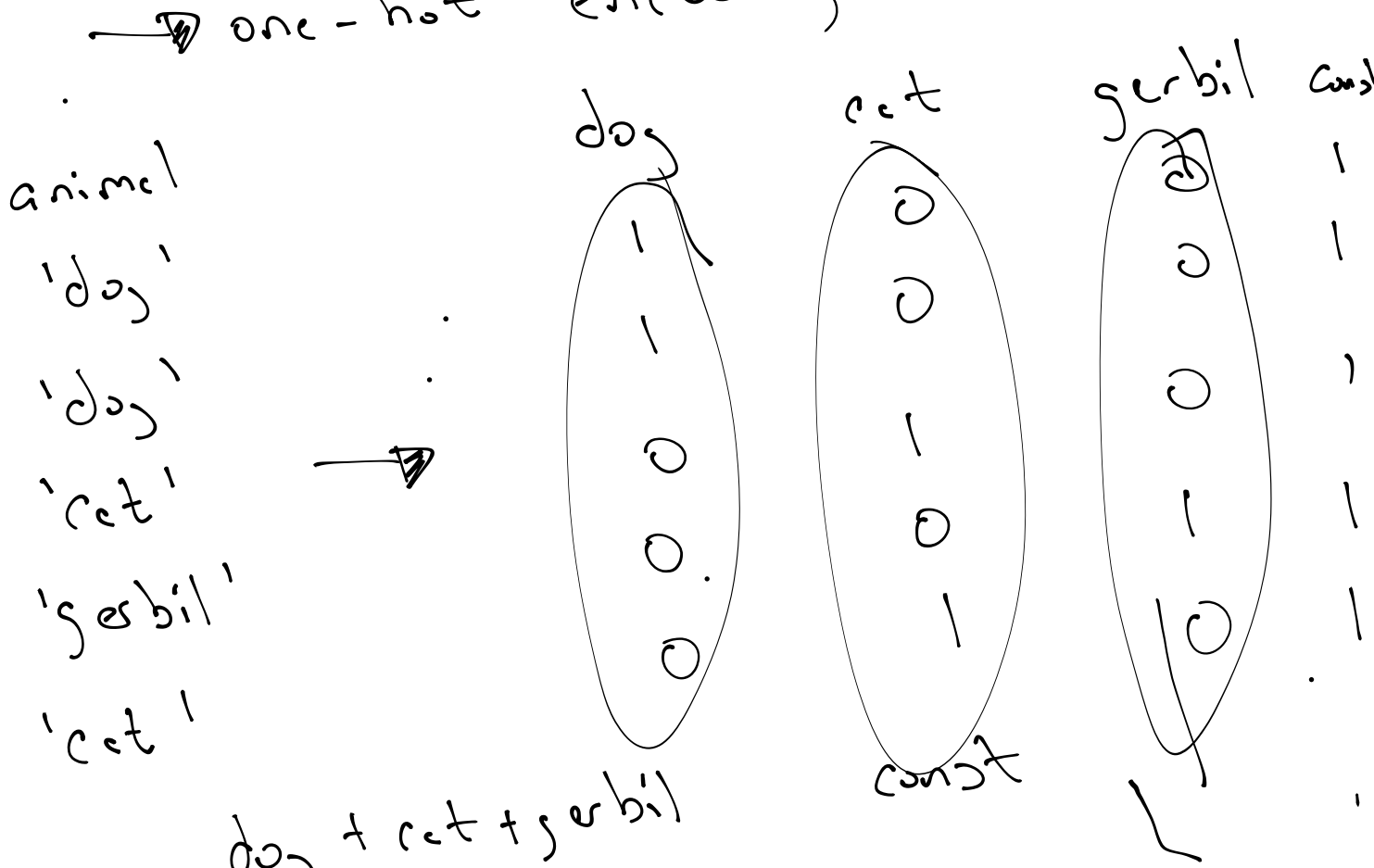
Continuous variables (features)

X continuous feature $\mathbb{R}$

categorical variable

'dog', 'cat', 'gerbil'

→ one-hot encoding



↳ columns aren't linearly independent

gender height weight

M	F	height
1	0	
1	0	
1	0	
1	0	
0	1	
0	1	
0	1	

ordinal variable

education → no HS 0
HS 1
some coll 2
college 3
grad degree 4

linear regression → categorical variables
one-hot encoding

tree-based models → handle ordinal variables

Classification

↳ binary classification

target data Y can be thought of as all 1s and 0s

↳ given some design matrix X,

I want to be able to predict the probability that $y_i = 1$ p_i

$$P(Y_i = 0) = 1 - p_i$$

machine learning: ① regression problems → continuous outcome

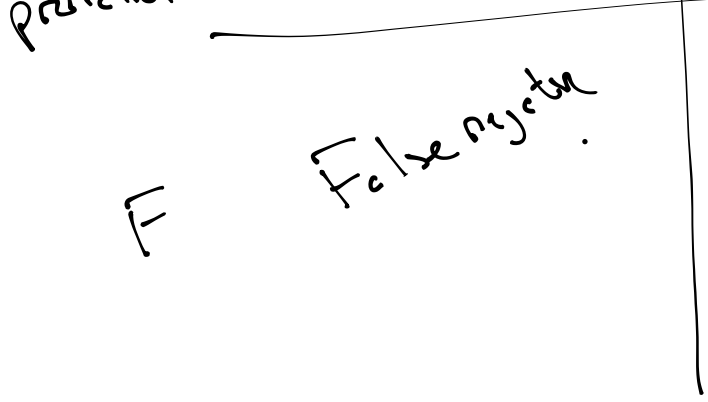
② classification → binary (categorical) outcome

evaluate regression:

$$\frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

evaluate classification:

	True value T	True value F
T	True positive	False positive



① Accuracy → th proportion of prediction that are correct

$$\frac{\#TP + \#TN}{N}$$

② Precision → if we predict positive, how often it is correct

$$\frac{\#TP}{\#\text{predicted positives}}$$

③ recall (sensitivity) → what percent of positives do we catch?

$$\frac{\#TP}{\#\text{positives}}$$

④ specificity = $\frac{\#TN}{\#\text{negatives}}$

Professor Dan hates all of these!

(5) Brier score

$$\sum_{i=1}^{n} (\hat{p}_i - y_i)^2 \longrightarrow \text{squared error with probabilities}$$

so of y_i's are 1

2.) of my data set is 1's

↳ largest predicted probability I would get

0.3

$\hat{y}_i = 0$