CMSC 473/673
Natural Language Processing

Instructor: Lara J. Martin (she/they)

TA: Duong Ta (he)

Slides modified from Dr. Frank Ferraro

Learning Objectives

Model classification problems using logistic regression

Define appropriate features for a logistic regression problem

Create a method to prepare data for a BoW model

Define an objective for LR modeling

Review: Classification Evaluation: the 2-by-2 contingency table

	What is the actual label?	
What label does our system predict? (↓)	Actual Target Class ("●")	Not Target Class ("○")
Selected/ Guessed ("●")	True Positive ● (TP) ● *Actual* *Guessed*	False Positive ○ (FP) ● *Actual* *Guessed*
Not selected/ not guessed ("○")	False Negative ● (FN) ○ *Actual* *Guessed*	True Negative ○ (TN) ○ *Actual* *Guessed*

Review: Classification Evaluation

Accuracy: % of items correct

$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

F-score: Weighted (harmonic) average of **P**recision & **R**ecall

$$F_1 = \frac{2 * P * R}{P + R} = \frac{2 * TP}{2 * TP + FP + FN}$$

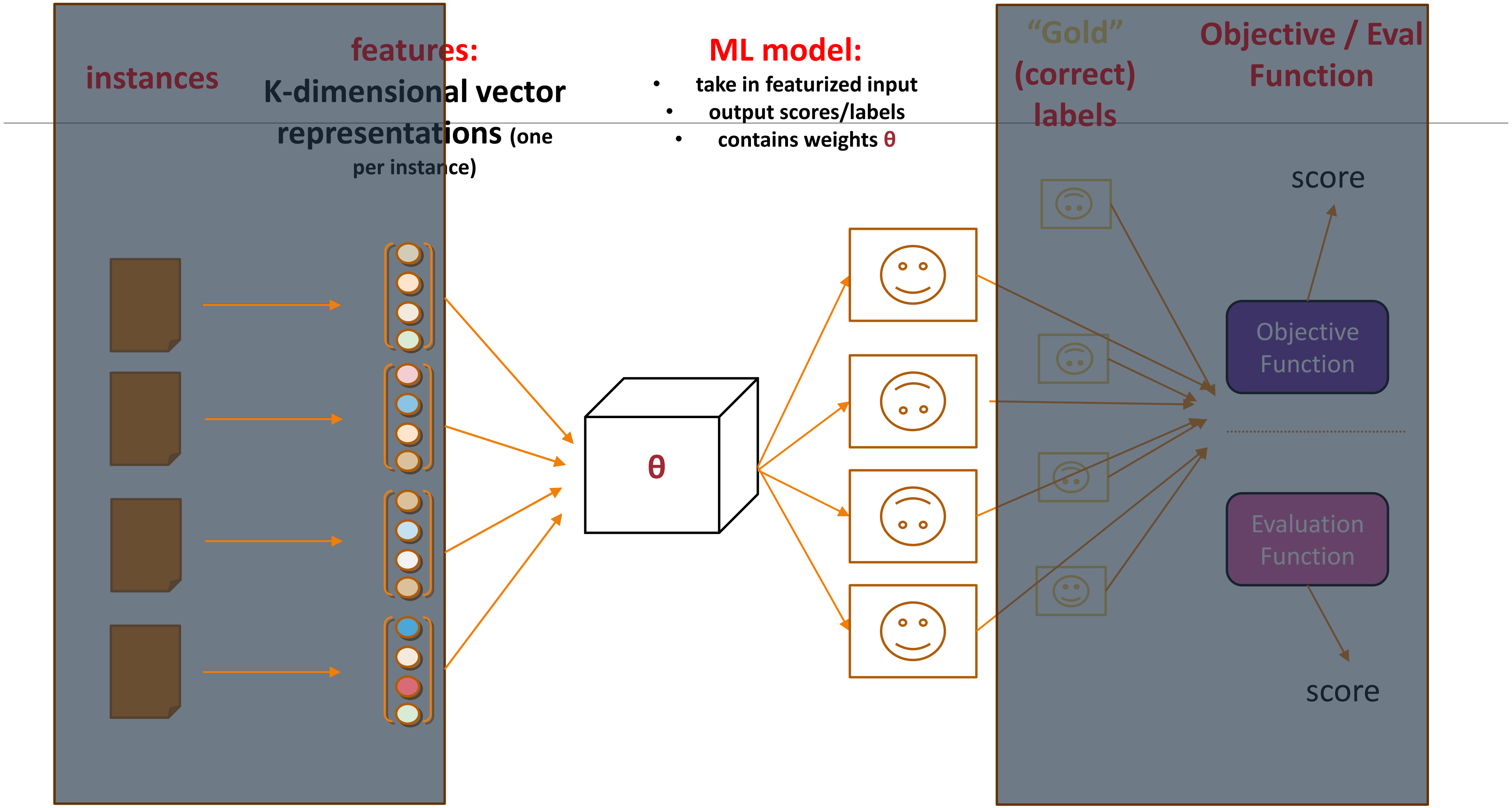
When would you want to use accuracy vs F1?

Accuracy works better if the dataset is <u>balanced</u>

Accuracy takes everything in consideration

F-Score is focused on TP

Defining the Model



Terminology

common NLP term	Log-Linear Models
as statistical regression	(Multinomial) logistic regression
	Softmax regression
based in information theory	Maximum Entropy models (MaxEnt)
a form of	Generalized Linear Models
viewed as	Discriminative Naïve Bayes
to be cool today :)	Very shallow (sigmoidal) neural nets

Maxent Models are Flexible

Maxent models can be used:

- to design discriminatively trained classifiers, or
- to create featureful language models

(among other approaches in NLP and ML more broadly)

Examining Assumption 3 Made for Classification Evaluation

Given X, our classifier produces a score for each possible label

p(●|X) vs. p(○|X)

Normally (*but this can be adjusted!)

$$\text{best label} = \arg\max_{\text{label}} P(\text{label}|\text{example})$$

Terminology: Posterior Probability

Posterior probability:

$$p(\bullet|X) \text{ vs. } p(\circ|X)$$

These are conditional probabilities

- If $\bullet$ and $\circ$ are the only two options:

$$p(\bullet|X) + p(\circ|X) = 1$$

and

$$p(\bullet|X) \geq 0, \; p(\circ|X) \geq 0$$

Bayes' Rule

$$\frac{\overbrace{P(X|Y)}^{\text{Likelihood}} \cdot \overbrace{P(Y)}^{\text{Prior}}}{P(X)}$$

Posterior: $P(Y|X)$

Posterior probability: probability of event Y with knowledge that X has occurred

NLP pg. 450

Terminology (with variables)

Posterior probability:

$$p(Y = label_1 \mid X) \text{ vs. } p(Y = label_0 \mid X)$$

Conditional probabilities:

$$p(Y = label_1 \mid X) + p(Y = label_0 \mid X) = 1$$

$$p(Y = label_1 \mid X) \geq 0,$$
$$p(Y = label_0 \mid X) \geq 0$$

Conditional probability: probability of event Y, assuming event X happens too

NLP pg. 449

💡 Key Take-away 💡

We will ***learn*** this

$$\boldsymbol{p(Y \mid X)}$$

Maxent Models for Classification: Discriminatively or …

Directly model the posterior

$$p(Y \mid X) = \mathbf{maxent}(X; Y)$$

Discriminatively trained classifier

Maxent Models for Classification: Discriminatively or Generatively Trained

Directly model the posterior

$$p(Y \mid X) = \mathbf{maxent}(X; Y)$$

Discriminatively trained classifier

Model the posterior with Bayes rule

$$p(Y \mid X) \propto \mathbf{maxent}(X \mid Y)p(Y)$$

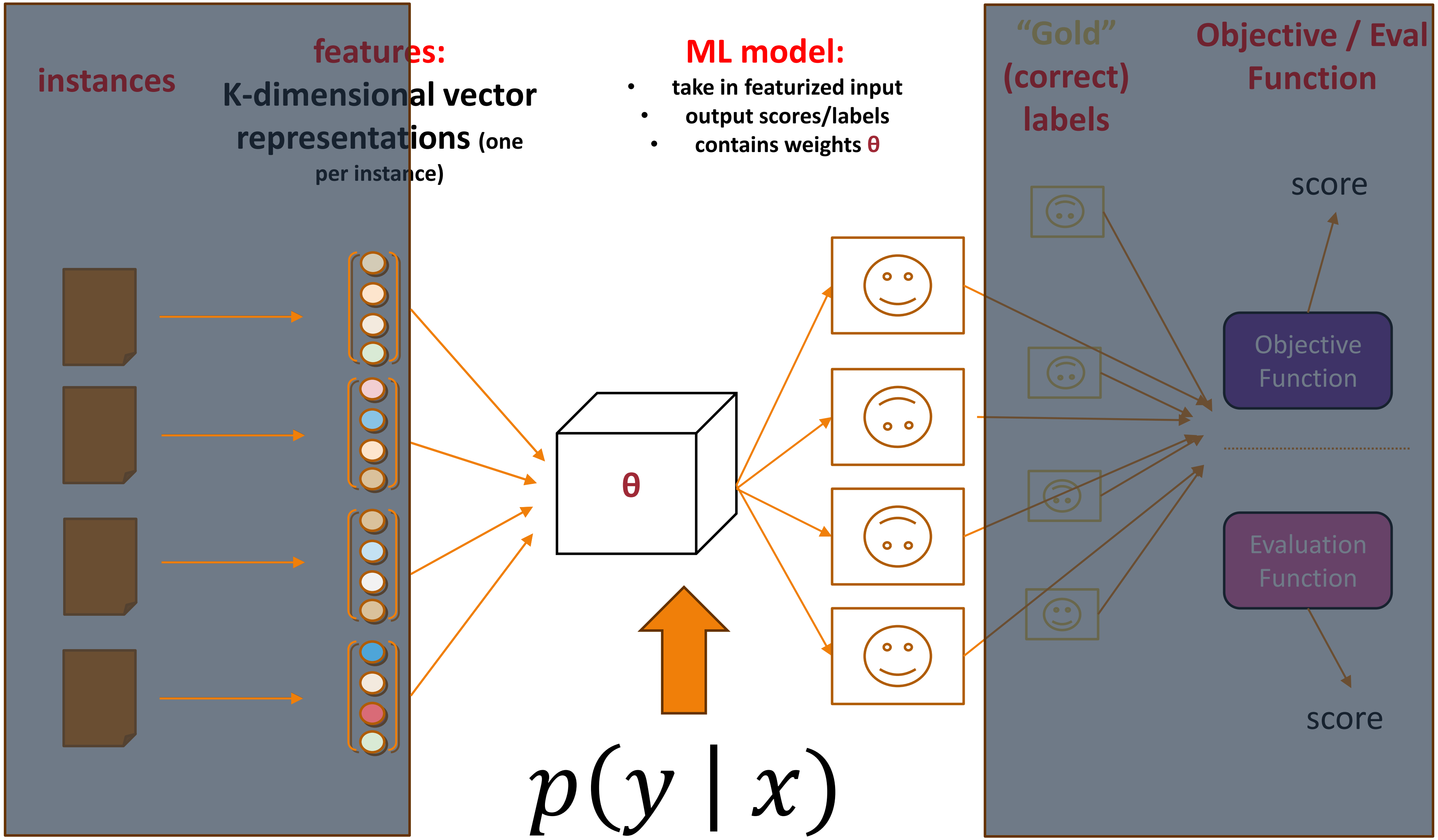
Generatively trained classifier with maxent-based language model

Maximum Entropy (Log-linear) Models For Discriminatively Trained Classifiers

(we'll start with this one)

$$p(y \mid x) = \text{maxent}(x, y)$$

discriminatively trained:
classify in one go



$$p(y \mid x) = \text{maxent}(x, y)$$

Core Aspects to Maxent Classifier p(y|x)

We need to define:

- **features** $f(x)$ from x that are meaningful;
- **weights** θ (at least one per feature, often one per feature/label combination) to say how important each feature is; and
- a way to **form probabilities** from f and θ

Overview of Featurization

Common goal: probabilistic classifier p(y | x)

Often done by defining **features** between x and y that are meaningful

- Denoted by a **general vector of K features**

$$f(x) = (f_1(x), \ldots, f_K(x))$$

Features can be thought of as "soft" rules

- **E.g., POSITIVE sentiments tweets *may* be more likely to have the word "happy"**

Review: Document Classification via Bag-of-Words Features (Example)

Electronic alerts have been used to assist the authorities in moments of chaos and potential danger: after the Boston bombing in 2013, when the Boston suspects were still at large, and last month in Los Angeles, during an active shooter scare at the airport.

TECH

NOT TECH

Core assumption: the label can be predicted from counts of individual word types

With V word types, define V feature functions $f_i(x)$ as

$f_i(x)$ =# of times word type *i* appears in document x

$$f(x) = \left(f_i(x)\right)_i^V$$

feature $f_i(x)$	value
alerts	1
assist	1
bombing	1
Boston	2
…	
sniffle	0
…	

Example Classification Tasks

GLUE
https://gluebenchmark.com/
🤗datasets: glue

GLUE Tasks

Name	Download
The Corpus of Linguistic Acceptability	
The Stanford Sentiment Treebank	
Microsoft Research Paraphrase Corpus	
Semantic Textual Similarity Benchmark	
Quora Question Pairs	
MultiNLI Matched	
MultiNLI Mismatched	
Question NLI	
Recognizing Textual Entailment	
Winograd NLI	
Diagnostics Main	

SuperGLUE T

Name	Identifier
Broadcoverage Diagnostics	AX-b
CommitmentBank	CB
Choice of Plausible Alternatives	COPA
Multi-Sentence Reading Comprehension	MultiRC
Recognizing Textual Entailment	RTE
Words in Context	WiC
The Winograd Schema Challenge	WSC
BoolQ	BoolQ
Reading Comprehension with Commonsense Reasoning	ReCoRD
Winogender Schema Diagnostics	AX-g

SuperGLUE
https://super.gluebenchmark.com/
🤗datasets: super_glue

Recognizing Textual Entailment (RTE)

Given a premise sentence s and hypothesis sentence h, determine if h "follows from" s

ENTAILMENT (yes):

NOT ENTAILED (no):

Recognizing Textual Entailment (RTE)

Given a premise sentence s and hypothesis sentence h, determine if h "follows from" s

ENTAILMENT (yes):

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.

h: The Bulls basketball team is based in Chicago.

NOT ENTAILED (no):

Recognizing Textual Entailment (RTE)

Given a premise sentence s and hypothesis sentence h, determine if h “follows from” s

ENTAILMENT (yes):

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.

h: The Bulls basketball team is based in Chicago.

NOT ENTAILED (no):

s: Based on a worldwide study of smoking-related fire and disaster data, UC Davis epidemiologists show smoking is a leading cause of fires and death from fires globally.

h: Domestic fires are the major cause of fire death.

RTE

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago.

ENTAILED

p(ENTAILED | s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships. h: The Bulls basketball team is based in Chicago.)

Discriminative Document Classification

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.

h: The Bulls basketball team is based in Chicago.

ENTAILED

Discriminative Document Classification

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.

h: The Bulls basketball team is based in Chicago.

ENTAILED

These extractions are all **features** that have **fired** (likely have some significance)

Discriminative Document Classification

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.

h: The Bulls basketball team is based in Chicago.

ENTAILED

These extractions are all **features** that have **fired** (likely have some significance)

Discriminative Document Classification

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.

h: The Bulls basketball team is based in Chicago.

ENTAILED

These extractions are all **features** that have **fired** (likely have some significance)

We need to *score* the different extracted clues.

extract_and_score$_{\text{Bulls, entailed}}$(🗎) **ENTAILED**

Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.

extract_and_score$_{\text{basketball, entailed}}$(🗎, ENTAILED)

h: The Bulls basketball team is based in Chicago.

extract_and_score$_{\text{Chicago, entailed}}$(🗎, ENTAILED)

Score and Combine Our Clues

$\text{score}_{1,\ \text{Entailed}}($🗎$)$

$\text{score}_{2\ ,\ \text{Entailed}}($🗎$)$

$\text{score}_{3\ ,\ \text{Entailed}}($🗎$)$

…

$\text{score}_{k\ ,\ \text{Entailed}}($🗎$)$

…

COMBINE

posterior probability of ENTAILED

Scoring Our Clues

score(s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago. , ENTAILED) =

(ignore the feature indexing for now)

$\text{score}_{1\text{, Entailed}}$(🗎) +

$\text{score}_{2\text{, Entailed}}$(🗎) +

$\text{score}_{3\text{, Entailed}}$(🗎) +

…

Turning Scores into Probabilities

score(s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago. , ENTAILED) > score(s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago. , NOT ENTAILED)

p(ENTAILED | s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago.) > p(NOT ENTAILED | s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago.)

KEY IDEA

Turning Scores into Probabilities (More Generally)

$$\text{score}(x, y_1) > \text{score}(x, y_2)$$

$$\Downarrow$$

$$p(y_1 | x) > p(y_2 | x)$$

KEY IDEA

Maxent Modeling

$$\underbrace{p\left(\text{ENTAILED} \,\middle|\, \begin{array}{l}\text{s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took}\\ \text{the Chicago Bulls to six National Basketball Association championships.}\\ \text{h: The Bulls basketball team is based in Chicago.}\end{array}\right)}_{\text{This must be a probability}} \propto$$

This could be any real number

Convert through function G? What is this function?

$$\overbrace{G}\left(\overbrace{\text{score}\left(\begin{array}{l}\text{s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the}\\ \text{Chicago Bulls to six National Basketball Association championships.}\\ \text{h: The Bulls basketball team is based in Chicago.}\end{array}, \text{ENTAILED}\right)}\right)$$

What function G…

operates on any real number?

is never less than 0?

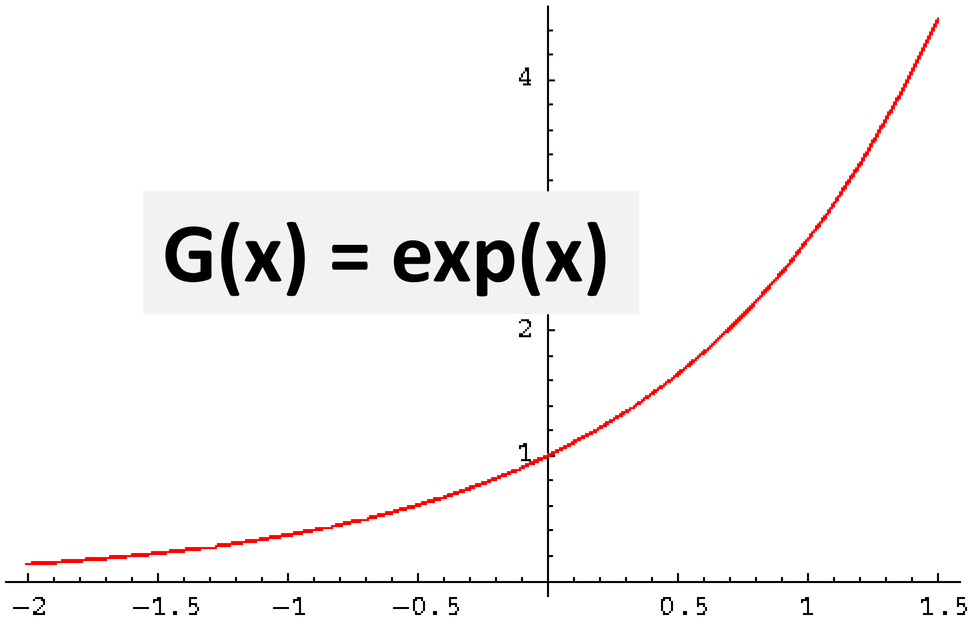
monotonic? ($a < b \Rightarrow G(a) < G(b)$)

What function G…

operates on any real number?

is never less than 0?

monotonic? (a < b ➔ G(a) < G(b))



Maxent Modeling

$p($ ENTAILED $|$ s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago. $) \propto$

$\exp(\mathrm{score}($ s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago. , ENTAILED $))$

Maxent Modeling

$$p(\textsc{Entailed} \mid \text{ s, h }) \propto$$

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago.

$$\exp\Big(\text{score}_{1,\,\text{Entailed}}(\,) + \text{score}_{2,\,\text{Entailed}}(\,) + \text{score}_{3,\,\text{Entailed}}(\,) + \ldots \Big)\Big)$$

Maxent Modeling

$p($ ENTAILED $|$

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago.

$) \propto$

$$\exp\Big(\text{weight}_{1,\,\text{Entailed}} * \text{applies}_1(🗎) + \text{weight}_{2,\,\text{Entailed}} * \text{applies}_2(🗎) + \text{weight}_{3,\,\text{Entailed}} * \text{applies}_3(🗎) + \ldots \Big)\Big)$$

Maxent Modeling

$p($ ENTAILED $|$ s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships. h: The Bulls basketball team is based in Chicago. $) \propto$

$\exp($ $\text{weight}_{1,\text{ Entailed}} * \text{applies}_1($🖹$) +$
$\text{weight}_{2,\text{ Entailed}} * \text{applies}_2($🖹$) +$
$\text{weight}_{3,\text{ Entailed}} * \text{applies}_3($🖹$) +$
…
$))$

K different weights… for K different features

Maxent Modeling

p(ENTAILED | s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships. h: The Bulls basketball team is based in Chicago.) ∝

exp(weight$_{1,\ \text{Entailed}}$ * applies$_1$(🖹) + weight$_{2,\ \text{Entailed}}$ * applies$_2$(🖹) + weight$_{3,\ \text{Entailed}}$ * applies$_3$(🖹) + …))

K different weights…

for K different features

multiplied and then summed

Maxent Modeling

p(ENTAILED | s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago.) $\propto$

exp(Dot_product of Entailed weight_vec feature_vec(🖹))

K different weights…

for K different features

multiplied and then summed

Maxent Modeling

$$p(\textsc{Entailed} \mid$$

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago.

$$) \propto$$

$$\exp\left(\theta_{\text{ENTAILED}}^{T} f(\;$$  $$\;)\right)$$

K different weights…

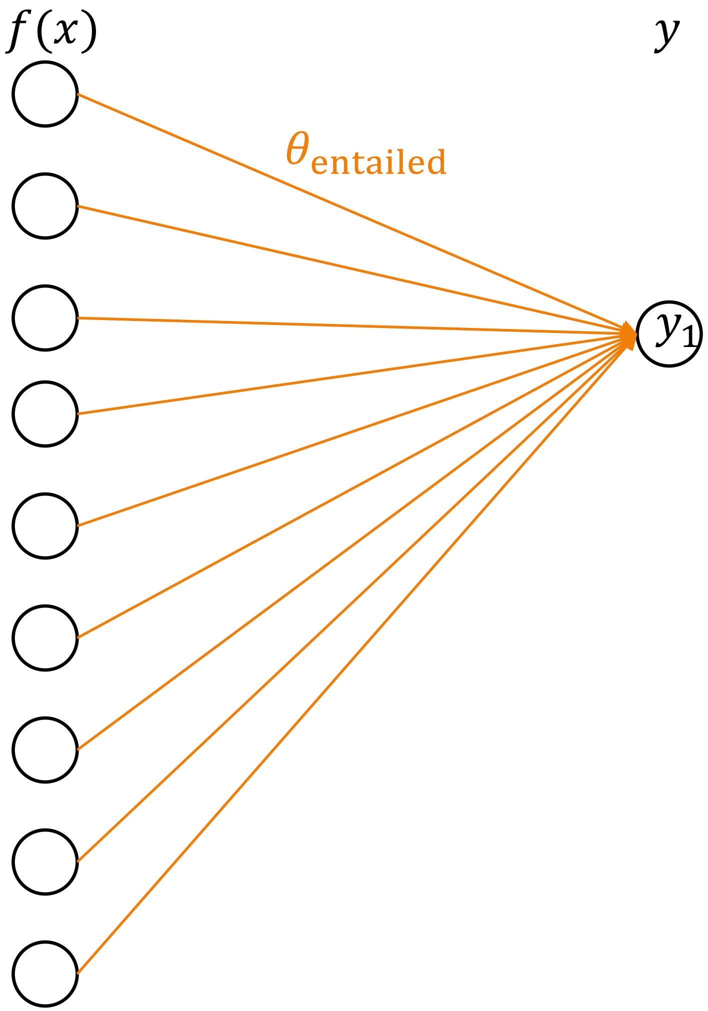
for K different features

multiplied and then summed

Knowledge Check: Data Prep

https://colab.research.google.com/drive/19yg0EUXQtHozBiSuO6cKOBhoSPzQHgug?usp=sharing

Maxent Classifier, schematically



$$p(y = \text{entailed}|x) \propto \exp(\theta_{\text{entailed}} f(x))$$

Maxent Modeling

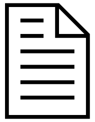
s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.
h: The Bulls basketball team is based in Chicago.

$$p(\text{ENTAILED} \mid \ldots) = \frac{1}{Z}\exp(\theta^{T}_{\text{ENTAILED}} f(\ldots))$$

Q: How do we define Z?

K different weights… for K different features… multiplied and then

Normalization for Classification

$$Z = \sum_{\text{label } j} \exp(\theta_j^T f($$  $$))$$

$p(y \mid x) \propto \exp(\theta_y^T f(x))$ *classify doc x with label y in one go*

Normalization for Classification (long form)

$$Z =$$

$$\sum_{\text{label } j} \exp\Big($$

weight$_{1,\, j}$ $*$ applies$_1$(🗎) +

weight$_{2,\, j}$ $*$ applies$_2$(🗎) +

weight$_{3,\, j}$ $*$ applies$_3$(🗎) +

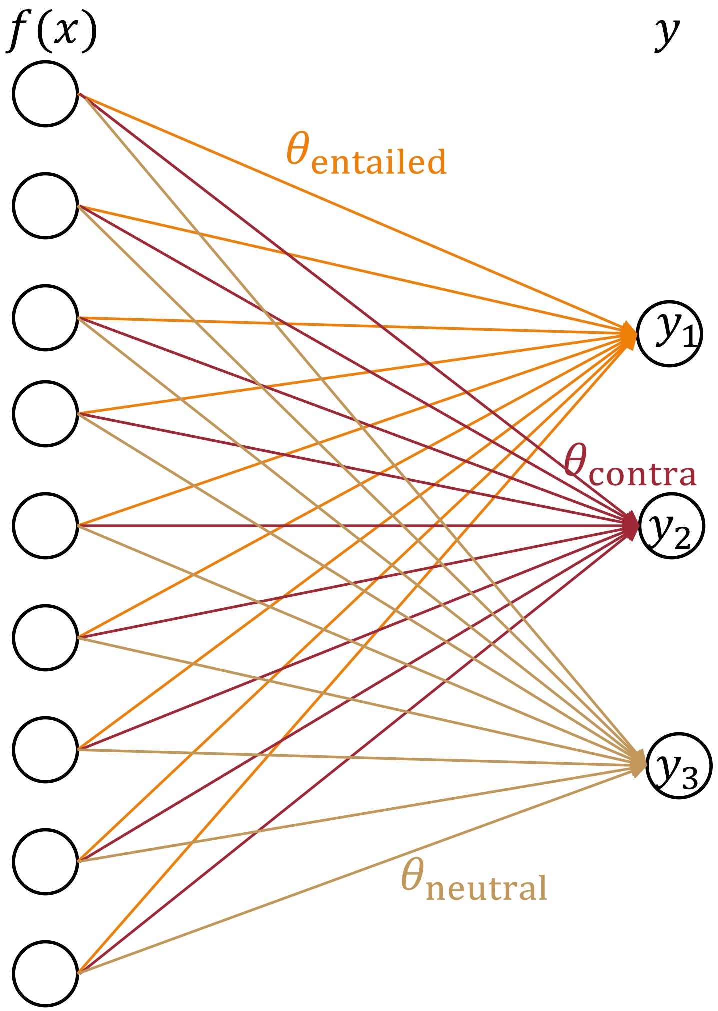
…

$\Big)$

$$p(y \mid x) \propto \exp(\theta_y^T f(x))$$

classify doc x with label y in one go

Maxent Classifier, schematically



Why would we want to normalize the weights?

$p(y = \text{entailed}|x) \propto \exp(\theta_{\text{entailed}} f(x))$

$p(y = \text{contra}|x) \propto \exp(\theta_{\text{contra}} f(x))$

$p(y = \text{neutral}|x) \propto \exp(\theta_{\text{neutral}} f(x))$

output:
i = argmax score_i
class *i*

Core Aspects to Maxent Classifier p(y|x)

features $f(x)$ from x that are meaningful;

weights θ (at least one per feature, often one per feature/label combination) to say how important each feature is; and

a way to **form probabilities** from f and θ

$$p(y \mid x) = \frac{\exp(\theta_y^T f(x))}{\sum_{y'} \exp(\theta_{y'}^T f(x))}$$

Different Notation, Same Meaning

$$p(Y = y \mid x) = \frac{\exp(\theta_y^T f(x))}{\sum_{y'} \exp(\theta_{y'}^T f(x))}$$

Different Notation, Same Meaning

$$p(Y = y \mid x) = \frac{\exp(\theta_y^T f(x))}{\sum_{y'} \exp(\theta_{y'}^T f(x))}$$

$$p(Y = y \mid x) \propto \exp(\theta_y^T f(x))$$

Different Notation, Same Meaning

$$p(Y = y \,|x) = \frac{\exp(\theta_y^T f(x))}{\sum_{y'} \exp(\theta_{y'}^T f(x))}$$

$$p(Y = y \mid x) \propto \exp(\theta_y^T f(x))$$

$$p(Y \mid x) = \text{softmax}(\theta f(x))$$

Defining Appropriate Features in a Maxent Model

Feature functions help extract useful features (characteristics) of the data

They turn *data* into *numbers*

Features that are not 0 are said to have fired

Binary-valued (0 or 1) or real-valued

Representing a Linguistic "Blob"

User-defined	Integer representation/one-hot encoding	Assign each word to some index i, where $0 \leq i < V$ Represent each word w with a V-dimensional **binary** vector e_w, where $e_{w,i} = 1$ and 0 otherwise
Model-produced	Dense embedding	Let E be some *embedding size* (often 100, 200, 300, etc.) Represent each word w with an E-dimensional **real-valued** vector e_w

Featurization is Similar but…

Vocab types (V) / embedding dimension (E) ➔ number of features (number of "clues")

"Linguistic blob" ➔ Instances to represent

Features are extracted on each instance

Review: Bag-of-words as a Function

Based on some tokenization, turn an input document into an array (or dictionary or set) of its unique vocab items

Think of getting a BOW rep. as a function f

input: Document

output: Container of size E, indexable by each vocab type v

Some Bag-of-words Functions

Kind	Type of f_v	Interpretation
Binary	0, 1	Did *v* appear in the document?
Count-based	Natural number (int >= 0)	How often did *v* occur in the document?
Averaged	Real number (>=0, <= 1)	How often did *v* occur in the document, normalized by doc length?
TF-IDF (term frequency, inverse document frequency)	Real number (>= 0)	How frequent is a word, tempered by how prevalent it is across the corpus (to be covered later!)
…		