Overview

Probability is the mathematical field concerned with reasoning under uncertainty.

The use of probabilities to describe the frequencies of repeatable events (like coin tosses) is fairly uncontroversial. In fact, frequentist scholars adhere to an interpretation of probability that applies **only** to such repeatable events. By contrast Bayesian scholars use the language of probability more broadly to formalize our reasoning under uncertainty.

Bayesian probability is characterized by two unique features: (i) assigning degrees of belief to non-repeatable events, e.g., what is the probability that the moon is made out of cheese?; and (ii) subjectivity. While Bayesian probability provides unambiguous rules for how one should update their beliefs in light of new evidence, it allows for different individuals to start off with different prior beliefs.

Statistics helps us to reason backwards, starting off with collection and organization of data and backing out to what inferences we might draw about the process that generated the data. Whenever we analyze a dataset, hunting for patterns that we hope might characterize a broader population, we are employing statistical thinking.

Basics

How to Interpret Probability

As mentioned above, we can record how many times something happened and everything is a counting problem. The sample space is only what has happened and all probabilities are assigned to events in accordance with their real frequencies. Or we can use Bayesian and interpret just as a number with our intuition.

Or we can jump out of time and imagine that we are timeless. So everything does indeed has a definite frequency in this universe. This can be seen as merging frequentism with Bayesian. But since we aren't timeless, we can only imagine and try extrapolate what we have seen to infinity and reason timelessly.

Remember that mathematics is the study of Form, not perception. So we can always argue purely on theoretical ground without relating sample space and events to real things that has or will happen, or probability to frequency and times of happening. They can just be ideas and everything is built from the axioms.

Statistics and Set Theory

- Outcome (w): a precise description of the state of the world (as far as our model is concerned) and cannot be split any further
- Outcome/Sample Space (Ω): **set** of all possible outcomes or results of that experiment.
- Event (A): **subset** of sample space (grouping of outcomes we are interested in)
- Indicator function (I): I_A(w)=1 if w∈A 0 otherwise

Operations of events follow operations of sets:

- Union of events (∪): A∪B={w∈Ω:w∈A or w∈B or w∈both}

- Intersection of events ($\cap$): $A\cap B=\{w\in\Omega : w\in A \text{ and } w\in B\}$
- Disjoint events: $A\cap B=\phi$

Laws of Boolean Algebra

A,B,C are sets and are each a subset of X

1. $A\cup\phi=A$, $A\cap\phi=\phi$
2. $A\cup X=X$, $A\cap X=A$
3. $A\cup A=A$, $A\cap A=A$
4. $A\cup B=B\cup A$, $A\cap B=B\cap A$
5. $A\cup(B\cup C)=(A\cup B)\cup C$, $A\cap(B\cap C)=(A\cap B)\cap C$
6. $A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$
7. $A\cup(X\setminus A)=X$, $A\cap(X\setminus A)=\phi$
8. $X(A\cup B)=(X\setminus A)\cap(X\setminus B)$

> Note that they are similar to boolean logic because the "unit" of statement here is $x\in A$, in set the only difference is to check this for every x in the set

Axioms of Probability

- σ-algebra F: a set of events such that 1) $\Omega\in F$ 2) F is closed under complements 3) F is closed under countable unions
- Probability Function/Probability/Probability Measure (P): mapping from F to [0,1] that satisfies the following axioms
- Kolmogorov Axioms: $P(Ai) >= 0\ \forall Ai$; $P(\Omega) = 1$; $A_i...An$ are disjoint → $P(A_i \cup A_j ... \cup An)=P(A_i)+P(A_j)+...+P(An)$
- Probability Space: (S,F,P) where S is sample space, F is a σ-algebra, P is a probability function satisfying certain axioms

> See Math notes about measure

> Also see Cox's theorem as an alternative axiom for Bayesian Probability

> Frequentism VS Bayesian: Frequentism defines an event's probability as the limit of its relative frequency in many trials (the long-run probability). Probabilities can be found (in principle) by a repeatable objective process (and are thus ideally devoid of opinion). In the classical interpretation, probability was defined in terms of the principle of indifference, based on the **natural symmetry** of a problem, so, e.g. the probabilities of dice games arise from the natural symmetric 6-sidedness of the cube. This classical interpretation stumbled at any statistical problem that has no natural symmetry for reasoning.

Propositions of Probability Axioms

- $P(\phi)=0$
- $P(A)\leq 1$ for all A
- $P(\sim A)=1-P(A)$. From 7th law of boolean algebra
- if $E \subseteq F$, $P(E) \leq P(F)$

- P(A∪B)=P(A)+P(B)-P(A∩B). https://math.stackexchange.com/a/3367724 the proof requires set theory, laws of boolean algebra, axioms of probability, and probability theory

Conditional and Joint Probability

- P(A|B): $\frac{P(A ∩ B)}{P(B)}$, probability of A in the universe of B
- P(A, B): P(A ∩ B)=P(A)·P(B|A) multiplication rule
- P(A)=$\frac{P(A)·P(B|A)}{P(B|A)}=\frac{P(A,B)}{P(B|A)}$
- P(A)=$\sum_{i}P(A, B_i)=\sum_{i}P(A|B_i)·P(B_i)$ Total probability theorem
- P(·|B) follows axioms of probability. $P(\cup A_i|B)=\Sigma P(A_i|B)$
- In general P(A|B)≠P(B|A)

> Note that A and B are events and P is probability measure function

> Other useful equations:

- P(A)=P(A∩B)+P(A∩~B) binary total probability
- P(A|B,C)·P(B|C)=P(A,B|C)
- P(A|B,C)·P(B|C)·P(C)=P(A,B,C)

(Conditional) Independence

- Independence: P(A,B)=P(A)·P(B) i.e. P(A|B)=P(A)
- Conditional independence: P(A|B,C)=P(A|C)->P(A,B|C)=P(A|C)·P(B|C)

> Proof of conditional independence: $P(A|B,C)=P(A|C)=\frac{P(A,B,C)}{P(B,C)}=\frac{P(A,B,C)}{P(C)·P(B|C)}=\frac{P(A,C)}{P(C)}\iff\frac{P(A,C)·P(B|C)}{P(C)}=\frac{P(A,B,C)}{P(C)}=P(A|C)·P(B|C)$

> Note:

- A and B independent iff A and ~B independent
- independence is not disjointness; Two disjoint events can never be independent, except in the case that one of the events is null.

Pairwise & Mutual Independence

Random Variable

A function X: Ω→R. An assignment of a value to every possible outcome. Formally, a mapping from sample space to real numbers (from headings of coin to 0 and 1). Technically, we can't map every outcome.

> Every probability model comes with its sample space (and a probability). It is often left out of the discussion because all the action is carried out by random variables, but it always lurks underneath.

> Note that a continuous random variable must have uncountably infinite sample space but an uncountably infinite sample space doesn't imply continuous random variable (geometric,poisson)

Random Variable VS Algebraic Variable

Why X+X≠2X while x+x=2x and f+f=2f?

Observe that X is a random variable which depends on a sample space Ω. However, when we have another X, it means the sample space now becomes Ω^2 since we now care about outcome of both. So X is (ω,ω')→R. Y=X+X is also (ω,ω')→R. This essentially means Y is a multivariate function. When we add f: R→R we don't have multivariate function because its domain is just R. A normal X actually has infinite sample space but we care about only a small subset.

The probability distribution of the sum of two independent random variables is the convolution of each of their distributions. Observe that $ P(Y=X+X=y) \ = P({(ω,ω')∈Ω^2|Y(ω,ω')=y}) \ = P({(ω,ω')∈Ω^2|X(ω)+X(ω')=y}) \ = P({ω∈Ω|X(ω)=x}∩{ω'∈Ω|X(ω')=y-x}) \ = \sum_xP(X=x)P(X=y-x) $ which is the exact definition of convolution.

Probability Distribution

A probability distribution is a mathematical description of the probabilities of events, subsets of the sample space.

However, when dealing with random variables, we often use the range of X as the sample space instead of domain of X. This way it's easier to understand probability of events that have been quantified and the sample space becomes a numerical set instead of arbitrary non-numerical values like heads or tails.

The most general descriptions is in the form of P: A->R where A is related to sample space Ω. So when using the general probability function/measure form, A is the σ-algebra of Ω. And in the case of using random variables which is more common, A is the image of Ω under random variable X which is a numerical set.

Key takeaway is that anything of the form P: A->R satisfying certain conditions are probability distributions since it describes probability of outcomes. A plain capital P mapping events to R, PMF and PDF are all probability distributions.

PMF & PDF & CDF

- PMF: Probability mass function is the probability distribution of a discrete random variable (X takes countably many values). It is the function p: R to [0,1] defined by $p_X(x)$ = P(X=x) which is P({ω∈Ω: X(ω)=x}) where P is a probability measure. It's normally written as p(x)
- PDF: $P(a \le X \le b)\=P({ω∈Ω\ s.t.\ a\le X(ω)\le b})\=\int_{a}^{b}f_X(x)dx\$ $f_X(x)·δ \approx P(x \le X \le x+δ) $
- CDF: F(x)=P(x≤X)

> !Note: Probability function maps **event** to values (real number [0,1]), RV maps **outcome** to values, PMF maps **values** to values.

> Note: PMF is lower case p while probability measure is capital P.

Joint & Conditional Distribution

- Joint PMF: $p_{X,Y}(x,y)\=P(X=x\ and\ Y=y)\=P({ω∈Ω\ s.t.\ X(ω)=x}∩{ω∈Ω\ s.t.\ Y(ω)=y})$
- Conditional PMF (Single RV): $p_{X|A}(x)\=P(X=x|A)\=\frac{P({ω∈A\ st\ X(ω)=x})}{P(A)}$

- Conditional PMF (2 RV): $p_{X|Y}(x|y)\=P(X=x|Y=y)\=P({ω∈Ω\ st\ Y(ω)=y})\=\frac{P({X=x}∩{Y=y})}{P(y)}$

> Note that X and Y should be defined on the same sample space. When pairing random variables, we create tuples of their outcomes as element of the new sample space (Though in reality everything is always in the same big sample space, we normally ignore most irrelevant states).

> Note that when you have multiple RV, you could essentially combine them to a single random vector/element with a complex outcome space but that entangles information and is not good for modeling and inference.

> Note that conditional probability/distribution is defined on joint probability/distribution. There are some other definitions.

Mean & Variance

- E(X): $∫x·f_X(x)dx=∫X(e)·P(e)de$ where e ∈ E
- Var(X): $E[(X-EX)^2]=E(X^2)-EX^2$
- Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)
- Chebyshev's inequality
- Covariance: $E[(X-E[X])·(Y-E[Y])]=E(XY)-E(X)E(Y)=∫∫f_{xy}(x,y)(x-μ_x)·(y-μ_y)dxdy$
- Random variables whose covariance is zero are called uncorrelated.
- Covariance Matrices: covariance can only be calculated between two variables, use covariance matrix represents covariance values of each pair of variables in multivariate data

> Mean & Variance are properties of the image of events under random variable X. Not properties of the events themselves. Thus to make these properties meaningful wrt the underlying random events which we truly want to measure, the mapping X should assign meaningful values to corresponding events.

Distributions

http://d2l.ai/chapter_appendix-mathematics-for-deep-learning/distributions.html#exponential-family

X~N(0,1) means random variable X follows a probability distribution N(0,1).

x~P(X) means a value x sampled from distribution P

Bernoulli

- Bernoulli trial/experiment: sample space has only two elements
- Bernoulli RV: maps the two elements to 0 and 1
- Bernoulli distribution/pmf of a Bernoulli RV: pX: [0,1]->[0,1] = P(X=x), ie pX(0)=P(X=0), pX(1)=P(X=1)

Binomial

The binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n **independent** Bernoulli experiment. A sequence of outcomes is called a Bernoulli process; for a single trial, i.e., n = 1, the binomial distribution is a Bernoulli distribution.

- Bernoulli process:
- Binomial RV: maps one sequence of outcomes from the Bernoulli process to a number that indicates number of successes (1s) in that sequence
- Binomial distribution: pX(x): [0,n]->[0,1] = (n choose x)*p^x*q^(n-x)
- Expectation: np. use MGF to get M=(q+pe^t)^n
- Var: npq

Poisson

Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

> The difference between binomial and poisson is that poisson views the experiment process as continuous (binomial with p->0 and n->infinity)

p(i)=P(X=i)=e^-λ(λ^i)/i! (Taylor approximation of e^x)

Exponential Family

Law of Large Numbers

Sample mean converges to population mean.

Convergence of Probability Distribution

$∃a∀ε,ε'>0∃δ∀n>δP(|Xn-a|>ε)<ε'$

> Note the example of convergence on 0 but E is 1 and variance goes to infinity (P(0)=1-1/n, P(n)=1/n)

Markov's Inequality

If X is a non-negative random variable, then, for any a > 0, then P(X>a)≤EX/a

Proof:

$ EX=Σxp(x)≥\sum_{x>a}xp(x)≥\sum_{x>a}ap(x)=aP(X>a) $

Alternatively

$ Define\ I_A=1 \iff X>a \ A \implies X>a \implies X/a > 1\ and\ X/a>0 \ X/a ≥ I_A \ E(X/a) ≥ E(I_A) \ E(X)/a ≥ P(X>a) $

Chebyshev's Inequality

If X is a random variable with mean μ and variance σ^2, then, for any a > 0, $P(|X-\mu|\geq a)\leq\sigma^2/a^2$

Proof: $ Let\ Y=(X-μ)^2 \ P(Y>a)≤EY/a \ P((X-μ)^2>a)≤E[(X-μ)^2]/a \ P(|X-μ|>\sqrt{a})≤σ²/a \ Let a'=\sqrt{a} \ P(|X-μ|>a')≤σ²/a'^2 $

Weak Law of Large Number

If X1, X2... are independent identically distributed random variables (i.i.d.) with mean μ and standard deviation σ, then for any ε>0, P(|(X1+X2+...)/n-μ|>ε)→0 as n→∞

Proof: $ \tilde{X}=(X_1+X_2+...+X_n)/n \ E(\tilde{X})=n\mu/n=\mu \ Var(\tilde{X})=n\sigma^2/n^2=\sigma^2/n \ P(|\tilde{X}-E(\tilde{X})|>ε)≤\sigma^2/nε^2 \ $

> How do we get iid RV? Suppose that the whole population is determined, then every individual is also determined. e.g. in counting vote, every person either say yes or no, no randomness involved. However, we are selecting these people RANDOMLY. Therefore, the randomness comes from our free choice which we believe to be iid.

Central Limit Theorem

Distribution of sample mean is normal

> That many natural phenomena are themselves normal can be interpreted as them being sample mean of many hidden events. e.g. brownian motion can be seen as result of interaction with many particles around it

> There are many other limit theorems that extend beyond iid rv. A rule of thumb is that in real life, as long as you have enough samples, distributions don't need to be identical nor too independent.

Intuition

$ Sn=X1+X2+...+Xn \ \frac{Sn}{n}\text{ converges to mean of X}\ \frac{Sn}{\sqrt{n}}=\frac{Sn}{n}\sqrt{n}\text{ is more spread out and has variance of X} \ \frac{Sn-nEX}{\sqrt{n}\sigma} \text{ has zero mean and unit variance } \ Similarly \ \frac{\hat{X}-μ_X}{\frac{\sigma}{\sqrt{n}}} \text{ has zero mean and unit variance } $

Probability Distribution in Frequency Domain

The only function that keeps its shape after self convolution is Gaussian.

MGF

Binomial Poisson & Normal

Sum of Bernoulli converges to normal. Sum of Bernoulli is Binomial. Binomial converges to poisson. So does Poisson converges to normal? No because if we discretize poisson with tiny slots with p=λ/n, p changes as n changes which doesn't satisfy CLT assumption. What's wrong?

Binomial converges to poisson in a sense that λ=np is fixed, while Bernoulli converges to normal in a sense that p is fixed.

When λ is big enough (>=10), poisson approaches normal.

Stochastic Process

A stochastic process {Xt: t ∈ T} is a collection of random variables defined on a common probability space (Ω, F, P). The variables Xt take values in some set X called the state space which must be measurable with respect to some σ-algebra Σ. The set T is called the index set and for our purposes can be thought of as time. The index set can be discrete T = {0, 1, 2, . . .} or continuous T = [0, ∞) depending on the application.

Bernoulli

Bernoulli process is a finite or infinite sequence of binary random variables, so it is a discrete-time stochastic process that takes only two values, canonically 0 and 1. The component Bernoulli variables Xi are identically distributed and independent.

Markov

A discrete-time Markov chain is a sequence of random variables X1, X2, X3, ... with the Markov property, namely that the probability of moving to the next state depends only on the present state and not on the previous states.

The possible values of Xi form a countable set S called the state space of the chain.

Remarks

A Markov process is just like a physical process but with uncertainty. In an ideal world (you are God), processes are deterministic. For example, S=vt is describing a sequence of random variables Si conditioned on v and t where probability of Si taking the value vi*t is 1 (Note that for degenerate distributions like this, we don't use conditional probability but parametrization). But we are not God. We can only make guesses given incomplete information/noise.

When the world ends and we count all Si for all i (every possible distance every element traveled at every possible time). Probability of a certain S becomes a deterministic fraction. But suppose I can double the timeline, do we get the same fractions? The answer is yes if we carefully design what to include in the state. S itself is not enough, but (S,v) is enough. If I know (S,v) at anytime anywhere, I know (S,v)' at next instant. Since we model v as a random variable, (S,v) is also random. But because the transition is time homogeneous, it becomes a probabilistic loop where if I go back around to a certain state, my options and their probabilities are the same, much like an ordinary loop where if I go to a certain place in the loop I will take the exact same move as before.

In conclusion, since we are mortal, we use stochastic processes to make structured guesses. If we can find a state and transition mechanism that's time homogeneous, we also have the wonderful property of a probabilistic loop you are certain about the probability of the process at a certain state.

Transition VS State

Classical example of dynamic vs static. There is a duality between the two and knowing one gives you the other. You can say the world is only motion or only state. I prefer only state. You extract the world as small grids not only in space but also in time. You label their state with time. Then you have a set to do normal probability stuff.

Convergence

- Sequence of states Xi: the sequence doesn't converge to a number but the distribution of Xn does. (n means i to infinity)
- State: probability of any element Si of state space S converges to a real number
- Transition from Si to Sj: P(Xn=Sj|Xm=Si) when n and m are far apart, P converges to πj=P(Sj)=P(Xn=Sj)

Steady state

In a deterministic physical process, say dropping a ball in a bowl, no matter where you drop it, it slides down and slides up a little and eventually settle at the bottom. Being in the bottom is steady in that it doesn't change over time once it's there. But in other processes, say temperature, it goes up in summer and drops in winter. As winter comes to its peak, the temperature drops to its lowest, but there are fluctuations. Nevertheless, the set of temperatures have a steady probability distribution.

Linearity

Because of time homogeneity, an element of the set of steady states receives incoming transition with a constant rate. The constant rate itself, by duality, is constant steady state probability.

Bayes

Bayes' Theorem

P(A|B)=P(B|A)P(A)/P(B)

Although Bayes' theorem is a fundamental result of probability theory, it has a specific interpretation in Bayesian statistics. In the above equation, A usually represents a proposition (such as the statement that a coin lands on heads fifty percent of the time) and B represents the evidence, or new data that is to be taken into account (such as the result of a series of coin flips). P(A) is the prior probability which expresses one's beliefs about A before evidence is taken into account. The prior probability may also quantify prior knowledge or information about A. P(B|A) is the likelihood function, which can be interpreted as the probability of the evidence B given that A is true. The likelihood quantifies the extent to which the evidence B supports the proposition A. P(A|B) is the posterior probability, the probability of the proposition A after taking the evidence B into account. Essentially, Bayes' theorem updates one's prior beliefs P(A) after considering the new evidence B.

> Note that we use probability and probability distribution interchangeably. In a single trial, we have P(A) as a distribution, B is fixed so P(B|A) maps parameters to values. The resulting P(A|B) is also parameter to value and requires a simple function multiplication.

Statistical Inference

Statistical inference is the process of using data analysis to infer properties of an underlying probability distribution. It is assumed that the observed data set is sampled from a larger population.

Statistical Inference & Projection

Our primary senses used for inference is vision and sense of time. But we transform the 4D senses into language and symbols. In other words, the world projects upon us data.

There are processes that are simpler in nature and can be modeled exactly with few variables and algebraic relationships, like F=ma.

Some processes are more complex, like weather. But the method is the same. We try to disentangle inputs that have more direct relationship with outputs. An extreme example would be DNN where every input contributes linearly to hidden variables.

All of these is based on the belief that our universe has structures such that its projections contain traces of such structures.

Basics

Parametric Model

$F=\{f(x;\theta):\theta\in\Theta\}$

Point Estimation

Point estimation refers to providing a single “best guess” $\hat{θ}$ of some quantity of interest θ. θ is fixed (in frequentist view) while $\hat{θ}$ depends on data and is random

Estimator

An "estimator" or "point estimate" is a statistic (that is, a function of the data) that is used to infer the value of an unknown parameter in a statistical model. The parameter being estimated is sometimes called the estimand.

Suppose a fixed parameter θ needs to be estimated. Then an "estimator" is a function that maps the sample space to a set of sample estimates. An estimator of θ is usually denoted by the symbol $\hat\theta$

> Frequentist and Bayesian difference is to view the parameter as a constant or random variable, similar to God place dice or not.

Maximum Likelihood Estimation (MLE)

$\hat{θ}=\argmax_θ p_X(x;θ)$

Likelihood function: $L(\theta): \theta \mapsto R = p(x; \theta)$where p is a parametric family of distributions and x is the observed data sample.

The goal of maximum likelihood estimation is to find the values of the model parameters that maximize the likelihood function over the parameter space

> Recall: a parametric family or a parameterized family is a family of objects (a set of related objects) whose differences depend only on the chosen values for a set of parameters

> Example: we have x1,x2,...xn observed iid data with exponential distribution (we set a specific parametric family) with parameter θ. We know that with a given θ, which we believe is not random, an observation of x is $θe^{-θx}$. Since they are independent, the probability of jointly/simultaneously observing x1 through xn is $Π_iθe^{-θxi}$. Maximizing this is equal to maximizing its log. Take derivative and get $\hat{θ}=\frac{n}{x_1+...+x_n}$. The estimator in this case is $\hat{Θ}_n=\frac{n}{X_1+X_2+...+X_n}$

Maximum a Posteriori Estimation (MAP)

$\hat{θ}=\argmax_θ p_{Θ|X}(θ|x)$

In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.

We choose prior p(θ) and the statistical model p(x|θ), then we can get posterior using bayes rule. Finally we get the estimator as the mode of the posterior distribution $f(θ|x) \propto f(x|θ)\cdot f(θ)$ (p(x) doesn't matter because it's fixed after observation)

> The only difference between MLE and MAP is in MAP we have a prior. If the prior is uniform then MAP is the same as MLE.

Least Mean Squared (LMS)

minimize E[(Θ-estimate)^2]. optimal estimate E(Θ) where mean squared error is Var(Θ).

When you have data x, apply Bayes rule to get pΘ|X(θ|x) and use E[Θ|X=x] as estimator

Statistical Models

Regression

Regression is a method for studying the relationship between a response variable Y and a covariate X. The covariate is also called a predictor variable or a feature. One way to summarize the relationship between X and Y is through the regression function

$r(x) = E(Y|X=x) = \int y\cdot f(y|x)dy$.

Our goal is to estimate the regression function r(x) from data of the form

(Y1, X1),...,(Yn, Xn) ~ $F_{X,Y}$.

Simple Linear Regression

$Yi=\beta_0+\beta_1Xi+\varepsilon$ where ε is zero mean with variance σ^2

ML

Naive Bayes

Digit classification. Have image x want P(y|x) where y is a number from 0-9. Problem: compute P(y|x).

1. compute P(y|x) for every x (basically collecting all possible 32x32 images and assign probability to each one). Too large and no learning.
2. argmaxP(y|x)=argmaxP(x|y)P(y). P(x|y)=P(x1|y)·P(x2|x1,y)·...P(xd|x1,...xd-1,y). Still too 2^d computation
3. assume that each pixel is independent of each other. P(x|y)P(y) becomes ΠP(x_i|y)p(y). This is just a dxn matrix where element at (d,n) is P(xd=1|y=n)

Now we can estimate P(x_i|y) for every pixel. We count the number of occurrences for each of the n digits and divide it by the total amount of data.

Variational Bayeisan

Hidden variables can be interpreted from a Bayesian Statistics framework as prior beliefs attached to the observed variables. For example, if we believe X is a multivariate Gaussian, the hidden variable Z might represent the mean and variance of the Gaussian distribution. The distribution over parameters P(Z) is then a prior distribution to P(X).

ELBO

- Motivation: maximize p(x) with a latent model but p(x) is intractable through marginal over all latents.
- Why is ELBO lower bound for logp(x): KL is non negative.
- Why maximize ELBO: p(x) is a constant, so maximizing ELBO simultaneously minimizes KL of posterior (which can't be minimized directly without ground-truth posterior). We want to learn posterior because it models underlying latent structure of observed data.
- Interpretation of 15: probability of x is p(x,z)/p(z|x). If we replace p(z|x) with q(z|x), then p(x) will drop. The amount of drop is KL of p(z|x) and q(z|x).

VAE

ELBO can be further decomposed into 2 terms which correspond to a reconstruction and prior matching term.

Hierarchical VAE

Diffusion

Information Theory

Information

Measurement of uncertainty/randomness. How much more I know about the thing or how surprized I am after an event happens.

Entropy

Suppose we have a set of possible events whose probabilities of occurrence are p1, p2,... pn. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?

If there is such a measure, say H(p1, p2,... pn), it is reasonable to require of it the following properties:

1. H should be continuous in the pi.
2. If all the pi are equal, pi = 1/n, then H should be a monotonic increasing function of n. With equally n likely events there is more choice, or uncertainty, when there are more possible events.
3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

Alternatively:

1. The information we gain by observing a random variable does not depend on what we call the elements, or the presence of additional elements which have probability zero.
2. The information we gain by observing two random variables is no more than the sum of the information we gain by observing them separately. If they are independent, then it is exactly the sum.
3. The information gained when observing (nearly) certain events is (nearly) zero.

The only H satisfying the three above assumptions is of the form: $H=-K\Sigma p_i log p_i$

Self-Information

I(x)=-log(P(X=x))

base 2: bit; base e: nat.

> Note that information is used to describe events and entropy for random variable/system

Information as Bits

To get a sense of relationship between information as digital bits and happenings of events, it's easier to use a probabilistic example. If a 4 face die has outcomes with following distribution: 1 (1/8); 2 (1/8); 3 (1/4); 4 (1/2) then first and second event has 3 bits information, the third 2, and fourth has 1. If we throw the die 1000 times we expect to record roughly 125 1s and 2s, 250 3s and 500 4s. The recording would take 1750 bits and that's exactly entropy of the die throwing (1.75) times 1000. On the other hand if we just represent the 4 events with 4 different binary numbers (2 bits), we are essentially treating every event as equally possible (1/4) and this requires 2000 bits.

This means that in practice, if a random variable/system has high entropy, it is more random, and it requires more space to record/transmit it. If an event associated with the random variable has low probability, it means then recording it once would take more space. Optimally, entropy * number of recording = number of bits to transmit.

Joint Entropy

H(X,Y) = Ep(X,Y)[log(p(x,y))]

Max(H(X), H(Y)) <= H(X,Y) <= H(X)+H(Y)

Conditional Entropy

Example:

> suppose there are 16 images of 4 categories (each category with 10 images). X is the random variable mapping image to a number or to pixel values and Y from category label to number. Both PMF are uniform so H(X)=4 and H(Y)=2. Now since each image has exactly 1 label, $p_{Y|X}$=1/0, H(Y|X)=0 and H(X,Y)=4. This agrees with H(Y|X)=H(X,Y)-H(X) and indicates that X contains all the information about Y. On the other hand, $p_{X|Y}$=1/4 and H(X|Y)=2. H(X|Y)=H(X,Y)-H(Y), indicating that knowing label gives some but not all information about X. As an aside, if we want X and Y to be independent, then each image must be equally likely assigned any label, creating a sample space of 16*4 outcomes that have equal non zero probability (compared with 16).

Mutual Information

- I(X,Y)=T(Y,X)
- I(X,Y)>=0
- I(X,Y)=0 if X and Y are independent i.e. H(X,Y)=H(X)+H(Y)=H(X|Y)+H(Y|X) which is analogous to p(x,y)=p(x)*p(y)=p(x|y)*p(y|x). Notice that log term in H changed multiplication to summation.
- I(X,Y)=H(X)=H(Y) if X and Y are completely dependent (X is an invertible function of Y).

Point wise mutual information:

$pmi(x,y)=log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}$

This can be seen as measuring amount of surprise of seeing two things happening together compared to seeing them happen separately. Note that pmi can be positive and negative.

Entropy and Information

Self information can be seen as amount of things that add to your understanding. For example, in a coin flip, if you see a head, that's more information you have before the flip. So in another way, information is the amount missing before you see.

Entropy is a macrostate variable that doesn't depend on specific outcomes. You can't say entropy of a head. But the flip itself contains entropy. A news arguably contains more information than a flip. Or it gives you more knowledge about some states of the world. Or it is more random and harder to guess since news can be billions of combinations of events but coin flip has only 2 outcomes.

KL (Kullback–Leibler)

$D_{KL}(P||Q)=E_{x\sim P}[log\frac{p(x)}{q(x)}]$

- KL(P,Q)!=KL(Q,P)
- KL(P,Q)>=0, KL(P,Q)=0 if P=Q
- if there is an x such that p(x)>0 and q(x)=0, KL=∞

Cross Entropy

$CE(P,Q)=-E_{x \sim P}[log(q(x))]$

$CE(P,Q)=H(P)+D_{KL}(P||Q)$

Jensen-Shannon Divergence

$JSD(P||Q)=\frac{D(P||M)+D(Q||M)}{2},\ where\ M=\frac{P+Q}{2}$

Relationships

The following are equivalent:

- Maximize predictive likelihood
- Minimize CE
- Minimize KL