Probability and Information Theory

Expanded upon and presented by Alexander Shim

Importance of probability

To evaluate models, we calculate the expected value (average) of the loss function over our data distribution.

$$E_{x\sim P(x)}[L(x)] = \int_{(X,\Sigma,P_X)} L(x)dP_X(x)$$

More generally, it allows us to work with non-deterministic (stochastic) systems.

Inherently stochastic systems, limitations in measurement, incomplete modeling.

Interpretations of probability

Axiomatic

Frequentist - long term observed average, removes *a priori* assumptions.

Classical - counting equal probability outcomes

Measure of belief / Bayesian

What the real world is like delves into philosophy, so we will focus on the axiomatic approach.

Required concepts

Probability

Random Variables / Random Vector Spaces

Expected Value

Probability Distributions / densities

Conditional Probability

Bayes’ Rule

Our ideal probability space

(R, $\mathcal{B}$, P)

The set is the real line. Open, closed, half-intervals are measurable subsets of R, with

$\int_a^b \mathbf{1} dx$ = b-a, for b > a.

Probability is defined by a density function p(x):

p((a,b)) = $\int_a^b p(x)dx$

Set Theory

Set of objects	$S = \{ a, b, c, \dots \}$
Subsets (events)	$S1 = \{ b, c \}$
Powerset, set of all subsets	$\wp(S) = \{ \varnothing, \{a\}, \{b\}, \{c\}, \{a,b\}, \{a,c\}, \{b,c\}, \{a,b,c\} \}$
Union, intersection	$\{a, b\} \cup \{b,c\} = \{a, b, c\}$. $\{a,b\} \cap \{b,c\} = \{b\}$
Complement	$\{a, b\}^c = \{c\}$

Set Theory

Sets can contain anything.

Numbers, objects, data-points, in probability, we consider sets of outcomes.

{ Team A wins, Team B wins}

Heads AND roll 2 of 6 AND generate number 0.15 in S

There can be many different sets of outcomes you can use to describe a system.

Probability is **not** defined for outcomes, but sets of outcomes (events).

Set Theory

Too much work to specify every subset we want to measure, so we will try to generate them instead.

algebra of sets: Non-empty collection of subsets of a set, includes $\varnothing$, closed under complement (contains the complement of any element in itself), closed under finite unions and intersections.   Often we generate them from a given collection of subsets of a set.

σ-algebra: Same as algebra but closed under countable unions and intersections. **σ**-algebra generated from a subset $\mathcal{E} \subseteq \wp(S)$ is the smallest possible **σ**-algebra containing $\mathcal{E}$, written as $\mathcal{M}(\mathcal{E})$

Set Theory

Why do we need to use an algebra instead of any subset of the powerset? Let’s consider what is contained in an algebra:

For A, B in S, A U B in S. From finite additivity.

A intersect B in S:

(A^c union B^c)^c = A intersect B. From De Morgan’s Law of set theory, axiom 2 for algebras.

A \ B in S: A \ B = A intersect B^c

This allows us to calculate the probability of any subsets of Omega that are unions or intersections of other sets where we know their probabilities.

Set Theory

Example: Consider *N* = { 1, 2, 3, … }, the set of natural numbers. Let's try to generate the algebra from the set of sets containing a single even number:

S1 = {{2}, {4}, {6}, … }

Algebra of S1 contains all sets containing finitely many even numbers ({4, 18}), all of S except for finitely many even numbers, and finite unions and complements of those. Note subsets of S1 have 0 odd elements or all the odd elements, and finitely many even numbers or all even numbers except a finite number.

Example: Consider *R*, the set of real numbers. We can generate the **σ**-algebra from the set of open intervals { (a,b) ⊆ R: a < b }. This is what's commonly used in Euclidean spaces.

Topology

For a given set, we want to define which subsets are open. Also allows us to define continuous functions.

The topology of a set S is a collection of subsets Σ of S, (subset of the powerset ℘(S)), such that:

- *∅, S ∈ Σ*
- *Σ is closed under arbitrary (including uncountable) unions and finite intersections.*

A topological space is a set along with its topology, and it's written as (set, topology)

Elements of Σ are called open (but relative to that topology).

Closed sets - complement of an open set.

Intuition: some elements should be "close" to each other, so if we look at open sets around x, we should tend to see the same "close" elements. See figure.

Topology

Example: Topology L on R of open intervals.

Intervals (1,3), (2,4) $\in$ L

(1, 3) $\cup$ (2, 4) = (1, 4) $\in$ L

(1, 3) $\cap$ (2, 4) = (2, 3) $\in$ L

Topology

Example: Discrete Topology (we can define using the powerset)

a	b	c
d	e	f
g	h	i

Every subset is open! And closed … If we tried to use our idea of distance, each element would be arbitrarily far apart.

Topology

The same set can have many different topologies. A function from one topological space to another may be continuous in one topology but not continuous in another.

Intuitively we can shrink or expand sets into smaller/larger ones by using functions that projecting or transforming their topologies.

a1, a2, a3	b1, b2, b3	c1, c2, c3
d1, d2, d3	e1, e2, e3	f1, f2, f3
g1, g2, g3	h1, h2, h3	i1, i2, i3

We can transform spheres into bricks or into pancakes.

Topology

We can use the topology to generate a **σ**-algebra. This is called the Borel **σ**-algebra, and we denote it by $\mathscr{B}_X$, where X is the topological space.

Includes open sets, closed sets, countable unions and intersections of open and closed sets, and it keeps going. Otherwise closed intervals wouldn't be considered measurable!

Example: **σ**-algebra of open intervals as before. This turns out to include closed intervals as well and half-intervals.

We can create a product space for two (even countably many) topologies and generate a larger space. It can be shown the **σ**-algebra of a countable product space $(\otimes X_n, \mathcal{M}(\otimes \mathcal{E}_n))$ is the product space of the **σ**-algebras $(\otimes X_n, \otimes \mathcal{M}(\mathcal{E}_n))$.

Measures

In calculus, for integration we have the Lebesgue measure: $\boldsymbol{\mu}((a,b)) = (b-a)$, where (a,b) is the open interval from a to b, where $b > a$. Same for the closed interval $[a,b]$, $[a,b)$, $(a,b]$.

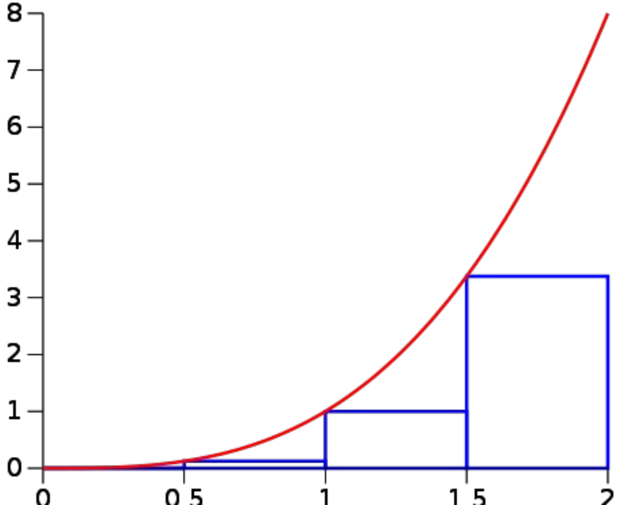
But why couldn't it be $2*(b-a)$, or $\int f(x)dx$?

Integration is dependent on the measure

We can use a function to alter a measure

Integration with a positive function can

be seen as creating a new measure.



Measure Theory definition

A measure $\boldsymbol{\mu}$ is defined on a set X and a **σ**-algebra of X.

Non-negative (exception: negative and complex measures, not applicable to cases we will consider): $\boldsymbol{\mu}$(E) >= 0

$\boldsymbol{\mu}$(∅) = 0 … though other subsets can have measure 0.

σ-additivity: for pairwise disjoint sets $\{E_i\}$, $\boldsymbol{\mu}$(union ()) = sum($\boldsymbol{\mu}$(E[i]), i).

A probability space is a measure space with a total measure of 1.

Many probability paradoxes are because different approaches may apply different probability measures.

Metric spaces

Our intuitive idea of open sets is open intervals, so we want a generalized idea of distance.

For n-dimensions, open n-balls (n-dimensional sets of points a distance $< p$ from a point) form a topology, and subsequently a borel σ-algebra.

We can generalize this by having a distance function $p: X \times X \to [0,\infty)$ which we define as a metric on our set X, with these properties:

$p(x,y) = 0$ iff $x = y$.

$p(x,y) = p(y,x)$ for all $x, y \in X$ (symmetry)

$p(x,z) \leq p(x,y) + p(y,z)$ for for all $x,y,z \in X$ (triangle inequality)

Metric spaces

Given a metric space (X, p), we can define the open ball of radius r about a point x

B(r,x) = { y ∈ X : p(x,y) < r)

These can be used as the basis of a topology for X, and going further we can define the Borel σ-algebra defined by this metric space.

Examples: R^n with the Euclidean metric (L2 distance)

(Total variation) $\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A} \in |\mathbb{P}_r - \mathbb{P}_g|$

Metric spaces

Metric spaces provide us a generalized idea of distance. Usually we work Euclidean distance, but not so for information theory, comparing distributions.

In math, metric spaces along with vector spaces allow us to perform calculations on lower dimensional manifolds, like a two dimensional object (say a sphere) in 3d space. Can also apply to natural images or natural words.

Later on, we will be working in parameter spaces to define the parameters of a models. But when we apply gradient descent to optimize the model, we are assuming the parameter spaces have a Euclidean metric. Once the metric changes, unadjusted gradient descent may not be efficient, and some tricks fix this.

Vector Spaces and Norms

A topological space is a topological vector space if we can represent addition and multiplication by “scalars”:

Let $\mathcal{X}$ = (X, M) be a topological space. It is a vector space if

there exists some continuous function X x X -> X (we represent it (x,y)->(x+y))

there exists a function K x X -> X, (we represent it as (λ, x) -> λ x) where K is a field (usually R or C)

Vector Space and Norms

We define a norm $\| \|$ on a vector space as a function from x->[0,infinity) such that

$\| x\| = 0$ iff x=0 (for semi-norms, this condition is removed)

$\| \lambda x\| = | \lambda |\| x\|$

$\| x + y\| <=\| x\| +\| y\|$

Note that a norm can be used as a metric on a vector space: $p(x,y) = \| y - x \|$. Of course such a metric is much more restricted!

Differentiation can be defined now!

Vector spaces and Norms

A directional derivative is defined as

lim | (f(x+h*v) - f(x) | / h

For topological vector spaces, if we generalize this by setting x, v in X, and h in K. It is Gateaux differentiable at x if it exists for all v in X.

Usually, we will be working with a norm on R^n. Just remember that this is usually assumed for differentiation (see probability density section).

Probability Space

A probability space (Ω, Σ, P) is defined by 3 terms:

Ω: The set of outcomes

Σ: A **σ**-algebra on Ω, that is, Σ subset ℘(Ω) satisfying these conditions:

Note: A subset of outcomes is called an event.

P: ℘(Ω) -> R that satisfies the Kolmogorov Axioms:

P(S) >= 0 for all S in ℘(Ω)

P(Ω) = 1

$$P(\bigcup_{i\in\mathbb{A}} S_i) = \sum_{i\in\mathbb{A}} P(S_i), \quad S_i \cap S_j = \varnothing \quad \forall i \neq j$$

There’s also a formulation with finite additivity, which if we think about our work above, we would replace **σ**-algebras with just algebras.

Probability Spaces - Common Distributions

For categorical/discrete distributions we can use the multinomial distribution (case k=2 is the Bernoulli distribution)

$$\binom{n}{j_1, \ldots, j_k} \prod_{i=1}^{k} p_i^{j_i}$$

In statistics, often assume data distributions match some common ones.

Gaussian

$$N(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Exponential

$$p(x; \lambda) = \lambda \mathbf{1}_{\mathbf{x} \geq \mathbf{0}} \exp(-\lambda x)$$

Laplace

$$Laplace(x; \mu, \gamma) = \frac{1}{2\gamma} \exp(-\frac{|x-\mu|}{\gamma})$$

Probability Spaces - Common Distributions

Mixture distributions

Latent variables, defining parameters such as μ,𝜎,ƛ to be used to define mixture distributions

Deep learning works with data distributions which we define implicitly through sampling from our data set.

Hierarchy of generalization

Note we can apply increasingly stricter conditions:

Σ is an algebra < Σ is a **σ**-algebra < Σ is a Borel **σ**-algebra of a topology < Σ is measurable < The topology (and followingly the Borel **σ**-algebra, there is a measure) is generated by a distance metric < Σ is a normed topological vector space (which induces a metric and a topology, and so on) -> Σ = R^n, is generated by the Lebesgue measure

An algebra is only closed for finite unions, whereas **σ**-algebras is closed for countable unions.

Random Variables

A random variable is a mapping from a probability space to the real numbers.

f: Ω -> R

Will any function work? Not necessarily.

We need a continuous mapping:

Topologically, a function is continuous if the pre-image of every open set is open:

Let V be open in R.

$g^{-1}(V)$ = U, in Ω. If V is open, U is open.

Technically, we must define a topology on R!

Random Variables

Now let's try to define a probability function on R, P(V) = P(f^{-1}(V)). However, if V is an element of the σ algebra of R, there's no guarantee that f^{-1}(V) is in the σ algebra of Omega. What are sufficient conditions for this to be satisfied?

We need f to be a measurable mapping:

> A function f: X->Y, where (X,script M), (Y, script N) is called (script M, script N) measurable if the preimage of any measurable set E in script N is measurable in script M, that is f^{-1}(E) in script M.

For open V in R, U=f^{-1}(V) is open in Omega, so it is in Σ. Following the above result, every continuous function is measurable on the Borel σ algebra.

Note: a function being continuous depends on the topologies of the domain and range. We can apply different topologies to the real number (such as the discrete topology or the trivial topology), and a function that is continuous for one may not be continuous for another.

Random Variables

If X and Y are topological spaces, every continuous function is (script B[X], script B[y]) measurable.

If X and Y are metric spaces, every continuous function is measurable.

Open sets are in the generated Borel set, so f is continuous iff $f^{-1}(V)$ is open in X for every open V in Y.

$f^{-1}(V)$ is open, and every open set is measurable.

The topology generates the σ algebra, and the measure is defined on that σ algebra. The preimage of open sets is open and the preimage of closed sets is closed, and those are subsets of the the set of open and closed sets, which generates the σ algebra, so they are measurable.

For a Borel set, it's generated by the open sets, so the σ algebra generated by the preimages of the open sets is a subset of σ algebra generated by the open sets, so they are measurable.

For metric spaces, we use open balls as a subset of the open sets which extends to all open sets.

Random Variables - Discrete Case

For the discrete case, every subset of Σ is open, so it is in the Borel set, so the probability distribution is defined on it and it is measurable (and hence P(s) is defined)

For a discrete topology (though there may be countably or uncountably many elements), every subset of Σ is open. From countable additivity,

So for a subset U of S, P(U) = sum(P(s), s in U)), if U has countably many elements.

Note that our classical probability definition follows from setting each P(s) to be equal (each outcome is equally likely).

Random Variables - Continuous Case

Cumulative distribution functions

F(x) = P[R](X <= x)

First we note that (-infinity, x] is in the Borel σ algebra generated by open intervals[1]. So assuming that R has the topology where open intervals are a basis for the open sets, and f is measurable to R with that topology and the Borel σ algebra, the preimage is a measurable subset of Omega, so

P[R](-X<=x) = P(f^{-1}((-infinity, x])

Random Variables - Probability Distribution Function

For continuous maps, we usually have P[R](X=x) = 0 for any specific point x. But we’ve seen probability distribution functions, which represent some idea of relative probability weights from one point compared to another, so what’s going on?

Moreover,

int(f(x),x=-infinity,infinity)=1.

Since the cumulative distribution function is well-defined, we define the derivative of the cumulative distribution function as the probability distribution function. Of course, the cumulative distribution function must be differentiable for this to work. This may not necessarily be true:

Probability Distribution Functions

Example:

Suppose some sample space maps to P[R] such that

P[R](x=0) = 1

P[R](x< 0) = 0

P[R](x> 0) = 0

This is conceivable, f(s) = 0 for all s in Omega. Notably, a probability distribution function would be infinite around x=0, so it would not be well-defined.

So let's assume F is differentiable.

By the first fundamental theorem of calculus, F(x') = int(dF/dx, x=-infinity,x'): note integrals require a measure.

We see this is a probability function on R, with the topology defined by open intervals:

P[R](R) = 1

The cumulative distribution function is increasing, so P[R](V)=0 for any measurable V in R.

Countable additivity for disjoint sets: the preimage of each are also disjoint, since they're the preimage.

Random Variables

Now let’s consider what a derivative is:

dF/dx = lim((F(x+h)-F(x))/h,h->0)

dF/dx = lim(P[f^(-1)(-infinity,x+h)-P[f^(-1)(-infinity,x)]/h, h->0)

dF/dx = lim(P[f^(-1)([x,x+h))]/h, h->0)

[x,x+h) is in the Borel σ algebra, so it is defined

So we can have a sequence of values of the probability of small intervals around x, divided by the size of the intervals, which converges under certain conditions , and that converged value is our probability distribution function, that is, it’s defined by how much the probability changes over infinitesimal intervals as we go along x.

Random Variables

For the continuous case, we can consider the probability distribution function to be a measure.

Specifically, normally we are integrating int(f(x), x=-infinity..infinity), which is using the Lebesgue measure. This means f is supported by sets measurable by the Lebesgue measure, so we can convert that to a measure mu, where f is the Radon-Nikodym derivative of mu, with mu << Lebesgue

$\int d\mu = \int f(x)dx$

Random Variables - Expected Value

Since a random variable defines a probability distribution on R, we can average functions of that random variable over that probability distribution.

Say our random variable maps our probability space to (R, μ)

We apply a function on the real numbers:
$g: R \to Y$, usually $Y = R$ with the same topology.

$E[g(X)] = \int g(x)\, d\mu(x)$

Random Variables - Expected Value

One common one is the mean

$E[X] = \int x \, d\mu(x) = \mu$ (this is a different μ)

If we subtract the mean and square it, that is well-known as the variance. That can be generalized to the nth central moment of a random variable

$E[(X-\mu)^n]$, and it has a counterpart, the moment $E[X^n]$ or $M_X(n)$

It turns out knowing all the moments of a distribution fully specifies the distribution, though in practice people tend to focus on the mean and variance.

Random Variables - Expected Value

However, we must be concerned with the same considerations that we applied to mapping our probability space into the random variables.

If g is not continuous, then the probabilities may not be defined.

g maps all events in script P which map to a subset of R to a subset of Y, so again we are losing information.

Even if g is continuous and measurable, f o g may not be Lebesgue measurable since a Lebesgue measurable function maps is a (script L, script B[R]) measurable function, not a (script L, script L) measurable function.

Random Variables - Changing distributions

Note in the previous case we utilized the same probability space. However, we can map one probability space into another probability space using a function. Specifically, we are considering g(x) in the probability space of R, instead of the probability space of Y. If we were to define our probability function as the preimage of g, then

E[y](g(x)) = E[x](1), since P[y](V)=P(g^{-1}(V), and moreover P[y](Y)=P(R)=1. That is, the induced probability function in Y would be defined merely by how g maps R into Y. As with our earlier discussion, you could consider g to be the Radon-Nikodym derivative of Y's probability measure with respect to R's. Using densities, this can be seen as

$$p_x(x) = p_y(g(x)) \left| \frac{\partial g(x)}{\partial x} \right|$$

Note p_x(x) dx and py(y) dy are the distributions relative to the Lebesgue measure. If we index the distribution of y using x, we have to apply the derivative of g relative to x (think the chain rule).

For higher dimensions, this is

$$p_x(x) = p_y(g(x)) \left| \det \frac{\partial g(x)}{\partial x} \right|$$

Random Vector Spaces

Instead of mapping sets to real numbers, we can map it to a vector space.

We can take the product space of random variable spaces, and define the set of measurable sets to be products of the measurable sets of the subspace.

Random Vector Spaces

When we have probability distribution functions for the subspaces, we can formulate a probability density, $p(x_1,...,x_n)$

Have independence when it's evenly distributed

$p(x_1,...,x_n) = p(x_1)...p(x_n)$

Requirement for independence is more than pairwise independence...

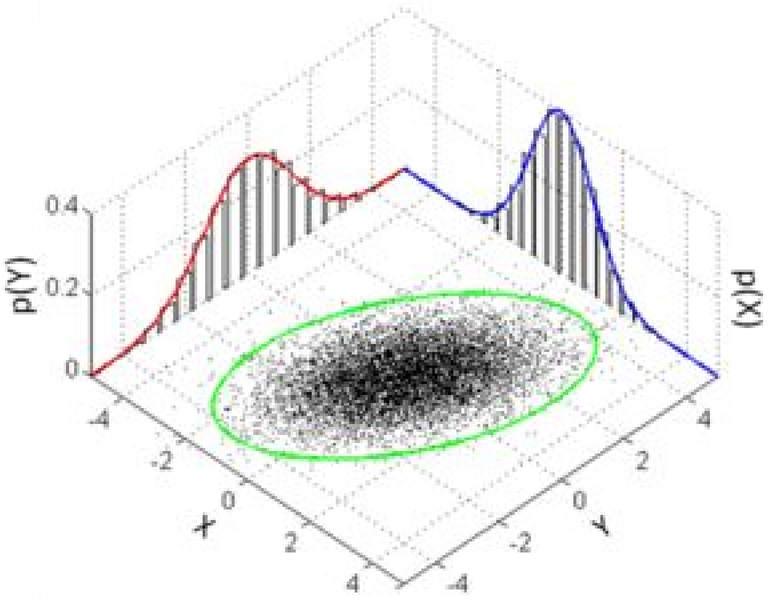
We have a specific function over a vector space:
$\text{Covariance}(f(x),g(y))=E_{XxY}[(f(x)-E_X[f(x)])(g(y)-E_y[g(y)])]$

Random Vector Spaces - Marginal Probabilities

We can recover the probability over each dimension through marginal probabilities

p(x)=∫ p(x,y)dy (can be generalized past Lebesgue measure support)

dμ = $\int_Y$ d(μ x ν)



Conditional Probabilities

Conditional probabilities

Let A, B be events in our probability space.

We define

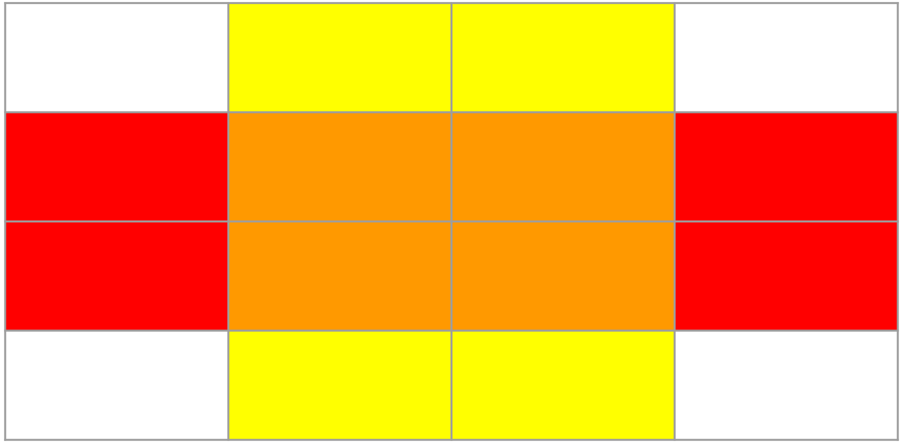
$P(A \mid B) = P(A \cap B) / P(B)$, assuming A and B are in Σ.

We are subsampling the probability space (Omega, Σ, P) to (B, Σ’, P_B)

where we define

C’ in Σ’ B when there exists C in Σ s.t. C intersect B = C’.

$P_B = P(A|B)$, for P(B) != 0.



We can see it as applying a characteristic function over B, with the probability function scaled down by the probability of B.

A special case is for the discrete case, where each value in Σ / R is an open and closed set, and the continuous case where it is the limit of open neighborhoods of that value.

Usually in random vector spaces we define the subset to be strictly defined by one variable.

Conditional Probabilities

Note that A and B must be from the same probability space. So if the probability space is an observation a first time, subsequent observations may change the probability space, most commonly the probability function. There's no reason this can't be a valid probability space.

Also, for random variables, technically we should look back to the probability space:

P[R](A | B) = P(f^(-1) (A intersect B)) / P(f^(-1) (B))

P[R](A | B) = P(f^(-1) (A) intersect f^(-1)(B)) / P(f^(-1) (B))

Since f^(-1) (A intersect B) = f^(-1) (A) intersect f^(-1)(B). This is easy to check in both directions, since f^(-1) (A \ B), f^(-1) (B \ A), and f^(-1) (A intersect B) are disjoint elements in Σ which map to disjoint subsets of R.

We will want to extend this definition to probability distribution functions, which technically would be undefined since pointwise the probability would be 0.

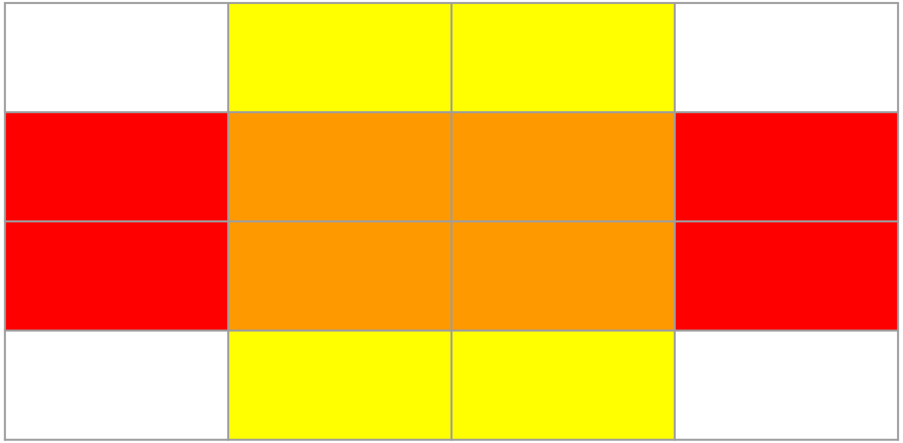
If we were concerned with an event in the probability space instead of the random variable probability space, we may not necessarily be able to consider the conditional probability of an event in script P and another event in R (see note at the end of the random variable section).

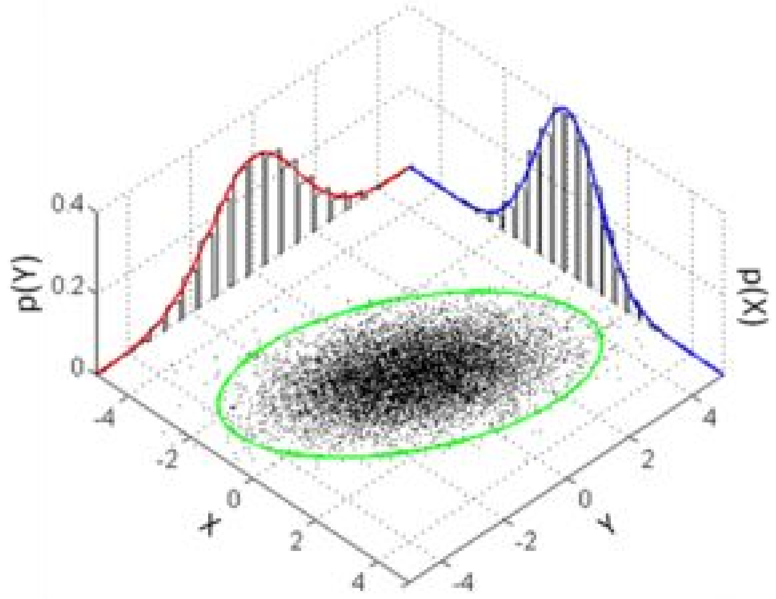
Unless we know/assume that subsequent observations have the same probability space, we need to create a new probability space for each observation, which may destroy information about how two observations may be related. We can also create a product probability space, see random vector spaces.

Conditional Probabilities - Bayes' Rule

$P(x \mid y) = P(x \text{ AND } y) / P(y)$

$P(x \mid y) = P(x)\, P(y \mid x) / P(y)$





Information Theory

Entropy: from statistical mechanics, is based around the number of configurations of states, Ω.

Entropy = k_B ln Ω

From our understanding of random variables (or possibly other measurable mappings), we've collected all the outcomes that lead to a certain value of an expected value (like the number of heads), so the natural question is to ask how many ways could we end up there?

Assuming each state is equally likely, the probability of any particular state is P=1/Ω,

Entropy = - $k_B$ln P

Information Theory

Technically this explanation only works for discrete, equally likely outcomes, but the logarithmic formulation allows us to extend the concept to unequal probability outcomes, and also to the continuous case:

Entropy = $E_x[- k_B \ln P_x]$, where x is a random variable value with probability P

Entropy = $\sum_i - k_B P_i \ln P_i$

Note we can change this from a natural log to another base by multiplying by some constant

Information Theory

Shannon’s formulation forms the base for information theory, though it was targeted at a different problem.

How much can an information source be compressed, so it can be transmitted, then decompressed, without losing information? And also considering noise.

Intuitively messages like words may be lower dimensional than their length due to certain relationships between the characters, so it can be compressed into a lower dimensional representation. But if it’s reduced to say n binary bits, we can’t code more than 2^n unique messages.

Information Theory

Three properties desired to measure information:

1. H should be continuous with respect to the probabilities
2. For n equally likely outcomes, H should be monotonically increasing with respect to n.
3. “If a choice can be broken down into successive choices, the original H should be the weighted sum of the individual values of H”

* If a message can be broken down into two independent events (splitting a message into two pieces), the information of the combined message should be the sum of the information of the pieces[1].

1-3 only satisfied by $H = \sum_i - k P_i \log P_i$.

Shannon/bit base 2, nat base e, hartley base 10.

Intuitively lower probability events should be weighted more heavily. If one outcome has probability 1, there’s no entropy.

Information Theory

When we try to model a distribution, we want to be able to compare it to other distributions. This can be done through metrics on spaces of distributions - or through a weaker version of metrics.

Divergences like a metric, but it does not need to be symmetric and does not need to satisfy the triangle inequality.

Probability distribution distances/divergences[1]

Total variation

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A} \in |\mathbb{P}_r - \mathbb{P}_g|$$

L_p distances for distributions absolutely continuous to Lebesgue measures

KL divergence

$$KL(\mathbb{P}_r||\mathbb{P}_g) = \int log(\frac{\mathbb{P}_r(x)}{\mathbb{P}_g(x)})\mathbb{P}_r(x)d\mu(x)$$

Jensen-Shannon divergence

$$JS(\mathbb{P}_r||\mathbb{P}_g) = KL(\mathbb{P}_r||\frac{\mathbb{P}_r + \mathbb{P}_g}{2}) + KL(\mathbb{P}_g||\frac{\mathbb{P}_r + \mathbb{P}_g}{2})$$

Wasserstein-1/Earth-Mover

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \prod(\mathbb{P}_r, \mathbb{P}_g)} E_{(x,y)\sim\gamma}[||x - y||]$$

Kolmogorov-Smirnov

$$D_{\mathbb{P}_r, \mathbb{P}_g} = \sup_{x} \left|CDF_{\mathbb{P}_r} - CDF_{\mathbb{P}_g}\right|$$

KL divergence

$$KL(\mathbb{P}_r||\mathbb{P}_g) = \int log(\frac{\mathbb{P}_r(x)}{\mathbb{P}_g(x)})\mathbb{P}_r(x)d\mu(x)$$

We’re comparing the ratio of Shannon information.

Must specify one of the distributions to be the probability distribution for the expected value.

$$KL(\mathbb{P}_r||\mathbb{P}_g) = \int \mathbb{P}_r log(\mathbb{P}_r)d\mu(x) - \int \mathbb{P}_r log(\mathbb{P}_g)d\mu(x)$$

Note the left side is entropy as we have defined it before. The right side is called the cross-entropy.

$$KL(\mathbb{P}_r||\mathbb{P}_g) = H(\mathbb{P}_r) - H(\mathbb{P}_r, \mathbb{P}_g)$$

Structured Probabilistic Models

Can be too complicated to specify the joint probability if there’s too many random variables.

Each random variable may only be dependent on a small number of other random variables.

We can express each random variable as a node in a graph. Deep learning packages built around dataflow graphs such as Torch and Tensorflow use these.