BASICS OF PROBABILITY

PURDUE UNIVERSITY

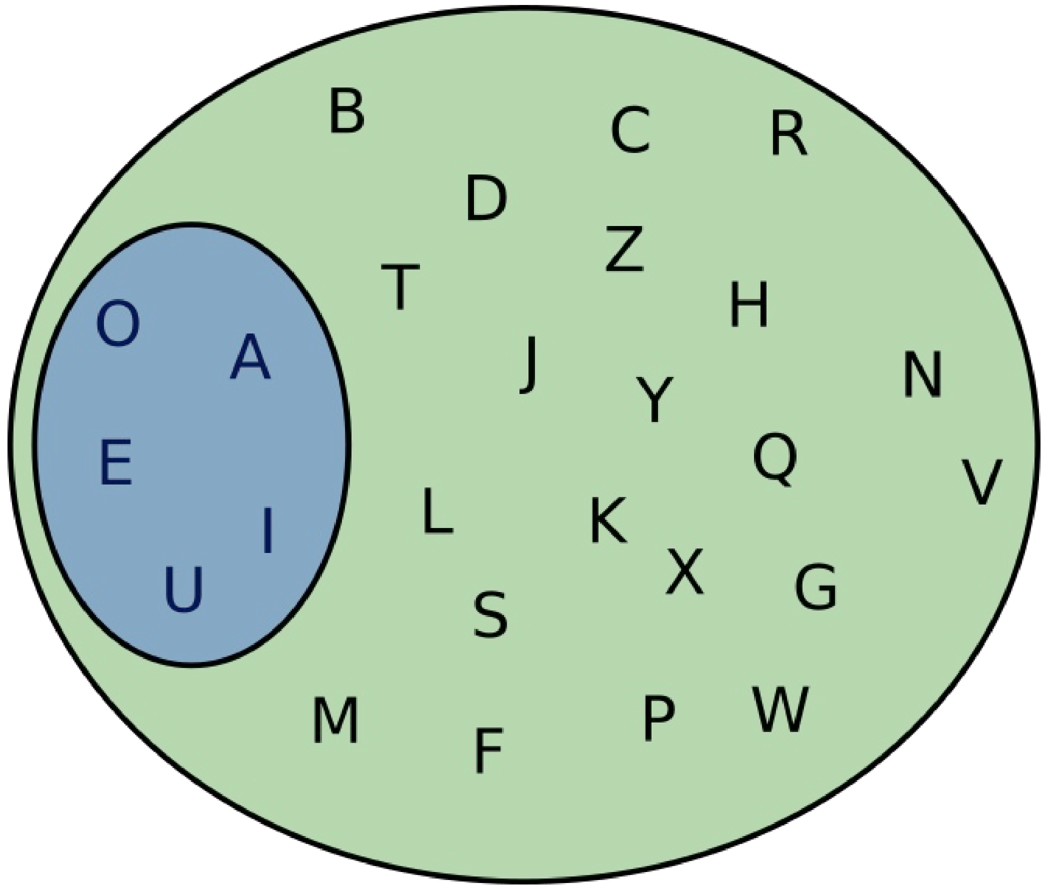
Randomness, Uncertainty and Probability

Working with data means dealing with uncertainty and randomness

- Datasets involve noisy measurements, so that repeated measurements can give different values
- Datasets involve missingness
- Dataset sizes are finite, so any conclusions about an underlying population involve uncertainty
- Deploying a trained model involves feeding it new unseen datapoints

Probability is a coherent language to quantify and manage such uncertainty

Random variables



Probability is a measure of how likely an outcome of a **random variable** is

- Is a number between 0 and 1, larger values meaning the outcome is more likely. Written as p(outcome), P(outcome) or Prob(outcome)

The collection of all possible outcomes is called the **sample space**

- For a coin flip, the sample space is {Heads, Tails}, for a die, it is {1,2,3,4,5,6}

The collection of associate probabilities is the probability distribution

- Usually we are more interested in the probability of a collection of outcomes (called an **event**). This is just the sum of probabilities of its constituent outcomes. E.g. p(Even number) = p(2) + p(4) + P(6)

- By convention, the sum of probabilities of all outcomes in the sample space is 1, though there is no loss of information if we multiply all probabilities by some constant

Conditional probabilities

A **conditional probability** $p(X|Y=y)$ gives the probability distribution of a random variable X **given** that another random variable Y has value y.

- You can think of this as a family of probability distributions over X, one for each possible value of Y
- Characterizes the statistical dependence of X on Y
- E.g. if X is the weather tomorrow and Y is today's weather, $p(X|Y=\text{'rainy'})$ is the probability distribution over tomorrow's weather given that it rained today
- E.g. if Y is the result of rolling two dice, and X is the outcome of one of them, then $P(Y|X=5)$ is the probability distribution of their sum, given one of them is 5

The joint probability of two variables is written as $p(X,Y)$ with $p(X,Y) = p(X)p(Y|X)$

Two random variables are independent if $p(X|Y) = p(X)$ (or equivalently, $p(X,Y) = p(X)p(Y)$)

Bayes' rule

Relates the conditional distribution P(X|Y) to P(Y|X)

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \propto P(Y|X)P(X)$$

$$Posterior \propto Likelihood \times Prior$$

Bayes rule allows you to take a model of observations, and 'invert' it, giving a probability over hidden quantities given observations

E.g. P(COVID | Loss of smell) $\propto$ P(Loss of smell | COVID) P(COVID)

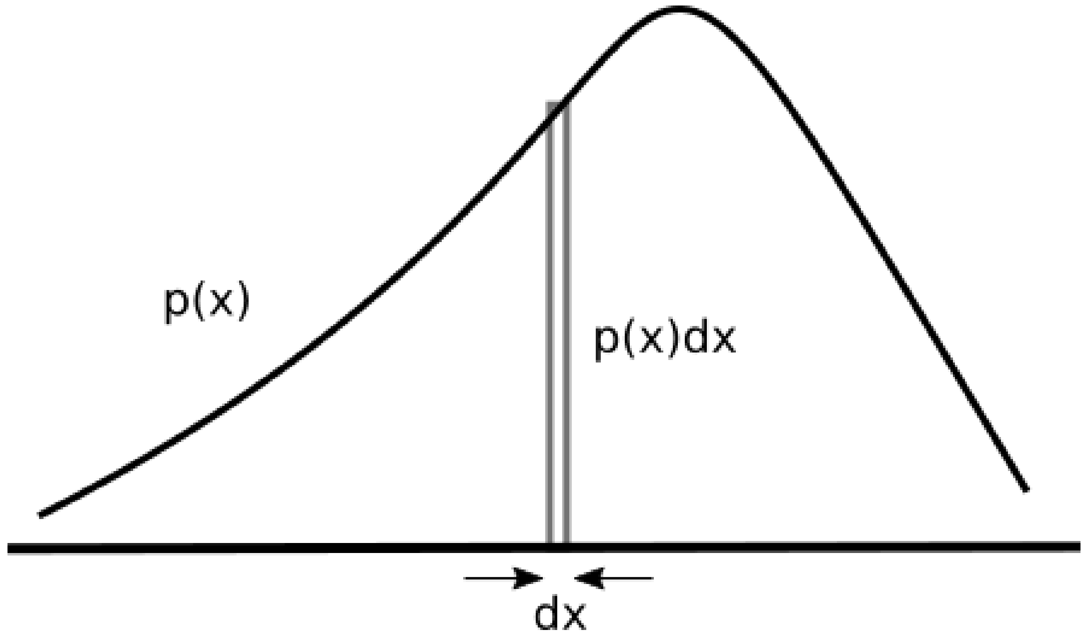
Continuous random variables

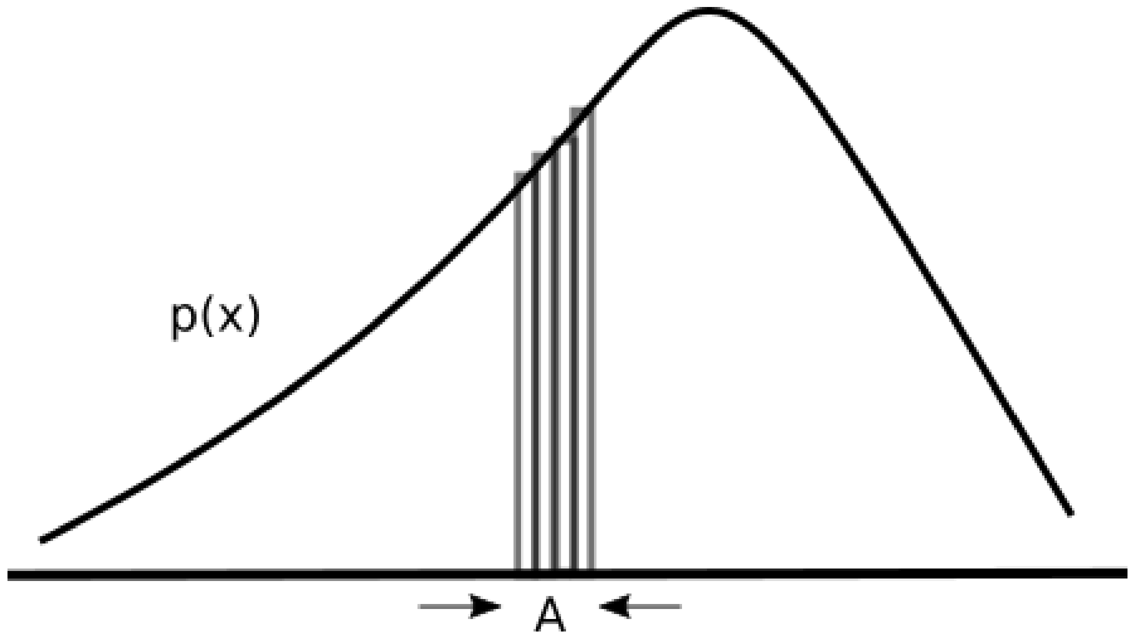
For continuous random variables, we use probability **densities** p(X)

Now the probability that X takes some value x is p(x)dx (which is 0!)

The probability of any event A is $\int_A p(x)\,dx$ (which need not be 0!)

We follow pretty much the same rules as before, except we replace probabilities with probability densities, and summations with integrals





Common probability distributions

- The Gaussian distribution
- The Bernoulli and binomial distributions
- The discrete and multinomial distributions
- The Poisson distribution
- ...