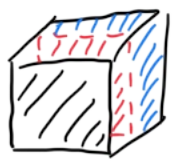
5) Math Preliminaries

Spectral Theories

- Idea: Complicated → simple → ... → simple but Complex is unintuitive to understand

- Deep learning is simple (math), complicated (many parts), complex (countless nonlinearities)

Math vs. Computer Terminologies

- Object ① 7 ② $\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$ ③ $\begin{bmatrix} -1 & 0 & 2 \\ 0 & 1 & 4 \\ 1 & 4 & 9 \end{bmatrix}$ ④ 

	math	scalar	Vector	Matrix	Tensor
		(geometric interpretation)	(1D List)	(2D spreadsheet)	(≥3D)
Numpy		Array	Array	ND array	ND array
PyTorch		Tensor	Tensor	Tensor	Tensor

- Ex) Grayscale image is a matrix of brightness intensity scalars

- Dummy-coding vs. One-hot encoding

Student	y
Pass	1
Pass	1
fail	0

$\Rightarrow y = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$

Genre	History	Scifi	kids
y_1	0	1	0
y_2	0	0	1
y_3	1	0	1

$\Rightarrow y = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$

Matrix Operations

① Transpose (A^T): change orientation (not information) of A

② Dot product: Given a & b are vectors/matrices of same shape

$$\alpha = a \cdot b = \langle a, b \rangle = a^T b = \sum_{i=1}^{n} a_i b_i$$

Ex) $\begin{bmatrix} 0 & 3 & 2 \\ -3 & -3 & 1 \\ 1 & 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 6 \\ 2 & -1 & 0 \\ 5 & 1 & 4 \end{bmatrix} = \begin{matrix} 0(1)+3(0)+2(6)+ \\ -3(2)-3(-1)+1(0)+ \\ 1(5)+0(1)+2(4) \end{matrix} = 22$ → reflects commonalities between 2 objects

③ Matrix Mult: a list of dot products

- Rule of validity: $A_{m\times r}\ B_{r\times n} = C_{m\times n}$ ⇒ Ex) $\begin{bmatrix}0&1\\2&3\end{bmatrix}\begin{bmatrix}a&b\\c&d\end{bmatrix}=\begin{bmatrix}0a+1c & 0b+1d\\ 2a+3c & 2b+3d\end{bmatrix}$

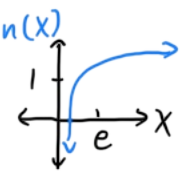
Softmax Function

- Formula: $\sigma_i = \frac{e^{z_i}}{\sum e^z}$ ⇒ Ex) Given $z=\{1,2,3\} \rightarrow e^z=\{2.72, 7.39, 20.01\}$
  ↳ $\sum e^z = 30.19 \rightarrow \sigma=\{0.09, 0.24, 0.67\}$ sum to 1

- Interpretation: $y_1, \dots, y_N \rightarrow$ [Softmax] $\rightarrow P(y_1), \dots, P(y_N)$ probability to occur } the softmax output is used for categorization

Logarithm Function

- Interpretation: log is montonic → $\arg\min x = \arg\min \log(x)$
  log stretches for small x → log works better for small #s

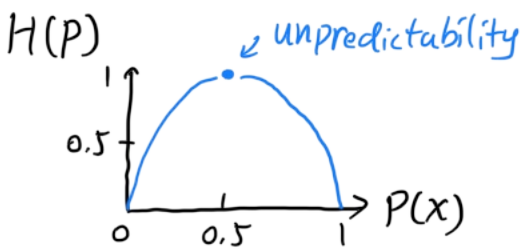


- So log is good for optimization of small inputs

Entropy

① Entropy: Amount of uncertainty about an output

$$H(P) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i),\ i = i\text{th possible event}$$



- Interpretation: low entropy → repeated value in dataset
  high entropy → high variability in dataset

② Cross Entropy: describes relationship b/t 2 probability distributions

$$H(P,Q) = -\sum_{i=1}^{n} P(x_i)\log Q(x_i),\ \text{where } P(x_i)=\sigma_{1i},\ Q(x_i)=\sigma_{2i}$$

Min/Max & Argmin/Argmax

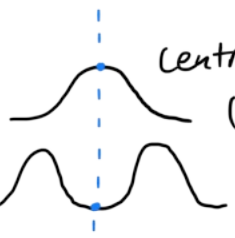
- arg means the position z where optimal values occur

Ex) If $f(x)=\{1,-1,3,0\}$, Then $z = \arg\max_x f(x) = 2$ (in python)

Ex2) In DL, arg is used to locate label that produces optimal probabilities

Mean & Variance

① Mean: $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$

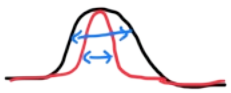


central tendency of data distribution (where most data lie)

note this only works well for normal distribution

↳ removes bias

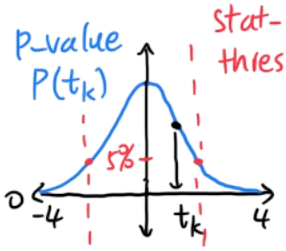
② Variance: $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$



dispersion of data distribution

t-Test

- Mechanism:



stat-diff threshold: i) $P(t_k) > 5\%$ → $|t_k|$ is small → no
ii) $P(t_k) < 5\%$ → $|t_k|$ is large → yes

t-value, $t_k = \frac{\bar{x} - \bar{y}}{s/\sqrt{n}} = \frac{\text{diff of means}}{\text{stdev}}$

· Yes interpretation: 2 samples of model prediction accuracies are very different
So 2 models can be evaluated by performance

Derivatives

- Interpretation: tells direction of increases/decreases in a location
Useful to move down the error function to optimize model

- Thus, useful for minimizing error function

Ex) Says $x = 0, 1, 2$ gives $f'(x) = 0 \Rightarrow$ "critical pts"

- Rules: ① $(f+g)' = f' + g'$ ② $(fg)' = f'g + fg'$ ③ $\frac{df}{dx} f(g(x)) = f'(g(x))\,g'(x)$