ECE594: Mathematical Models of Language

Spring 2022

Lecture 2: Word-level Models

Logistics

- Sign up sheet will be available before the weekend
- Resources

Text as Signal

- What are units of text?
- What is a word?

What is a word?

- Count the words:

Bob's handyman is a do-it-yourself kinda guy, isn't he?

Slides credit: Yulia Tsvetkov, CMU

Words

- Orthographic definition
  - strings separated by white spaces
  - problems: *Bob's handy man is a do-it-yourself kinda guy, isn't he?*
  - unwritten languages, languages that don't use white spaces, etc.
- Prosodic definition
  - words have one main stress and longer words may have a secondary stress
  - problems: function words, clitics
- Semantic definition
  - words are units that describe a single idea or a semantic concept
  - problem: many semantic concepts span phrases or sentences and don't have a corresponding word
- Syntactic definition
  - words are the syntactic building blocks of sentences

Slides credit: Yulia Tsvetkov, CMU

Morphology

Wordform: inflected word as it appears in text

Number (singular/plural), Tense (present, past, future)

banks	bank
sung	sing
duermes	dormir

Turkish	English
kork(-mak)	(to) fear
korku	fear
korkusuz	fearless
korkusuzlaş (-mak)	(to) become fearless
korkusuzlaşmış	One who has become fearless
korkusuzlaştır(-mak)	(to) make one fearless
korkusuzlaştırıl(-mak)	(to) be made fearless
korkusuzlaştırılmış	One who has been made fearless
korkusuzlaştırılabil(-mek)	(to) be able to be made fearless
korkusuzlaştırılabilecek	One who will be able to be made fearless
korkusuzlaştırabileceklerimiz	Ones who we can make fearless
korkusuzlaştırabileceklerimizden	From the ones who we can make fearless
korkusuzlaştırabileceklerimizdenmiş	I gather that one is one of those we can make fearless
korkusuzlaştırabileceklerimizdenmişçesine	As if that one is one of those we can make fearless
korkusuzlaştırabileceklerimizdenmişçesineyken	when it seems like that one is one of those we can make fearless

Words have Grammatical Functions

- Nouns, Adjectives, Verbs, Adverbs
  - Parts of speech

- Who did what to whom and how?
  - Subject, verb, object, manner

Vocabulary vs. Lexicon

- Vocabulary–set of words
- Lexicon–words and meaning
  - Base forms (lemma)
  - Optionally derived forms

Vocabulary vs. Lexicon

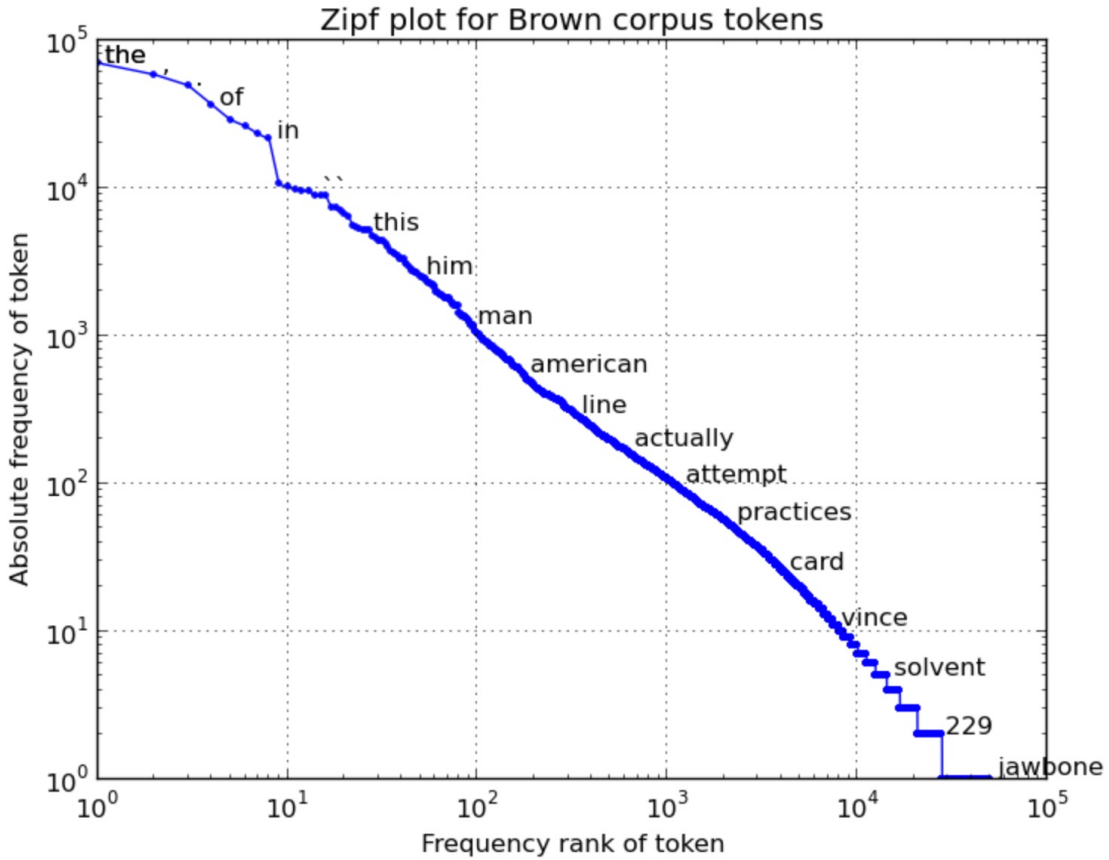
- Vocabulary is discrete but not finite
  - Can you think of new words that may not have existed 10 years ago?

  Deepfake - *a fake, digitally manipulated video or audio file produced by using deep learning, an advanced type of machine learning, and typically featuring a person's likeness and voice in a situation that did not actually occur.*

- Furthermore, the distribution over words resembles that of a **power law** (Zipf, 1949)

Zipf's Law

- Zipf's law of word frequency distributions
  - frequency of the *k*th most frequent word is proportional to 1/*k*



Text Classification

- What level of information about language/text can we derive using knowledge of words alone?

Is this Spam?

FROM THE DESK OF MR. CHARLES FRANCIS FEENEY
EMAIL: charlesfeeney55@gmail.com
UNITED NATION AUTHORIZE OFFICE.
DONATION FUNDS PERMIT DEPARTMENT OFFICE.

Attention.

This is to acknowledge the receipt of your details in my payment desk and to confirm your reliability on this donation which I have sent for a deposit a few Months ago. What is going on between you and Mr. Alan Smith' executive manager Citibank New York City regarding your approval of donation funds of $12.9 million us dollars which has been permitted to your name? I have decided to use my Bank Citibank (388 Greenwich St, New York, NY 10013, United States) to disburse this donation funds to your position.

Therefore, I have issued out a bank draft of [$12.900,000 USD MILLION DOLLARS] in your name through my attorney Mrs. Hilda Williams, which has been sent for deposit with my bank in Citibank (388 Greenwich St, New York, NY 10013, United States), the Bank is authorized to conclude the disbursement of the funds via Online Banking or ATM CARD or Bank Draft Check or Bank Wire Transfer, so kindly contact Citibank Email: citib918@gmail.com

You are hereby advised to keep this donation strictly confidential until your claim has been fully recovered before you can share the good news. Kindly contact the bank for instructions of the wire

What is the subject of this medical article?

MEDLINE Article

?

MeSH Subject Category Hierarchy

- Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

14

Positive or negative movie review?

\+ *...zany characters and richly applied satire, and some great plot twists*

\− *It was pathetic. The worst part about it was the boxing scenes...*

\+ *...awesome caramel sauce and sweet toasty almonds. I love this place!*

\− *...awful pizza and ridiculously overpriced...*

Why sentiment analysis?

- *Movie*: is this review positive or negative?
- *Products*: what do people think about a new product?
- *Public sentiment*: how is consumer confidence?
- *Politics*: what do people think about a candidate or issue?

Text Classification: Definition

- *Input*:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
- *Output*: a predicted class $c \in C$

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “you have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Supervised Machine Learning

- *Input:*
  - a document *d*
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of *m* hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$, *i.i.d*
- *Output:*
  - a learned classifier $\gamma:d \rightarrow c$

Linear Classification

- Naïve Bayes (Generative classifier)
- Logistic regression (Discriminative classifier)
- Classification decision based on weighted sum of individual features
  - Word counts are features

Naïve Bayes Classifier

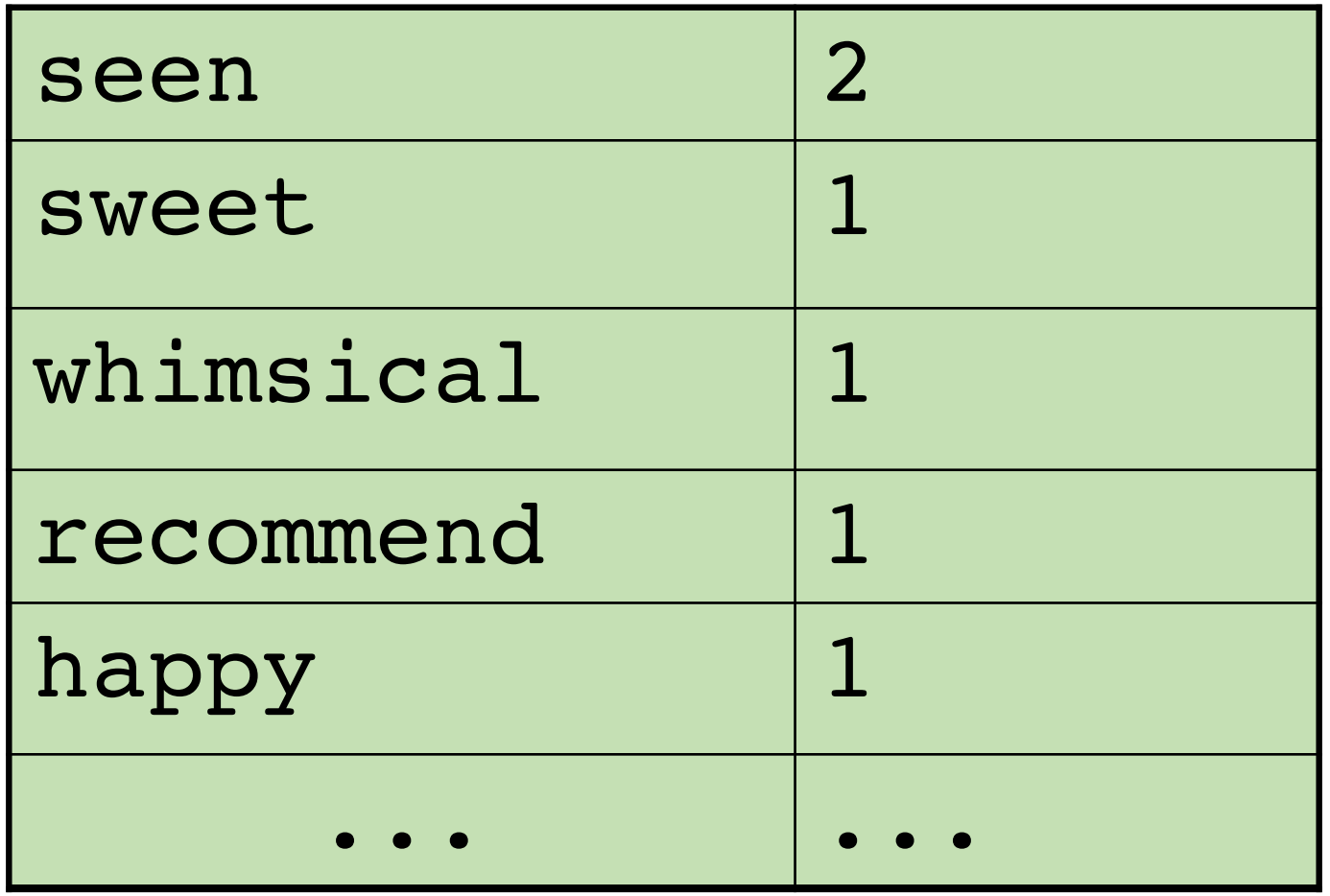
- Simple ("naïve") classification method based on Bayes' rule
- Relies on very simple representation of document
  - **Bag of words**

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Classification using bag of words

$\gamma($ 

seen	2
sweet	1
whimsical	1
recommend	1
happy	1
...	...

$) = c$

Bayes' Rule Applied to Documents and Classes

- For a document *d* and a class *c*

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Naive Bayes Classifier

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

Document d represented as features x1..xn

Multinomial Naive Bayes Independence Assumptions

$$P(x_1, x_2, \ldots, x_n \mid c)$$

- **Bag of Words assumption**: Assume position doesn't matter
- **Conditional Independence**: Assume the feature probabilities $P(x_i|c_j)$ are independent given the class c.

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Multinomial Naive Bayes Classifier

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

Applying Multinomial Naive Bayes Classifiers to Text Classification

positions $\leftarrow$ all word positions in test document

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

Computed in log space

Instead of this: $$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

We have $$c_{\mathrm{NB}} = \underset{c_j \in C}{\operatorname{argmax}} \left[\log P(c_j) + \sum_{i \in \mathrm{positions}} \log P(x_i | c_j) \right]$$

Notes:

1) Taking log doesn't change the ranking of classes!
   The class with highest probability also has highest log probability!
2) It's a linear model:
   Just a max of a sum of weights: a **linear** function of the inputs

Learning the Multinomial Naive Bayes Model

- First attempt: maximum likelihood estimates
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Problem with Maximum Likelihood

- What if we have seen no training documents with the word ***fantastic*** and classified in the topic **positive** (***thumbs-up)***?

$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{count(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$c_{MAP} = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Smoothing for Naïve Bayes

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) \right) + 1)}$$

$$= \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c) \right) + |V|}$$

Multinomial Naïve Bayes: Learning

- From training corpus, extract *Vocabulary*

- Calculate $P(c_j)$ terms
  - For each c_j in C do

    $docs_j \leftarrow$ all docs with class $= c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Unknown words

- What about unknown words that appear in test but not in training data?

- We **ignore** them

Other Features

- Some systems ignore stop words
  - **Stop words:** very frequent words like *the* and *a*.
    - Sort the vocabulary by word frequency in training set
    - Call the top k words the **stopword list**.
    - Remove all stop words from both training and test sets
- But removing stop words doesn't usually help
  - Most NB algorithms use **all** words
- Alternatives to word frequency
  - word occurrence (absence or presence)
  - TF-IDF score

Main Drawback of NB

- Simplistic BoW assumption
  - Disregard word order
    - Two documents with the same words are considered similar
  - Disregard semantic similarity between words
    - Smart, clever, book

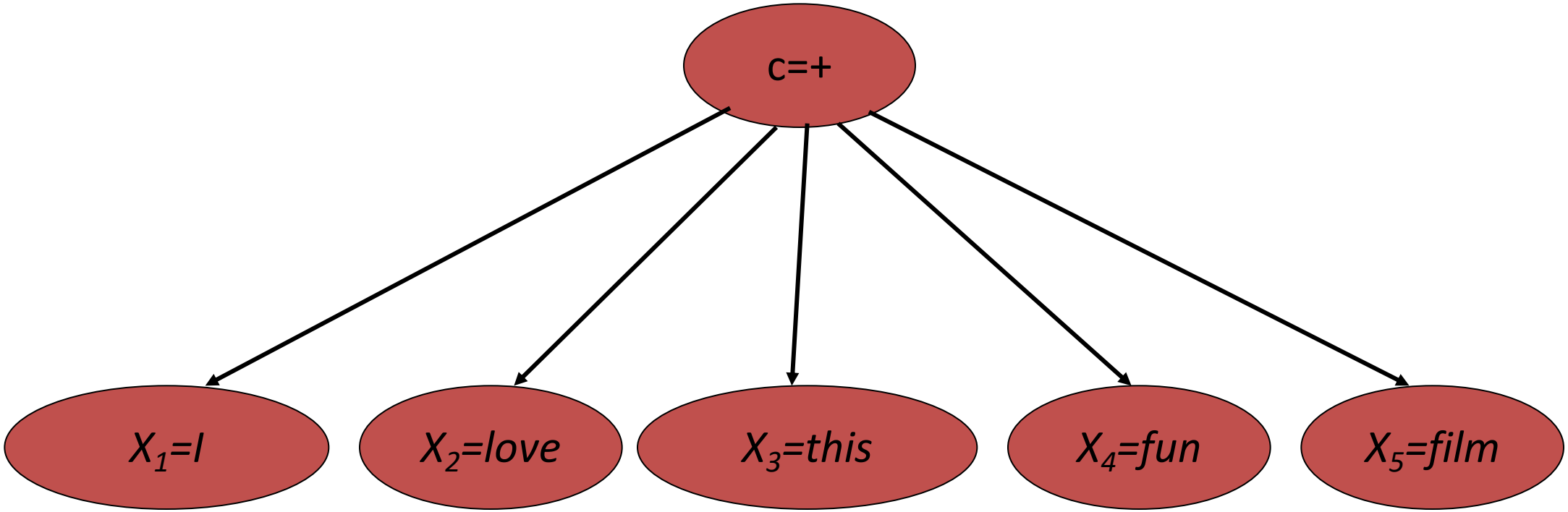
Summary: Naive Bayes is Not So Naive

- Very fast, low storage requirements
- Works well with very small amounts of training data
- Robust to irrelevant features
  Irrelevant features cancel each other
- Very good in domains with many equally important features

- Optimal if the independence assumptions hold
- A good dependable baseline for text classification
  - There are other better performing classifiers

Slide from Chris Manning

Generative Model for Multinomial Naïve Bayes



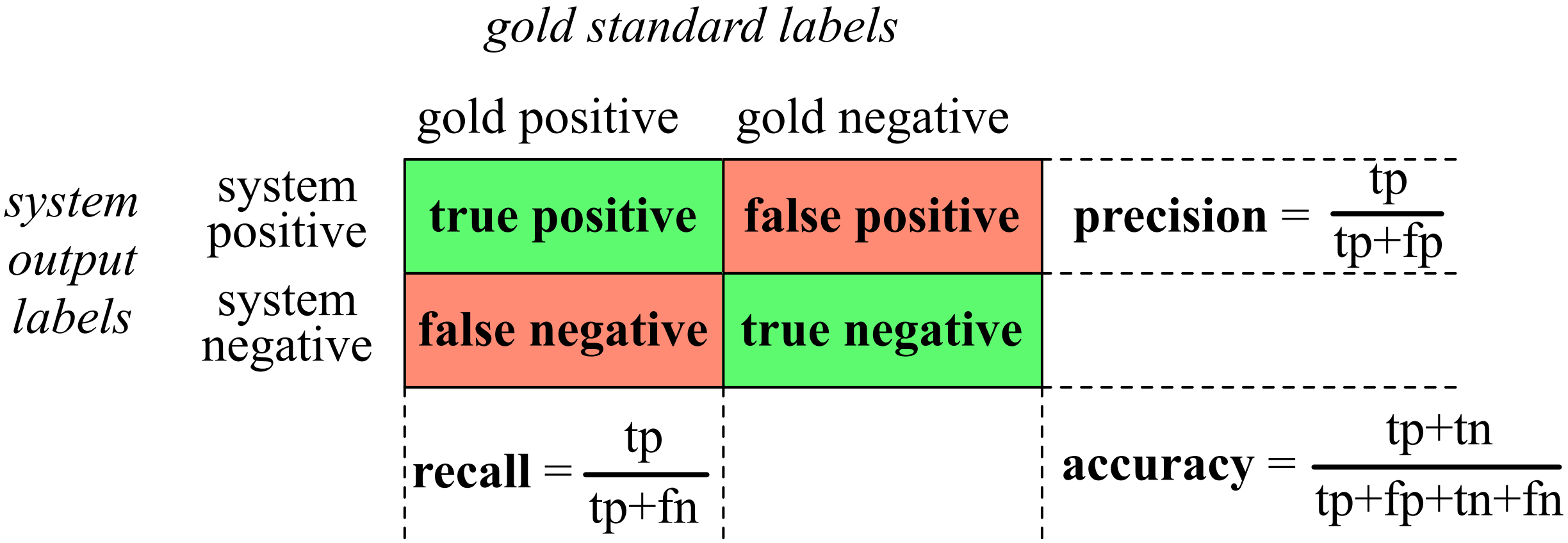
Evaluation

Evaluation

- Let's consider just binary text classification tasks

- Imagine you're the CEO of Delicious Pie Company
  - You want to know what people are saying about your pies
  - You build a "Delicious Pie" tweet detector
    - Positive class: tweets about Delicious Pie Co
    - Negative class: all other tweets

The 2-by-2 confusion matrix



Evaluation

- How good is **accuracy** as our metric?
- Imagine we saw 1 million tweets
  - 100 of them talked about Delicious Pie Co.
  - 999,900 talked about something else
- We could build a dumb classifier that just labels every tweet "not about pie"
  - 99.99% accuracy!
  - But useless! Doesn't return comments of interest
  - Instead use **precision** and **recall**

Evaluation: Precision

- % of items the system detected (i.e., items the system labeled as positive) that are in fact positive (according to the human gold labels)

$$\textbf{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Evaluation: Recall

- % of items actually present in the input that were correctly identified by the system.

$$\textbf{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Why Precision and Recall

- Our dumb pie-classifier
  - Just label nothing as "about pie"

    Accuracy=99.99% but Recall = 0 (doesn't get the Pie tweets)

- Precision and recall, unlike accuracy, emphasize true positives:
  - find the things that we are supposed to be looking for

A combined measure: F

- F measure: a single number that combines P and R:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- We almost always use balanced F_1 (i.e., $\beta = 1$)

$$\mathrm{F}_1 = \frac{2PR}{P + R}$$

Generative and Discriminative Classifiers

- Naive Bayes is a **generative** classifier

- Logistic regression is a **discriminative** classifier

Generative and Discriminative Classifiers

Suppose we're distinguishing cat from dog images

imagenet

imagenet

Generative Classifier:

- Build a model of what's in a cat image
  - Knows about whiskers, ears, eyes
  - Assigns a probability to any image:
    - how cat-y is this image?

Also build a model for dog images

Now given a new image:

Run both models and see which one fits better

Discriminative Classifier

Just try to distinguish dogs from cats

Oh look, dogs have collars!
Let's ignore everything else

Finding the correct class c from a document d: Generative vs Discriminative Classifiers

- Naive Bayes

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overbrace{P(d|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

- Logistic Regression

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overbrace{P(c|d)}^{\text{posterior}}$$

Next Lecture: Logistic Regression