Logistic Regression Models

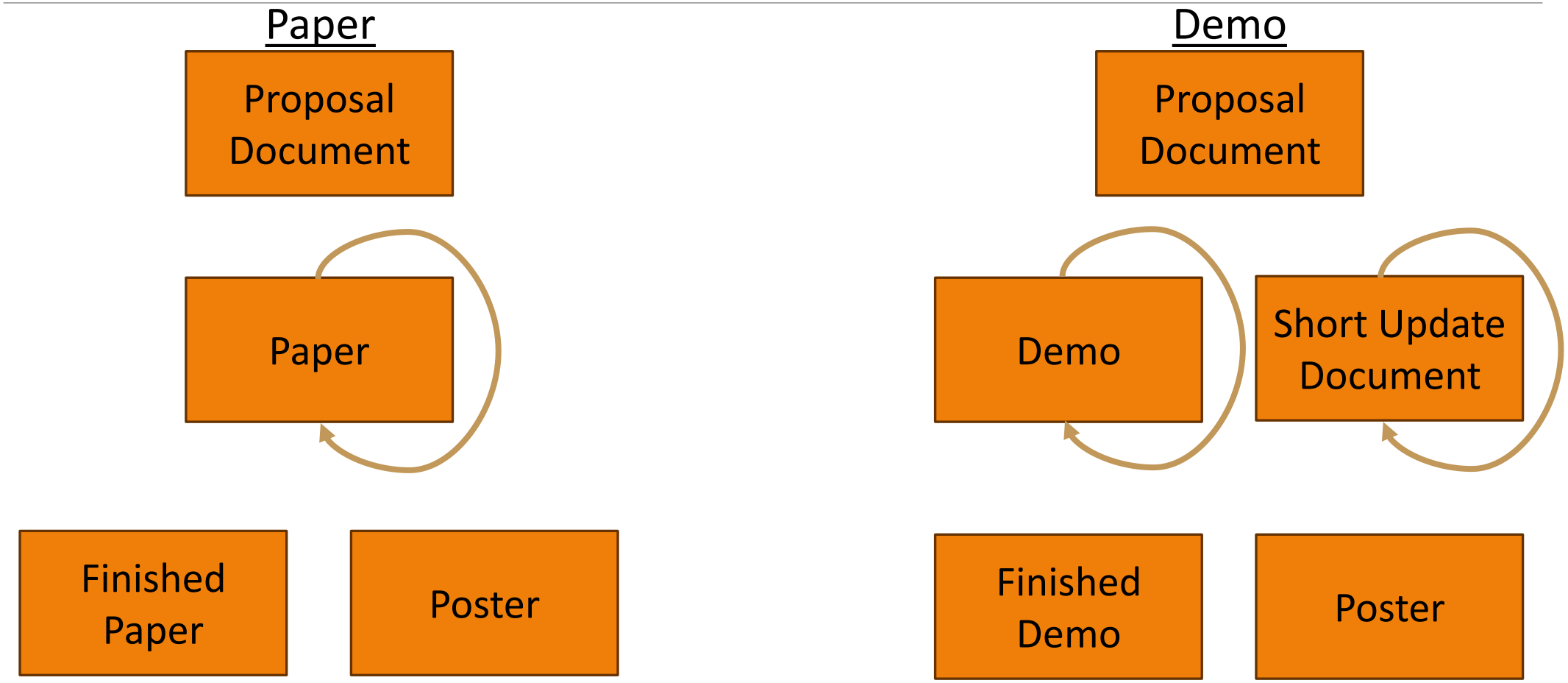
Instructor: Lara J. Martin (she/they)

TA: Omkar Kulkarni (he)

https://laramartin.net/NLP-class/

Slides modified from Dr. Frank Ferraro

Project Flow



Learning Objectives

Model classification problems using logistic regression

Define appropriate features for a logistic regression problem

Review: F1 (or F-score)

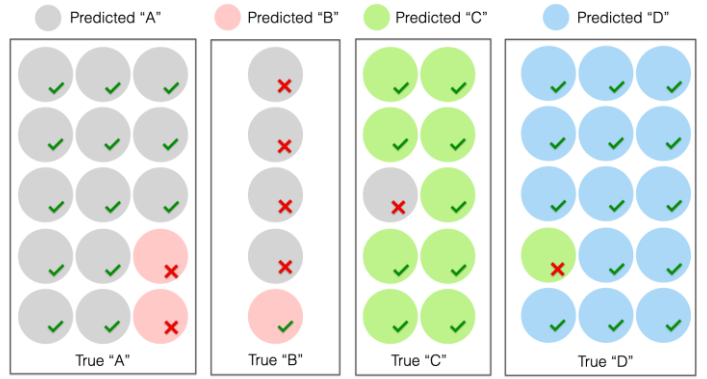
Weighted (harmonic) average of **P**recision & **R**ecall

F1 measure: equal weighting between precision and recall

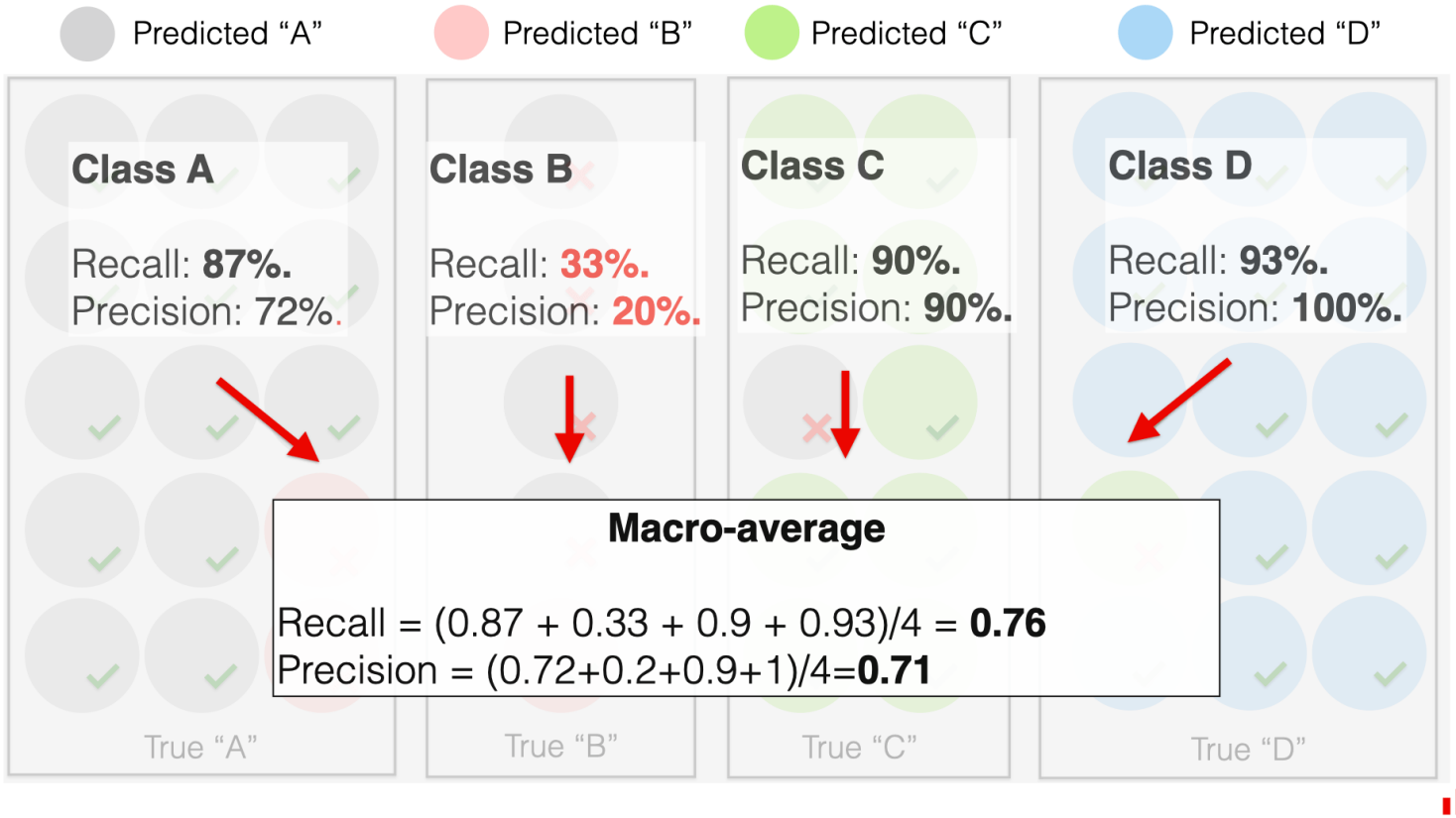
$$F_1 = \frac{2 * P * R}{P + R} = \frac{2 * TP}{2 * TP + FP + FN}$$

(useful when $P = R = 0$)

Each *class* has equal weight



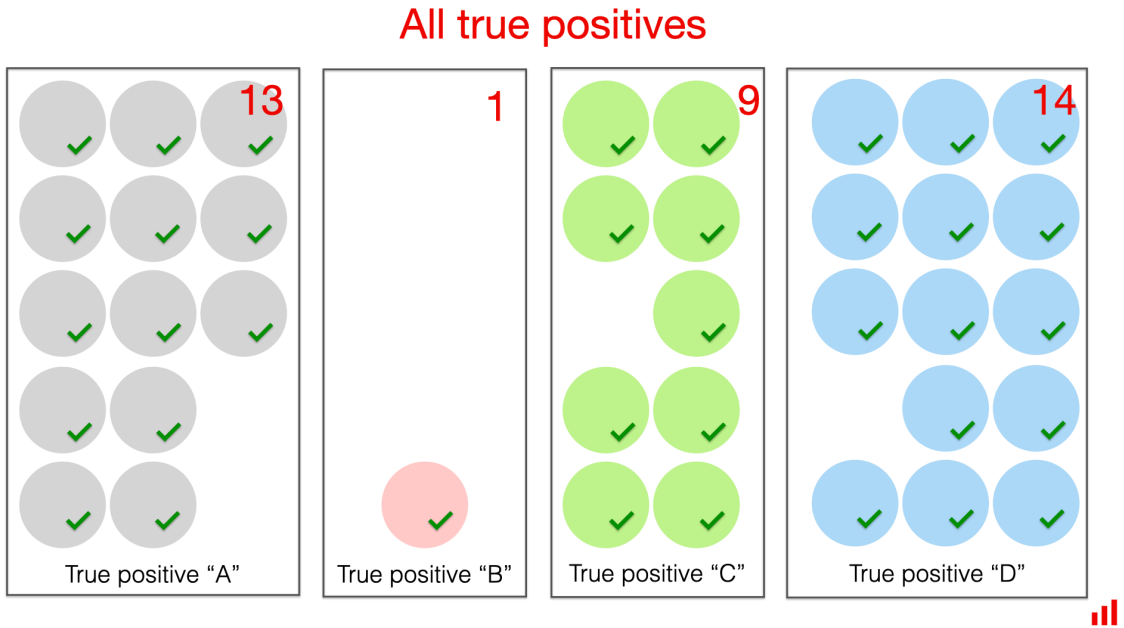
Review: Macro-Average

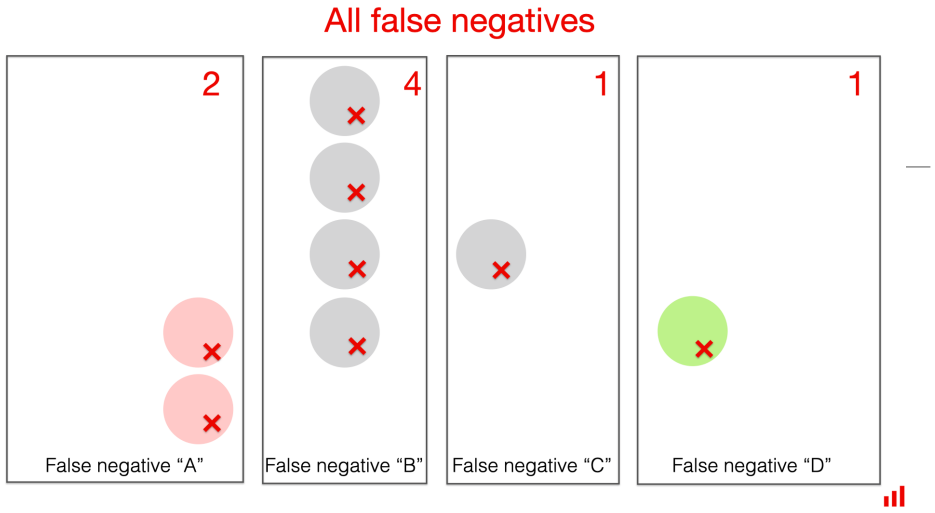


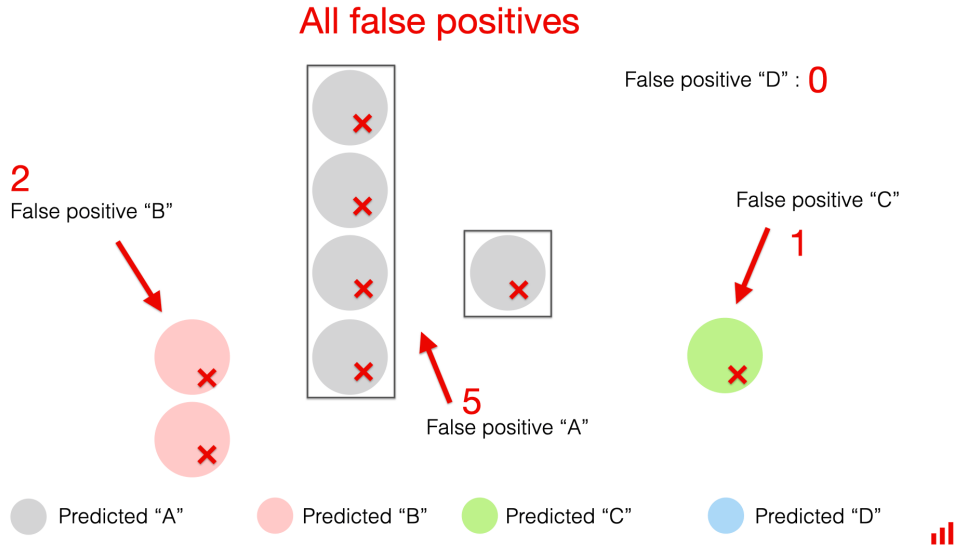
https://www.evidentlyai.com/classification-metrics/multi-class-metrics

Each *instance* has equal weight

Review: Micro-Average







Total TP | Total FP | Total FN

$$\text{Precision}_{\text{Micro-average}} = \frac{13 + 1 + 9 + 14}{13 + 1 + 9 + 14 + \boxed{2 + 5 + 1 + 0}} = 0.82$$

$$\text{Recall}_{\text{Micro-average}} = \frac{13 + 1 + 9 + 14}{13 + 1 + 9 + 14 + 2 + 4 + 1 + 1} = 0.82$$

https://www.evidentlyai.com/classification-metrics/multi-class-metrics

Types of Classification Metrics

AUC https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html

F-score https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Accuracy https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

Confusion matrix https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

Precision (can specify macro/micro average) https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html

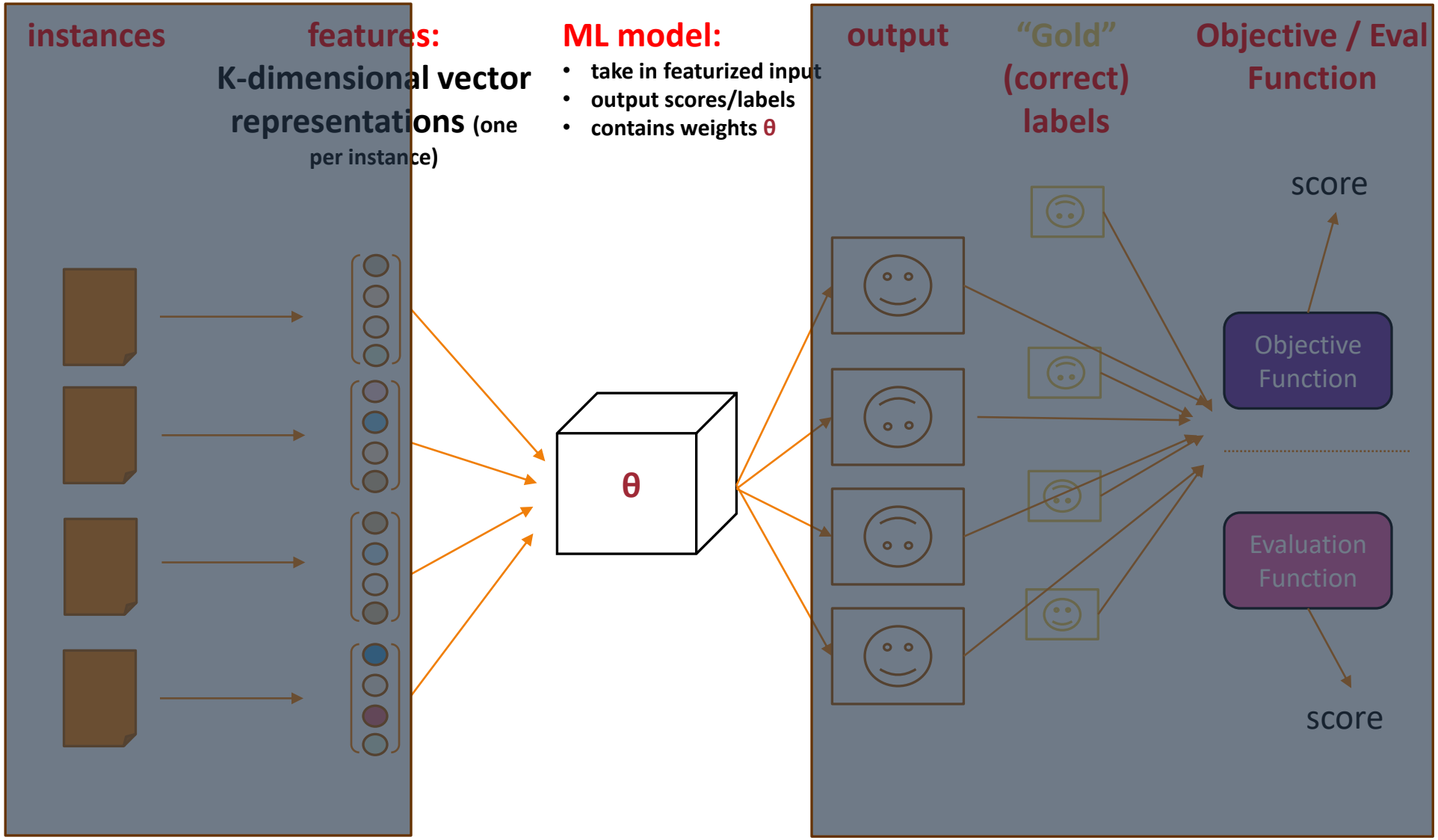
Recall (can specify macro/micro average) https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html

Outline

Maximum Entropy classifiers

- Defining the model
- Defining the objective
- Learning: Optimizing the objective

Defining the Model



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

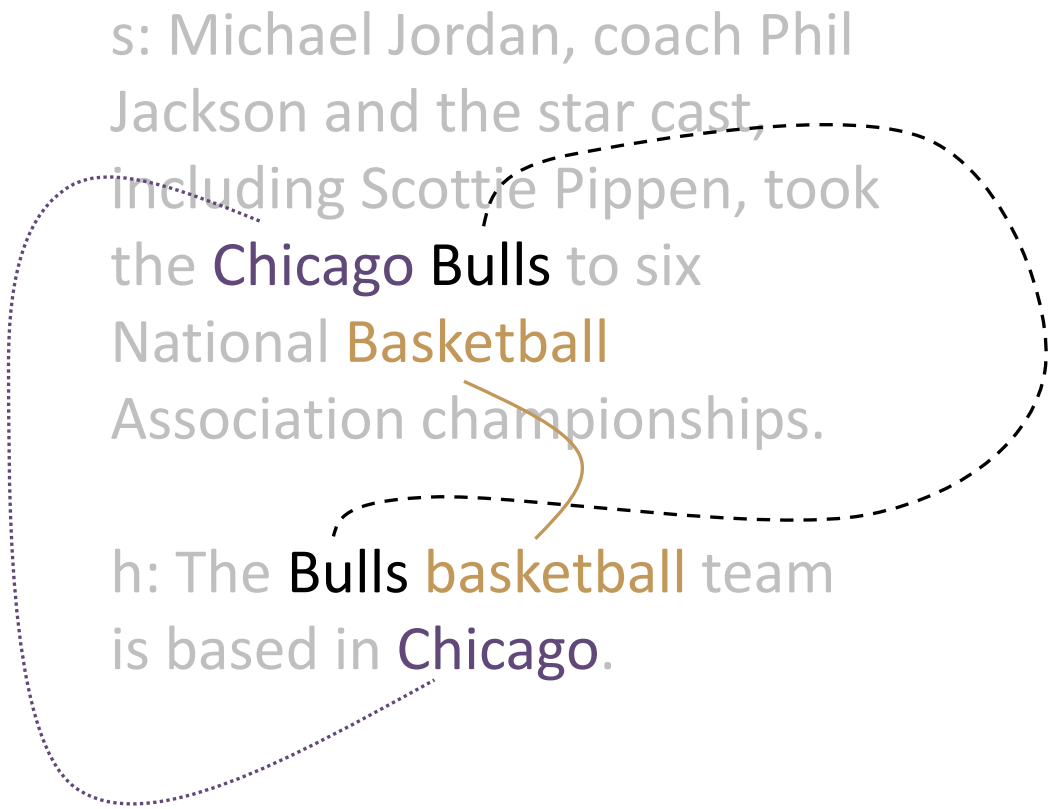
$$p(Y \mid X) \propto \mathbf{maxent}(X \mid Y)p(Y)$$

Generatively trained classifier with maxent-based language model

Review: Discriminative Model using Document Classification Example

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

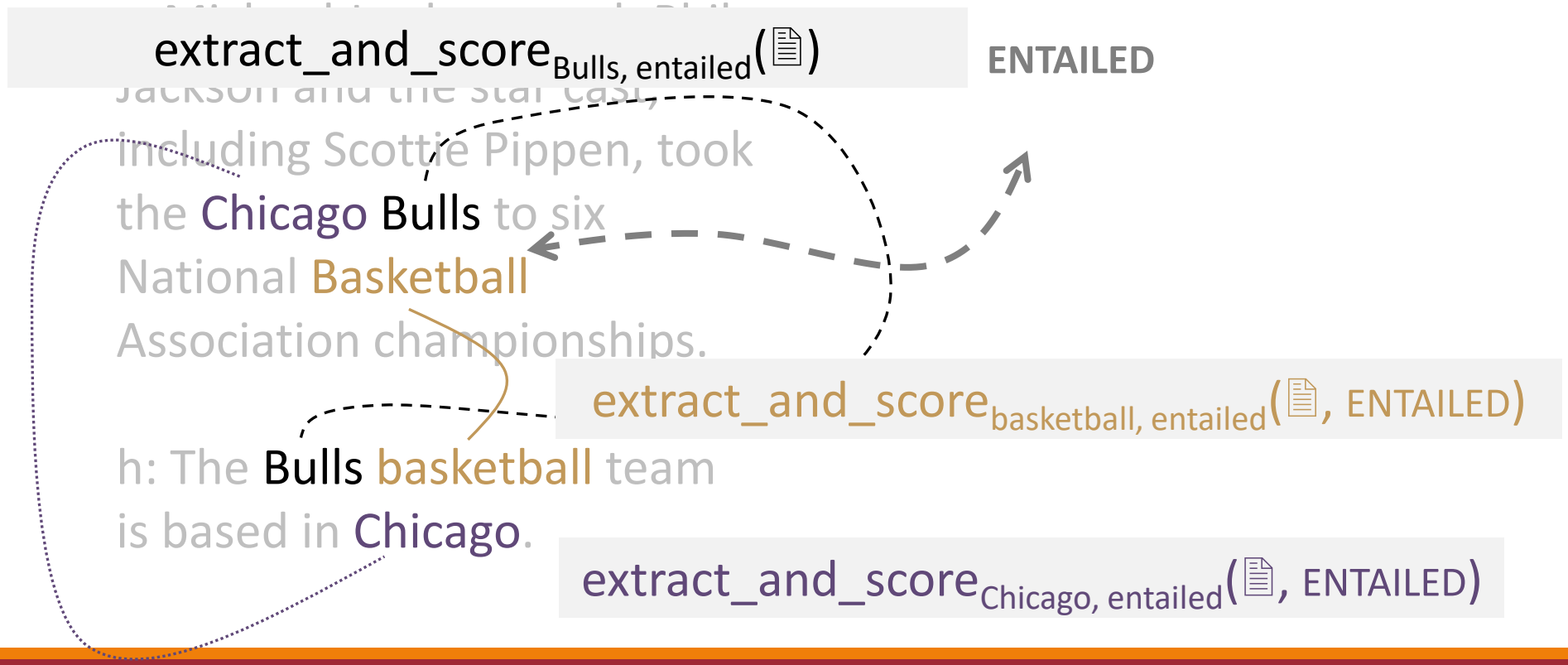
Review: Extracting Features



ENTAILED

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

We need to *score* the different extracted clues.



Review: Scoring Our Clues

score(, 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.

(ignore the feature indexing for now)

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

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

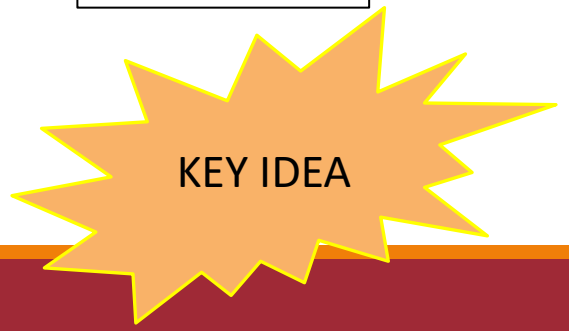
$\text{score}_{3\ ,\ \text{Entailed}}$(📄) ✚

…

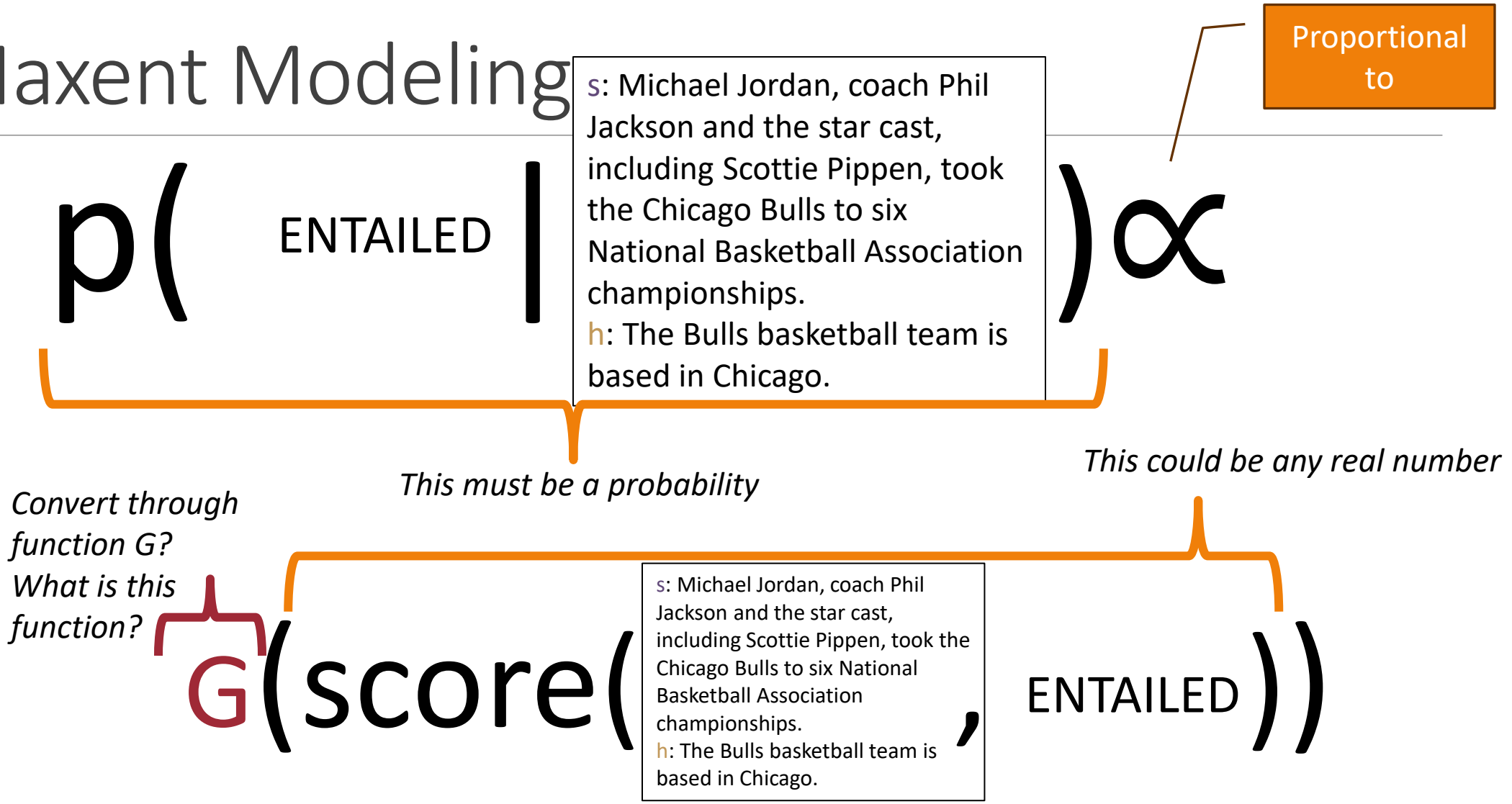
Review: Turning Scores into Probabilities

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.)



Maxent Modeling



What function G…

operates on any real number?

is never less than 0?

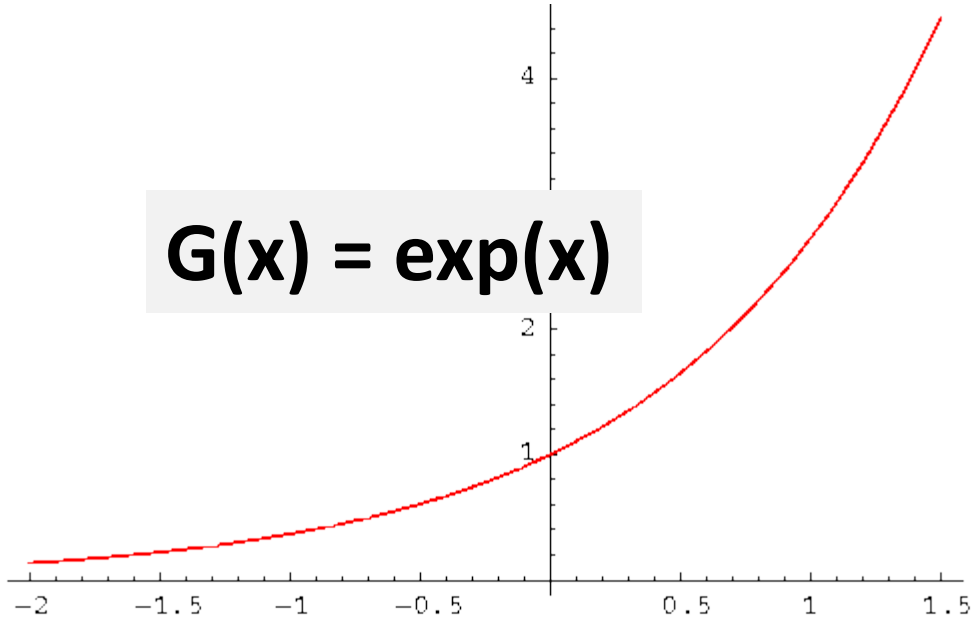
is monotonic? (if a < b, then G(a) < G(b))

What function G…

operates on any real number?

is never less than 0?

is monotonic? (if a < b, then G(a) < G(b))



Maxent Modeling

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

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

Maxent Modeling

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

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

Maxent Modeling

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

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

Maxent Modeling

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

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

K different weights…

for K different features

θ

.31
-.5
.1
.002
.522
…

$f(x)$

1
1
1
2
0
…

Maxent Modeling

$$p(\text{ENTAILED} \mid \text{s, h}) \propto \exp\Big(\sum_{K} \text{weight}_{K,\text{Entailed}} * \text{applies}_{K}(\text{s, h})\Big)$$

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.

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

K different weights…

for K different features

multiplied and then summed

Maxent Modeling

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

$$\exp(\text{Dot\_product of Entailed weight\_vec feature\_vec}(\text{📄}))$$

K different weights…

for K different features

multiplied and then summed

Maxent Modeling

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

$$\exp(\theta^T_{\text{ENTAILED}} f(\text{}))$$

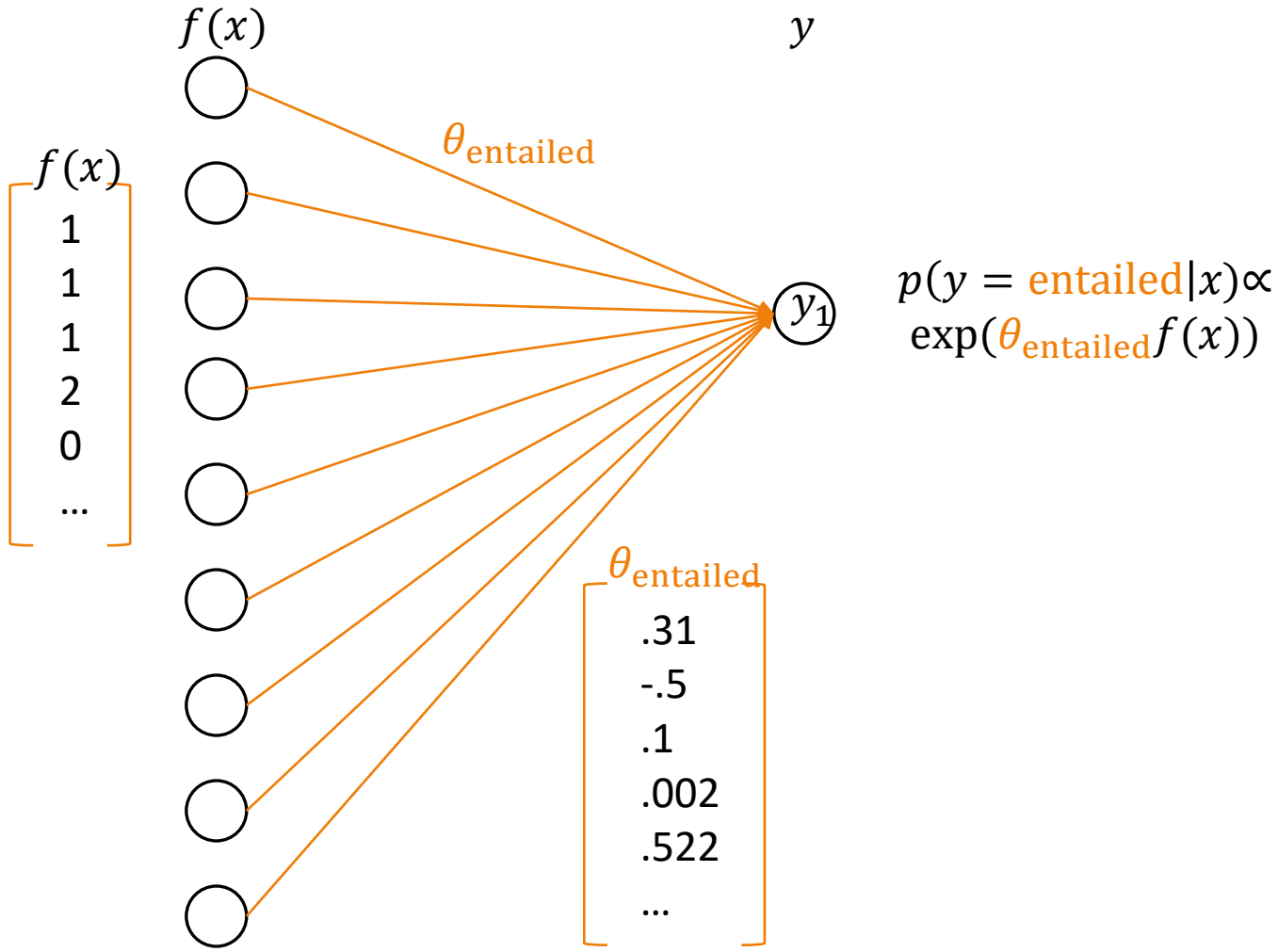
$$[.31\ -.5\ .1\ .002\ .522\ \ldots] \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 0 \\ \ldots \end{bmatrix}$$

K different weights…

for K different features

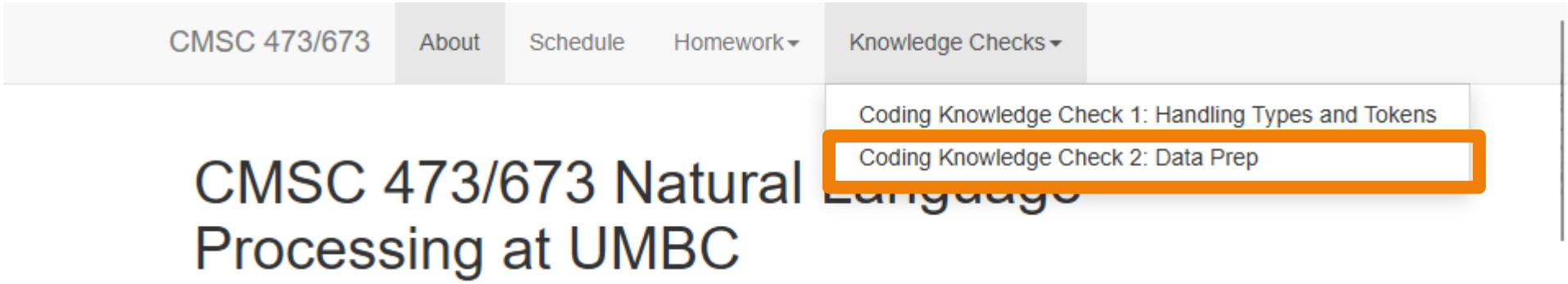
multiplied and then summed

Maxent Classifier, schematically



Knowledge Check: Data Prep

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



Maxent Modeling

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

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

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. $)=$

How do we define Z?

$$\frac{1}{Z}\exp(\quad \theta_{\text{ENTAILED}} f(\text{📄}) \quad)$$

Normalization for Classification

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

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

classify doc x with label y in one go

Normalization for Classification (long form)

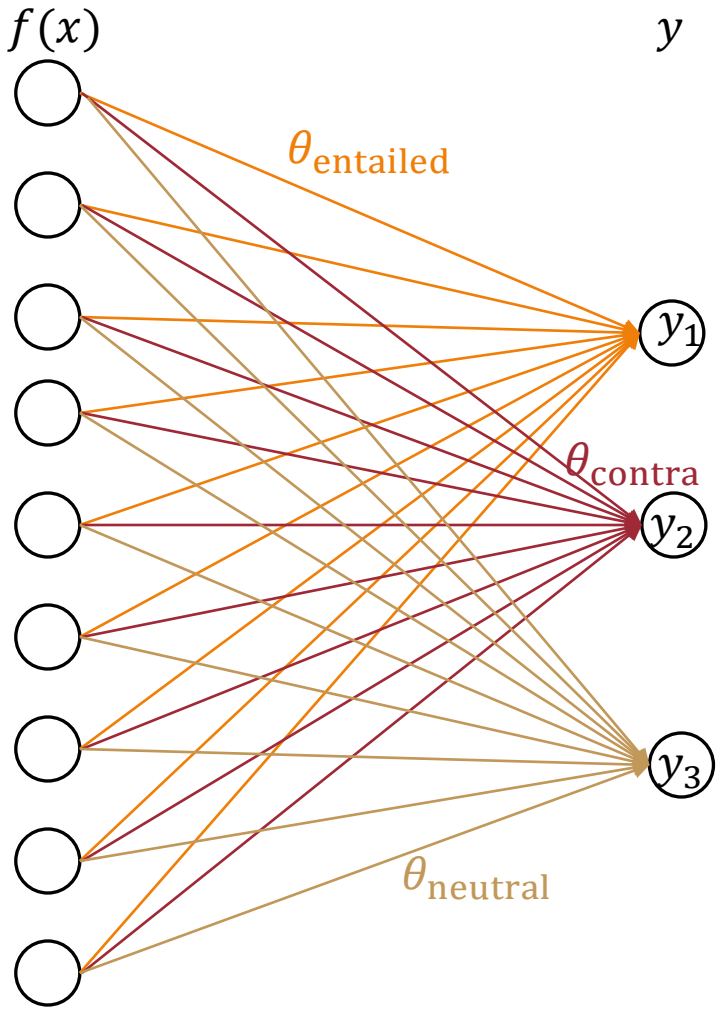
$$Z = \sum_{\text{label } j} \exp\Big(\text{weight}_{1,j} * \text{applies}_1(\text{📄}) + \text{weight}_{2,j} * \text{applies}_2(\text{📄}) + \text{weight}_{3,j} * \text{applies}_3(\text{📄}) + \dots \Big)$$

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

classify doc x with label y in one go

Multiclass 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*

Final Equation for Logistic Regression

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

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

Different Notation, Same Meaning

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

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

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

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

Generally *templated*

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!)
…		

Q: Is this a reasonable representation?

Q: What are some tradeoffs (benefits vs. costs)?