BINARY CLASSIFICATION & LOGISTIC REGRESSION

Narges Norouzi

BINARY CLASSIFICATION

APPLICATION

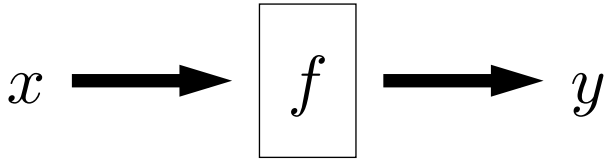
- Input: x = email message
- Output: $y \in \{spam, non-spam\}$
- Objective: obtain a predictor f

```
From:  pliang@cs.stanford.edu
Date:  September 26, 2018
Subject:  CS221 announcement

Hello students,
 I've attached the answers to homework 1...
```

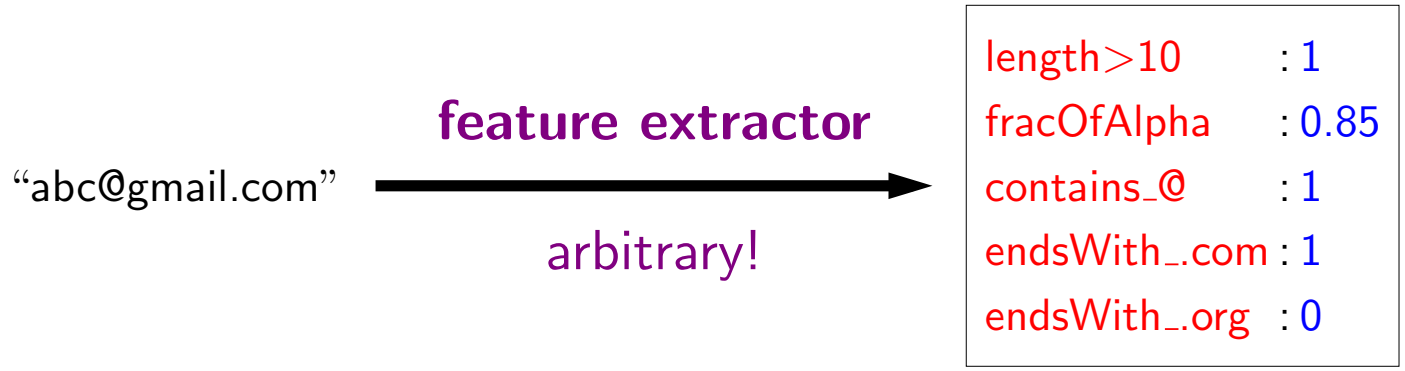
```
From:  a9k62n@hotmail.com
Date:  September 26, 2018
Subject:  URGENT

Dear Sir or maDam:
 my friend left sum of 10m dollars...
```



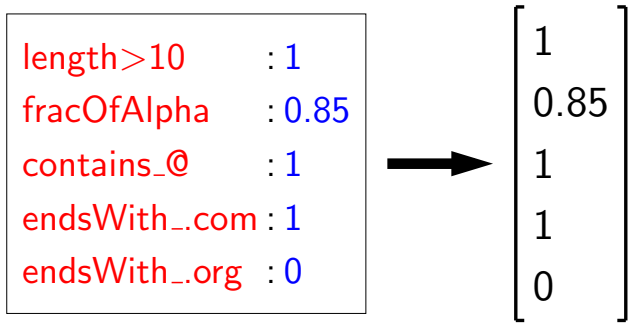
FEATURE EXTRACTION

- Example task: predict y knowing x (email address)
- Question: what properties of x might be relevant in predicting y?
- Feature Extractor: Given input x, output a set of (feature name, feature value) pairs.



FEATURE VECTOR NOTATION

- Mathematically, feature vector does not need feature names



Definition: Feature Vector

For an input x, its feature vector is

$$\phi(x) = [\phi_1(x), \ldots, \phi_d(x)]$$

Think of $\phi(x) \in \mathbb{R}^d$ as a point in the high-dimensional space.

WEIGHT VECTOR

- For each feature j, have real number θ_j representing contribution of feature to prediction

length>10	:-1.2
fracOfAlpha	:0.6
contains_@	:3
endsWith_.com	:2.2
endsWith_.org	:1.4
...	

LINEAR PREDICTORS

- Score or prediction: weighted combination of features

Weight vector $\theta \in \mathbb{R}^d$

length>10	:-1.2
fracOfAlpha	:0.6
contains_@	:3
endsWith_.com	:2.2
endsWith_.org	:1.4

Feature vector $\phi(x) \in \mathbb{R}^d$

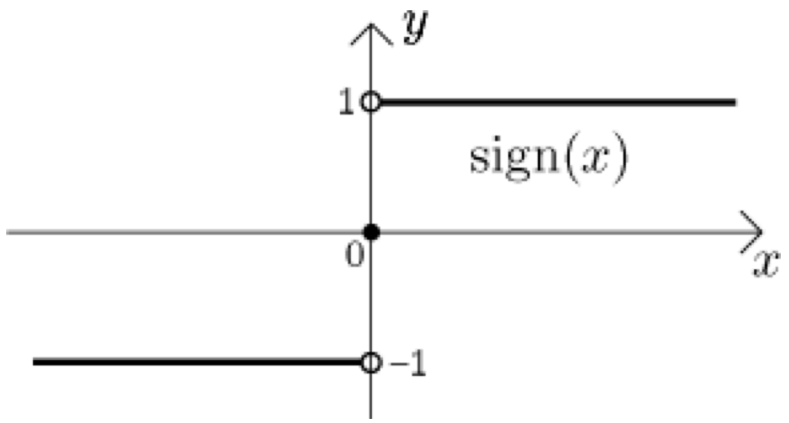
length>10	:1
fracOfAlpha	:0.85
contains_@	:1
endsWith_.com	:1
endsWith_.org	:0

$$\theta.\phi(x) = \sum_{j=0}^{d} \theta_j \phi_j(x) \qquad \phi_0(x) = 1$$

Example: $-1.2(1) + 0.6(0.85) + 3(1) + 2.2(1) + 1.4(0) = 4.51$

BINARY CLASSIFIER

$$f_{\mathbf{w}}(x) = \text{sign}(\mathbf{w} \cdot \phi(x)) = \begin{cases} +1 & \text{if } \mathbf{w} \cdot \phi(x) > 0 \\ -1 & \text{if } \mathbf{w} \cdot \phi(x) < 0 \\ ? & \text{if } \mathbf{w} \cdot \phi(x) = 0 \end{cases}$$



BINARY CLASSIFICATION
PROBABILISTIC APPROACH

BINARY OUTCOME VARIABLES

- This comes up often in science.
- We might want to predict whether:
  - Patients will live or die
  - Species will thrive or go extinct
  - Structures will fail or stay standing
- … based on their scores on a set of potential predictor variables.

FROM PROBABILITY TO ODDS

- Another way of thinking about probabilities is to transform them using the function:

$$odds = \frac{p}{1 - p}$$

- This is the probability of something happening divided by the probability of it not happening.
- Similarly, if we were told that the odds of an event E are x to y, then

$$odds(E) = \frac{x}{y}$$

Which means

$$p(E) = \qquad\qquad , p(E') =$$

FROM PROBABILITY TO ODDS

- Odds are commonly used in gambling.
  - "9 to 1 against", meaning a probability of 0.1.
  - "Even odds", meaning $p = 0.5$.
  - "3 to 1 on" meaning $p = 0.75$.
- What's the range of possible values for the odds ratio?

THE LOGIT FUNCTION

- With one more transformation we can get a value that is unbounded over the real numbers.
- This is the logit function, it takes a value between 0 to 1 and maps it to a value between $-\infty$ and $+\infty$:

$$y =$$

LOGISTIC FUNCTION

- Logit function:

$$z = \log_e\left(\frac{p}{1-p}\right)$$

$$e^z = \frac{p}{1-p}$$

- Logistic function (inverse logit function):

$$p =$$

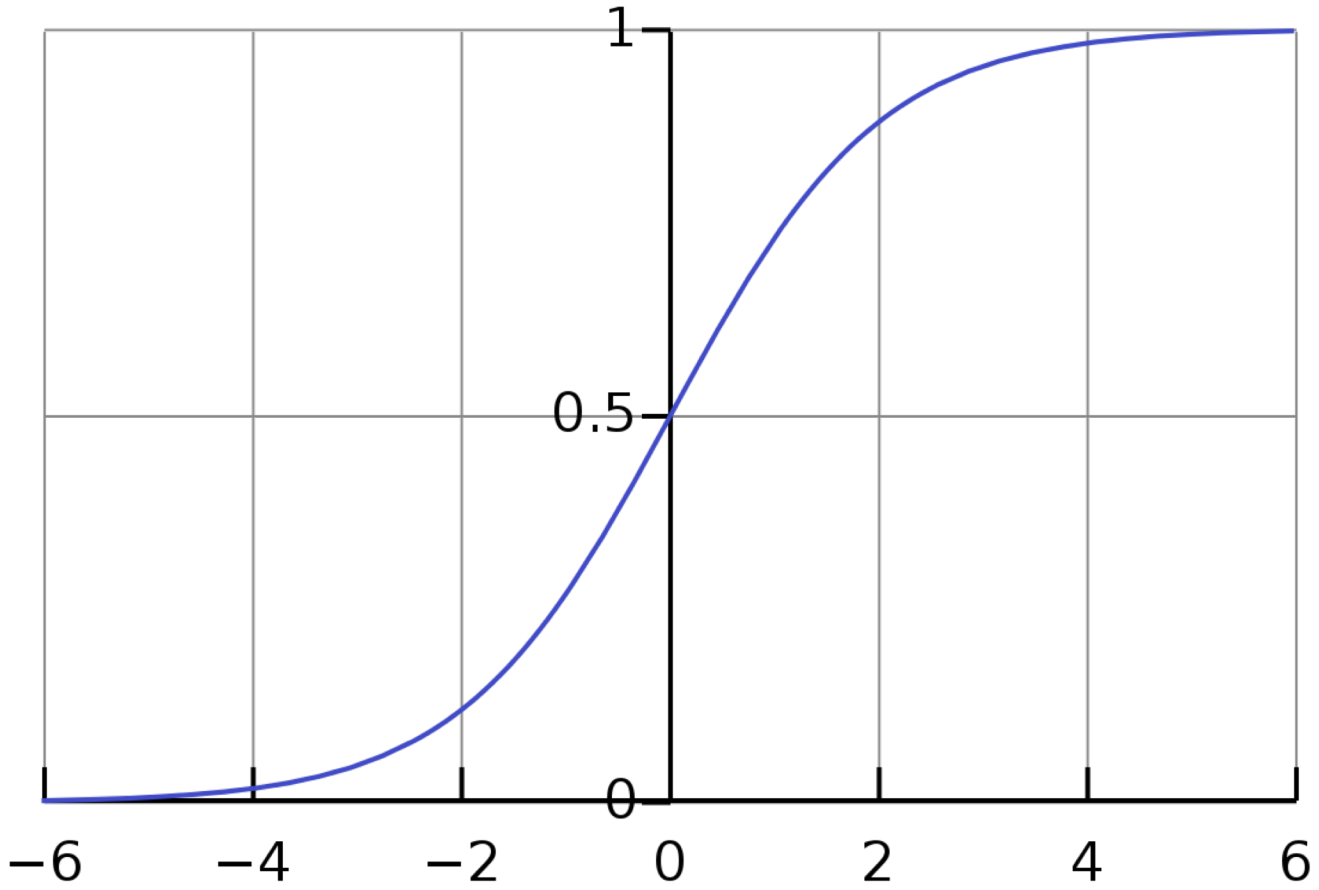
- The logistic function takes a value between $-\infty$ and $+\infty$ and maps it to a value between 0 and 1.

DIFFERENT WAYS OF EXPRESSING PROBABILITY

- Consider a two-outcome probability space, where:
  - $p(O_1) = p$
  - $p(O_2) = 1 - p = q$
- Can express probability O_1 as:

	notation	Range Equivalents		
Standard probability	p	0	0.5	1
Odds	$\frac{p}{q}$	0	1	$+\infty$
Log odds (logit)	$\log(\frac{p}{q})$	$-\infty$	0	$+\infty$

LOGISTIC FUNCTION OR SIGMOID



LOGISTIC REGRESSION

LOGISTIC REGRESSION

- Name is somewhat misleading. Really a technique for classification, not regression
  - "Regression" comes from fact that we fit a linear model to the feature space.
- Involves a more probabilistic view of classification.

USING A LOGISTIC REGRESSION MODEL

- Model consists of a vector θ in (d+1)-dimensional feature space
- For a point x in feature space, project it onto θ to convert it into a real number z in the range in the range $-\infty$ to $+\infty$

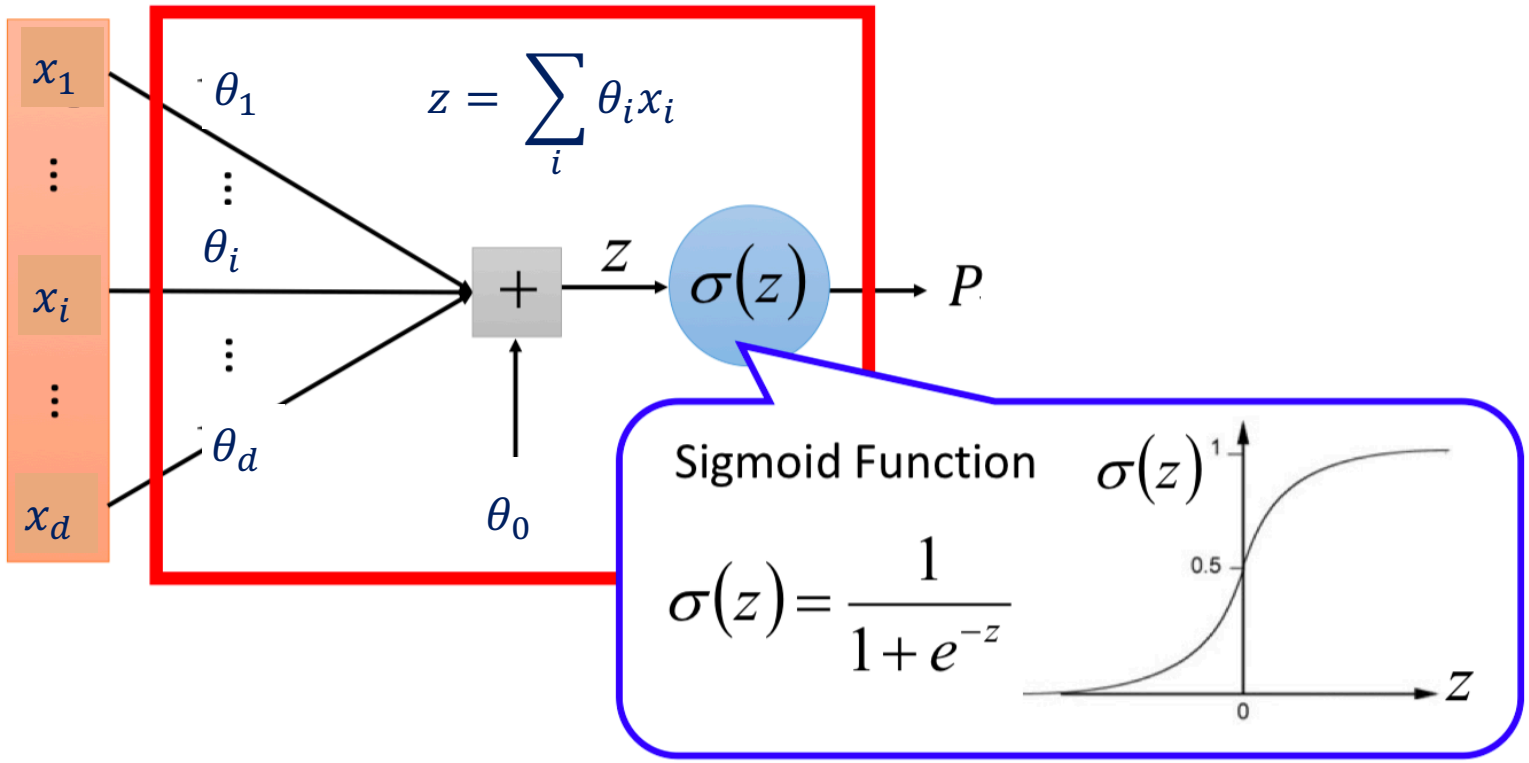
$$z = \theta_0 + \theta_1 x_1 + \ ... + \theta_d x_d$$
$$z = \theta . x = \theta^T x$$

- Map z to the range 0 to 1 using the logistic function (sigmoid function)

$$p = y(x) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

- Overall, logistic regression maps a point in d-dimensional space to a value in the range 0 to 1.

LOGISTIC FUNCTION

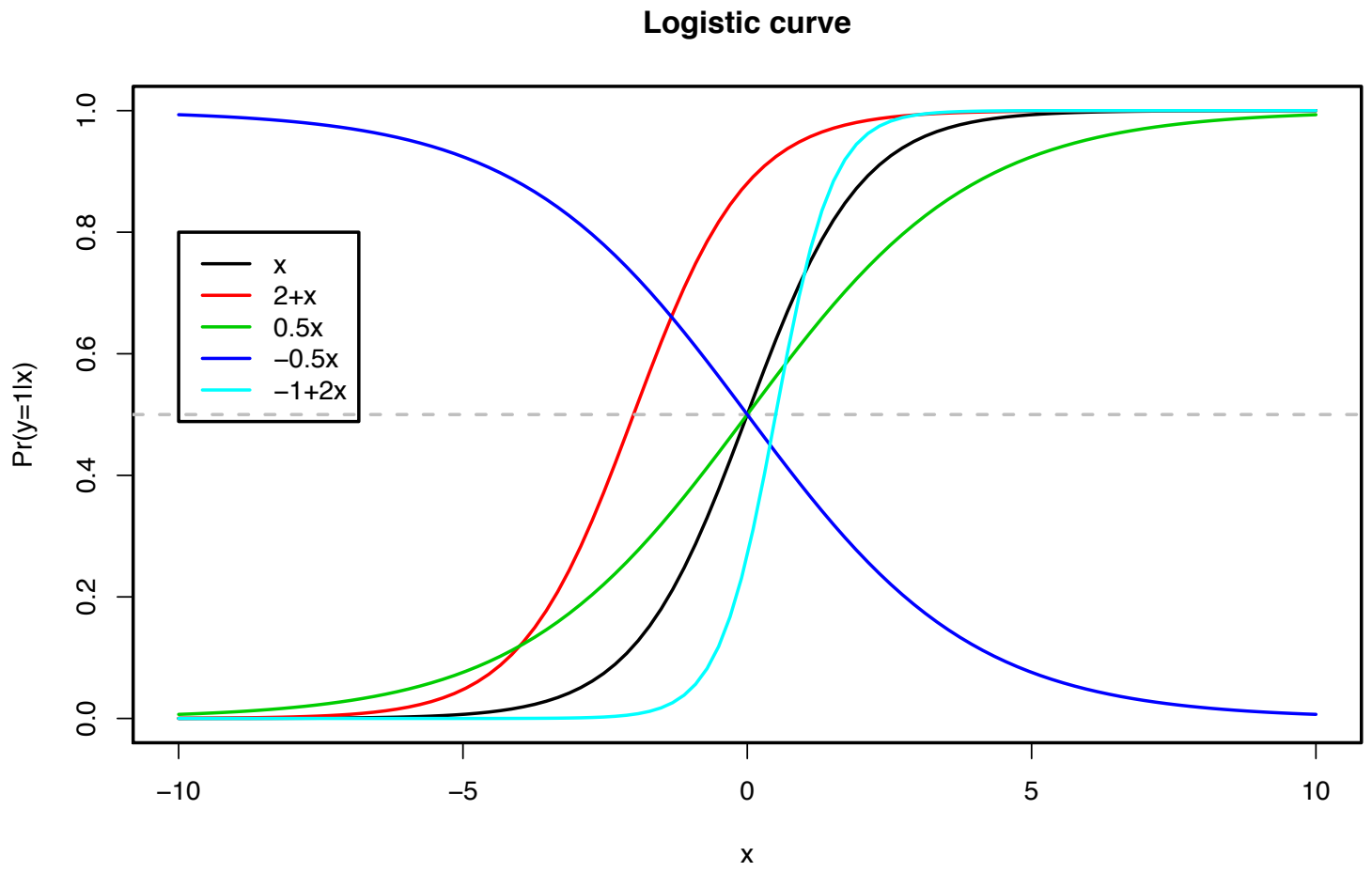


https://walkccc.github.io/CS/ML/5/

PROPERTIES OF LR

- One parameter per data dimension (feature) and a bias
- Features can be discrete or continuous
- Output of the model $y \in [0, 1]$
- Allows for gradient-based learning of parameters

SHAPE OF LOGISTIC FUNCTION



PROBABILISTIC INTERPRETATION

- If we have a value between 0 and 1, let's use it to model class probability:

$$p(C = 1|x) = \sigma(\theta^T x) \quad with\ \sigma(z) = \frac{1}{1 + e^{-z}}$$

- Substituting we have

$$p(C = 1|x) = \frac{1}{1 + e^{-\theta^T x}}$$

- Suppose we have two classes, how can I compute $p(C = 0|x)$?
- Use the marginalization property of probability

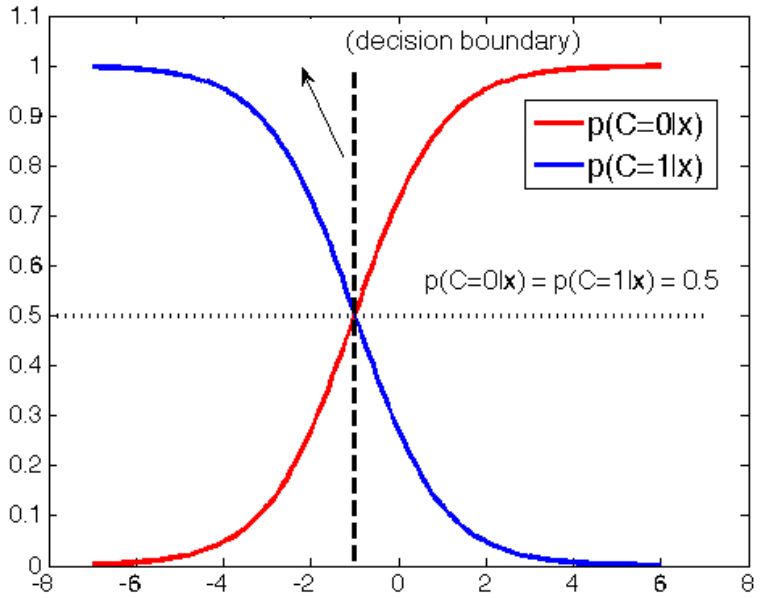
$$p(C = 0|x) + p(C = 1|x) = 1$$

- Thus

$$p(C = 0|x) = 1 - \frac{1}{1 + e^{-\theta^T x}} = \frac{e^{-\theta^T x}}{1 + e^{-\theta^T x}}$$

DECISION BOUNDARY FOR LR

- What is the *decision boundary* for logistic regression?
- $p(C = 1|x, \theta) = \frac{1}{1+e^{-\theta^T x}} = 0.5$
- $p(C = 0|x, \theta) = \frac{e^{-\theta^T x}}{1+e^{-\theta^T x}} = 0.5$
- Decision boundary: $\theta^T x = 0$
- Logistic regression has a linear decision boundary



QUESTIONS?