CS 215- Data Interpretation and Analysis
(Post Midsem)

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay
25sep23

A Perspective on Machine Learning

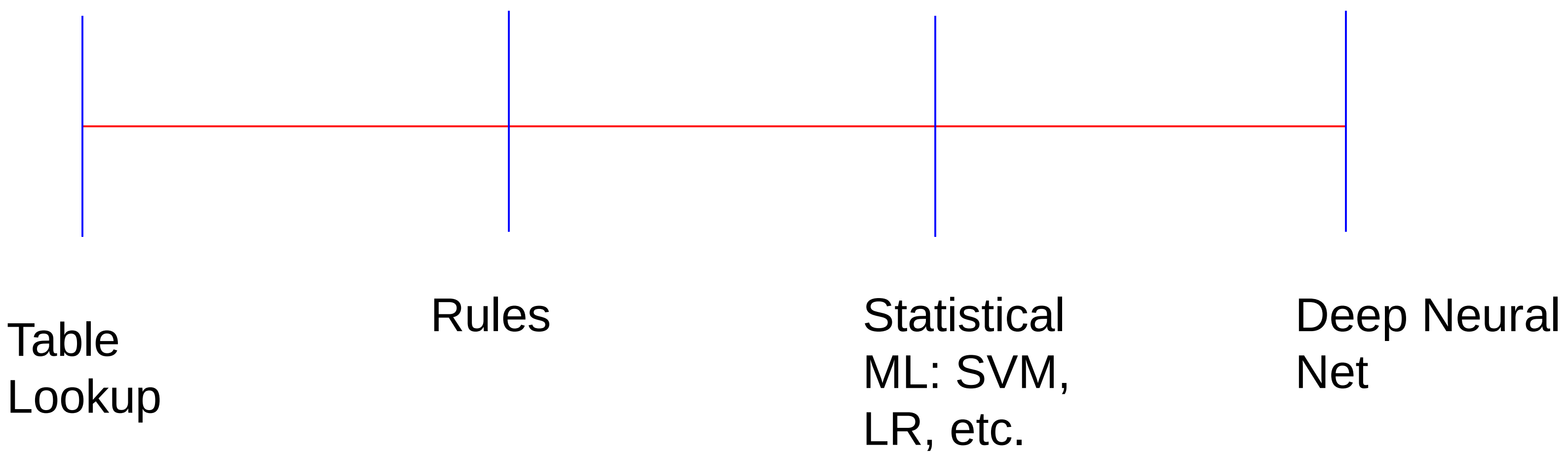
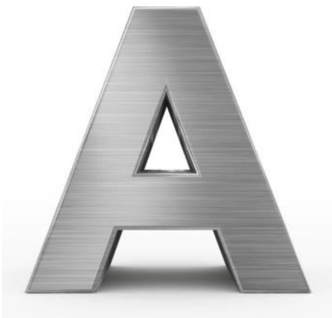


Table Look up



How many to store?

What is the essential “Aness”?

Rules

- Letter ‘A’ is formed from **two inclined straight lines, meeting at a point** with a **horizontal straight line cutting across**
  - Exception: need not be straight lines; need not meet; the 3rd line need not be horizontal, need not be straight

- Leads to false negative- ERROR OF OMMISSION

From Exact to Approximate, 100% to X% (X< 100)

- Very, very, hard to eliminate completely- false positives and false negatives
- Even humans cannot achieve that performance in most complex tasks
- Decision making under uncertainty, under error bound

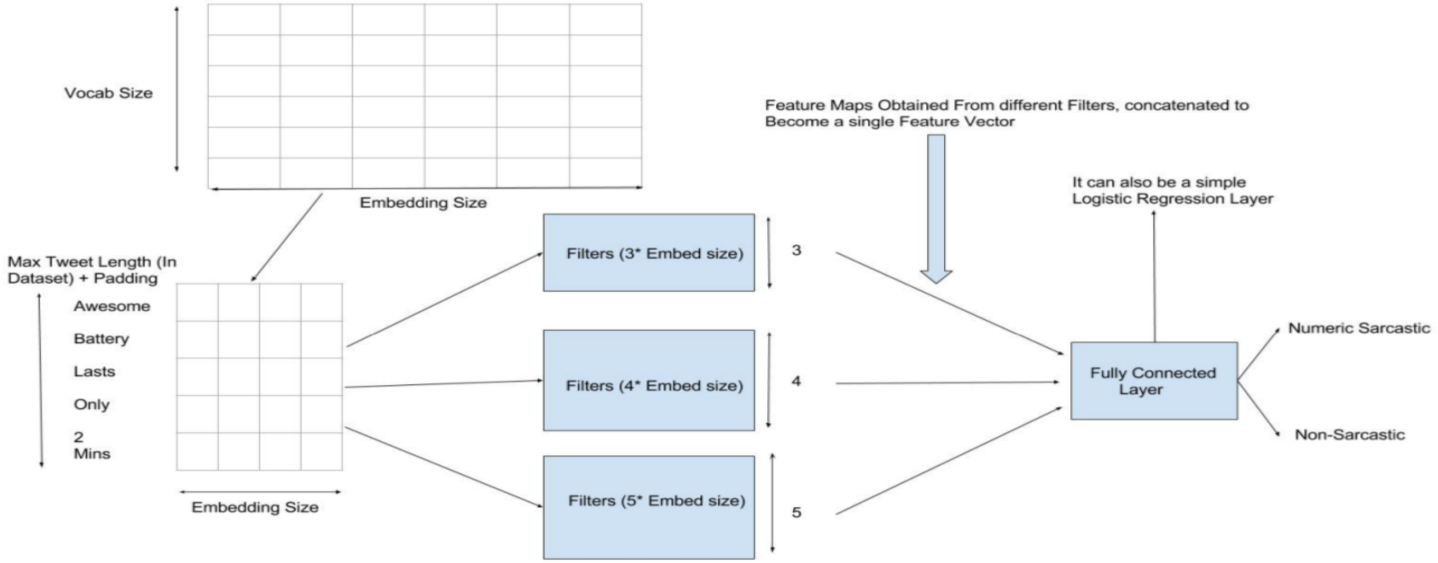
LEARN from Data with Probability Based Scoring

- **Data + Classifier > Human decision maker !!**
- With LOTs of data, learn with
  - High precision (small possibility of error of commission)
  - High recall (small possibility of error of omission)

- But depends on human engineered features, i.e., capturing essential properties

Reduce human dependency: DEEP LEARN

- End to end systems; essential properties learnt at intermediate layers



LEARNING *vs.* KNOWING, and the role of probability (1/2)

- If we KNOW, we do not need data and ML
- RULES capture the underlying phenomenon
- But if we do not KNOW, we need data and probability
- For example, laws of physics
- 2nd law of motion: $F = d(MV)/dt = ma$

LEARNING *vs.* KNOWING, and the role of probability (2/2)

- Breaks in “special” situations, such as those involving
  - high speeds, strong gravitational fields, and quantum-scale interactions,
  - where more advanced theories like relativity and quantum mechanics are required
- However, it was data that got distilled into laws! Story of Tyco Brahe→ Kepler's Laws of Motion

Another situation of probability and ML: Incompleteness of Information is Inevitable

- Input to output is not unique
- Given the complete information the mapping is unique, but it IMPOSSIBLE to get ALL the information most of the time!!
- E.g., sentence meanings
- *I saw the boy with a telescope* has two meanings: I have the telescope or the boy has the telescope
- *I dropped the telescope, when I was seeing the boy with a telescope*

A Practical Problem

- A bridge is being built. The weight it can tolerate has a distribution with $\mu=400$ and $\sigma=40$. A car that goes on the bridge has weight distribution given by $\mu=3$ and $\sigma=0.3$. We want the probability of damage to the bridge to be less than 0.1. How many cars can we allow to go on the bridge?

When does the bridge break?

$$W_{total} > W_{tolerance}$$

Deterministic

- Damage if

$$3N=400$$

$$\Rightarrow N=133$$

Deterministic, but with bounds (1/2)

- Strongest bridge and lightest car
- Bridge withstand 440 and car weight 2.7
- Most liberal situation also most risky!

ceiling (2.7N=440)

$\Rightarrow$ N=163 !!

Deterministic, but with bounds (2/2)

- Weakest bridge and heaviest car
- Bridge withstand 360 and car weight 3.3
- Most conservative situation and safest
- But resource wise most inefficient!!

floor(3.3N=360)

⇒ N=109 !!

Lets look at these numbers for a while

- Most liberal, 163 nos.
- Most conservative, 109 nos.
- What should be the ACTUAL NO. of cars to be allowed?
- This is an OBJECTIVE DECISION
- A precise no. has to be allowed
- How much is that?

Depends on the priority: safety the only consideration

- As an Administrator, I want to PLAY VERY SAFE
- No risk
- Then only 109 cars
- Bridge will never break
- I am safe

Point of view and priority: earning first, throughput first, efficiency first

- I want to have maximum utilization of the bridge
- Maximum earning from toll
- Maximum movement across river
- Maximum economic activity
- Maximum interaction
- People happy ☺

But risk is higher!

- The bridge will VERY LIKELY cross the tolerance limit
- Bridge breaks
- Lives lost
- Property damaged
- People unhappy ☹

Relate to covid-19 situation?

- Yes
- Do not go out
- Do not interact
- Very safe
- But no economic and social activity
- How to sustain?
- How to break monotony

Need balance, sweet spot is somewhere in between, MIDDLE PATH

How to get the sweet spot? The middle path?

- Answer

PROBABILITY

Back to the bridge

- MOO: Multi-objective Optimization
- Many objectives to be satisfied
  - Safety
  - Utilization of facility
  - Earning
  - People satisfaction
  - *Etc.*

Bring in probability

- #cars = N
- Each car's weight is normal with $\mu=3$ and $\sigma=0.3$
- Invoke Central Limit Theorem

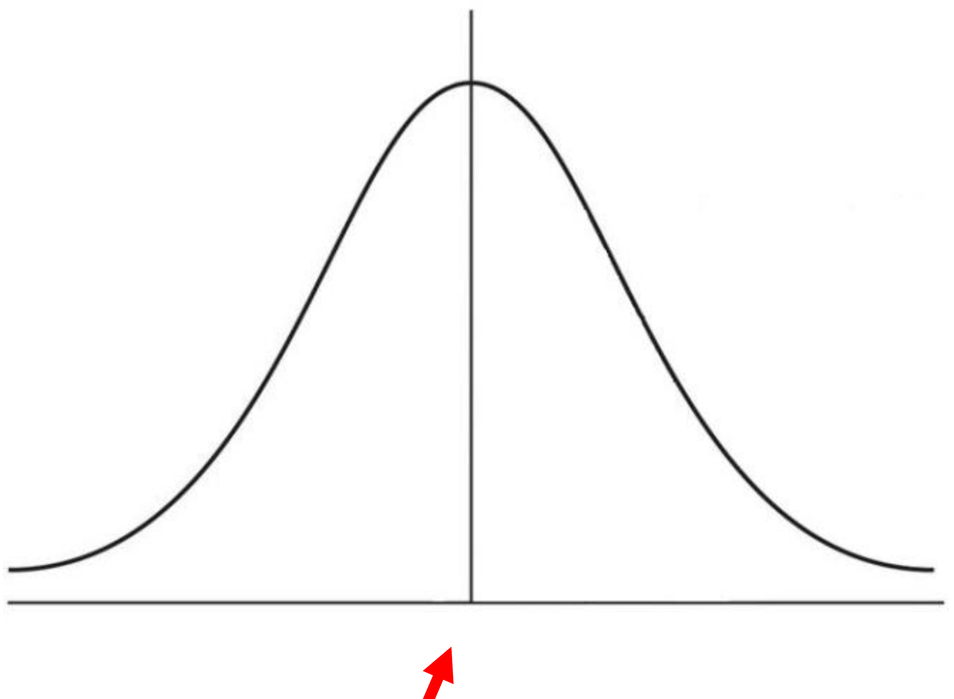
Apply CLT

- By central limit theorem, the sum of Gaussian Random Variables is Gaussian with mean and variance being sums of individual means and variances

$\Rightarrow$

total weight of N cars is normal with μ=3N and σ^2=0.09N

$W_{tolerance}$ looks like this…



μ=400 and σ=40

We allow some risk

- Bridge is damaged when

- $W_{total} > W_{tolerance}$

- *i.e.,* $W_{total} - W_{tolerance} > 0$

Allowing Risk…

- Why allow risk?
- Remember 109 cars will be completely safe
- But that will not utilize the **RESOURCE** optimally
- Allow more cars
- Take some **RISK**

RISK-RESOURCE Trade Off

- We want to take some risk
- To utilize resource optimally
- But guarantee that the RISK is NOT TOO MUCH!!
- What instrument do we have?

PROBABILITY

We want

- What no. of cars will cause the probability to exceed 0.1?

$$\textit{Probability}(W_{total} - W_{tolerance}) > 0.1$$

LHS is a function of N

W_{total} is a function of N by CLT

Meaning of $\boldsymbol{Probability(W_{total}\text{-}W_{tolerance}) > 0.1}$

- Let N_{unsafe} be the limit on the number of cars allowed on the bridge

- Out of 1000 cases of the bridge allowing N_{unsafe} cars to pass over it, in more than 10 cases the bridge will break

Range of N_{unsafe}

$$109 <= N_{safe} <= 163$$

Min Risk
Min utilization

Max utilization
Max Risk

Bring N, the number of cars in picture

- Central Limit Theorem applied again
- $W_{total} - W_{tolerance}$ is a random variable
- Follows Normal Distribution
- *Mean= 3N-400*
- *Variance= 0.09N^2+1600*

Convert to Standard Normal Form

$$z \equiv \frac{(W_{total} - W_{tolerance}) - (3N - 400)}{\sqrt{0.09N + 1600}}$$

We want this event…

$$(W_{total} - W_{tolerance}) > 0$$

$$\Rightarrow \frac{(W_{total} - W_{tolerance}) - (3N - 400)}{\sqrt{0.09N + 1600}} > \frac{-(3N - 400)}{\sqrt{0.09N + 1600}}$$

$$\Rightarrow z > \frac{-(3N - 400)}{\sqrt{0.09N + 1600}}$$

When will this Probability exceed 0.1

$$P\left(z > \frac{-(3N-400)}{\sqrt{0.09N+1600}}\right) > 0.1$$

Solving this gives N <= 117

How?

Use Standard Normal Form Table

$$P(z < V) = \int_{-\infty}^{V} \frac{1}{\sqrt{2\pi}} \exp\left(-y^2/2\right) dy$$

$$Now\ P(z > V) = 1 - P(z < V)$$

$$Since\ we\ want\ P(z > V) > 0.1$$

$$\Rightarrow 1 - P(z < V) > 0.1$$

$$\Rightarrow P(z < V) <= 0.9$$

V=1.28, consulting the table

V=1.28

Standard Normal Probabilities

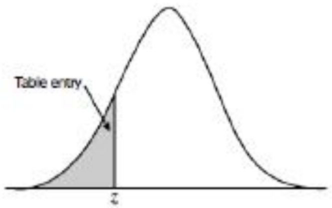


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

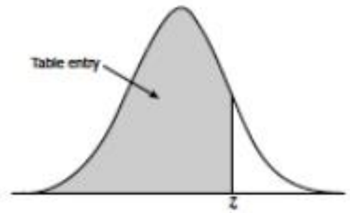


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Get N from…

$$1.28 = \frac{-(3N - 400)}{\sqrt{1600 + 0.09N}}$$

N= ~117

Conclusion

If we allow more than 117 cars on the bridge, then in 10 out 1000 such cases the BRIDGE WILL BREAK!!