Lab 3
CS114 Spring 2018
Probability Theory I: (Tabular) Probability Distribution Function: Joint, Conditional, and Marginal Probabilities

Computational linguistics is now mostly dominated by probabilistic methods (as opposed to rule-based deterministic relationships). Probability is the key framework you will need to understand the modern techniques in computational linguistics and natural language processing. In this lab, we will review and cover not-so-basic probability theory that actually gets used a lot in computational linguistics.

Probability distribution function

It is a function that shows how the probability cake is distributed among all of the possible values. For example, let X be a random variable that models the part of speech of an English word. The probability distribution function (PDF) of X might look as follows:

X	**P(X)**
Verb	0.4
Noun	0.3
Preposition	0.1
Adjective	0.1
Adverb	0.1

One must note the properties of probability distributions:

1. All of the possible values must be mutually exclusive.
2. The probability values must be between 0 and 1.
   Formally, $0 \leq P(X=x) \leq 1 \; ; \forall \, x$
3. The sum of the probabilities over all possible values must be 1.
   Formally, $\sum_{x}^{\square} P(X=x) = 1$

In short, no crazy values can go in these tables.

Joint probability distribution function

We are still trying to model an English word. Now we add one more variable to each word and model the two variables jointly. Let Y be a binary random variable that indicates whether the word is capitalized or not.

X	**Y**	**P(X,Y)**
Verb	Capitalized	0.1
Noun	Capitalized	0.25
Preposition	Capitalized	0.04
Adjective	Capitalized	0.03
Adverb	Capitalized	0.04
Verb	Not	0.3
Noun	Not	0.05
Preposition	Not	0.06
Adjective	Not	0.07
Adverb	Not	0.06

The properties are the same, but we have to consider ALL POSSIBLE VALUES instead.

1. All of the possible values must be mutually exclusive.
2. The probability values must be between 0 and 1.
   Formally, $0 \leq P(X=x, Y=y) \leq 1; \forall x \forall y$
3. The sum of the probabilities over all possible values must be 1.
   Formally, $\sum_{x}^{\square} \sum_{y}^{\square} P(X=x, Y=y) = 1$

Again, no crazy values can go in these tables.

Note that crazy things can happen: number of parameters

X has 5 possible values, and Y has two possible values. The number of parameters for P(X,Y) is 5 x 2. So the table can get really big and unwieldy. We will see later that it will cause us trouble.

Suppose we are modeling four random variables: S, T, X, and Y. Each variable has five possible values. How many parameters does P(S,T,X,Y) have? It's 5 x 5 x 5 x 5 = 625. It gets huge fast.

Marginal distribution

Marginalization is when you want to get rid of one of the variables in your joint PDF. To continue with example above, if you want to extract P(Y) from P(X,Y), you have to marginalize out X.

Sum up the yellow part and sum up the blue part.

X	Y	P(X,Y)
Verb	Capitalized	0.1
Noun	Capitalized	0.25
Preposition	Capitalized	0.04
Adjective	Capitalized	0.03
Adverb	Capitalized	0.04
Verb	Not	0.3
Noun	Not	0.05
Preposition	Not	0.06
Adjective	Not	0.07
Adverb	Not	0.06

The resulting P(Y) is below:

Y	P(Y)
Capitalized	0.1 + 0.25 + 0.04 + 0.03 + 0.04
Not	0.3 + 0.05 + 0.06 + 0.07 + 0.06

Simple enough right? Formally, it's written as $P(Y=y)=\sum_{x}^{\square} P(X=x,Y=y)$. You sum up all of the rows that have Y=y.

Great and where's the useful part?: Conditional Probability

Suppose we observe that a word is capitalized. What's the probability that the word is a verb? Conditional probability gives us the ability to infer the unknown (e.g. part of speech) from the information you have (capitalization or not). Formally,

$$P(X=verb|Y=capitalized)=\frac{P(X=verb,Y=capitalized)}{P(Y=capitalized)}$$

It is easy to enough to plug in the formula and get the answer for this. But let's look at it a little more closely. What's the difference between P(X|Y) and P(X,Y)?

Dropping stuff behind the bar: Independence

If two random variables are independent, it means that knowing one does not tell us anything about the other. They are not correlated. More formally, if X and Y are independent,

$$P(X|Y)=P(X)$$

It does not matter whether we know the value of Y or not, the probability of X is the same. Once you assume independence, you can drop stuff from the bar.

Another super useful part: Chain Rule of probability + Independence

The chain rule makes modeling easy because we don't need to deal with a huge probability table. To illustrate this, suppose we are modeling four random variables: S, T, X, and Y. If we don't assume independence at all, and S, T, X, and Y each have five different possible values, then we have 625 rows in the table. That's big…. Here's how you apply the chain rule.

1. Split the variables into two groups called *front* and *back*
   Suppose *front* = {S, Y} and *back* = {T, X}
2. Then P(S,T,X,Y) = P(*front* | *back)* P(*back*) = P(S,Y | T,X) P(T,X)

You can further apply the Chain Rule to P(S,Y|T,X) by not moving the variables behind the bar

1. Split S,Y only into two groups as usual. You are not allowed to touch T, X because they are behind the bar already.
   Suppose *front* = S and *back* = Y
2. then P(S,Y| T,X) = P(*front* | *back*, T, X) P(*back* | T, X) = P(S| Y,T,X) P(Y|T,X)

Putting it all together

$$P(S,T,X,Y)=P(S,Y|T,X)P(T,X)=P(S|Y,T,X)P(Y|T,X)P(T,X)$$

It's still a lot of stuff. If we assume that S is independent of Y, T, and X then $P(S|Y,T,X)=P(S)$ and we can make the joint distribution simpler.