Entropy and Cross Entropy.

Review

Tensor Flow - tensors

multidimensional data arrays

examples

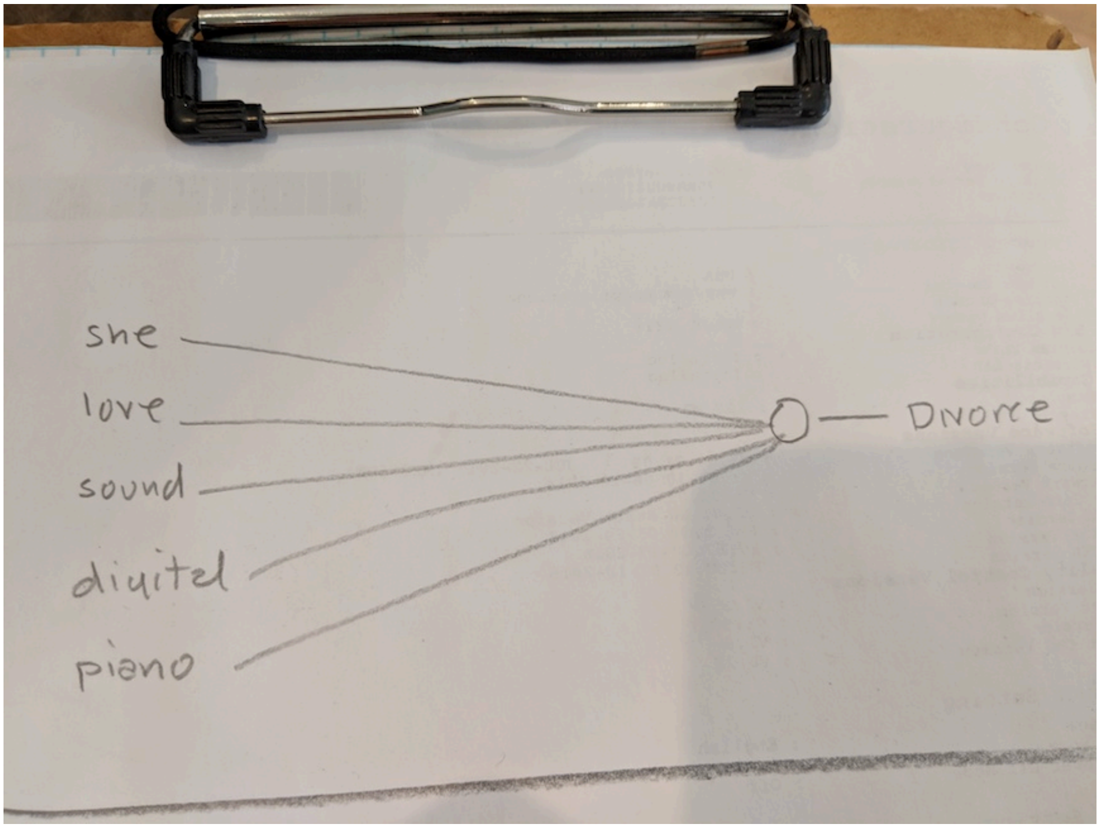
Divorce

Category	She	Love	Sound	Digital	Piano
Divorce	5	4	1	0	1

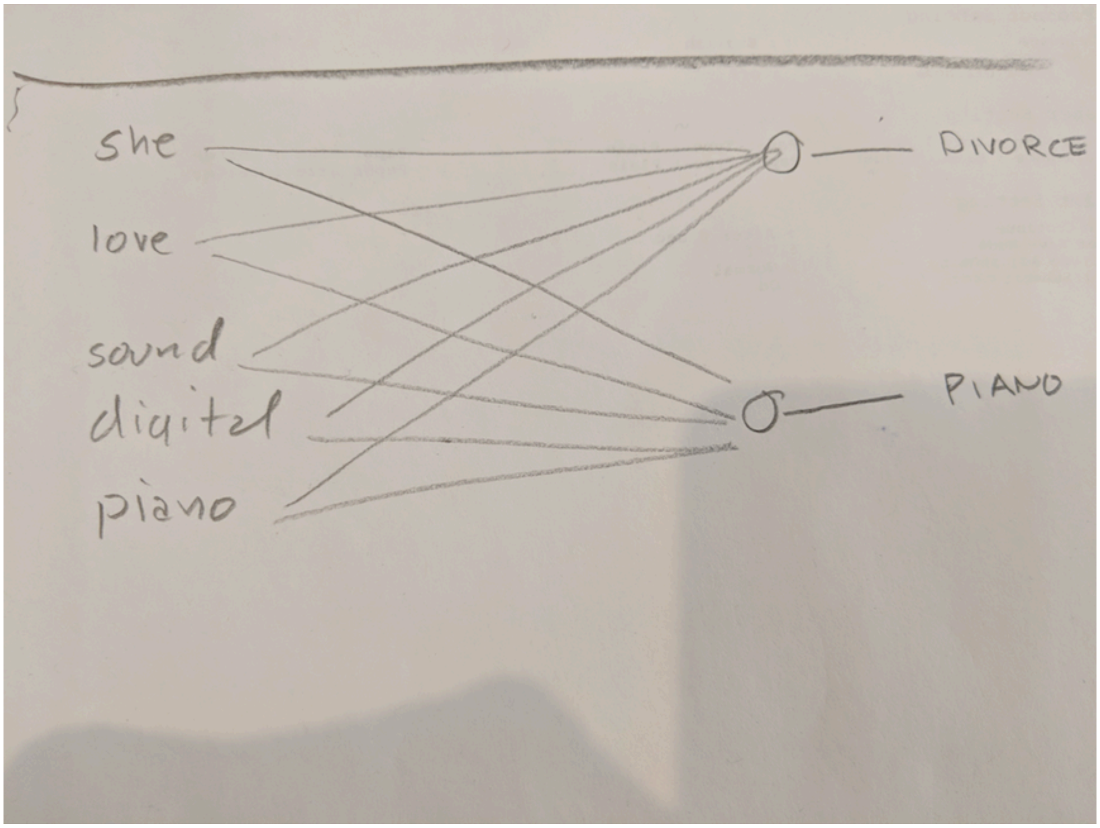
The training set

Category	She	Love	Sound	Digital	Piano
Divorce	5	4	1	0	1
Divorce	1	4	1	0	1
Not Divorce	0	1	3	1	1
Divorce	3	2	1	0	0

If our goal is to identify whether the article is about divorce or not:



Goal to classify divorce or piano



Team - what would design be for the following:

(show sample datasets)

- athletes
- diabetes
- titanic
- detect emails from Chris

definitions

we call the values of these features for each training instance

```
tf.placeholder
```

so what we might call a variable that holds all the data in a row is called a placeholder.

```
X = tf.placeholder(tf.float32, [None, 28, 28, 1])
```

None means it can take any number of rows

we would also have a placeholder for the labels.

Y_ = tf.placeholder(tf.float32, [None, 10])

that parses to

image	0	1	2	3	4	5	6	7	8	9
0	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0
2	0	0	1	0	0	0	0	0	0	0

weights the w0 etc. are called

```
tf.Variable

W = tf.Variable(tf.zeros([784, 10]))
# biases b[10]
b = tf.Variable(tf.zeros([10]))
```

to make further progress understanding the code

need to back up a bit.

hand out quad copter data

entropy and decision trees

covering different algorithms.

0R

0 stands for zero and R stands for rule so zero rule – no rules.

If you had to make a decision without asking any questions, what would your decision be?

Looking at the quad table. We are trying to decide whether to fly or not.

How do we do this?

Got that?????

why is 0r important.

Suppose we code a fancy Deep Learning TensorFlow algorithm to detect a rare disease that is 99.99% accurate. Is that good? It seems good.

Suppose we use a random set of people in the general population for our test. And suppose the disease only affects 1 in 10,000 people. then the 0R case is 99.99% accurate.

1R

Now we are going to go for a more difficult approach. 1R. Any guesses as to the meaning of that?

The goal here is to pick the best rule. Let's say I have two tests for breast cancer detection: xray and ultrasound. Xray has an accuracy of 95% ultrasound 99% – in every other way, cost, etc. they are equal. Which would I pick if I could only choose one test?

WHY?

Accuracy. So if we are picking 1R we go with the most accurate.

Makes sense doesn't it?

Look at the tennis case. I am going to construct a rule for each attribute—each column:

Case outlook: sunny: →?? overcast:→ ?? rainy: →??

How do I fill in the question marks. In the first case should I say if oulook is sunny should I play tennis or not?

Count 5 sunny's for 3 of them the answer is no and for 2 of them the answer is yes.

I would choose ????? no. And how many errors does that part of the rule make? 2

###Case outlook:

```
sunny →  no         2/5 errors
overcast: → ??
rainy: → ??
```

fill in the rest. Don't look in book. It is worth going through the example

Case outlook:

```
sunny: no         2/5 errors
overcast: yes   0/4
rainy: yes    2/5
```

So the total errors for this rule is 4/14

Then we do it again for the temperature attribute

Case temp

```
hot → ?
mild → ?
cold → ?
```

how many errors does this make? Teams – compute this and for all other attributes. Put on board

Case temp 5/14 errors

```
hot → no    2/4
mild → yes  2/6
cold →  yes 1/4
```

humidity has 4/14 errors

and windy has 5/14 errors

so we pick one of the 4/14 cases.

If age → young none 4/8 AGE errors 8/24 age → pre-presbyopic 3/8 age → presbyopic 1/5

Presription → myope

Question: Why don't I use the day column? I would have zero errors?

SUMMARY 1R –> nothing special makes sense

Bits

Before moving on I am going to diverge a bit from our path.

Teams review what are bits.

Suppose I have a wire from my office to the computer lab. In the lab I have a display:

Ron Z is …. IN / OUT

In my office I have some switch. Say when I sit in my chair.

Every 5 minutes, That office chair is going to send out a code—a sequence of bits—saying whether I am in or not.

How many bits long does that code need to be?

Suppose we are in some spy operation—We are watching a house. Give me a professor in the computer science dept. Same wire deal

state	prob
empy	x
A in	X
B in	X
A & B in	x

Every 5 minutes I send a message to headquarters telling them that status.

On average how many bits do I need per message? 2 bits

Suppose I give you additional info.

state	prob
empy	1/2
A in	1/8
B in	1/8
A & B in	1/4

½ the time the house is empty.

¼ the time both Suspect A and Suspect B are in the house

1/8 of the time only Suspect A is in the house 1/8 of the time only Suspect B is in the house

Can we devise a coding scheme where the average message length is less than 2 bits?

state	prob	code
empy	1/2	0
A in	1/8	110
B in	1/8	111
A & B in	1/4	10

```
0   no one home
10  both suspects home
110 only suspect A
111 only suspect B
```

Critical to have a code that we can divide

1 bit * .50 + 2 bits * .25 + 3 bits * .125 + 3 bits * .125

```
.1  + .5 + .75  = 1.75
.2
```

When you compress a file with zip what does it do?

Demo

entropy

The word entropy as used in computer science and statistics means how many bits on average do I need to encode the information.

How many bits do I needto report the rolling of a fair 8 sided dice?

3 bits.

You can also think of entropy as answering a game of twenty questions. On average how many yes/no questions do I need to ask to get the info.

```
1   in the case above.
```

We compute entropy as follows

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

BASE 2

Let's go back to our spy example

```
 - (½ log ½  + ¼ log ¼ + 1/8 log 1/8  + 1/8 log 1/8)
```

= - (½ (-1) + ¼ (-2) + 1/8(-3) + 1/8(-3) = -(-1/2 + - ½ + -.375 + -.375 = 1.75

TEAM WORK

Simplified Polynesian

```
p 1/8
t  ¼
k 1/8
a ¼
I 1/8
u 1/8
```

What is the per letter entropy of this?

4 * 1/8 log 1/8 + 2 * ¼ log ¼ = 2.5 bits

This idea of entropy – information theory – is key for many areas of computer science. Machine translation, speech recognition. END OF DIVERSION

Decision Trees

Show visual

How does this apply to deep learning?

Recall when I was trying to recognize car images.

Image	Side	Backcorner	front corner	back	front	wheel
1	1	0	0	0	0	0
2	0	0	1	0	0	0
3	0	1	0	0	0	0

and my program gave

Image	Side	Backcorner	front corner	back	front	wheel
1	0.74899	0.15299	0.07086	0.01863	.00348	0.00505
2	0.30985	0.09308	0.4094	0.07918	0.10087	0.01209
3	0	1	0	0	0	0

I can use **entropy** to see how good my answer is. This is known as **cross entropy**

So we measure how well we are doing by comparing 2 vectors The formula is

So **L** is the actual label - the real value and **S** is our guess

$$D(S, L) = -\sum_i L_i \log(S_i)$$

remember

$$D(S, L) \neq D(L, S)$$

not symmetric

Let's do this for image 1:

Image	Side	Backcorner	front corner	back	front	wheel
1	1	0	0	0	0	0
1	0.74899	0.15299	0.07086	0.01863	.00348	0.00505

$$D(S, L) = -1 \log(0.74899) = 0.417347$$

that's our cross entropy for one training example, but hard to judge on one example. So we will look at more examples and take the average.

$$L = -\frac{1}{N} \sum_i D(S(WX_i + b), L_i)$$

This is the loss. which is the average cross entropy.