COMP 554 / CSDS 553 Advanced NLP

Faizad Ullah

Basic Probability

Probability

- Fair Coin Toss:
  - Probability of heads: ½ → $P(H) \rightarrow \mathbf{0.5}$
  - Probability of tails: ½ → $P(T) \rightarrow \mathbf{0.5}$

- Fair Coin Toss universe has only two outcomes. There is no other possibility.

Probability

- Fair Dice roll
- Probability of getting a 6: 1/6 → P('6') = 0.16666666666
- All possible outcomes in the current universe are 6.

Joint Probability

- Joint probability refers to a statistical measure that calculates the likelihood of two events occurring together and at the same point in time.

- Suppose we throw a white and black die simultaneously. What is the probability that the outcome would sum to 3?

- (1,2) and (2,1) are the only two out of 36 possibilities that sum to 3.

- So: $P(sums\ to\ 3)$ = 2/36

Conditional Probability

- Conditional probability is known as the possibility of an event or outcome happening, based on the existence of a previous event or outcome.
- Now let us suppose we have already thrown the black dice and got a 2.
- What is the probability of “sums to 3” given this event?
- Only one possibility out of 6 possible outcomes remains.
- So: $P(sums\ to\ 3 \mid already\ a\ 2\ on\ black\ dice) = 1/6$

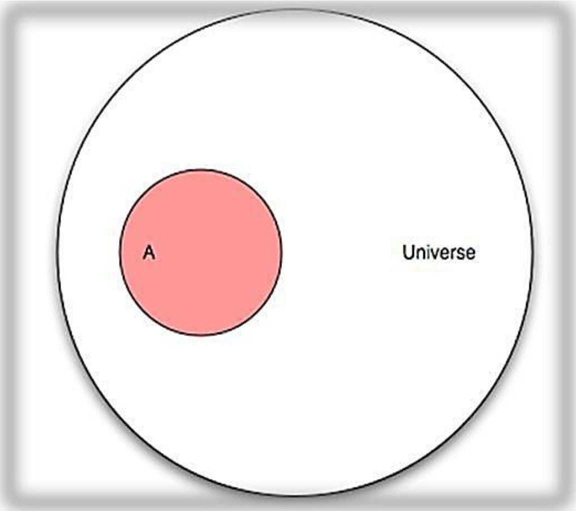
Conditional Probability

- A Universe with all possible outcomes
- Interested in some subset of them (some event)
- Assume we are studying diabetes:
  - We observe people and see whether they have diabetes or not
  - If we take as our Universe, all the people participating in our study, then there are two possible outcomes for any individual: Either they have diabetes, or they do not have diabetes
- We can then split our universe in two events:
  - The event "people with diabetes" (designated as A)
  - The event "people with no diabetes" (designated as $\sim A$)

Conditional Probability

- So, what is the probability that a randomly chosen person has diabetes?
  - The number of elements in A divided by the number of elements in U (universe)
  - We denote the number of elements of A as |A| (the cardinality of A)
  - We define the probability of A, $P(A)$ as:

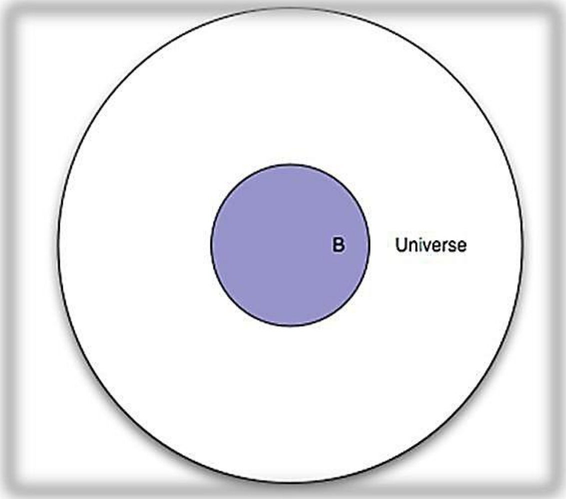
$$\boxed{P(A) = \frac{|A|}{|U|}}$$



- Since A can have at most the same number of elements as U, the probability $P(A)$ can be at most 1.

Conditional Probability

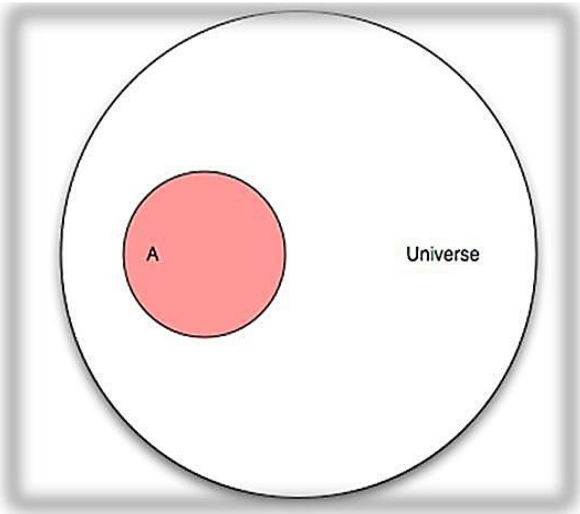
- Let’s say there is a new screening test that is supposed to measure something
- That test will be “positive” for some people, and “negative” for others.
- If we take the event B to be “people for whom the test is positive”
- What is the probability that the test will be “positive” for a randomly selected person?

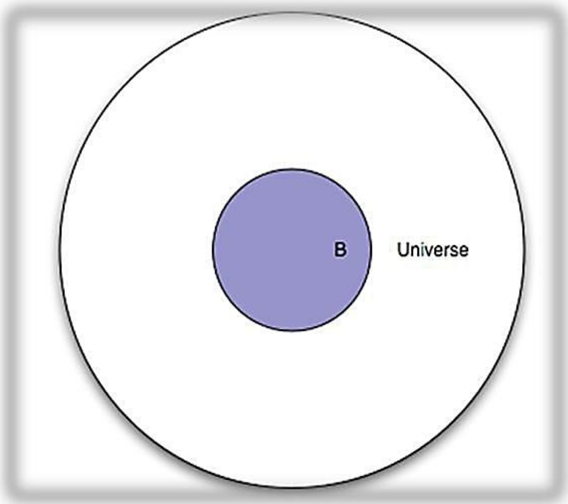


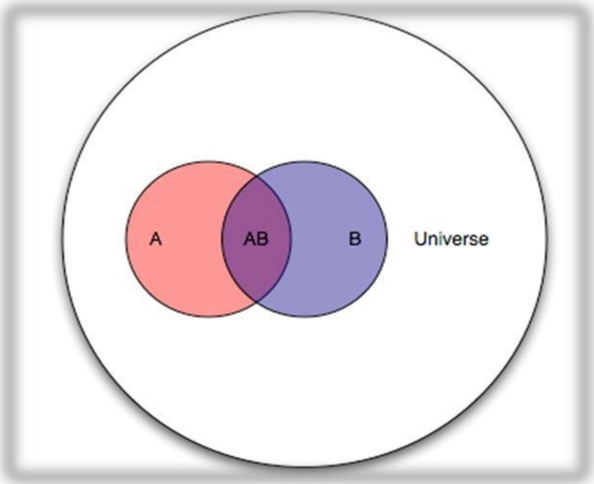
$$\boxed{P(B) = \frac{|B|}{|U|}}$$

The Two Events Jointly

- What happens if we put them together?





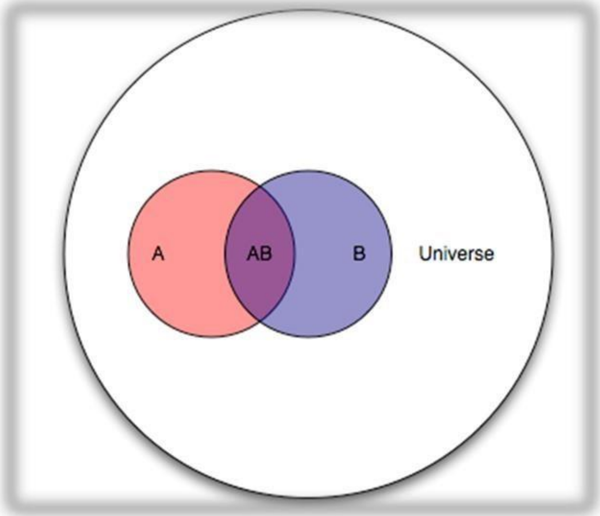


- So, we can compute the probability of both events occurring as:

$$\boxed{P(AB) = \frac{|A \cap B|}{|U|}}$$

The Two Events Jointly

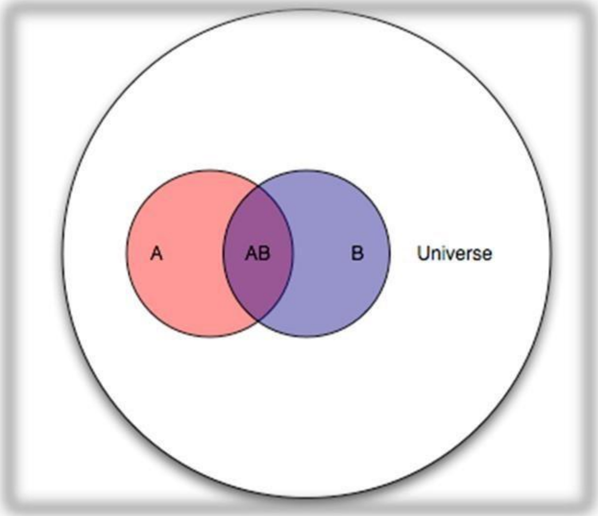
- We are dealing with:
  - An entire Universe (all people)
  - The event A (people with cancer)
  - The event B (people for whom the test is positive)
- There is also an overlap, the event AB ($A \cap B$)
  - "People with diabetes and with a positive test result".
- There is also the event $B - AB$:
  - "People with a positive test result and without diabetes"
- And the event $A - AB$:
  - "People with diabetes and with a negative test result"



Conditional Probability

- “Given that the test is positive for a randomly selected individual, what is the probability that said individual has diabetes?”
- In terms of our Venn Diagram:
  - Given that we are in region B, what is the probability that we are in region AB?

$$\boxed{P(A|B) = \frac{|A \cap B|}{|B|}}$$

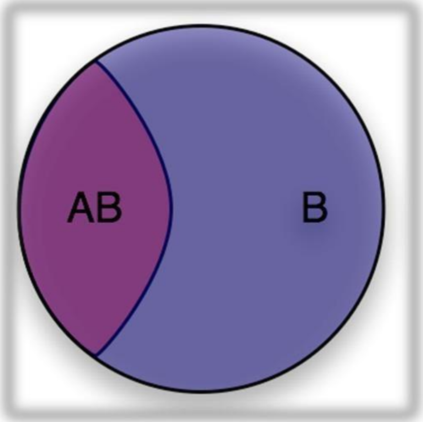


- Or stated differently:
- “If we make region B our new Universe, what is the probability of A?”
- The notation for this is $P(A|B)$ (Probability of A given B)

Conditional Probability

- Let us convert the counts to probabilities
- Dividing both the numerator and denominator by $|U|$, we get:

$$\boxed{P(A|B) = \frac{\frac{|A \cap B|}{|U|}}{\frac{|B|}{|U|}} = \frac{|A \cap B|}{|B|}} = P(AB)/P(B) \rightarrow \text{Equation 1}$$



- What we've effectively done is change the Universe from U (all people) to B (people for whom the test is positive), but we are still dealing with probabilities defined in U

Conditional Probability

- Now let’s ask the converse question:
  - “given that a randomly selected individual has cancer (event A), what is the probability that the test is positive for that individual (event AB)?

$$\boxed{P(B|A) = \frac{|A \cap B|}{|A|}} = P(AB)/P(A) \rightarrow \text{Equation 2}$$

The Bayes Theorem

- Now we have everything we need to derive Bayes theorem, putting equation 1 and 2 together, we get:

$$\boxed{\boldsymbol{P(A|B)P(B) = P(B|A)P(A)}}$$

- Which is to say $P(A \cap B)$ is the same whether you're looking at it from the point of view of A or B.

$$\boxed{\boldsymbol{P(A|B)P(B) = P(B|A)P(A)}}$$

$$\boxed{\boldsymbol{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}}$$

The Bayes Theorem

Posterior

Likelihood

Prior

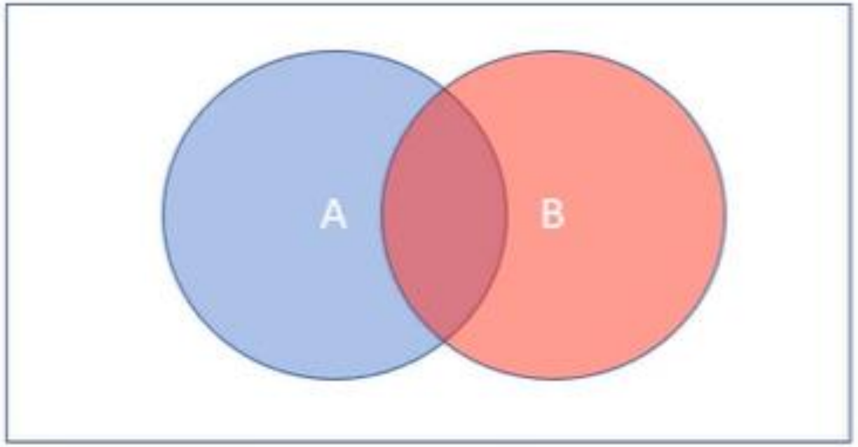
$$P(h\,|D) = \frac{P(D|h) \times P(h)}{P(D)}$$

Evidence of Predictive Probability

Independence

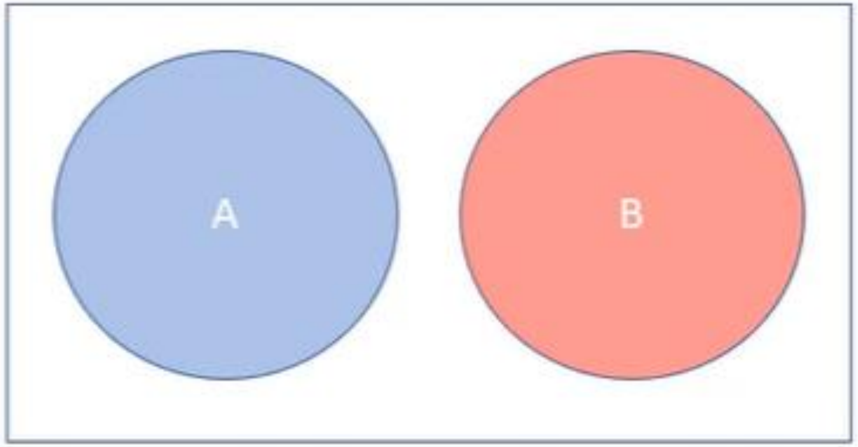
- If the probability of occurrence of an event A is not affected by the occurrence of another event B, then A and B are said to be independent events.
  - A = “Today is Friday”
  - B = “Heads on fair coin”
- If A and B are independent:
  - $P(A \cap B) = P(A)P(B)$
- Or stated a bit differently:
  - $P(A|B) = P(A)$ if $P(B) > 0$ and $P(B|A) = P(B)$ if $P(A) > 0$
  - $P(A|B) = P(A \cap B) / P(B)$ is not defined when $P(B) = 0$
  - $P(A|B) = P(A \cap B) / P(A)$ is not defined when $P(A) = 0$

Independence and Mutual Exclusion



Not Independent

$$P(A|B) \neq P(A)$$
$$P(B|A) \neq P(B)$$
$$P(A \cap B) = P(A).P(B|A)$$
$$P(A \cap B) = P(B).P(A|B)$$
$$P(A \cap B) \neq P(A).P(B)$$



Independent? / Mutually exclusive?

$$P(A \cap B) = 0$$

For independent events:

$$P(A \cap B) = P(A).P(B)$$

Mutually exclusive events A and B are independent if and only if

$$P(A) = 0 \; or \; P(B) = 0$$

Otherwise, A and B are **not independent**

Also,

$$P(A|B) = 0 \neq P(A), and \; P(B|A) = 0 \neq P(B)$$

Summary

- For independent events A and B:

$$P(AB) = P(A)P(B)$$
$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

- For independent events A b and C:

$$P(ABC) = P(A)P(B)P(C)$$
$$P(AB|C) = P(AB) = P(A)P(B)$$
$$P(BC|A) = P(BC) = P(B)P(C)$$

- For dependent event A and B:

$$P(AB) = P(A|B).P(B) = P(B|A).P(A)$$
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- For dependent events A, B and C:

$$P(ABC) = P(A)P(B|A)P(C|AB)$$

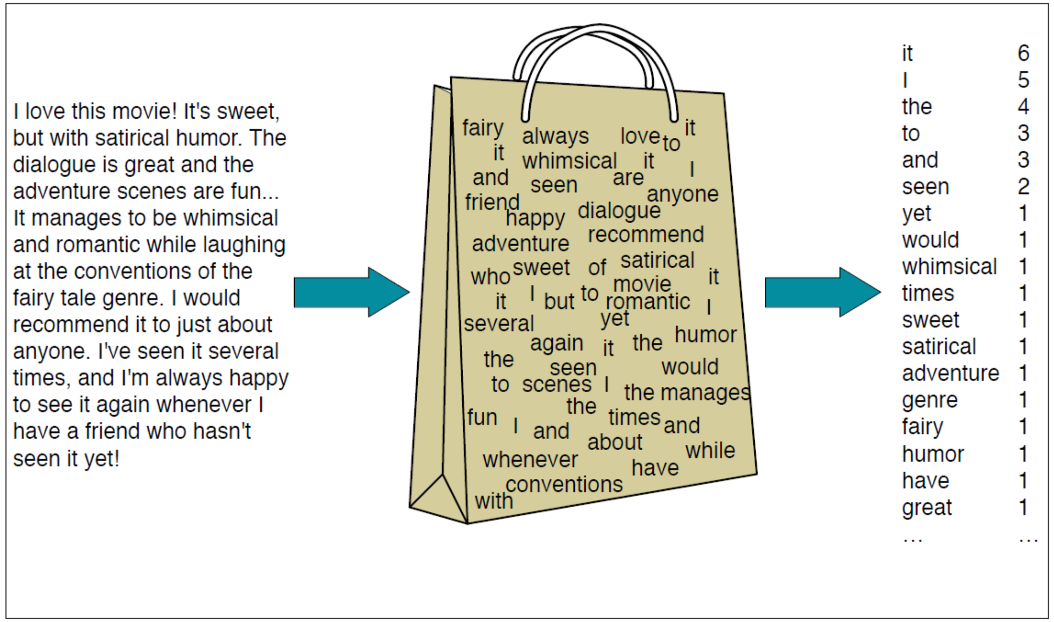
Naïve Bayes Classifier

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
…	…

Naïve Bayes Classifier

- Naïve Bayes is a probabilistic classifier, meaning that for a document d, out of all classes c \in C the classifier returns the class \hat{c} which has the maximum posterior probability given the document d.



Naïve Bayes Classifier

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

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

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

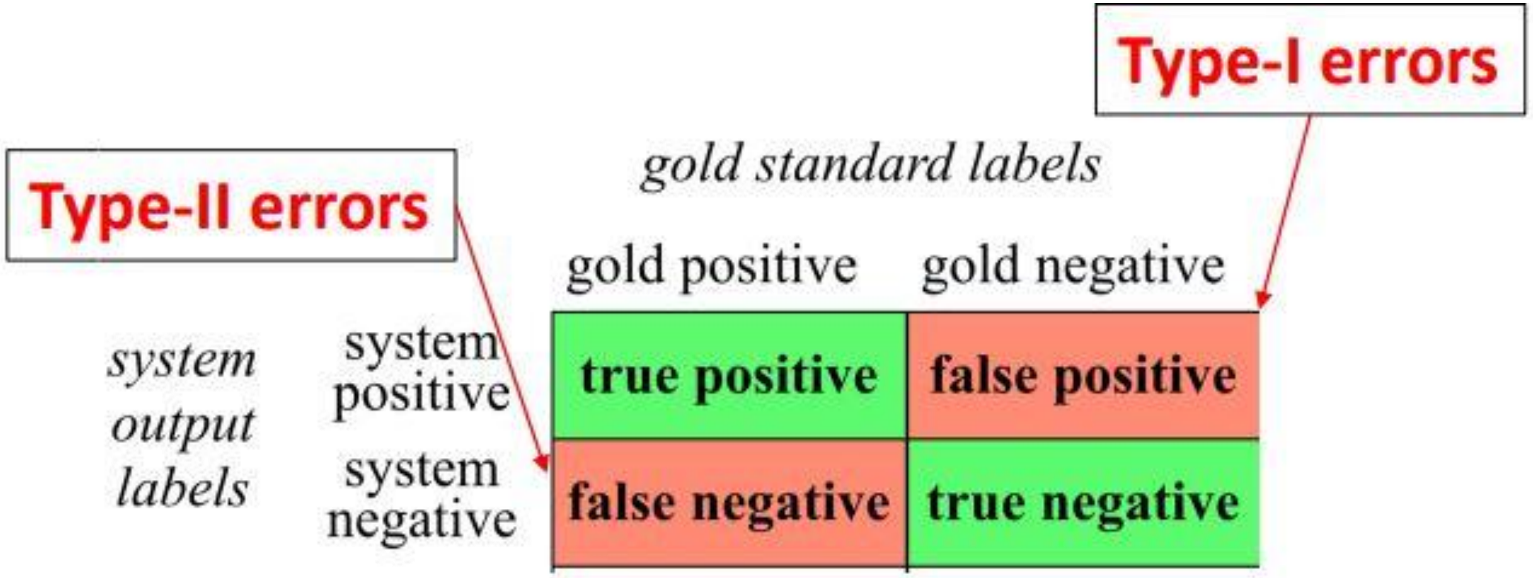
Confusion Matrix

Sentiment Analysis/Classification Task

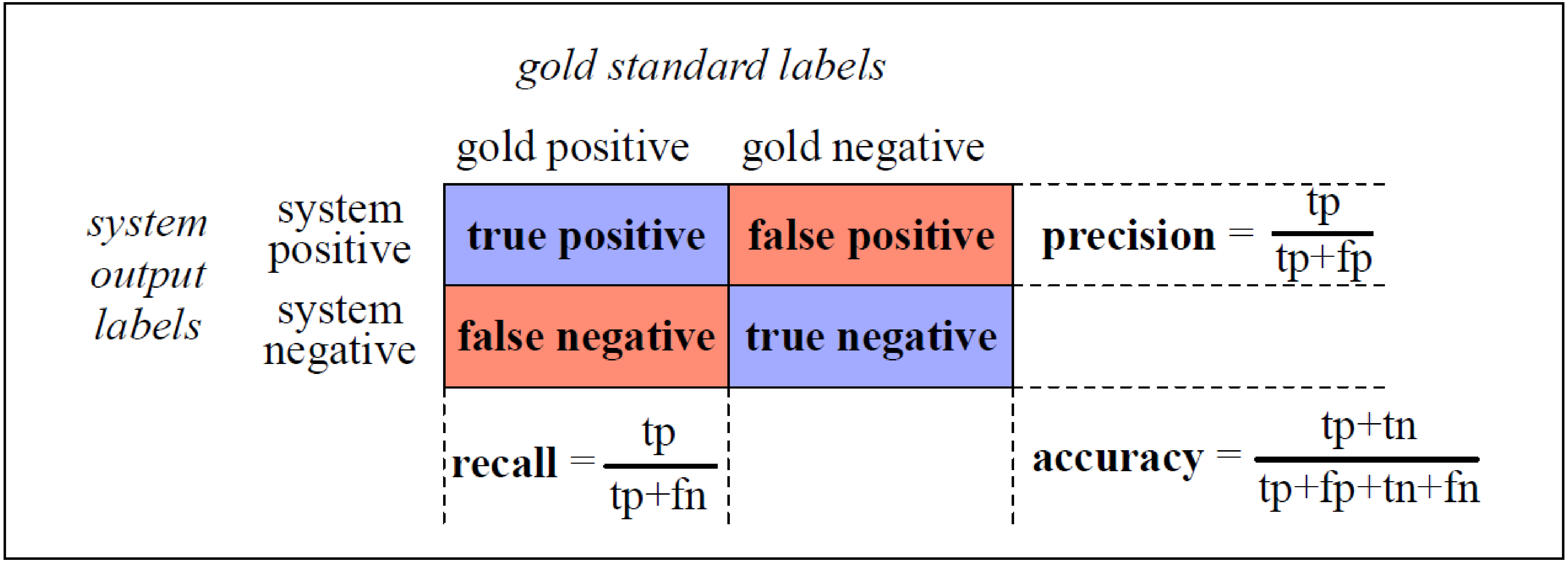
Sentiment Analysis Example:

Tweet	Actual Label	Predicted label

- Type I Error = False Positive
- Type II Error = False Negative


Type-I errors
Type-II errors
gold standard labels
gold positive
gold negative
system output labels
system positive
system negative
true positive
false positive
false negative
true negative

Confusion Matrix



- Accuracy: $\frac{\textit{My Correct Answers}}{\textit{All Questions}} = \frac{tp + tn}{tp+tn+fp+fn}$
    - What fraction of time am I correct in my classification
- Precision $\frac{\textit{True Positives}}{\textit{My Positives}} = \frac{tp}{tp+fp}$
    - How much should you trust me when I say that something tests positive
    - What fraction of my positives are true positives
- Recall = Sensitivity $\frac{\textit{True Positives}}{\textit{Real Positives}} = \frac{tp}{tp+fn}$
    - How much of the reality has been covered by my positive output?
    - What fraction of the true positives is captured by my positives?
- Specificity $\frac{\textit{True Negatives}}{\textit{Real Negatives}} = \frac{tn}{tn+fp}$
    - How much of the reality has been covered by my negative output?
    - What fraction of the true negatives is captured by my negatives?

Accuracy

$$Accuracy \frac{tp + tn}{tp + fp + tn + fn}$$

Correct predictions over all predictions

A Real Example 1 (Spam vs. Not Spam)

Issues with Precision and Recall

- One possible way may be to combine both.
- But, how to combine Precision and Recall?
- Average?

x0	x1	x2	x3	x4	x5	x6	x7	x8	AM	GM	HM
1	2	3	4	5	6	7	8	9	5.00	4.15	3.18
2	4	8	16	32	64	128	256	512	113.56	32.00	9.02
5	5	5	5	5	5	5	5	5	5.00	5.00	5.00
5	5	5	5	5	5	5	5	**10**	5.56	5.40	5.29
5	5	5	5	5	5	5	5	**100**	15.56	6.97	5.59
5	5	5	5	5	5	5	5	**1000**	115.56	9.01	5.62
5	5	5	5	5	5	5	5	**10000**	1115.56	11.63	5.62
5	5	5	5	5	5	5	5	**100000**	11115.56	15.03	5.62
5	5	5	5	5	5	5	100000	**100000**	22226.11	45.16	6.43
5	5	5	5	5	100000	100000	100000	**100000**	44447.22	407.89	9.00
5	100000	100000	100000	100000	100000	100000	100000	**100000**	88889.44	33274.21	44.98
100000	100000	100000	100000	100000	100000	100000	100000	**100000**	100000.0	100000.0	100000.0

F-1-MEASURE

- The harmonic mean of P and R:
  - Is high when both P and R are high.
  - Is low when even one of P and R is low.
- A combined measure that assesses the P/R tradeoff is the F-measure (weighted harmonic mean of precision and recall)

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R}$$

Precision	Recall	F-1
0	1	0
0.1	0.9	0.18
0.2	0.8	0.32
0.3	0.7	0.42
0.4	0.6	0.48
0.5	0.5	0.5
0.6	0.4	0.48
0.7	0.3	0.42
0.8	0.2	0.32
0.9	0.1	0.18
1	0	0
1	1	1
0.5	1	0.666667
1	0.5	0.666667
0.1	1	0.181818
1	0.1	0.181818

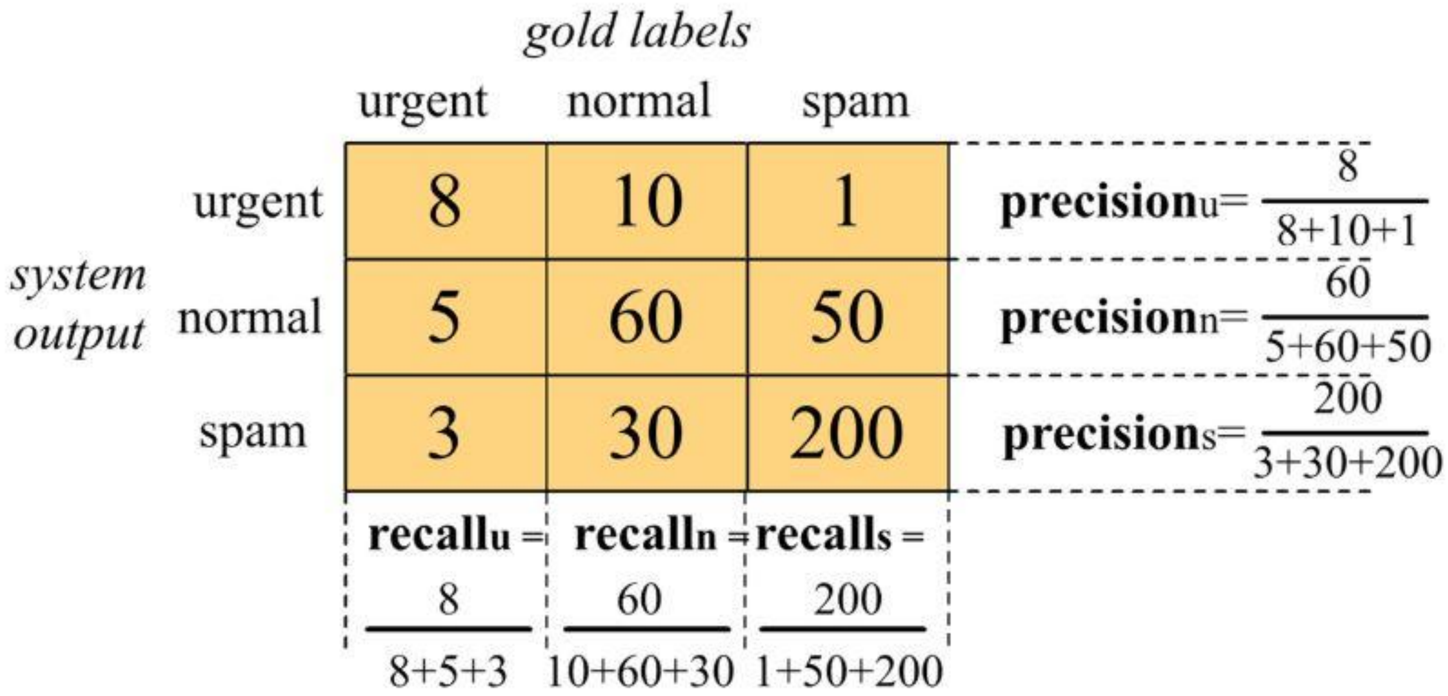
More than 2 Classes

More than two classes

- Lots of classification tasks in language processing have more than two classes:
  - Sentiment analysis (positive, negative, neutral),
  - Part-of-speech tagging (|POS tags|)
  - Emotion detection (|emotions|)

Evaluation

- one-of email categorization decision (urgent, normal, spam)



Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?
    1. Macro-averaging: Compute performance for each class, then average.
    2. Micro-averaging: Collect decisions for all classes, compute contingency table, evaluate

		gold labels urgent	*gold labels* normal	*gold labels* spam	
system output	urgent	8	10	1	**precision**u= $\frac{8}{8+10+1}$
system output	normal	5	60	50	**precision**n= $\frac{60}{5+60+50}$
system output	spam	3	30	200	**precision**s= $\frac{200}{3+30+200}$
		recallu = $\frac{8}{8+5+3}$	**recall**n = $\frac{60}{10+60+30}$	**recall**s = $\frac{200}{1+50+200}$	

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Evaluation

- A micro-average is dominated by the more frequent class (in this case spam)
  - The counts are pooled
- The macro-average better reflects the statistics of the smaller classes
  - More appropriate when performance on all the classes is equally important.

Sources

- https://web.stanford.edu/~jurafsky/slp3/2.pdf
- https://web.stanford.edu/~jurafsky/slp3/3.pdf
- **Machine Learning for Intelligent Systems**, Kilian Weinberger, Cornell, Lectures 3-6, https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecture note03.html
- **Prof. Mitesh M. Khapra** (https://www.cse.iitm.ac.in/~miteshk/) on NPTEL's (http://nptel.ac.in/) Deep Learning course (https://onlinecourses.nptel.ac.in/noc18_cs41/preview)
- **Perceptrons. An Introduction to Computational Geometry. Marvin Minsky and Seymour Papert. M.I.T. Press, Cambridge, Mass., 1969.** https://science.sciencemag.org/content/165/3895/780