9/22

Today

- CS AG update
- Homework ? GitHub
- Review from Last year
  - Prob/stats
  - NN
  - NLP

NLP

- Translation
- Summarization
- Sentiment Analysis

① Representations of Language / Feature Extraction

- The movie was great   +, 1 Label
- POS tagging   the - article
  movie - noun ⇒ NN

{ great, good .. } ← +
{ bad, terrible ... } ← −

The movie was not great   −   + ?

Bag of Words : $\bar{x}$: [The movie was great]

Corpus : V ← $f(\bar{x})$: [the 1, movie 1, ..., spectacular 0, ..., great 1]

|V| x d vectors

- Sparse vectors
- Dense vectors
  - Counter
    'the' : 1
    'movie' : 1
    ⋮

great | great
bad | GREAT

Preprocessing

① tokenizing → sentence → tokens
great great!

② stopwords : a, the ~~X~~

③ lowercasing : great GREAT
↳ ✓

④ Rare words : Counter ← 10,000 words

Lowercasing :
The movie was GREAT → the movie was great

---

nltk

- Domain-specific tasks

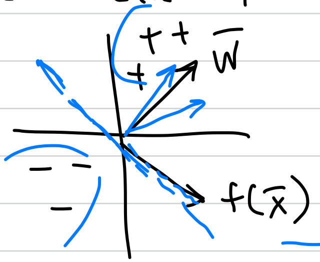
$f(\bar{x}) \rightarrow$ dense feature vector

$\bar{w} \rightarrow$ weight vector

$\bar{w} \cdot f(\bar{x})$ $> 0 \Rightarrow +$ $< 0 \Rightarrow -$

$$\begin{bmatrix} 4 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 4 + 6 = 10$$

Linear classification



$\rightarrow 0$

Fact: dot product of orthogonal vectors is 0

Neural Nets

$\rightarrow$ linear classification

– (really, good) → Counter (not, good) → : 1

– (not, good)

The movie was not good

Bigrams

Single words: Unigrams

	LR	
- Unigram	81%	Pang et al. (2002)
- Bigrams	77.3%	
- Unigrams + Bigrams	80.6%	

Counter
- "the" : 1
- "(the, movie)" : 1

Kim (2014)

CNN

~83%

---

update weights in training

- SGD : stochastic gradient descent

```
for t in range(0, epochs):
    for i in range(0, D)    ← D is # of labeled data points
        sample j ~ {0, 1, ... D-1}
        - predict
```

$$\bar{w} \leftarrow \bar{w} - \alpha \cdot \frac{\partial loss}{\partial \bar{w}}$$

step size / learning rate

Probability / Statistics Review

Definitions

(R.V.) Random Variable (X): A variable whose outcomes depend on randomness

Probability ($P(X)$): chance of event X occuring

Expectation ($E[X]$): Expected outcome over ∞ trials

Variance ($Var[X]$): Expected squared deviation of R.V. from $E[X]$ (or mean)

Standard Deviation (σ): Expected deviation from the mean

Covariance ($Cov(X, Y)$): The joint variability between two R.V.

Correlation ($Corr(X, Y)$): A measure of the statistical relationship between two R.V

Probability Density Function (PDF): function that shows the distribution of values of R.V.

Cummulative Distribution Function (CDF): Integral of PDF, tells us $P(a < X \leq b)$

Formulas

X is a random variable with sample space Ω

- $E[X] = \sum_{\omega \in \Omega} P(X=\omega) \cdot \omega$

  $E[X+Y] = E[X] + E[Y]$

  $E[aX] = aE[X]$

- $Var[X] = E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X^2]]$

  $= E[X^2] - 2(E[X])^2 + (E[X])^2$

  $= E[X^2] - (E[X])^2$

  $Var[aX] = a^2 Var[X]$

  $Var[X+Y] = Var[X] + Var[Y] + 2Cov(X,Y)$
  $= Var[X] + Var[Y]$ (if $Cov = 0$)

  $Var[X+b] = Var[X]$

- $\sigma = \sqrt{Var(X]}$

- $Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$
  $= E[XY] - E[X]E[Y]$

- $Corr(X,Y) = \dfrac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$

Concepts / Theorems

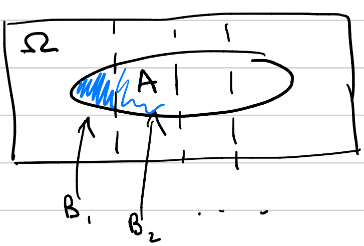
- Common PDFs: (Discrete)
  - Bernoulli: $p = P(X=1)$ $\quad q = 1-p = P(X=0)$
  - Binomial: $P(x) = \binom{n}{x} p^x q^{n-x}$

    $n = \#$ trials
    $x = \#$ success
    $p = P(X=1)$
    $q = P(X=0) = 1-p$
  - Uniform: $P(X=x) = \frac{1}{n}$ $\quad n = \#$ outcomes

- (continuous)
  - Uniform: $X \sim U(a,b)$ $\quad f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$

    $P(c < x < d) = \frac{c-d}{b-a}$
  - Normal (Gaussian): $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

- Law of Total Probability:

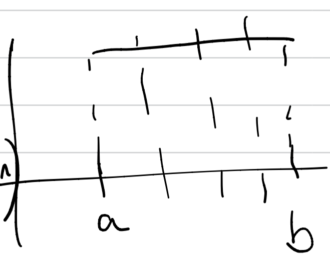


$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots$

$P(A) = \sum_{i=0}^{k} P(A|B_i) \cdot P(B_i)$

$P(\text{Umbrella}) = P(\text{Umbrella}|\text{Rains}) + P(\text{Umbrella}|\text{no rain})$

Rains or it doesn't



- Bayes Theorem: $P(A|B) = \dfrac{P(B|A) \cdot P(A)}{P(B)}$

- Law of Large Numbers: As sample size increases the sample mean approaches the true population mean

- Central Limit Theorem: Any distribution with well defined mean and variance can be transformed into a new distribution that approaches a normal distribution.

3 coins

HT	HH	TT
$1/3$	$1/3$	$1/3$

1st flip is heads

$$P(HH \mid 1^{st}\text{ flip heads}) = \frac{\overbrace{P(1^{st}\text{ flip is heads} \mid HH)}^{1} \cdot \overbrace{P(HH)}^{1/3}}{P(1^{st}\text{ flip heads})}$$

$$\frac{1/3}{1/2} = 2/3$$

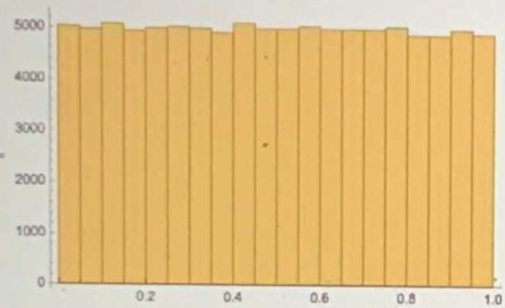
$$P(\text{Heads}) = P(\text{Heads} \mid HT) \cdot P(HT) + P(\text{Heads} \mid HH) \cdot P(HH) + P(\text{Heads} \mid TT) \cdot P(TT)$$

$$= 1/2 \cdot 1/3 + 1 \cdot 1/3 + 0 = 1/2$$

CLT

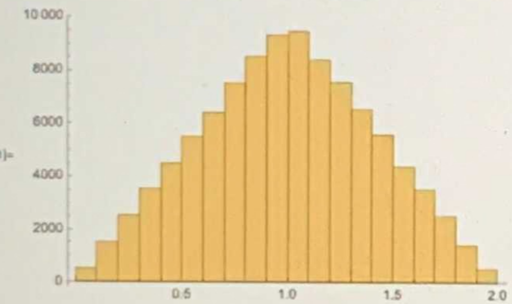
```
In[148]:= r := RandomVariate[UniformDistribution[]]
In[150]:= Table[r, {100 000}] // Histogram
```

Out[150]=



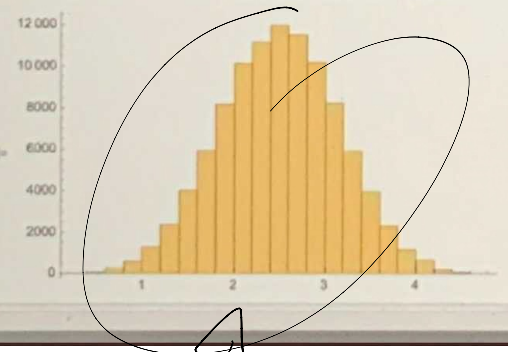

```
In[151]:= Table[r + r, {100 000}] // Histogram
```

Out[151]=

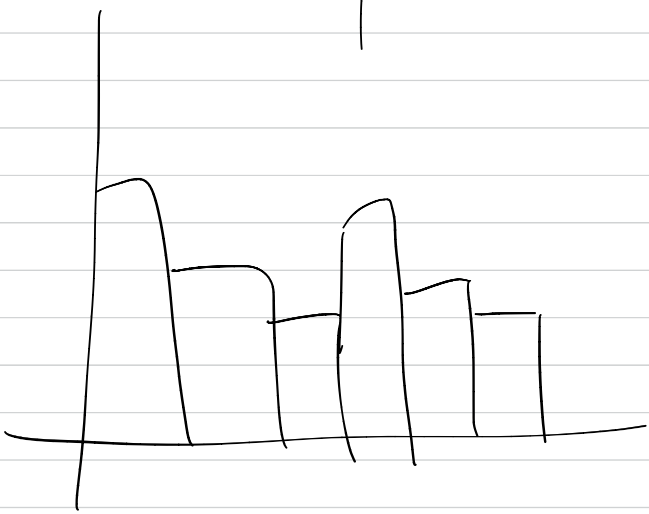



```
In[153]:= Table[r + r + r + r + r, {100 000}] // Histogram
```

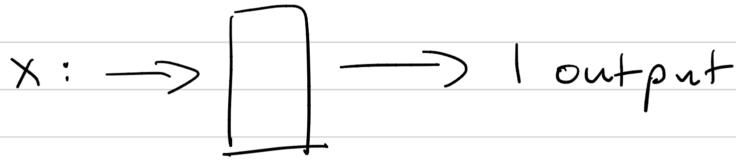
Out[153]=



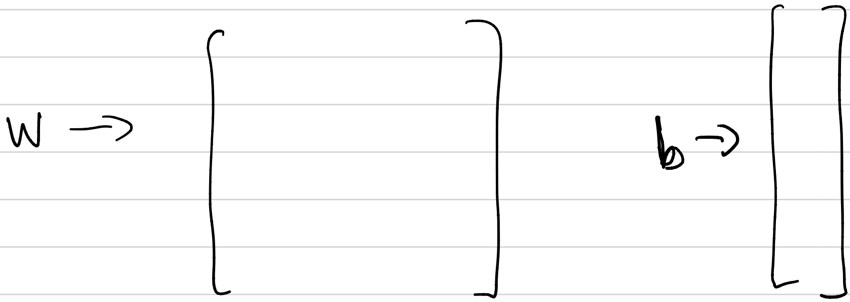
Examples



Neural Nets



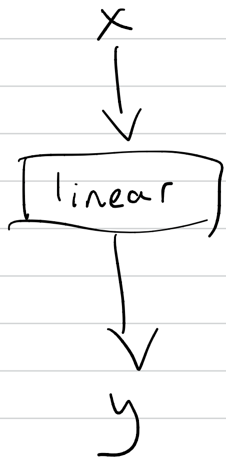
W , b

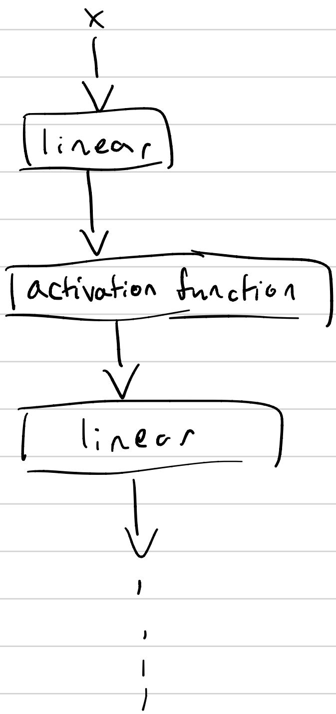


$$y = Wx + b$$

$$x = [\ 10 \quad \text{values}\ ]$$

$$y = {}_{1}\overset{10}{\begin{bmatrix} \quad & \quad \end{bmatrix}}\ {}_{10}\overset{1}{\begin{bmatrix} \quad \\ \quad \end{bmatrix}} + {}_{1}\overset{1}{\begin{bmatrix} \quad \end{bmatrix}}$$


x
linear
y


x
linear
activation function
linear

$\text{Relu}(x) = \max(0, x)$

tanh

sigmoid