ML Evaluation → Classification

CMSC 473/673 - NATURAL LANGUAGE PROCESSING

Slides modified from Dr. Frank Ferraro

Learning Objectives

Develop an intuition about precision & recall

Extend P/R to multi-class problems

Identify when you might want certain evaluation metrics over others

Model classification problems using logistic regression

Define appropriate features for a logistic regression problem

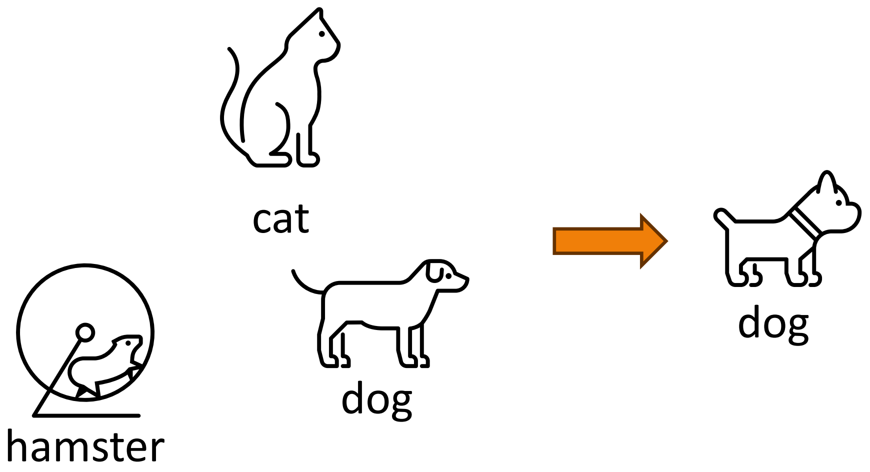
Review: Classification Evaluation: the 2-by-2 contingency table

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

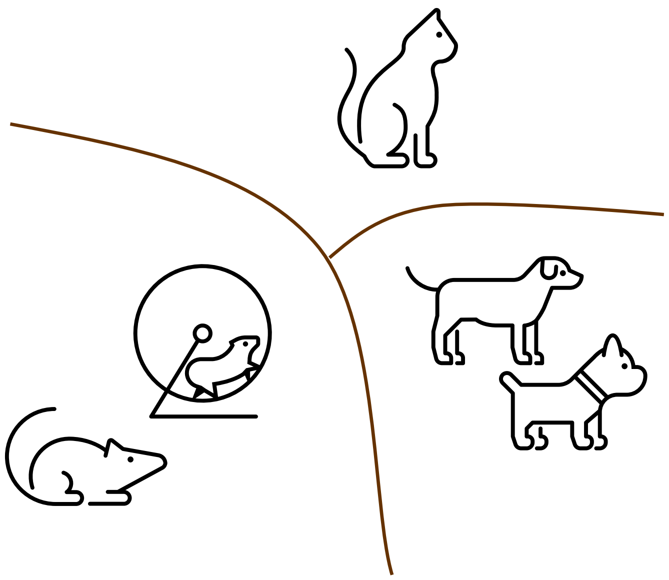
Construct this table by *counting* the number of TPs, FPs, FNs, TNs

Review: Types of Learning

SUPERVISED LEARNING

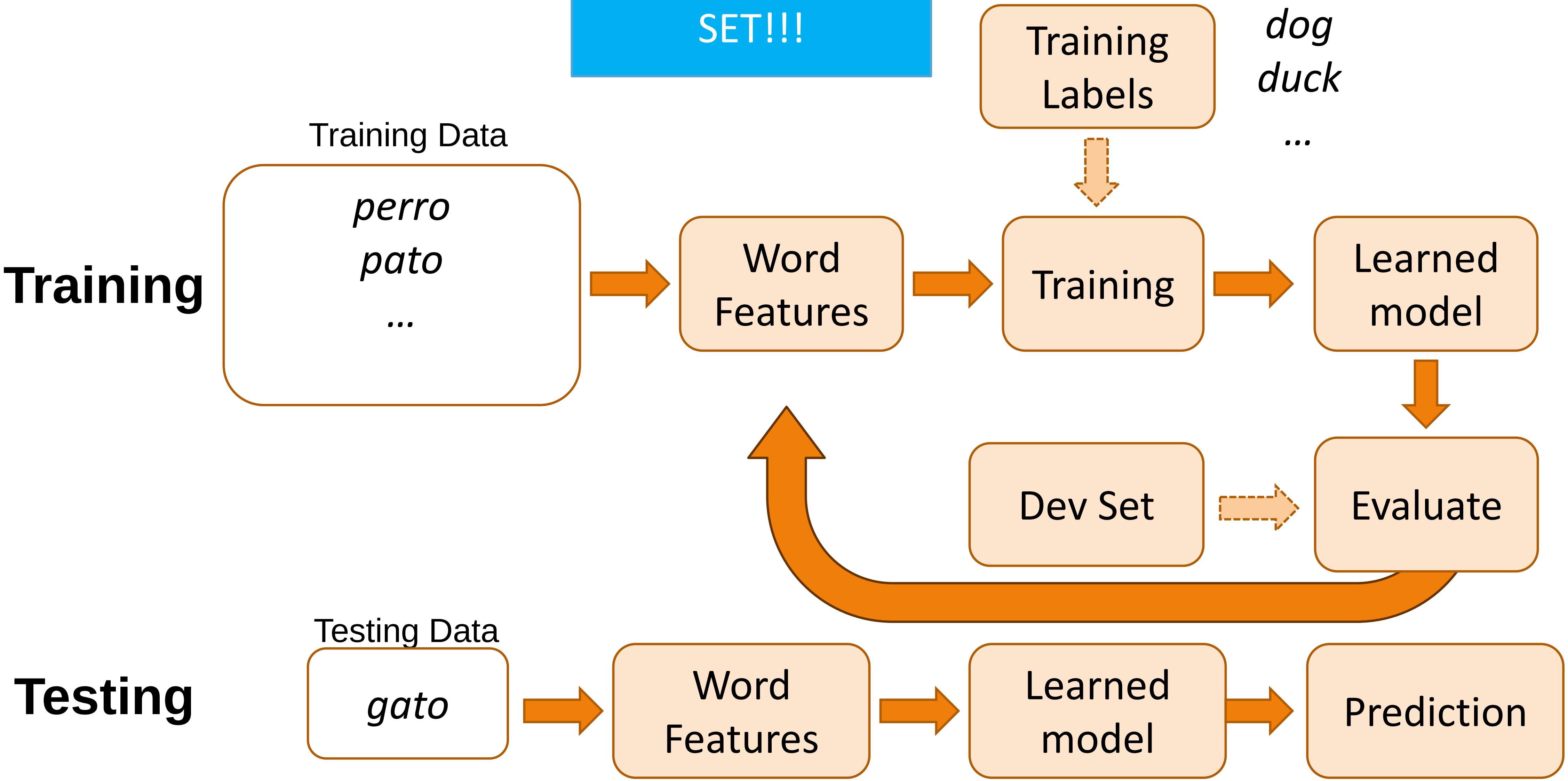


UNSUPERVISED LEARNING

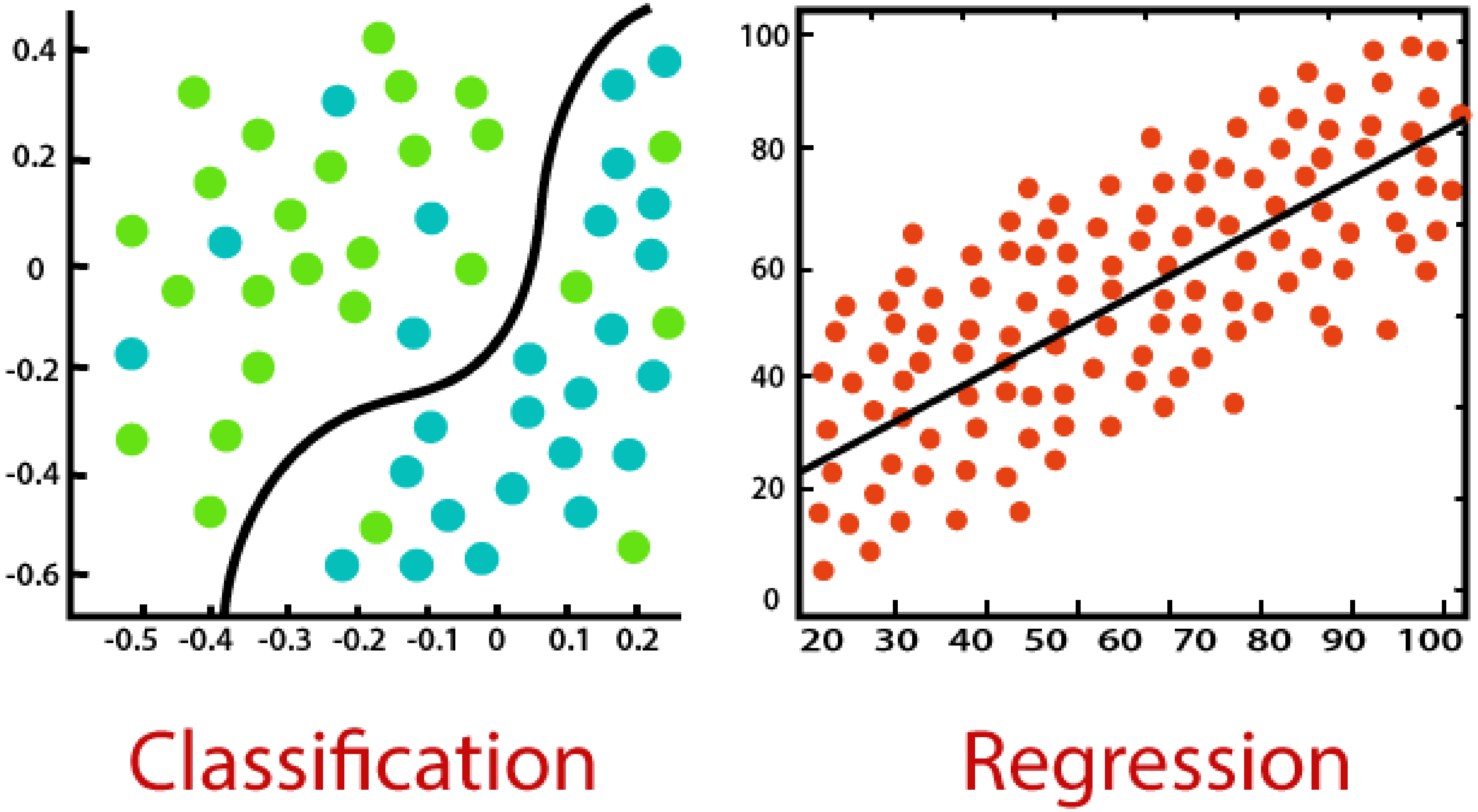


Review: Steps

DO NOT ITERATE ON THE TESTING SET!!!



Review: Types of models



https://medium.com/unpackai/classification-regression-in-machine-learning-7cf3b13b0b09

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*

Contingency Table (out of table form)

Query: Articles about dogs

FN

Adopting an Animal at the Pound

Why Huskies Howl

Who's a good boy?

TP

10 Tips for Grooming your Dog

What Foods to Hide from Your Dog During the Holidays

FP

It's raining cats and dogs

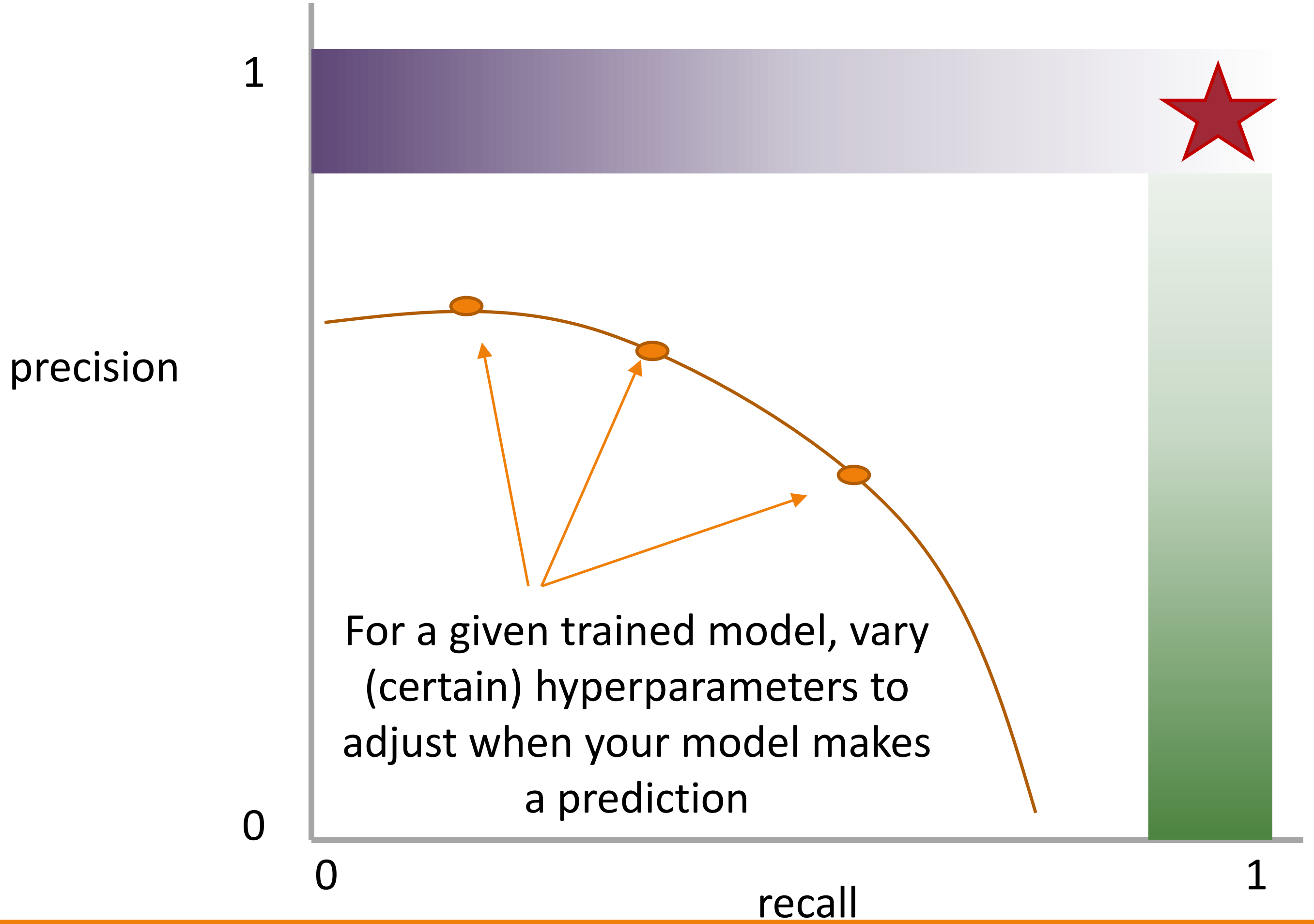
TN

15 Weird Animals You Can Find in North America

When Cats Rule the World

Meme from: *https://www.reddit.com/r/AdviceAnimals/comments/ck8xh0/yo_dawg_i_heard_you_like_old_memes/*

Review: Precision and Recall Present a Tradeoff



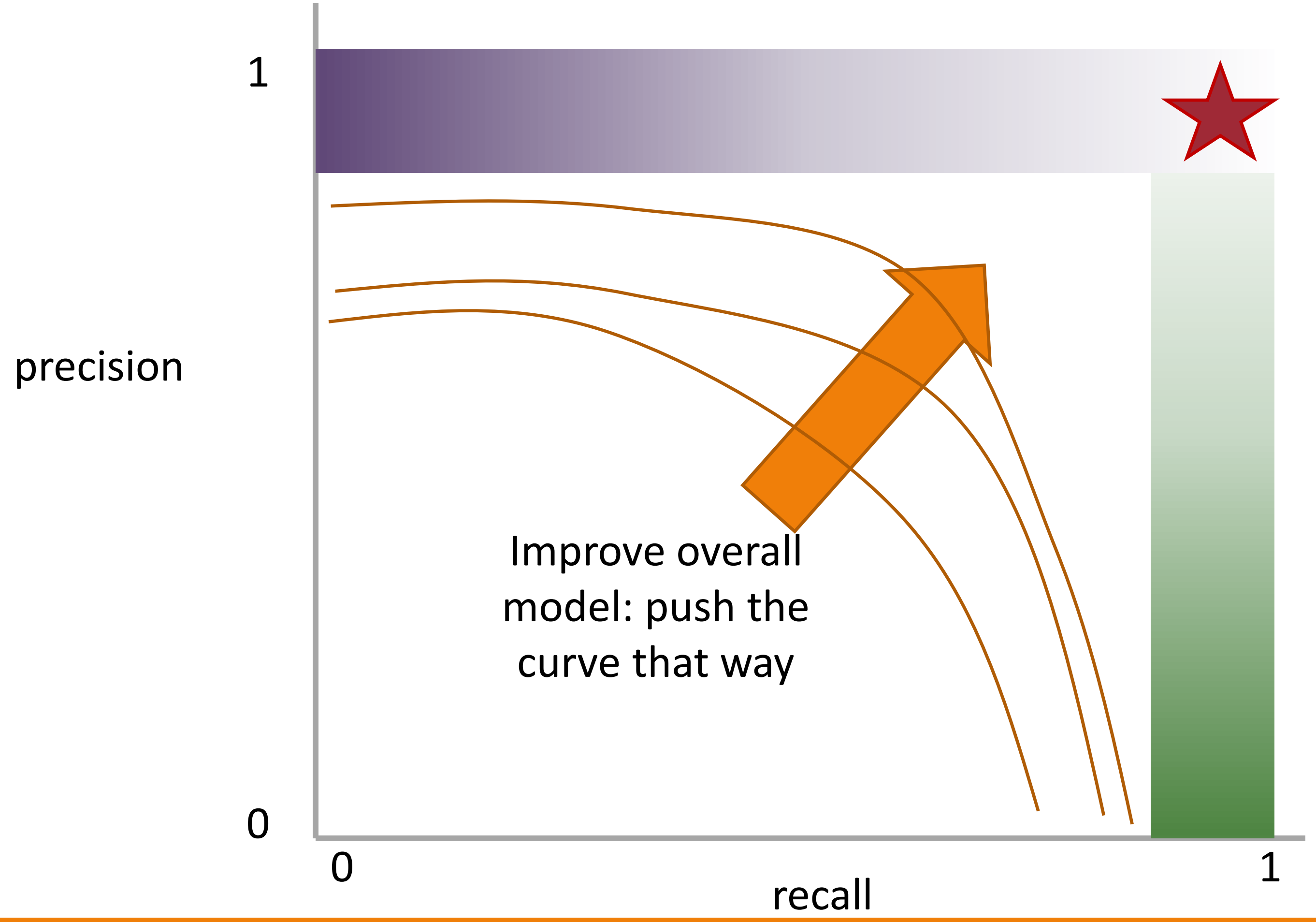
Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Q: You have a model that only make correct predictions. Where on this graph is it?

Idea: measure the tradeoff between precision and recall

Review: Precision and Recall Present a Tradeoff



Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Q: You have a model that only make correct predictions. Where on this graph is it?

Idea: measure the tradeoff between precision and recall

Review: A combined measure: 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 $P = R = 0$)

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

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

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

Accuracy takes everything in consideration

F-Score is focused on TP

When would you want to use accuracy vs F1?

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

	Actually Target	Actually Not Target
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

P/R/F in a Multi-class Setting: Micro- vs. Macro-Averaging

Macroaveraging: Compute performance for each class, then average.

$$\text{macroprecision} = \frac{1}{C}\sum_{c}\frac{\mathrm{TP_c}}{\mathrm{TP_c}+\mathrm{FP_c}} = \frac{1}{C}\sum_{c}\text{precision}_c$$

$$\text{macrorecall} = \frac{1}{C}\sum_{c}\frac{\mathrm{TP_c}}{\mathrm{TP_c}+\mathrm{FN_c}} = \frac{1}{C}\sum_{c}\text{recall}_\mathrm{c}$$

when to prefer macroaveraging?

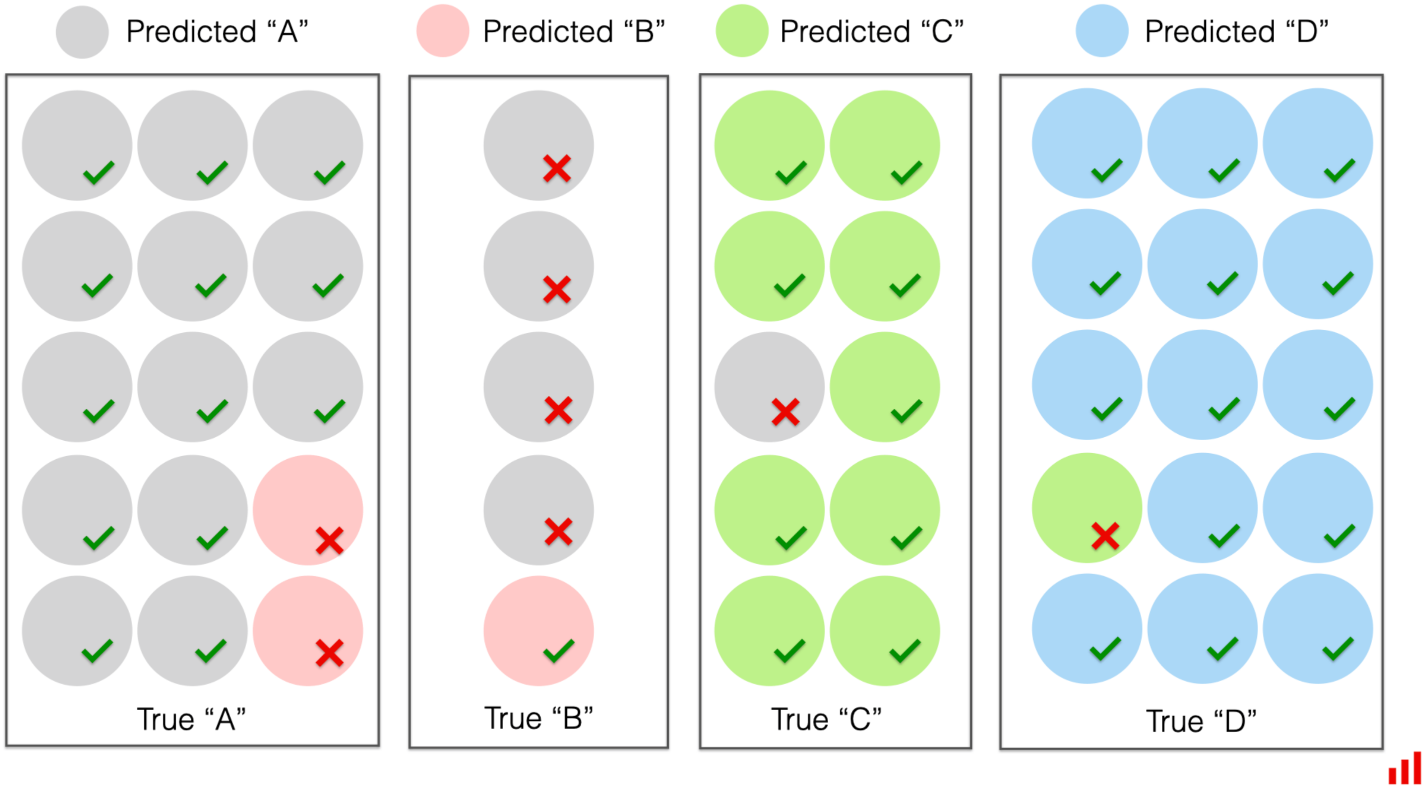
Microaveraging: Collect decisions for all classes, compute contingency table, evaluate.

when to prefer microaveraging?

$$\text{microprecision} = \frac{\sum_\mathrm{c}\mathrm{TP_c}}{\sum_\mathrm{c}\mathrm{TP_c}+\sum_\mathrm{c}\mathrm{FP_c}}$$

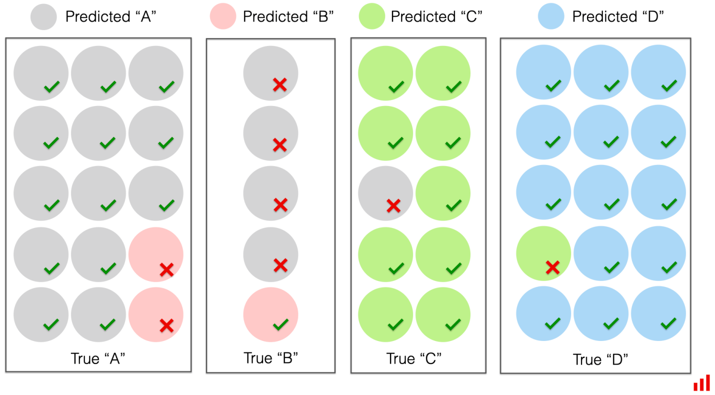
$$\text{microrecall} = \frac{\sum_\mathrm{c}\mathrm{TP_c}}{\sum_\mathrm{c}\mathrm{TP_c}+\sum_\mathrm{c}\mathrm{FN_c}}$$

Macro/Micro Example

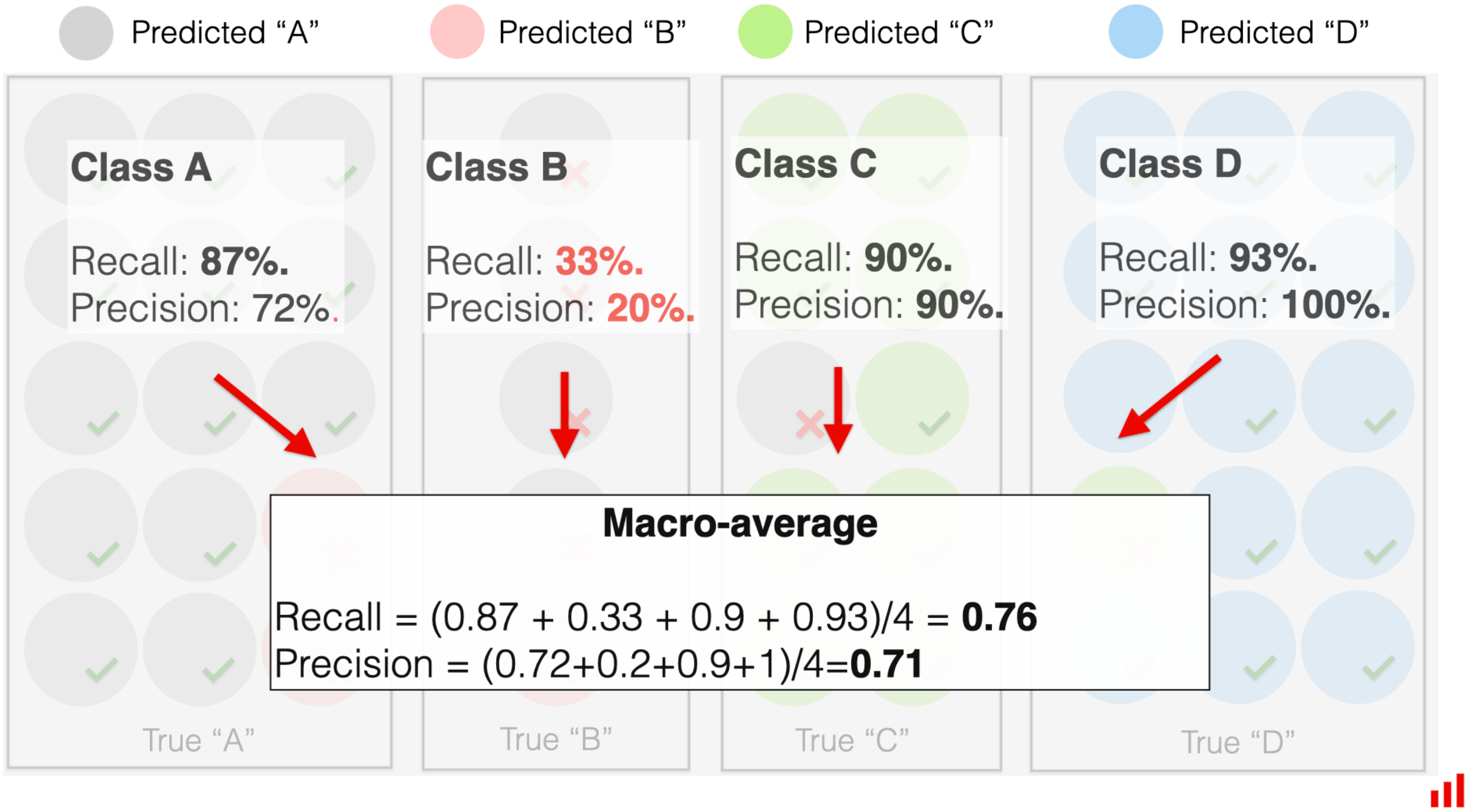


https://www.evidentlyai.com/classification-metrics/multi-class-metrics

Each *class* has equal weight



Macro-Average

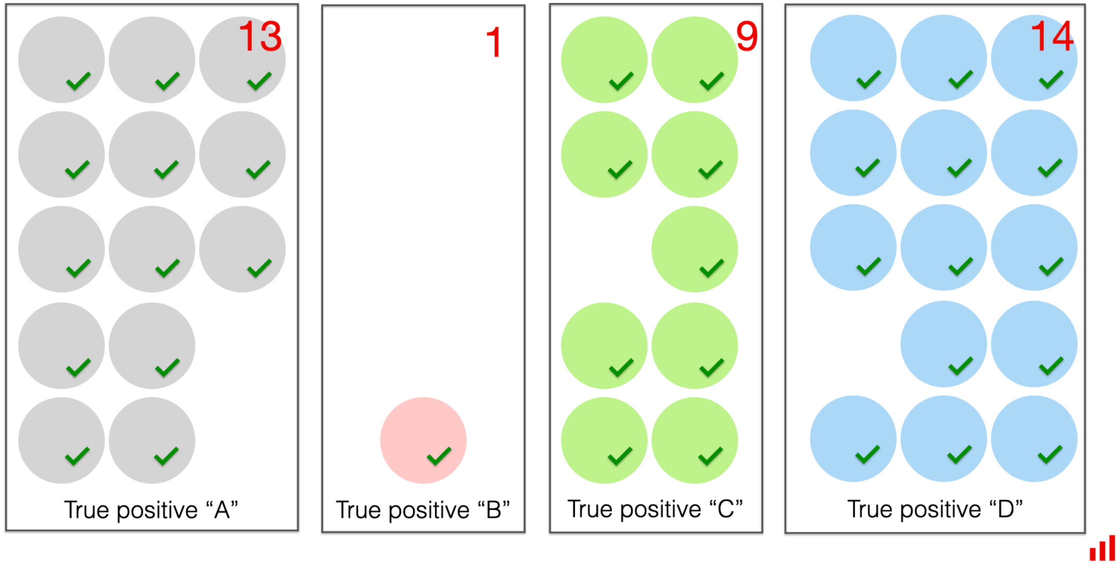


https://www.evidentlyai.com/classification-metrics/multi-class-metrics

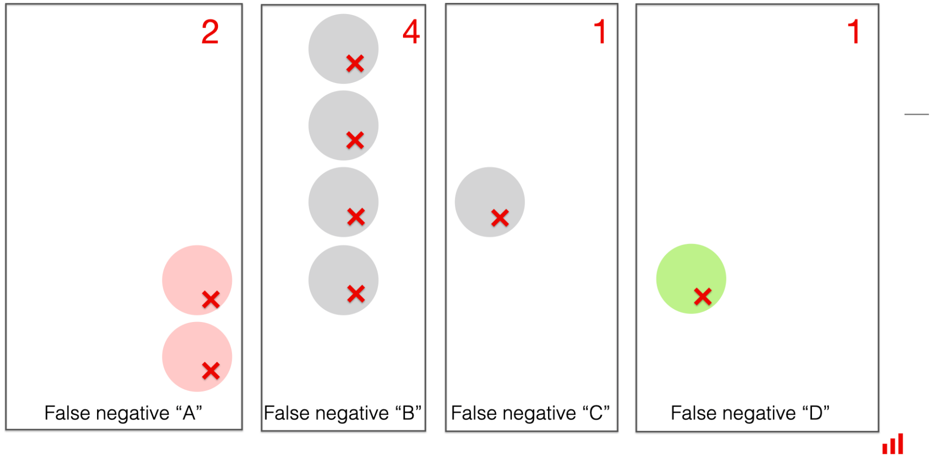
Each *instance* has equal weight

Micro-Average

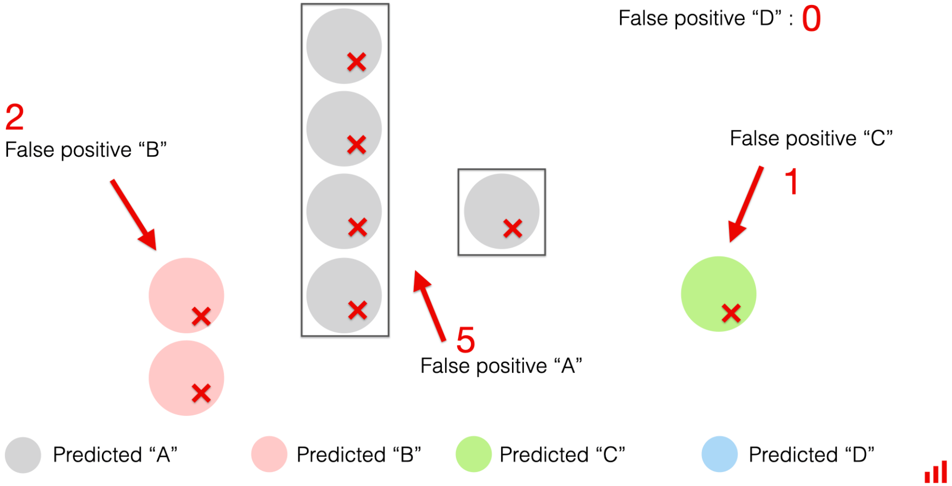
All true positives



All false negatives



All false positives



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

Micro- vs Macro-Average

So when would we want to prefer micro-averaging vs macro-averaging?

$$\text{macroprecision} = \frac{1}{C}\sum_{c}\frac{\mathrm{TP_c}}{\mathrm{TP_c} + \mathrm{FP_c}} = \frac{1}{C}\sum_{c}\text{precision}_{c}$$

$$\text{macrorecall} = \frac{1}{C}\sum_{c}\frac{\mathrm{TP_c}}{\mathrm{TP_c} + \mathrm{FN_c}} = \frac{1}{C}\sum_{c}\mathrm{recall_c}$$

$$\text{microprecision} = \frac{\sum_{\mathrm{c}}\mathrm{TP_c}}{\sum_{\mathrm{c}}\mathrm{TP_c} + \sum_{\mathrm{c}}\mathrm{FP_c}} \qquad \text{microrecall} = \frac{\sum_{\mathrm{c}}\mathrm{TP_c}}{\sum_{\mathrm{c}}\mathrm{TP_c} + \sum_{\mathrm{c}}\mathrm{FN_c}}$$

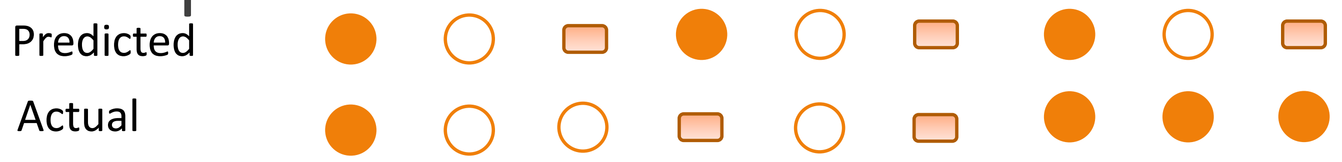
But how do we compute stats for multiple classes?

We already saw how the “polarity” affects the stats we compute…

Two main approaches. Either:

1. Compute “one-vs-all” 2x2 tables. OR
2. Generalize the 2x2 tables and compute per-class TP / FP / FN based on the diagonals and off-diagonals

1. Compute "one-vs-all" 2x2 tables

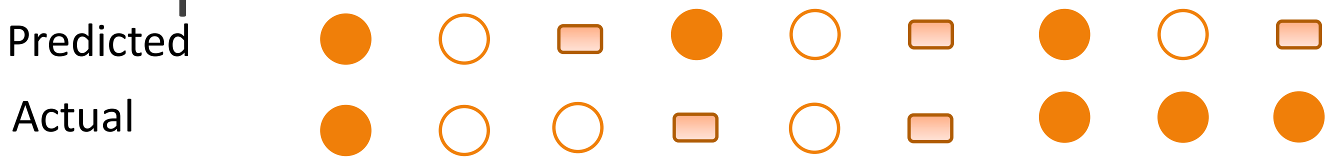


Look for	Actually Target	Actually Not Target
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

Look for	Actually Target	Actually Not Target
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

Look for	Actually Target	Actually Not Target
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

1. Compute “one-vs-all” 2x2 tables

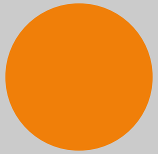
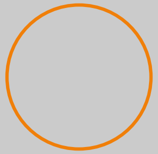
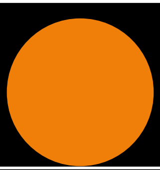
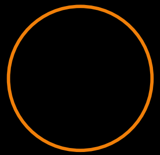


Look for	Actually Target	Actually Not Target
Selected/Guessed	2	1
Not select/not guessed	2	4

Look for	Actually Target	Actually Not Target
Selected/Guessed	2	1
Not select/not guessed	1	5

Look for	Actually Target	Actually Not Target
Selected/Guessed	1	2
Not select/not guessed	1	5

2. Generalizing the 2-by-2 contingency table

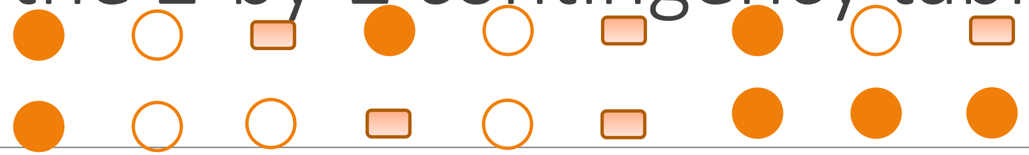
		Correct Value		
				
Guessed Value		#	#	#
		#	#	#
		#	#	#

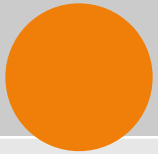
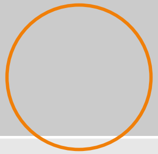
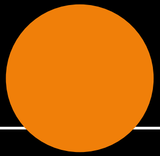
This is also called a **Confusion Matrix**

2. Generalizing the 2-by-2 contingency table

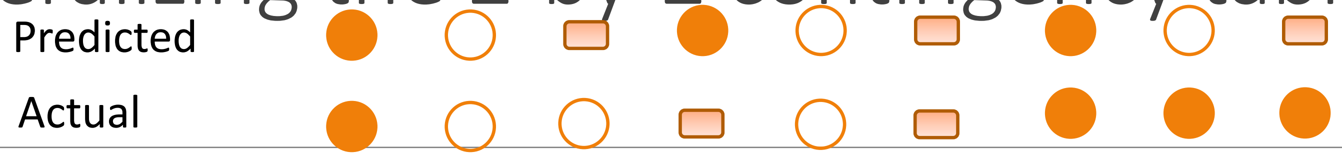
Predicted

Actual



		Correct Value		
				
Guessed Value		#	#	#
		#	#	#
		#	#	#

2. Generalizing the 2-by-2 contingency table



		Correct Value		
		●	○	▭
Guessed Value	●	2	0	1
	○	1	2	0
	▭	1	1	1

2. Generalizing the 2-by-2 contingency table

Predicted

Actual

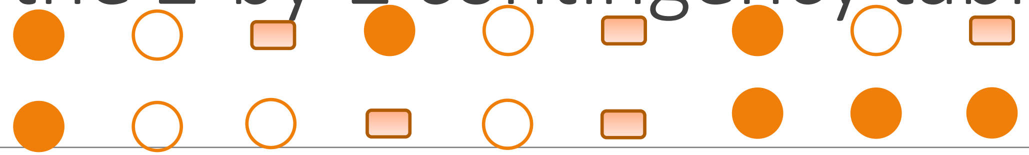
		Correct Value		
Guessed Value		A 2	B 0	C 1
		D 1	E 2	F 0
		G 1	H 1	I 1

How do you compute TP ?

2. Generalizing the 2-by-2 contingency table

Predicted

Actual



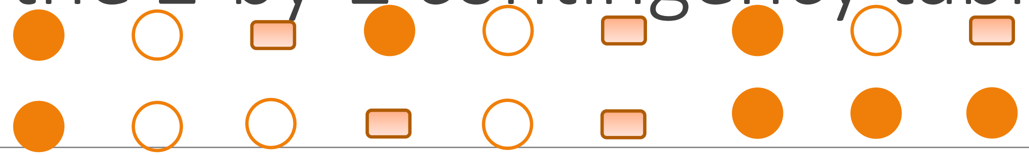
		Correct Value		
Guessed Value		A 2	B 0	C 1
		D 1	E 2	F 0
		G 1	H 1	I 1

How do you compute TP ?

2. Generalizing the 2-by-2 contingency table

Predicted

Actual



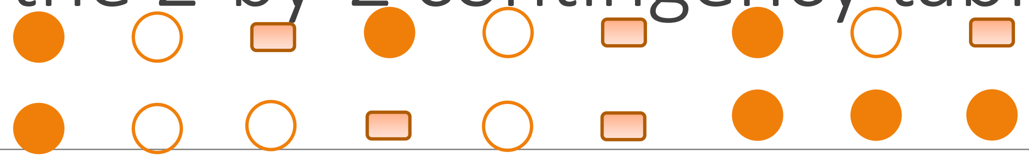
		Correct Value		
Guessed Value		A 2	B 0	C 1
		D 1	E 2	F 0
		G 1	H 1	I 1

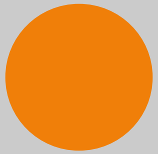
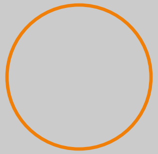
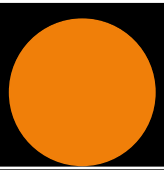
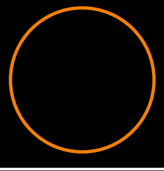
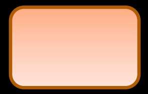
How do you compute FN ?

2. Generalizing the 2-by-2 contingency table

Predicted

Actual



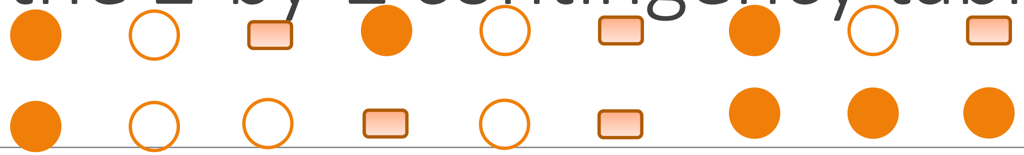
		Correct Value		
				
Guessed Value		A 2	B 0	C 1
		D 1	E 2	F 0
		G 1	H 1	I 1

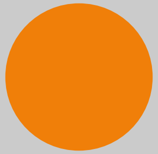
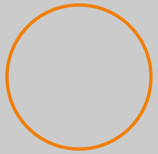
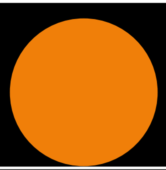
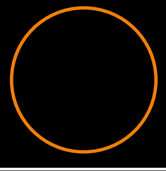
How do you compute FN?

2. Generalizing the 2-by-2 contingency table

Predicted

Actual



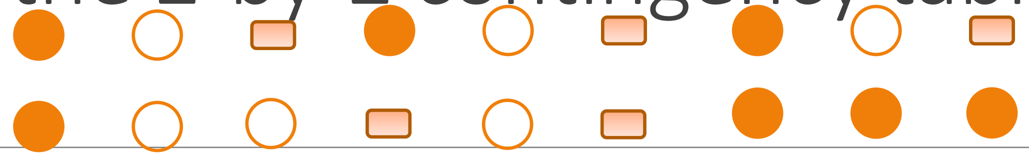
		Correct Value		
				
Guessed Value		A 2	B 0	C 1
		D 1	E 2	F 0
		G 1	H 1	I 1

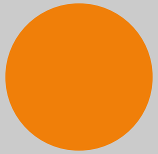
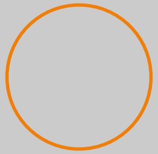
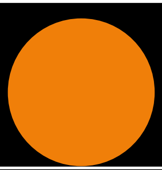
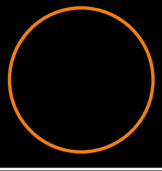
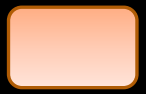
How do you compute FP ?

2. Generalizing the 2-by-2 contingency table

Predicted

Actual



		Correct Value		
				
Guessed Value		A 2	B 0	C 1
		D 1	E 2	F 0
		G 1	H 1	I 1

How do you compute FP ?

Generalizing the 2-by-2 contingency table

Q: Is this a good result?

		Correct Value		
Guessed Value		80	9	11
		7	86	7
		2	8	9

Generalizing the 2-by-2 contingency table

Q: Is this a good result?

		Correct Value		
Guessed Value		30	40	30
		25	30	50
		30	35	35

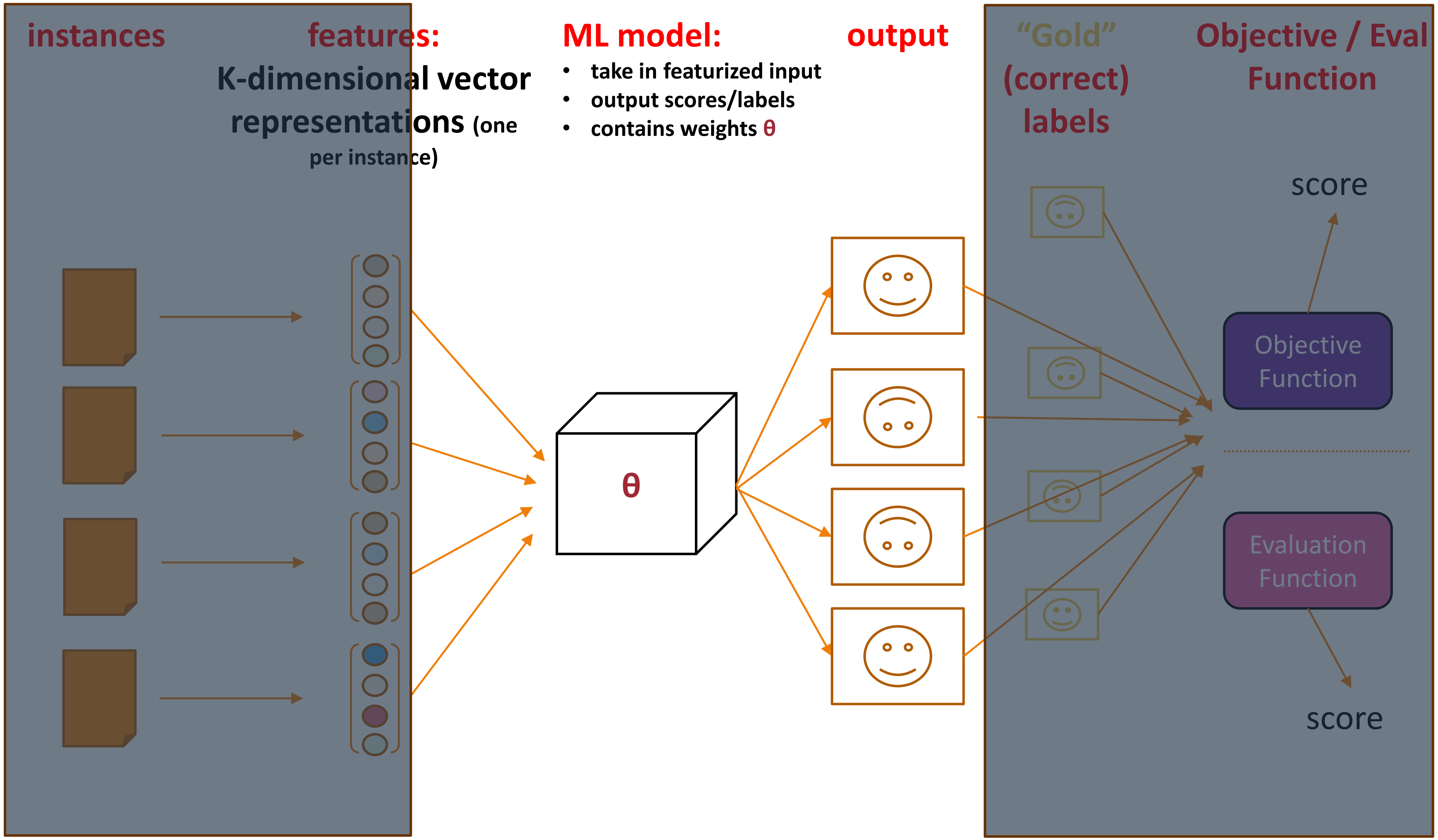
Generalizing the 2-by-2 contingency table

Q: Is this a good result?

		Correct Value		
		●	○	▭
Guessed Value	●	7	3	90
	○	4	8	88
	▭	3	7	90

Classification

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

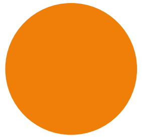
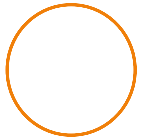
Maxent models can be used:

- 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 <u>knowledge that X has occurred</u>

NLP pg. 450

Terminology (with variables)

Posterior probability:

$$p(\ Y = \text{label}_1\ |X) \text{ vs. } p(Y = \text{label}_0\ |X)$$

Conditional probabilities:

$$p(Y = \text{label}_1 \mid X) + p(Y = \text{label}_0 \mid X) = 1$$

$$p(Y = \text{label}_1 \mid X) \geq 0,$$
$$p(Y = \text{label}_0 \mid X) \geq 0$$

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

NLP pg. 449

💡 Key Take-away 💡

We will ***learn*** this

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

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

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

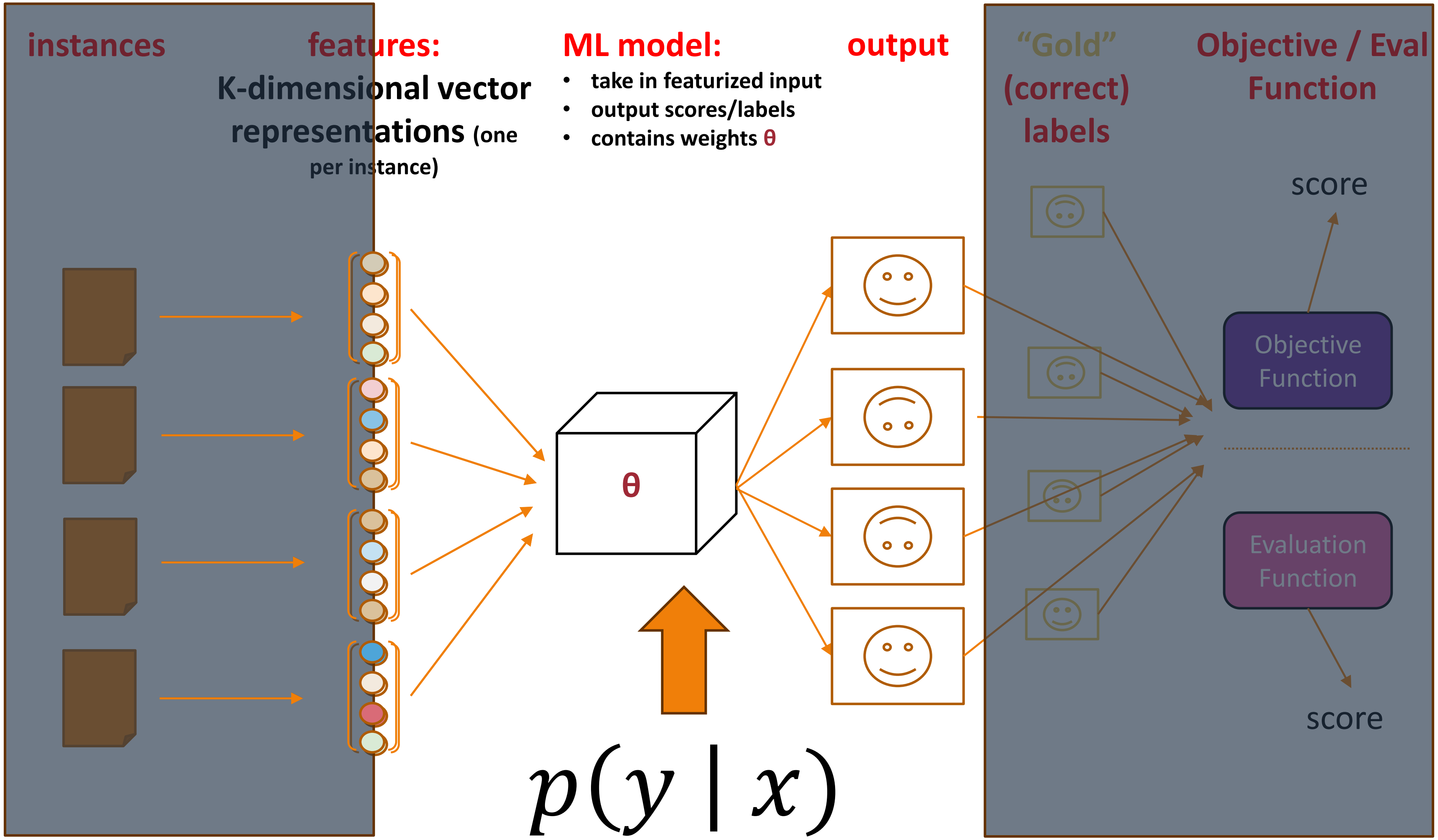
Generatively trained classifier with maxent-based language model

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

(we'll start with this one)

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

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**

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

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)

Amazon acquired MGM in 2022, taking over a sprawling library that includes more than 4,000 feature films and 17,000 television shows. The tech behemoth also earned the rights to distribute all the Bond movies, but the new deal solidifies the company's oversight of Bond's big-screen future.

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
Amazon	1
acquired	1
behemoth	1
Bond	2
…	
sniffle	0
…	

Adapted from https://www.nbcnews.com/pop-culture/movies/amazon-taking-control-james-bond-movie-franchise-rcna192967

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.

$\text{extract\_and\_score}_{\text{Bulls, entailed}}(\ )$

ENTAILED

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

$\text{extract\_and\_score}_{\text{basketball, entailed}}(\ \text{, ENTAILED})$

h: The Bulls basketball team is based in Chicago.

$\text{extract\_and\_score}_{\text{Chicago, entailed}}(\ \text{, ENTAILED})$

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

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

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.

This must be a probability

This could be any real number

Convert through function G? What is this function?

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

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.

What function G…

operates on any real number?

is never less than 0?

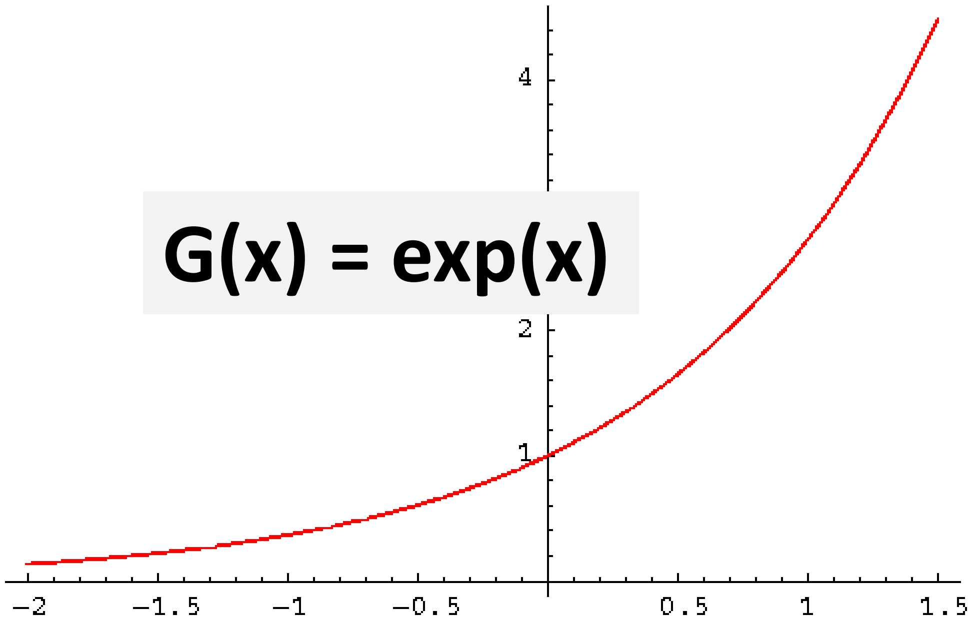
monotonic? ($a < b \Rightarrow G(a) < G(b)$)

What function G…

operates on any real number?

is never less than 0?

monotonic? (a < b ➔ G(a) < G(b))



Maxent Modeling

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

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.

$$) \propto$$

$$\exp(\mathrm{score}($$

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.

$$, \textsc{Entailed}))$$

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.

$) \propto$

$$\exp\left(\begin{array}{l} \text{score}_{1,\,\text{Entailed}}(\ ) + \\ \text{score}_{2,\,\text{Entailed}}(\ ) + \\ \text{score}_{3,\,\text{Entailed}}(\ ) + \\ \quad\ldots \end{array} \right))$$

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

Maxent Modeling

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

$$\exp\left(\begin{array}{l} \text{weight}_{1,\ \text{Entailed}} * \text{applies}_1(🗎) + \\ \text{weight}_{2,\ \text{Entailed}} * \text{applies}_2(🗎) + \\ \text{weight}_{3,\ \text{Entailed}} * \text{applies}_3(🗎) + \\ \quad \ldots \end{array}\right)$$

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

K different weights…

for K different features

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.) ∝

exp(Dot_product of Entailed weight_vec feature_vec(🖹))

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

Maxent Modeling

$$p(\text{ENTAILED} \mid \quad) \propto$$

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.

$$\exp(\theta_{\text{ENTAILED}}^{T} f(\quad) \quad)$$

K different weights…

for K different features

multiplied and then summed

💻 Implementation Check

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

Assume we can compute $\mathrm{x} = f(\text{🖹})$

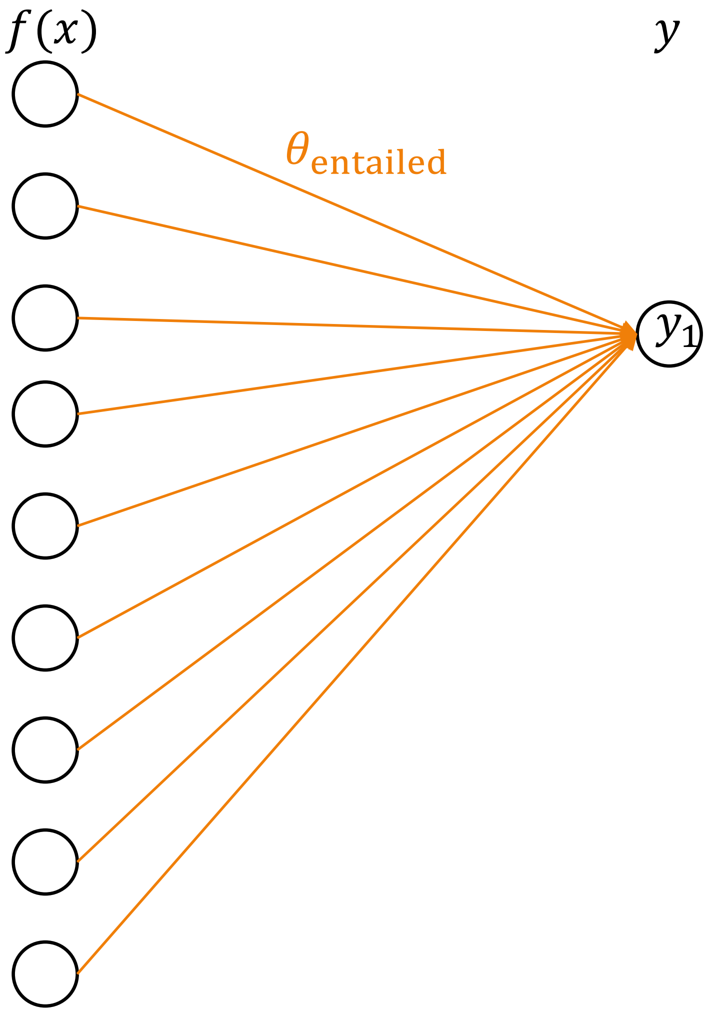
(a one-dimensional tensor)

How can we implement the above computation in Pytorch?

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

$p($ 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. $) =$

Q: How do we define Z?

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

K different weights…

for K different features…

multiplied and then

Normalization for Classification

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

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

classify doc x with label y in one go

Normalization for Classification (long form)

Z =

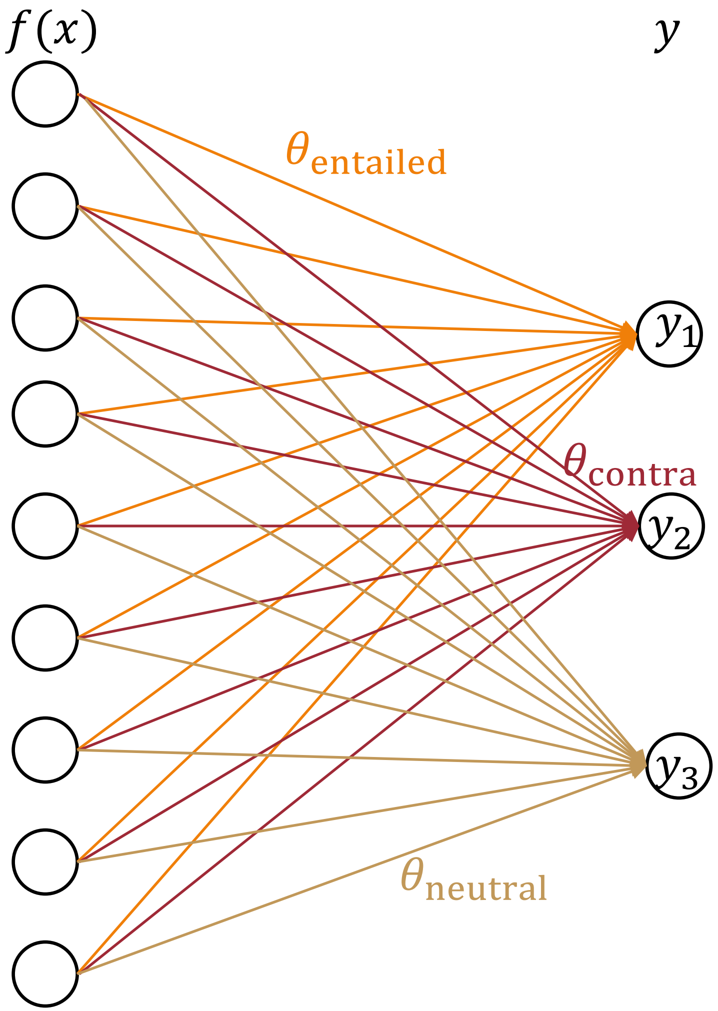
$$\sum_{\text{label } \color{red}{j}} \exp\left(\begin{array}{l} \text{weight}_{1,\color{red}{j}} * \text{applies}_1(\;) \;+ \\ \text{weight}_{2,\color{red}{j}} * \text{applies}_2(\;) \;+ \\ \text{weight}_{3,\color{red}{j}} * \text{applies}_3(\;) \;+ \\ \quad\quad \ldots \end{array} \right)$$

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

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

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

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

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

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

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

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)?

Useful Terminology: n-gram

Within a larger string (e.g., sentence),
a contiguous sequence of n items (e.g., words)

Colorless green ideas sleep furiously

n	Commonly called	History Size (Markov order)	Example n-gram ending in "furiously"
1	unigram	0	furiously
2	bigram	1	sleep furiously
3	trigram (3-gram)	2	ideas sleep furiously
4	4-gram	3	green ideas sleep furiously
n	n-gram	n-1	$w_{i-n+1} \ldots w_{i-1}\, w_i$

Templated Features

Define a feature fclue(🖹) for each clue you want to consider

The feature fclue fires if the clue applies to/can be found in 🖹

Clue is often a target phrase (an n-gram)

Maxent Modeling: Templated Binary Feature Functions

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(
$\text{weight}_{1\text{, Entailed}} * \text{applies}_1(\text{🗎})$ +
$\text{weight}_{1\text{, Entailed}} * \text{applies}_2(\text{🗎})$ +
$\text{weight}_{1\text{, Entailed}} * \text{applies}_3(\text{🗎})$ +
…
))

$$\text{applies}_{\text{target}}(\text{🗎}) = \begin{cases} 1, \text{target } matches \text{ 🗎} \\ 0, \quad \text{otherwise} \end{cases}$$

binary

Example of a Templated Binary Feature Functions

$$\text{applies}_{\text{target}}(\text{🖹}) = \begin{cases} 1, \text{target } \mathit{matches} \text{ 🖹} \\ 0, \qquad \text{otherwise} \end{cases}$$

$$\text{applies}_{\text{ball}}(\text{🖹}) = \begin{cases} 1, \text{ball } \mathit{in} \text{ both s and h of 🖹} \\ 0, \qquad \text{otherwise} \end{cases}$$

Example of a Templated Binary Feature Functions

$$\text{applies}_{\text{target}}(\text{🗎}) = \begin{cases} 1, \text{target } \mathit{matches} \text{ 🗎} \\ 0, \quad \text{otherwise} \end{cases}$$

$$\text{applies}_{\text{ball}}\,(\text{🗎}) = \begin{cases} 1, \text{ball } \mathit{in} \text{ both s and h of 🗎} \\ 0, \quad \text{otherwise} \end{cases}$$

Q: If there are V vocab types and L label types:

1. How many features are defined if unigram targets are used (w/ each label)?

Example of a Templated Binary Feature Functions

$$\text{applies}_{\text{target}}(\text{🗎}) = \begin{cases} 1, \text{target } matches \text{ 🗎} \\ \quad 0, \qquad \text{otherwise} \end{cases}$$

$$\text{applies}_{\text{ball}}(\text{🗎}) = \begin{cases} 1, \text{ball } in \text{ both s and h of 🗎} \\ \quad 0, \qquad \text{otherwise} \end{cases}$$

Q: If there are V vocab types and L label types:

1. How many features are defined if unigram targets are used (w/ each label)?

A1: VL

Example of a Templated Binary Feature Functions

$$\text{applies}_{\text{target}}(🗎) = \begin{cases} 1, \text{target } matches \text{ } 🗎 \\ 0, \quad\quad \text{otherwise} \end{cases}$$

$$\text{applies}_{\text{ball}}(🗎) = \begin{cases} 1, \text{ball } in \text{ both s and h of } 🗎 \\ 0, \quad\quad \text{otherwise} \end{cases}$$

Q: If there are V vocab types and L label types:

1. How many features are defined if unigram targets are used (w/ each label)?

A1: VL

2. How many features are defined if bigram targets are used?

Example of a Templated Binary Feature Functions

$$\text{applies}_{\text{target}}(\text{🗎}) = \begin{cases} 1, \text{target } matches\ \text{🗎} \\ 0, \qquad \text{otherwise} \end{cases}$$

$$\text{applies}_{\text{ball}}(\text{🗎}) = \begin{cases} 1, \text{ball } in \text{ both s and h of } \text{🗎} \\ 0, \qquad \text{otherwise} \end{cases}$$

Q: If there are V vocab types and L label types:

1. How many features are defined if unigram targets are used (w/ each label)?

A1: VL

2. How many features are defined if bigram targets are used (w/ each label)?

A2: V^2L

Example of a Templated Binary Feature Functions

$$\text{applies}_{\text{target}}(🗎) = \begin{cases} 1, \text{target } \textit{matches} \text{ 🗎} \\ 0, \quad \text{otherwise} \end{cases}$$

$$\text{applies}_{\text{ball}}(🗎) = \begin{cases} 1, \text{ball } \textit{in} \text{ both s and h of 🗎} \\ 0, \quad \text{otherwise} \end{cases}$$

Q: If there are V vocab types and L label types:

1. How many features are defined if unigram targets are used (w/ each label)?

A1: VL

2. How many features are defined if bigram targets are used (w/ each label)?

A2: V^2L

3. How many features are defined if unigram and bigram targets are used (w/ each label)?

Example of a Templated Binary Feature Functions

$$\text{applies}_{\text{target}}(🖹) = \begin{cases} 1, \text{target } matches \text{ 🖹} \\ 0, \quad\quad \text{otherwise} \end{cases}$$

$$\text{applies}_{\text{ball}}(🖹) = \begin{cases} 1, \text{ball } in \text{ both s and h of 🖹} \\ 0, \quad\quad \text{otherwise} \end{cases}$$

Q: If there are V vocab types and L label types:

1. How many features are defined if unigram targets are used (w/ each label)?

A1: VL

2. How many features are defined if bigram targets are used (w/ each label)?

A2: V^2L

3. How many features are defined if unigram and bigram targets are used (w/ each label)?

A2: $(\mathrm{V} + V^2)L$