).

ANSWER: The bad case is *possible*, but not **probable**.

Probability to the Rescue: Hoeffding's Inequality

Hoeffding/Chernoff proved that, most of the time, ν cannot be too far from μ:

$$\mathbb{P}\left[|\nu-\mu|>\epsilon\right] \leq 2e^{-2\epsilon^2 N}, \qquad \text{for any } \epsilon > 0.$$
$$\mathbb{P}\left[|\nu-\mu|\leq\epsilon\right] \geq 1-2e^{-2\epsilon^2 N}, \qquad \text{for any } \epsilon > 0.$$

box it and memorize it ☺

We get to select any ϵ we want.

newsflash: $\nu \approx \mu \implies \mu \approx \nu$ ☹.
$\mu \approx \nu$ is *probably approximately correct* (PAC-learning)

Probability to the Rescue: Hoeffding's Inequality

$$\mathbb{P}\left[|\nu-\mu|>\epsilon\right] \leq 2e^{-2\epsilon^2 N}, \qquad \text{for any } \epsilon > 0.$$
$$\mathbb{P}\left[|\nu-\mu|\leq\epsilon\right] \geq 1-2e^{-2\epsilon^2 N}, \qquad \text{for any } \epsilon > 0.$$

box it and memorize it ☺

Example: $N = 1,000$; draw a sample and observe ν.

99% of the time $\quad \mu - 0.05 \leq \nu \leq \mu + 0.05 \quad (\epsilon = 0.05)$
99.9999996% of the time $\quad \mu - 0.10 \leq \nu \leq \mu + 0.10 \quad (\epsilon = 0.10)$

What does this mean? If I repeatedly pick a sample of size 1,000, observe ν and claim that

$$\mu \in [\nu - 0.05, \nu + 0.05], \qquad \text{(the error bar is } \pm 0.05\text{)}$$

I will be right 99% of the time. On any particular sample you **may** be wrong, but not often.

We learned *something*. From ν, we reached outside the data to μ.

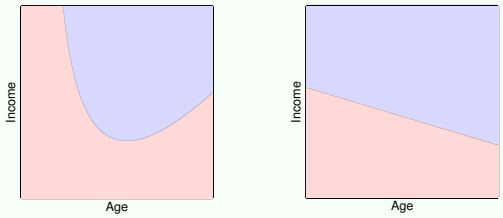
How Did Probability Rescue Us?

- Key ingredient samples must be *independent*.
  If the sample is constructed in some arbitrary fashion, then indeed we cannot say anything.
  Even with independence, ν can take on arbitrary values; but some values are way more likely than others.
  This is what allows us to learn *something* – it is likely that $\nu \approx \mu$.
- The bound $2e^{-2\epsilon^2 N}$ does not depend on μ or the size of the bin
  The bin can be infinite.
  It's great that it does not depend on μ because μ is unknown; and we mean *unknown*.
- The key player in the bound $2e^{-2\epsilon^2 N}$ is N.
  If $N \to \infty$, $\mu \approx \nu$ with very very very ... high probabilty, *but not for sure*.
  Can you live with 10^{-100} probability of error?

We should *probably* have said *"independence to the rescue"* ☺

Relating the Bin to Learning

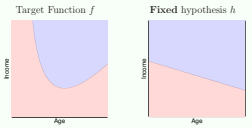
Target Function f | **Fixed** hypothesis h



UNKNOWN | KNOWN

In learning, the unknown is an entire function f; in the bin it was a single number μ.

Relating the Bin to Learning - The Error Function



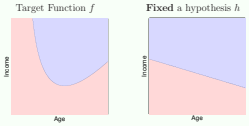
green: $h(\mathbf{x}) = f(\mathbf{x})$
red: $h(\mathbf{x}) \neq f(\mathbf{x})$

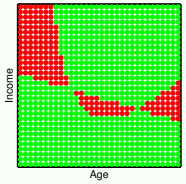
$E(h) = \mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})]$
("size" of the red region)

$P(\mathbf{x})$

UNKNOWN

Relating the Bin to Learning - The Error Function





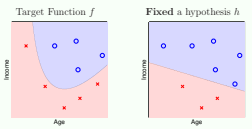
green "marble": $h(\mathbf{x}) = f(\mathbf{x})$
red "marble": $h(\mathbf{x}) \neq f(\mathbf{x})$
BIN: $\mathcal{X}$

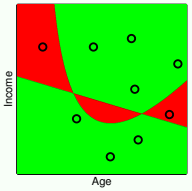
$E_{\text{out}}(h) = \mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})]$

out-of-sample

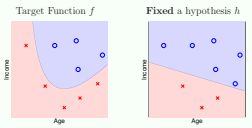
UNKNOWN

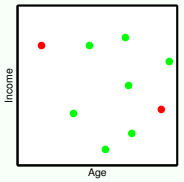
Relating the Bin to Learning - the Data





Relating the Bin to Learning - the Data





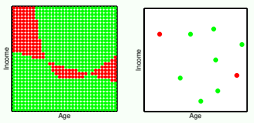
green data: $h(\mathbf{x}_n) = f(\mathbf{x}_n)$
red data: $h(\mathbf{x}_n) \neq f(\mathbf{x}_n)$

$E_{\text{in}}(h)$ = fraction of red data

in-sample misclassified

KNOWN!

Relating the Bin to Learning



Unknown f and $P(\mathbf{x})$, **fixed** h

Learning

input space $\mathcal{X}$

$\mathbf{x}$ for which $h(\mathbf{x}) = f(\mathbf{x})$

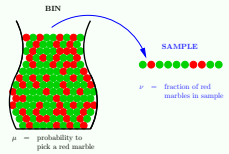
$\mathbf{x}$ for which $h(\mathbf{x}) \neq f(\mathbf{x})$

$P(\mathbf{x})$

data set $\mathcal{D}$

Out-of-sample Error: $E_{\text{out}}(h) = \mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})]$

In-sample Error: $E_{\text{in}}(h) = \frac{1}{N}\sum_{n=1}^{N} [\![h(\mathbf{x}) \neq f(\mathbf{x})]\!]$



Bin Model

Bin

green marble

red marble

randomly picking a marble

sample of N marbles

μ = probability of picking a red marble

ν = fraction of red marbles in the sample

Hoeffding says that $E_{\text{in}}(h) \approx E_{\text{out}}(h)$

$$\mathbb{P}\left[|\boldsymbol{E_{\text{in}}(h) - E_{\text{out}}(h)}| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}, \qquad \textbf{for any } \epsilon > 0.$$

$$\mathbb{P}\left[|\boldsymbol{E_{\text{in}}(h) - E_{\text{out}}(h)}| \leq \epsilon\right] \geq 1 - 2e^{-2\epsilon^2 N}, \qquad \textbf{for any } \epsilon > 0.$$

E_{in} is **random, but known**; E_{out} **fixed, but unknown**.

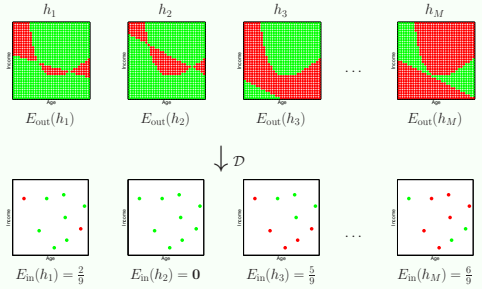
- If $E_{\text{in}} \approx 0 \implies E_{\text{out}} \approx 0$ (with high probability), i.e. $\mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})] \approx 0$;
  We have learned something about the *entire* f: $f \approx h$ over $\mathcal{X}$ (outside $\mathcal{D}$)
- If $E_{\text{in}} \gg 0$, we're out of luck.
  But, we have still learned something about the entire f: $f \not\approx h$; it is not very useful though.

Questions:

Suppose that $E_{\text{in}} \approx 1$, have we learned something about the entire f that *is* useful?

What is the worst E_{in} for inferring about f?

Real Learning – Finite Learning Models



Pick the hypothesis with minimum E_{in}; will E_{out} be small?

Hoeffding says that $E_{\text{in}}(g) \approx E_{\text{out}}(g)$ for Finite $\mathcal{H}$

$$\mathbb{P}\left[|\boldsymbol{E}_{\text{in}}(\boldsymbol{g}) - \boldsymbol{E}_{\text{out}}(\boldsymbol{g})| > \epsilon\right] \leq \boldsymbol{2}|\mathcal{H}|\boldsymbol{e}^{-\boldsymbol{2}\epsilon^{\boldsymbol{2}}\boldsymbol{N}}, \qquad \textbf{for any } \epsilon > 0.$$

$$\mathbb{P}\left[|\boldsymbol{E}_{\text{in}}(\boldsymbol{g}) - \boldsymbol{E}_{\text{out}}(\boldsymbol{g})| \leq \epsilon\right] \geq 1 - \boldsymbol{2}|\mathcal{H}|\boldsymbol{e}^{-\boldsymbol{2}\epsilon^{\boldsymbol{2}}\boldsymbol{N}}, \qquad \textbf{for any } \epsilon > 0.$$

We don't care how g was obtained, *as long as it is from* $\mathcal{H}$

Some Basic Probability
Events A, B

Implication
If $A \implies B$ ($A \subseteq B$) then $\mathbb{P}[A] \leq \mathbb{P}[B]$.

Union Bound
$\mathbb{P}[A \text{ or } B] = \mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$.

Bayes' Rule
$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A] \cdot \mathbb{P}[A]}{\mathbb{P}[B]}$

Proof: Let $M = |\mathcal{H}|$.

The event "$|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon$" implies
"$|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon$" OR ... OR "$|E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon$"

So, by the implication and union bounds:

$$\begin{aligned}\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] &\leq \mathbb{P}\left[\bigvee_{m=1}^{M} |E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon\right] \\ &\leq \sum_{m=1}^{M} \mathbb{P}[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon], \\ &\leq 2Me^{-2\epsilon^2 N}.\end{aligned}$$

(The last inequality is because we can apply the Hoeffding bound to each summand)

Interpreting the Hoeffding Bound for Finite $|\mathcal{H}|$

$$\mathbb{P}\left[|\boldsymbol{E}_{\text{in}}(\boldsymbol{g}) - \boldsymbol{E}_{\text{out}}(\boldsymbol{g})| > \epsilon\right] \leq \boldsymbol{2}|\mathcal{H}|\boldsymbol{e}^{-\boldsymbol{2}\epsilon^{\boldsymbol{2}}\boldsymbol{N}}, \qquad \textbf{for any } \epsilon > 0.$$

$$\mathbb{P}\left[|\boldsymbol{E}_{\text{in}}(\boldsymbol{g}) - \boldsymbol{E}_{\text{out}}(\boldsymbol{g})| \leq \epsilon\right] \geq 1 - \boldsymbol{2}|\mathcal{H}|\boldsymbol{e}^{-\boldsymbol{2}\epsilon^{\boldsymbol{2}}\boldsymbol{N}}, \qquad \textbf{for any } \epsilon > 0.$$

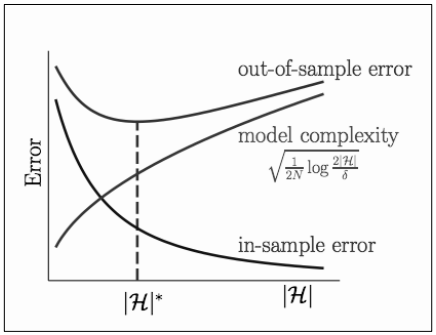
Theorem. With probability at least $1 - \delta$,

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{1}{2N}\log\frac{2|\mathcal{H}|}{\delta}}.$$

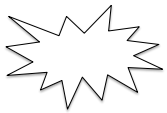
We don't care how g was obtained, *as long as* $g \in \mathcal{H}$

Proof: Let $\delta = 2|\mathcal{H}|e^{-2\epsilon^2 N}$. Then
$\mathbb{P}\left[|E_{\text{in}}(g) - E_{\text{out}}(g)| \leq \epsilon\right] \geq 1 - \delta$.
In words, with probability at least $1 - \delta$,
$|E_{\text{in}}(g) - E_{\text{out}}(g)| \leq \epsilon$.
This implies
$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \epsilon$.
From the definition of δ, solve for ϵ:
$\epsilon = \sqrt{\frac{1}{2N}\log\frac{2|\mathcal{H}|}{\delta}}$.

Meaning of all these theoretical proofs -



Statistical Language



- **So learning possible (statistical sense), but how does it help with actual learning algorithms?**
- **How does it address the Black Swan events? What requirements to be put on dataset?**
- **What other implicit "outside" assumptions (constancy and continuity???)**

Learning From Data

A SHORT COURSE

Yaser S. Abu-Mostafa
California Institute of Technology

Malik Magdon-Ismail
Rensselaer Polytechnic Institute

Hsuan-Tien Lin
National Taiwan University

Problem 1.8 The Hoeffding Inequality is one form of the *law of large numbers*. One of the simplest forms of that law is the *Chebyshev Inequality*, which you will prove here.

(a) If t is a non negative random variable, prove that for any $\alpha > 0$, $\mathbb{P}[t \geq \alpha] \leq \mathbb{E}(t)/\alpha$.

(b) If u is any random variable with mean μ and variance σ^2, prove that for any $\alpha > 0$, $\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$. *[Hint: Use (a)]*

(c) If $u_1, \cdots, u_N$ are iid random variables, each with mean μ and variance σ^2, and $u = \frac{1}{N}\sum_{n=1}^{N} u_n$, prove that for any $\alpha > 0$,

$$\mathbb{P}[(u-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}.$$

Notice that the RHS of this Chebyshev Inequality goes down linearly in N, while the counterpart in Hoeffding's Inequality goes down exponentially. In Problem 1.9, we develop an exponential bound using a similar approach.

SOME STAT WITH EMPHASIS ON INEQUALITY

Markov Inequality

Let X be a random variable such that $X \in \mathbb{R}^+$. The **Markov Inequality** is states that

$$P(X > a) \leq \frac{E[X]}{a}. \tag{1.1}$$

Proof: We start with the expected value and show the inequality holds as follows

$$\begin{aligned} E[X] &= \int_0^\infty x f_X(x)\,dx \\ &= \int_0^a x f_X(x)\,dx + \int_a^\infty x f_X(x)\,dx \\ &\geq \int_a^\infty x f_X(x)\,dx \\ &\geq \int_a^\infty a f_X(x)\,dx \\ &\geq a\int_a^\infty f_X(x)\,dx = a\,P(X > a) \end{aligned}$$

Thus, we have

$$P(X > a) \leq \frac{E[X]}{a}$$

as desired.

1.8

(b) u is a random variable.

Then $(u-\mu)^2$ is a non negative random variable

So according to Markov's Inequality

$$P\left((u-\mu)^2 \geq \alpha\right) \leq \frac{E\left[(u-\mu)^2\right]}{\alpha}$$

But $E\left[(u-\mu)^2\right] = E\left[u^2 - 2u\mu + \mu^2\right] = E\left[u^2\right] - \mu^2 = \sigma^2$

$$\therefore P\left((u-\mu)^2 \geq \alpha\right) \leq \sigma^2/\alpha$$

(c) Similarly, for random variable $u = \frac{1}{n}\sum_{i=1}^{n} u_i$

$(u-\mu)^2$ is again non negative random variable.

$$\text{so } P\left[(u-\mu)^2 \geq \alpha\right] \leq \frac{E\left[(u-\mu)^2\right]}{\alpha} = \frac{\sigma^2}{n\alpha}$$

$$(u-\mu)^2 = u^2 - 2u\mu + \mu^2$$

$$u = \frac{1}{n}\sum_i u_i \quad u^2 = \frac{1}{n^2}\left[\sum_i u_i^2 + \sum_{i \neq j} u_i u_j\right]$$

$$E\left[u^2\right] = \frac{1}{n^2}\left[n(\sigma^2+\mu^2) + n(n-1)\mu^2\right]$$

$$= \frac{1}{n^2}\cdot\left[n\sigma^2 + n^2\mu^2\right]$$

$$E\left[2u\mu\right] = 2\mu^2$$

$$E\left[\mu^2\right] = \mu^2$$

$$E\left[(u-\mu)^2\right] = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}$$

Problem 1.9 In this problem, we derive a form of the law of large numbers that has an exponential bound, called the *Chernoff bound*. We focus on the simple case of flipping a fair coin, and use an approach similar to Problem 1.8.

(a) Let t be a (finite) random variable, α be a positive constant, and s be a positive parameter. If $T(s) = \mathbb{E}_t(e^{st})$, prove that

$$\mathbb{P}[t \geq \alpha] \leq e^{-s\alpha} T(s) .$$

[Hint: e^{st} is monotonically increasing in t.]

(b) Let $u_1, \cdots, u_N$ be iid random variables, and let $u = \frac{1}{N}\sum_{n=1}^{N} u_n$. If $U(s) = \mathbb{E}_{u_n}(e^{su_n})$ (for any n), prove that

$$\mathbb{P}[u \geq \alpha] \leq \left(e^{-s\alpha} U(s)\right)^N .$$

(c) Suppose $\mathbb{P}[u_n = 0] = \mathbb{P}[u_n = 1] = \frac{1}{2}$ (fair coin). Evaluate $U(s)$ as a function of s, and minimize $e^{-s\alpha}U(s)$ with respect to s for fixed α, $0 < \alpha < 1$.

(d) Conclude in (c) that, for $0 < \epsilon < \frac{1}{2}$,

$$\mathbb{P}[u \geq \mathbb{E}(u) + \epsilon] \leq 2^{-\beta N} ,$$

where $\beta = 1 + (\frac{1}{2}+\epsilon)\log_2(\frac{1}{2}+\epsilon) + (\frac{1}{2}-\epsilon)\log_2(\frac{1}{2}-\epsilon)$ and $\mathbb{E}(u) = \frac{1}{2}$. Show that $\beta > 0$, hence the bound is exponentially decreasing in N.

(a)

$$P(t \geq \alpha) = P\left(st \geq s\alpha\right) \quad [s = \text{positive}]$$

$$= P\left(e^{st} \geq e^{s\alpha}\right) \quad [e^x \text{ monotonic and positive}]$$

$$\leq \frac{E\left[e^{st}\right]}{e^{s\alpha}}$$

$$= e^{-s\alpha} E\left[e^{st}\right]$$

(b) Similarly

$$\mathbb{P}[u \geq \alpha] \leq \frac{E_u[e^{nuS}]}{e^{n\alpha s}} = \left[e^{-\alpha s} E_{u_i}(e^{su_i})\right]^n$$

$$e^{nus} = e^{\left(\frac{1}{n}\sum u_i\right)s} = e^{n\frac{s}{n}\sum u_i} = \prod_i e^{\frac{s}{n}u_i}$$

$$E_u[e^{us}] = \prod_i E_{u_i}[e^{\frac{s}{n}u_i}] = \left[E_{u_i}(e^{su_i})\right]^n$$

(c) Suppose $\mathbb{P}(u_n=0)=\frac{1}{2}$, $\mathbb{P}(u_n=1)=\frac{1}{2}$

$$U(s) = E_u(e^{su_n}) = \frac{1}{2}+\frac{1}{2}e^s$$

$$e^{-s\alpha}U(s) = e^{-s\alpha}\left(\frac{1}{2}+\frac{1}{2}e^s\right) = \frac{1}{2}e^{-s\alpha}+\frac{1}{2}e^{s(1-\alpha)}$$

$$\frac{\partial}{\partial s}\left[e^{-s\alpha}U(s)\right]=0 \Rightarrow e^{s^*}=\frac{\alpha}{1-\alpha}$$

$$\text{So } \left[e^{-s\alpha}U(s)\right]_{\min} = e^{-\alpha s^*}\cdot\frac{1}{2}\left[1+\frac{\alpha}{1-\alpha}\right]$$

$$= \left(\frac{\alpha}{1-\alpha}\right)^{-\alpha}\frac{1}{2}\left(\frac{1}{1-\alpha}\right) = \frac{1}{2}\frac{\alpha^{-\alpha}}{(1-\alpha)^{1-\alpha}}$$

(d) For $E(u)=\frac{1}{2}$

$\alpha = E(u)+\epsilon = \frac{1}{2}+\epsilon$, $1-\alpha = \frac{1}{2}-\epsilon$

So for $0<\epsilon<\frac{1}{2}$: $1>\alpha>0$, $1-\alpha>0$

$$\mathbb{P}\left(u>\tfrac{1}{2}+\epsilon\right) \leq \left[e^{-s\alpha}U(s)\right]_{\min}^N = 2^{N\log_2[\,]_{\min}} = 2^{\beta N}$$

$$\beta = \log_2[\,]_{\min} = -1-\left(\tfrac{1}{2}+\epsilon\right)\log_2\left(\tfrac{1}{2}+\epsilon\right)-\left(\tfrac{1}{2}-\epsilon\right)\log_2\left(\tfrac{1}{2}-\epsilon\right)$$

For $0\leq\epsilon\leq\frac{1}{2}$: $\frac{1}{2}\leq\frac{1}{2}+\epsilon\leq 1$

$\left[\log_2\left(\frac{1}{2}+\epsilon\right)\right]\left[\frac{1}{2}+\epsilon\right]$ $\quad -\frac{1}{2} \leq \left(\frac{1}{2}+\epsilon\right)\log_2\left(\frac{1}{2}+\epsilon\right) \leq 0$

$\frac{1}{2}\log_2\frac{1}{2} = -\frac{1}{2}$

$\left(\frac{1}{2}+\frac{1}{2}\right)\log_2\left(\frac{1}{2}+\frac{1}{2}\right) = 0$ $\quad \beta<0$

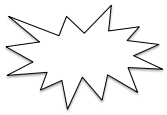
$2^{\beta N}$ exponentially decaying

$\epsilon=0$: $\frac{1}{2}\log_2\frac{1}{2} = -\frac{1}{2}$

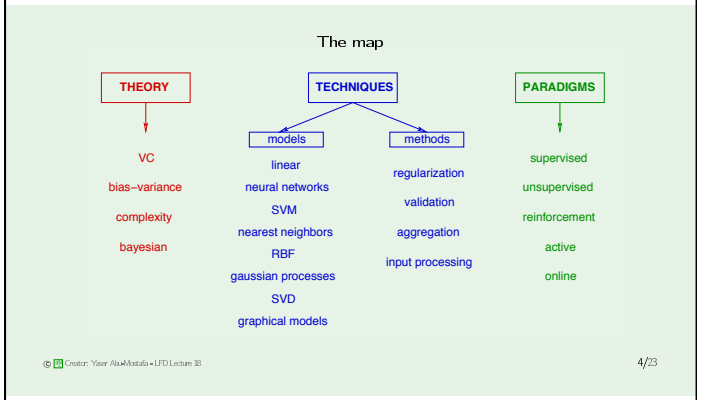
$\epsilon=\frac{1}{2}$: $0\log_2 0 = 0$

$-\frac{1}{2} \leq \left(\frac{1}{2}-\epsilon\right)\log_2\left(\frac{1}{2}-\epsilon\right) \leq 0$

Road Map?



- **Better drawing on Road Map (GAN? Memory? Phase Transition?)**
- **Reading assignment needs to be so clear (which pages)**
- **Hard/easy learning problem (more precise definition?)**



Lec1 Part-1 Introduction
Lec2 Part-1 ML Types (examples)
Lec3 Part-2 Linear Model - Regression
Lec4 Part-2 Bias-Var
Lec5 Part-2 Regularization and Kernel
Lec6 Part-2 Functional Space
Lec7 Part-3 Classification
Lec8 Part-3 Decision Tree
Lec9 Part-3 Entropy vs ML
Lec10 Part-3 Random Forrest
Lec11 Part-3 SVM Kernel
Lec12 Part-3 SVM Kernel
Lec13 Part-3 Project Enhancer
Lec14 Part-4 Graph Models
Lec15 Part-4 Exact Inference
Lec15 Part-4 Inference as Traditional Statistics (vs Bayesian)
Lec16 Part-4 Deterministic Approx KL to LBP
Lec17 Part-4 Deterministic Approx KL to MF
Lec18 Part-4 Stochastic Approx MCMC
Lec19 Part-4 Bayesian Model Selection
Lec19 Part-5 PCA
Lec20 Part-5 Latent Variable
Lec21 Part-6 Neural Network (Universal)
Lec22 Part-6 Neural Network (BP)
Lec23 Part-6 Hopfield
Lec24 Part-6 Boltzmann Machine
Lec25 Part-6 BM and RBM Training
Lec26 Part-6 DL DCA RNA

AI

Easy for people to describe, but Hard for people to do
(1) Clear Features?
(2) Large-scale simulations (Newton, Einstein)?

Easy for people to do, but Hard for people to describe
(1) Feature selection?
(2) Boltzmann machine?
(3) Each Layer is very confusing (necessity?)

Questions: 1. Weather forecast? Which category?
2. Inverse-engineering Physics Law? Which category?

Intuitive Problems –
The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones. If we draw a graph showing how these concepts are built on top of each other, the graph is deep, with many layers. For this reason, we call this approach to AI deep learning.