Probabilities and Language Models

Kevin Duh

Lecture Notes for Natural Language Processing, Fall 2019

1 Probability Basics

Natural language processing involves ambiguity resolution. Probability and statistics are effective frameworks to tackle this. Here, we will define some basic concepts in probability required for understanding language models and their evaluation.

Definitions

- Event space $\mathcal{E}$: the set of all possible outcomes
- Event $E \subseteq \mathcal{E}$: a subset of the event space
- Probability $P(E)$: a non-negative number assigned to an event, indicating it's probability of occurrence

	Event Space $\mathcal{E}$	Event E	Probability $P(E)$
The output of a coin flip	{Head, Tail}	{Head} {Tail}	$P(\{Head\}) = 0.5$ $P(\{Tail\}) = 0.5$
The output of a dice roll	{1, 2, 3, 4, 5, 6}	{1} {1,3,5}	$P(\{1\}) = 1/6$ $P(\{1,3,5\}) = 0.5$
The word after "Hi, my name..."	{is, was, ...} all vocabulary	{is} {was}	$P(\{is\}) = 0.7$ $P(\{was\}) = 0.1$
A random sentence from Shakespeare	all sentences		P({"To be or not to be"}) = 1E-5 P({"That's all, folks!"}) = 1E-9

Table 1: Example of events and their probability

Axioms of Probability

1. $P(E) \geq 0$ for all events E. Each subset of event space is assigned some probability of occurrence (may be zero).
2. $P(\mathcal{E}) = 1$.

3. If two E_1 and E_2 are disjoint, then $P(E_1 or E_2) = P(E_1) + P(E_2)$. Dice roll example: $P(\{1,3,5\}) = P(\{1\}) + P(\{3\}) + P(\{5\}) = 1/6 + 1/6 + 1/6 = 0.5$. Extends to $P(\cup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$ for mutually-exclusive E_i's.

Random Variable A random variable X is a variable whose possible values are the outcomes of some random trial, e.g. a coin flip. This can be discrete (e.g. Head or Tail) or continuous (e.g. any real number between 0 and ∞).

A probability mass function $p(x)$ is a function that assigns probability when a discrete random variable X is equal to some value x.

$$p(x) \stackrel{\text{def}}{=} P(X = x) \tag{1}$$

Dice roll example: Let x be 1. $p(1) = P(X = 1) = 1/6$; this is the probability we assign to the event $X = 1$. We also call $p(x)$ the probability distribution. For continuous random variables, we will be using a probability density function $f(x)$ to characterize the distribution; it defines probability on intervals: $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Later in the course, we will be exploring many ways to "learn" a probability distribution from data. This distribution might be a multinomial distribution (which can be viewed as a table, with $p(x)$ values for each x), or a some log-linear parameterization, or some output of a softmax layer after a neural network. In all cases, we'll be using notation $x \sim Distribution(\theta)$ which means x is a sample drawn from the distribution with parameters θ.

Joint Probability Sometimes we care about the probability of two or more variables, and they might interact. This can be modeled by a joint probability of e.g. two variables X and Y:

$$p(x, y) \stackrel{\text{def}}{=} P(X = x, Y = y) \tag{2}$$

Say we flip two coins, where X represents the output of the first coin and Y represents that of the second coin. Our event space is:

$$\mathcal{E} = \{(Head, Head), (Head, Tail), (Tail, Tail), (Tail, Head)\} \tag{3}$$

and the probabilities are each $1/2 \times 1/2 = 1/4$, e.g. $p((Head, Tail)) = 1/4$. In this case, the two events do not influence each other, so the probability of both occurring is simply the product of their probabilities: $p(x, y) = p(x)p(y)$. We say that X and Y are **independent**.

Now consider a weather example, where we have two variables: X for rain and Y for cloud. Each event has two possible outcomes: {True, False}, so there are four possible outcomes in the joint event space. Suppose we observe, out of 12 days, the outcomes shown in Table 2.

We can compute the marginal probability from the joint probability by summing out the variable that is not considered: $p(x) = \sum_y p(x, y)$ and $p(y) = \sum_x p(x, y)$. In the weather example: P(Rain=True)

= P(Rain=True and Cloud=True) + P(Rain=True and Cloud=False) = 3/12 + 0/12 = 3/12.

Joint Probability	Cloud=True	Cloud=False	Marginal p(Rain)
Rain=True	3/12	0/12	3/12
Rain=False	1/12	8/12	9/12
Marginal p(Cloud)	4/12	8/12	

Table 2: Weather example: Joint and Marginal probabilities for Rain and Cloud

Definition of Independence Two variables are independent if and only if (iff) their joint probability is a product of the marginal probabilities. We see this is not the case for the weather example, e.g. P(Rain=True) $\times$ P(Cloud=True) $= 3/12 \times 4/12 \neq$ P(Rain=True and Cloud=True). Naturally, rain can't happen without clouds.

Conditional Probability Continuing from the weather example, we can ask "What's the probability of rain, given that we already know it is cloudy?" This is a conditional probability, which we would write as P(Rain=True | Cloud=True). Formally,

$$p(x \mid y) \stackrel{\text{def}}{=} \frac{p(x,y)}{p(y)} \tag{4}$$

One can interpret this definition as restricting the sample space to situations where the conditioning factor occurs.[1] So P(Rain=True | Cloud=True) = P(Rain=True and Cloud=True) $\div$ P(Cloud=True) $= 3/12 \div 4/12 = 3/4$.

Note that if two variables are independent, the $p(x \mid y) = \frac{p(x)p(y)}{p(y)} = p(x)$. This means knowledge of y does not tell us anything about x.

Chain Rule The conditional probability definition (Eq. 4) can be extended to more variables. First, start with $p(x \mid y, z) \stackrel{\text{def}}{=} \frac{p(x,y,z)}{p(y,z)}$. Also, $p(y \mid z) \stackrel{\text{def}}{=} \frac{p(y,z)}{p(z)}$. Putting the two together, we get:

$$p(x, y, z) = p(x \mid y, z)p(y \mid z)p(z) \tag{5}$$

In general, the chain rule says that a joint probability can be decomposed into a series of conditional probabilities like the form above.

[1] Take as another example the discussion of random sentence generation by PCFG in previous lectures. We have a grammar that generates many types of sentences, $s_1, s_2, s_3, \ldots$. At the same time, the sentence s_1 may have been generated by two different trees t_a and t_b. What if we want to know the probability of a specific tree t_a given s_1? We would calculate the conditional probability $p(t_a|s_1) = \frac{p(t_a,s_1)}{p(s_1)} = \frac{p(t_a,s_1)}{p(t_a,s_1)+p(t_b,s_1)}$. The division by $p(s_1)$ re-normalizes the probabilities to focus on cases where s_1 occurred. $p(t_a, s_1)$ and $p(t_b, s_1)$ can be obtained from our PCFG by multiplying the rule probabilities.

Bayes Rule Bayes Rule says:

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} \tag{6}$$

This can be derived from our definition of conditional probability. First, from Eq. 4, we get $p(x, y) = p(x|y)p(y)$ and $p(y, x) = p(y|x)p(x)$. Next, we know $p(x, y) = p(y, x)$, so $p(x|y)p(y) = p(y|x)p(x)$. Finally, assuming non-zero probabilities, we can divide through by $p(y)$ to get: $p(x|y) = p(y|x)p(x)/p(y)$.

Bayes Rule can be viewed as a way to link $p(x \mid y)$ and $p(y \mid x)$. In this case, we'll call $p(x)$ the prior, $p(y \mid x)$ the likelihood, and $p(x \mid y)$ the posterior. Now, can you compute P(Cloud=True | Rain=True) using Bayes Rule?

2 Entropy and Perplexity

Definition of Information Let E be some event with occurs with probability $P(E)$. If I told you that E occurred, then I have given you

$$I(E) = -\log_2 P(E) \tag{7}$$

bits of information. Why does this make sense to define information in terms of probabilities? Suppose you know that it almost never rains but I tell you that it just rained. That's is a valuable information. But if it always rains and I tell you it just rained, that's little information for you. One can think of information as the amount of surprise when observing E, or alternatively, the amount of bits needed to communicate about E.

More examples. How much information is in...

- the outcome of a fair coin flip? $-\log_2 1/2 = 1$ bit. I can use bit 0 to tell you that the outcome it's Head and bit 1 to tell you the outcome is Tail.
- the outcome of two fair coin flips? $-\log_2 1/4 = 2$ bits. Note that information is additive.
- the outcome of a dice roll? $-\log_2 1/6 = 2.59$ bits.
- drawing a random word from a 10000 vocab? $-\log_2 1/10000 = 13.29$ bits.

Entropy Now, suppose we have a distribution $p(x)$ with K possible values $x_1, x_2, \ldots, x_K$ in its event space. What is the **average amount of information** in observing the samples from this distribution?

$$H(p) = \sum_{k=1}^{K} p(x_k)I(x_k) = -\sum_{k=1}^{K} p(x_k)\log_2 p(x_k) \tag{8}$$

This is called entropy in information theory, and it is a function fully characterized by the distribution $p(x)$. One can think of entropy as the average amount of surprise of a distribution. Note that entropy is by definition non-negative.

What kind of distribution has higher entropy: one with uniform probabilities, or one with skewed probabilities?

Cross-Entropy Let's say we don't know the true distribution (denoted $p^*(x)$) that generated our data. We have some model (denoted $p(x)$) that attempts to approximate this distribution. We want to know how good the model is. Analogous to our definition of entropy, we can answer this by asking how surprised our model is on average when tested on data generated by $p^*(x)$.

$$H(p^*, p) = \lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} p^*(x_m) I(x_m) = \lim_{M\to\infty} -\frac{1}{M} \sum_{m=1}^{M} p^*(x_m) \log_2 p(x_m) \tag{9}$$

This is the cross-entropy of p on p^*. In practice, we can't sample from p^* forever ($M \to \infty$), so we approximate with M finite samples $x_1, x_2, \ldots, x_M$ drawn randomly from p^*, leading to $H(p^*, p) \approx -\frac{1}{M} \sum_{m=1}^{M} \log_2 p(x_m)$. In the context of language models, this means preparing some test sentences and measuring our language model probabilities on them.

Perplexity Perplexity is simply defined as $2^{cross-entropy}$. In other words, given some test set of $x_1, x_2, \ldots, x_M$ samples, the perplexity (PPL) of a model p is calculated by:

$$PPL = 2^{-\frac{1}{M} \sum_{m=1}^{M} \log_2 p(x_m)} \tag{10}$$

A nice interpretation of perplexity is that it measures the "branching factor." Cross-entropy measures uncertainty in terms of bits, but when exponentiated it measures the size of an equally weighted distribution with uniform probability. Consider an unfair dice with entropy of 2.32 bits rather than $-\log_2 1/6 = 2.59$ bits; its perplexity is $2^{2.32} \approx 5$, which means it has the same amount of information as a 5-sided dice with equal probabilities ($-\log_2 1/5 = 2.32$).