Mathematical Foundations

Probability Theory

Probability is the measure of how likely an event is to occur.

- A fair coin lands on Heads 50% of the time, Tails every other time.
- A fair pair of dice rolls a seven 1/6th of the time.

We can generalize this idea so that it applies to more than just coins and dice.

- An process is **random** if it can result in one or more possible outcomes under the same conditions.
- The **sample space** (Ω) is a set containing all the possible outcomes.
- An **event** is a subset of the sample space.
    - In the case of a 6-sided die, {2, 4, 6} is the event "the result is even"
- An **experiment** or **trial** is the process in which we observe the result of a random process.
- And a **probability function** (*P*) is defined on every subset of the sample space.
    - Every outcome has a probability between 0 and 1 inclusively.
    - The sample space itself is also an event and has probability 1.
        - This is the probability of the event "something happens"

Probability Theory (Continued)

- Events are sets of outcomes and can naturally be used with set operators.
  - A and B are events
    - $A \cup B$ is the event "Either A, B, or both happen"
    - $A \cap B$ is the event "Both A and B happen", sometimes simply written as A, B
      - $P(A, B) = P(A \cap B)$ and is called the **joint probability**
- **Prior** probabilities are the probabilities of the events before we consider additional knowledge.
  - Denoted $P(A)$, Read as "The probability of A"
- **Posterior** probabilities are the probabilities after we consider additional knowledge.
  - Denoted $P(A \mid B)$, Read as "The probability of A given B"
    - The probability of an event A given that we know whether or not B happened.
  - Commonly called **conditional** probability

Conditional Probability

We can define conditional probability of events *A* and *B* intuitively.

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

If we know whether or not *B* happens, how likely is *A* to happen?

- $P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2|A_1)\ldots P(A_n|A_1, A_2, \ldots A_{n-1})$

This equation is very important to Statistical NLP.

Independence

Two events *A* and *B* are **independent** of each other if

$P(A) = P(A \mid B)$

If the posterior probability of *A* doesn't change once we know whether or not *B* happens.

Those two events are **conditionally independent** of a third event *C* if and only if

$P(A,B \mid C) = P(A \mid C)P(B \mid C)$

$A \cap C$ *and* $B \cap C$ are independent.

A and *B* are independent under the condition *C*

Bayes' Theorem

Bayes' Theorem relates the conditional probabilities of events *A* and *B* by the order in which we observe them.

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

This is a direct result of the chain rule.

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

Random Variables

A **random variable** is a function that takes values in the sample space and gives real number values.

Not actually a conventional 'variable'

Outcomes are not inherently numerical in a random process.

Flipping a coin three times can give us the outcome H, H, T.

Random variables act as the bridge from outcomes to numbers.

The random variable "number of heads" takes H, H, T and gives us 2.

$$P(\text{sum of dice}) = \frac{|6 - |\text{sum of dice } - 7||}{36}$$

This allows us to define the **probability density function** as a function of real numbers instead of abstract representations of events.

Probability Distributions

The **expectation** or **expected value** is the **mean** or average of a random variable if it were to measured over an arbitrarily large number of trials.

This notion is defined on random variables, not on outcomes.

The **variance** of a random variable is a measure of how far apart the values of the random variable are.

Again, this is defined on random variables, not outcomes.

These are called **moments** of a random variable's **distribution**.

Every unique distribution has a unique collection of moments (which include but are not limited to expectation and variance).

Two distributions are **identically distributed** if they have the same exact moments.

Each distribution has a **cumulative probability function** $P(X < x)$ which is the sum of all probabilities when its random variable takes values less than x.

Distributions

Bernoulli Distribution [*p*]

P(x) represents the probability of a binary result: P(X = 1) = *p* P(X = 0) = *1 - p*

Binomial Distribution [*n, p*]

Characterized by a number of trials *n* and a probability of occurrence *p*.

P(r) represents the probability of a *p*-probability event occurring *r* times in *n* trials.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\binom{n}{k}$ is a **binomial coefficient**.

Multinomial Distribution [*n, p1, … , pk*]

Generalizes the Binomial to more than two outcomes, depending on a set of probability parameters.

$$P(X_1 = x_1, \ldots, X_k = x_k) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k}$$

"What is the probability that *A* happens *x* times, *B* happens *y* times, and *C* happens *z* times, when we pick a letter from the alphabet *n* times?"

Normal Distribution

Normal Distribution [μ, σ]

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal distributions are often called **Gaussian** distributions.

They are a **continuous distribution**, their cumulative probability function is continuous.

- There are no sharp jumps if we follow the graph of the function from left to right.
- For a continuous distribution, the probability of any one outcome *x* happening is 0, but the probability that an event containing *x* may be positive.

For such distributions, we define a probability density function *P(x)* which gives us a probability-like density of the events *x*.

Normal Distribution (Continued)

If we believe data data to be **normally distributed**, any normal distribution is characterized by two parameters: mean μ and variance/standard deviation σ.

A normal random variable is said to a **standard normal random variable** if it's associated distribution has $\mu = 0$ and $\sigma = 1$.

The distribution is then referred to as the **standard normal distribution**.

Any normal random variable X can be mapped to the standard normal via the following formula.

$$\frac{X - E(X)}{Var(X)}$$

This process is called **normalization**

It is also called the **standard score** when applied to random variables in general.

Distributions in Practice

In practice, the distributions of language events are unknown.

- In general, we don't know how likely the word "apple" is compared to every other word.

Estimating the distribution is the next best thing.

There are two schools of thought to estimating the distribution.

- **Parametric**: The distribution falls somewhat nicely into a well-known family of distributions.
- **Non-parametric**: Treat all distributions generally; do not assume it has a well-known family.

Frequentist Statistics

Frequentist statistics is a standard interpretation of probability in which probabilities are defined from repeatable objective processes.

The probability of an event is the frequency at which it occurs as the number of trials over which it is tested goes to infinity.

> Said another way, as the number of trials increases, the relative frequency **converges** to the true probability.

As a result, when the sample size is "large enough", the relative frequency approximates the true probability.

The estimated probability distribution of a sample dataset will therefore converge to the actual probability distribution as we get more and more data.

Parametric Estimation

Parametric estimation seeks to identify the distribution by learning parameters of an assumed model.

Relative frequency of an event can be used in-place of a distribution mean by performing sufficiently many trials.

Relative frequency approaches the true average of a distribution over very many trials.

This is called the **Law of Large Numbers**.

Sample variance is similarly calculated using the same trials as relative frequency and can be used as an estimate of the true variance.

Parameter Estimation

- Estimating a Bernoulli random variable's parameter *p*:
  - Run a large number of trials *N*.
  - Count the occurrences of the positive outcome *C*.
  - *p* is approximately *C*/*N*
    - This also happens to by the random variable's relative frequency.
- Estimating a Normal random variable's parameters:
  - μ is assigned the relative frequency.
  - σ is assigned the sample variance.
- For the other distributions, we need a stronger process to estimate their parameters.

Maximal Likelihood Estimation

Maximal Likelihood Estimation is the process by which the parameters of a distribution as estimated by selecting a parameter set which maximizes the probability of observed data.

If we observe the results of a sequence of trials *s* and believe the distribution to have parameters *p1, p2, … pk*

$$MLE = \arg \max_{p_1,\ldots,p_k} P(s \mid p_1, \ldots, p_k)$$

This parameter set is the one most likely to correspond to the distribution that produced those outcomes.

Bayesian Statistics

Bayesian Statistics is another standard interpretation of probability in which a probability is assigned to a present state of belief or knowledge.

Bayesian statistics is based on the notion of modeling **prior belief** and updating beliefs given new **evidence**.

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

If you recall, this is Bayes' Theorem:

It is an integral part of Bayesian methodology.

Bayesian Statistics: Updating Beliefs

An **a priori** probability distribution is a probability distribution in which we start with no knowledge of the probability density function of the associated random variable.

A common *a priori* distribution is to assume a uniform distribution.

Not possible over continuous random variables.

Starting from an *a priori* probability distribution, we can update our knowledge about the random variable using Bayes' Theorem.

The probability of our **r**esult given our **e**vidence $P(\boldsymbol{R} \mid \boldsymbol{E})$

$(R \cap E)$ becomes our new event, and we proceed until we find new evidence.

Bayesian Statistics: Decision-making

Using Bayesian Statistics, one can solve a binary decision problem given evidence.

Looking at the **likelihood ratio** between P(yes | evidence) and P(no | evidence), we know which answer is more likely.

This can be abstracted to multi-category decision making.

When making this decision, Bayes' Theorem can be simplified.

$$P(\text{Decision} \mid E) = \frac{P(E \mid \text{Decision})P(\text{Decision})}{P(E)}$$

For each decision, the same denominator of $P(E)$ appears.

When calculating likelihood ratios, it disappears and therefore doesn't need to be calculated for the purpose of the decision problem.

Entropy

Entropy is a measure of the amount of inherent uncertainty in a random process measured in bits.

Alternatively, it's the level of information held within it.

Consider an 8-sided die,

$$H(X) = -\sum_{x} p(x) \log_2 (p(x))$$

The entropy of a dice roll is 3.

It takes 3 bits to uniquely describe each outcome.

This further acts as a measure of uncertainty.

I have three chances to be wrong about any one bit of the outcome.

$$H(1d8) = -8\left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) = 3$$

Sometimes, we reference the entropy of a random variable by the entropy of its associated distribution *p*.

$$H(p) = H(X)$$

Joint Entropy and Conditional Entropy

Joint entropy is the amount of information to uniquely describe the outcomes of a pair of random processes.

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log(p(x,y))$$

This notion can be generalized to larger-sized tuples using the full sums of their joint probabilities.

Conditional entropy is the amount of information needed to uniquely describe an outcome when there is already some known information about it.

$$H(X|Y) = -\sum_x \sum_y p(x,y) \log(p(y|x))$$

Back to the dice example,
if we know the 3rd bit (parity bit),
we only need to know two more.

$$H(1d8 \mid \text{roll is even}) = 2$$

Chain Rule for Entropy

The chain rule for entropy is as follows:

$$H(X_1, X_2) = H(X_1) + H(X_2 \mid X_1)$$

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_n, \ldots, X_1)$$

There is some potential **shared information** between all the X_i variables, so to avoid using too many bits, we consider one variable at a time.

We put aside enough bits for that variable and consider how many new bits are needed to represent the next as well, given that we know the previous.

This process repeats until we have enough bits to represent every variable together.

Mutual Information

The shared information between random variables X and Y is as follows

$$H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

Often denoted as $I(X;Y)$ or equivalently the reverse.

Through some expansion and simplification, we can provide a direct formula for it.

$$I(X;Y) = -\sum_{x}\sum_{y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Again, this is the information shared between two random variables.

If they are independent, the shared information is 0 bits.

Cross Entropy

Cross entropy is defined between distributions over the same outcome space rather than the random variables themselves.

Not to be confused with joint entropy, cross entropy is defined as follows:

$$H(p, q) = -\sum_{x} p(x) \log(q(x))$$

Cross entropy is defined on distributions, joint entropy is defined on random variables.

It is the measure of the average number of bits needed to encode an outcome if we expected outcomes to conform to the distribution defined by *q* when they actually conform to *p*.

Relative Entropy - Kullback-Leibler Divergence

Kullback-Leibler Divergence or **relative entropy** is the measure of how different two distributions are.

The **KL divergence** is the number of bits wasted by attempting to encode events from their true distribution *p* according to a different distribution *q*.

The formula for KL divergence is as follows

$$D(p||q) = \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

$$H(p,q) = H(p) + D(p||q)$$

KL divergence can also be related back to entropy and cross entropy by

Perplexity

Perplexity is a measure of how well a probability model or distribution predicts a sample.

It can be measured directly in terms of the entropy of the distribution.

$$\text{Perplexity(p)} = 2^{H(p)}$$

This is the de facto measure for evaluating language models.

It produces numbers on a non-logarithmic scale that can be easily distinguished from each other by their difference.