Basic Probability Theory

Anjali Bhavan

07-10-2020

1 Basic concepts

- Some terms: possible outcomes, sample space etc.
- Random variables
- Cond. Probability

2 Some Terms

- outcome: possible result of an experiment
- sample space: set of all possible outcomes
- event: a set of outcomes of an experiment (subset of sample space)
- event space: set of all possible events (power set of sample space?)

3 Random Variables

Definition 1 (Random variables). *A random variable is a function from the sample space to R:* $\Omega \rightarrow \mathbb{R}$

There are two types of random variables: discrete and continuous.

- Discrete:
  (a) Uniform: all outcomes have same probability
  (b) Bernoulli: two possible outcomes
  (c) Binomial: two possible outcomes, n trials
  (d) Multinomial: k possible outcomes, n trials
  (e) Poisson: big n, small p approx. of binomial
- Continuous:
  (a) continuous uniform distribution
  (b) Gaussian (normal distribution)

Definition 2 (Random vector). *Finite dimensional vector of random variables:* $X = [X_1, ..., X_k]$.
$P(x) = P(x_1, ..., x_n) = P(X_1 = x_1, ...X_n = x_n))$

4 Probability

Three types:

- Joint Probability: prob. of $X = x$ and $Y = y$ happening together
- Conditional Probability: prob. of $X = x$ given $Y = y$
- Marginal Probability: prob. of $X = x \forall Y$.

Chain rule: Calculate joint prob from marginal and condl. prob

$$P(A, B) = P(A) * P(B|A) = P(B) * P(A|B) \tag{1}$$

Calculating marginal prob from joint prob:

$$P(A) = \sum_B P(A, B) \tag{2}$$

Bayes' Rule:

$$P(B|A) = \frac{P(A, B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} \tag{3}$$

A will be the constant factor in the question i.e. whose prob doesnt change.
Important Bayes rule eqn (used in machine translation):
(arg max over y means whichever y gives max value of expression)

$$y^* = \arg\max P(y|x) = \arg\max \frac{P(x|y)P(y)}{P(x)} = \arg\max P(x|y)P(y) \tag{4}$$

(in the third step $P(X)$ is removed because constant term. We are interested in best value of y.)

5 Independence

Definition 3 (Conditional Independence)**.** *Once we know C, the value of A doesn't affect B and vice versa.*
$P(A, B|C) = P(A|C)P(B|C)$
$P(A|B, C) = P(A|C)$
$P(B|A, C) = P(B|C)$

Why do we make independence assumption in language models despite it not being true? because space and time of model. it becomes large and unwieldy.