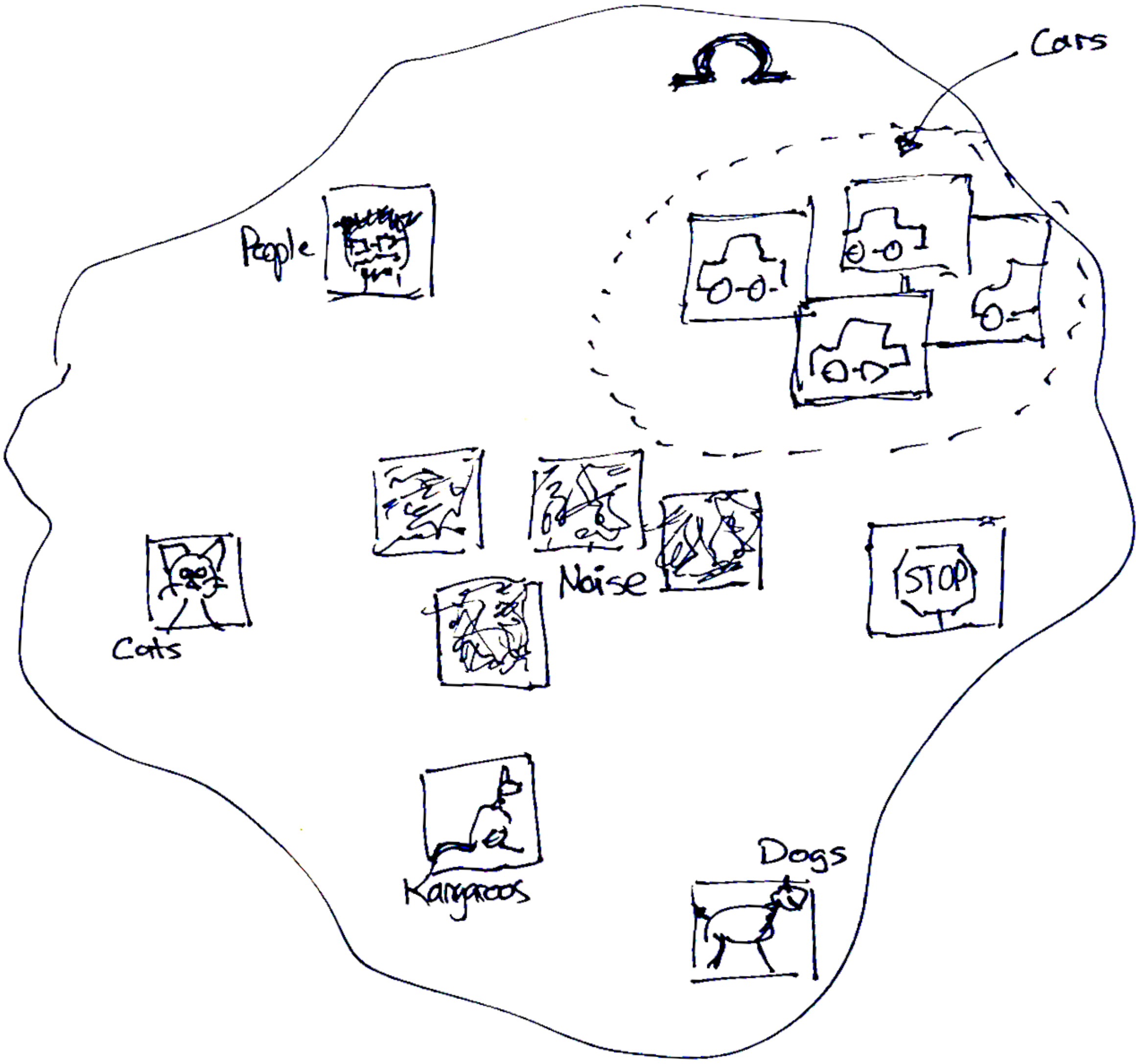
Probability Spaces

A *probability space* is a little different from a vector space. It is denoted by a *triple* $(\Omega, \mathcal{F}, P)$ where Ω is a set of all the "atomic" outcomes that can happen. Then $\mathcal{F}$ is also a set, but it is a σ-algebra based on Ω. That is a pretty technical thing, but what you have to know is $\mathcal{F}$ contains subsets of Ω that we call *events*. The subsets are chosen so that the σ-algebra behaves consistently. A common choice when Ω is finite is the powerset of Ω. Finally the space has a probability function (called a measured in most literature) that assigns a probability to each event in $\mathcal{F}$ in such a way that these are all valid probabilities.

Here's an example where the "outcomes" are images, and, for instance, the events might include subsets like the group of car images.



This is a very standard idea in stochastic modelling. It is the foundation of the ideas of probability, so it seems intuitively obvious that this should work, but there are some thinking points you should consider.

1. The implication that there is a probability distribution comes with baggage. Probabilities must obey a set of axioms, with many subsequent implications. These are great mathematically, but why should images obey these axioms. Why should an image have a probability? There are a very large number of variants of images of a single car (shifts and rotations) – how do I define a meaningful probability to all those variants?
2. The space is big. To quote Douglas Adams: "Space 'the Hitchhiker's Guide says,' is big. Really big. You just won't believe how vastly hugely mind-bogglingly big it is. I mean, you may think it's a long

way down the road to the chemist, but that's just peanuts to space." Consequently the probabilities we deal with here are correspondingly tiny, and hence almost impossible to work with.

3. Most of the space is "noise". Most outcomes in the space are not recognisable images. The images of real-world things are in an almost imperceptible volume of that space.

That said, its still a really useful conceptual model. One of the uses we come across often is (again) looking at distances.

Distances

We are used (now) to think about distances between objects in a space. For instance, the distance between a dog image, and a cat image in the space above.

However, now we want to think about distances **between spaces**. That is, given two probability spaces $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ we want to compute a distance between them. Commonly (but not universally) we assume that they have the same set of outcomes and events, so we only need to look at the distance between the probability measures P_1 and P_2.

There are many such "distances", but I want to focus on one (which is not actually a metric) called the *Kullback-Leibler (KL) divergence*. It is not a metric because it is not symmetric (it also does not satisfy the triangle inequality). But it is a useful measure of *divergence* of one probability measure P from another Q and we write it as $D_{KL}(P||Q)$.

The divergence comes from the area of information theory, where it is also called *relative entropy* and it has many clever interpretations. Its formal definition is

$$D_{KL}(P||Q) = \sum_{x \in \Omega} P(x) \log \left(P(x)/Q(x) \right).$$

Some notes:

- There is an implicit assumption here that $Q(x) > 0$ for all $x \in \Omega$. We can calculate it where $P(x) = 0$, because we use the convention that $0 \log 0 = 0$ but Q is a problem.
- The base of the log doesn't matter, but it implies units (base-2 implies units of bits).
- It is conventional (in the ML literature) to write expressions such as the one above in terms of expectations, *e.g.,*

$$D_{KL}(P||Q) = \mathbb{E}_{x \sim P} \log P(x) - \mathbb{E}_{x \sim P} \log Q(x).$$

The notation means:

1. the expectation is treated like an *operator* so we (often) omit brackets around the operand;
2. There are two probability distributions here, so we have to say which the expectation refers to: here the expectation $\mathbb{E}_{x \sim P}$ means we take the expectation with respect to the distribution P; and
3. You have to get used to people expanding (or contracting) logs and leaving out brackets.
4. The first term $\mathbb{E}_{x \sim P} \log P(x)$ is just the minus the entropy of distribution P, *i.e.,* $-H(P)$, which, for a given distribution P is fixed, so often when this is incorporated into an optimisation problem on Q, this term will be dropped because it is a constant.
5. The second term $-\mathbb{E}_{x \sim P} \log Q(x)$ is *cross entropy* between P and Q.

The KL divergence comes up again and again and again in machine learning and statistical estimation theory. It links into maximum-likelihood estimation, maximum entropy estimation, machine learning objectives, and many other places either directly or as an approximation. For instance, in my case it has come into problems as a regularisation term.

There are many other alternative measures of dissimilarity between probability spaces. I like the Jensen-Shannon (JS) divergence, which is a symmetrised version of KL divergence.

$$D_{JS}(P, Q) = D_{KL}(P||M) + D_{KL}(Q||M),$$

where $M = (P + Q)/2$. The JS divergence works better when the *support* of the two distributions (the set on which the probabilities are non-zero) doesn't entirely overlap.

Why does it matter?

There are many many places where such distances are needed. For instance, they can be used as loss functions to judge the quality of an output distribution compared to an input.

But a key reason that the above is important is that your training data will come from a distribution Q, but real data comes from a real-world distribution P. To some extent the best quality you can achieve for learning about P from Q will depend on the size of $D_{KL}(P||Q)$ (or visa versa depending on context).

Links

- https://danmackinlay.name/notebook/probability_metrics.html
- https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained
- *** https://dibyaghosh.com/blog/probability/kldivergence.html
- https://www.stat.cmu.edu/~cshalizi/754/2006/notes/lecture-28.pdf