ML and Numerical Software Development

Probability and Information Theory

Organon Analytics

November 14, 2019

We must know, we will know

- Natural laws are expressed with the language of mathematics (algebra, geometry, analysis)
- Classical Mechanics: Newton's equations of motions
- Electro-Magnetism: Maxwell's equations
- Quantum Mechanics: Schroedinger's equations
- Fluid Dynamics: Navier-Stokes' equations
- Natural laws $\Leftrightarrow$ Mathematical equations (models)
- Extremely Successfull $\Rightarrow$ PHYSICS ENVY

Determinism

- The equations of physics (the laws of physics) are deterministic
- Re-run the experiment with the same parameters and conditions: The results are the same: NO SURPRISE
- What about the other disciplines involving humans?
- Economics, Finance, Marketing, Sociology, Psychology, etc.
- Models are everywhere
- However, they are not very precise. UNCERTAINTY creeps
- It appears in Physics as well: Quantum phenomena are NOT deterministic
- The tool to model uncertainty and randomness: Probability Theory

Mathematical Modelling

- A model takes input(s) and (may) produce output(s)
- The model itself involves, constants, parameters, and mathematical formulas (algebraic equations, logical equations, differential equations, etc.)
- Input(s) and output(s) are called DATA
- The mathematical formula $F(\cdot)$ is called the MODEL
- The goals of Machine Learning (in fact, Science in general) are
  - *i* Estimation: To LEARN model from FINITE data
  - *ii* Understanding: To UNDERSTAND and EXPLAIN the model
  - *iii* Generalization: To make new PREDICTIONS (via model) under new circumstances

Mathematical Modelling examples for ML

- Finance: Output: Default event, Inputs: All you know about customer
- Marketing: Which customer will buy what. Output: Purchase event, Inputs: All you know about customer
- Computer Vision: Output: Object category, Inputs: Image pixel values
- Healthcare: Output: Disease or not, Inputs: Patient medical history
- Retail: Output: Sales per product, Inputs: Historical sales, calendar, customer data

Sources of uncertainty

1. **Incomplete observation**: The data is generated by the system $D = X_1 + X_2$. But you can only observe $\{D, X_1\}$.
2. **Measurement errors**: Because of the measurement method, errors are introduced:
   - The system produces the data S (i.e. the signal)
   - You measure $S + N$ (i.e. (signal + noise))
3. **True randomness**
   - Quantum phenomena
   - Systems where data is generated by the independent actions of many agents (e.g., motion of particles suspended in a fluid, the price of a stock)

Probability: Sample Space & Events

Sample Space: Set of all outcomes that are generated by a process(experiment).
Examples:

a Flipping a coin: S = H, T

b Rolling a dice : S = 1, 2, 3, 4, 5, 6

c Lifetime of a car: S = [0, infinity)

d Flipping two coins: S = (H,H), (H,T), (T,H), (T,T)

Event: Any subset of the sample space S is called an event
Examples:

a Even numbers on a rolled dice: E = 2, 4, 6

b Observing at least one head on two flipped coins: E = (H,H), (H,T), (T,H)

c Lifetime of a car: E = [2,6]. Event that the car lasts between two and six years

Probability: Sample Space & Events cont'd

Events are sets of outcomes. So we can talk about:

a Their unions: $E \cup F$ (E OR F).

b Their intersections: $E \cap F$ (E AND F)

c Their differences $E \setminus F$ (E but not F)

d Their complements: E^c (NOT E

If $E \cap F = 0$, events are **mutually exclusive**:

E: sum of the numbers on dice even

F: sum of the numbers on dice odd

E: sum of the numbers on dice even

F: sum of the numbers on dice odd

E: customer defaults on the credit account

F: customer pays in full

Axioms of probability

The probability is a function defined on **event space** obeying the following axioms:

1. $0 \leq P(E) \leq 1$ (0: impossibility, 1: certainty)
2. $P(S) = 1$ (What is observed is an outcome)
3. For any sequence of events E_1, E_2, E_3 that are mutually exclusive

$$P(\cup E_n) = \sum_n P(E_n)$$

Examples:

- Fair coin: P(H) = 1/2, P(T) = 1/2
- Biased coin: P(H) = 2/3, P(T) = 1/3
- Loaded dice: P(1) = 1/4, P(6) = 1/12, P(E) = 1/6 for $E \in \{2, 3, 4, 5\}$

Important properties of probability

1. $P(E^c) = 1 - P(E)$. e.g. $P(head) = 1 - P(tail)$
2. $P(S) = 1$: At least one of the outcomes is observed.
3. For any sequence of events $E_1, E_2, \cdots, E_n$ that are mutually exclusive

$$P(\cup E_n) = \sum_n P(E_n)$$

Examples:

- Coin flip: $E = \{\text{at least one head}\}$, $E^c = \{\text{all tails}\}$
- Rolling dice: $E_i = \{sum = i\}, i \in [2, 12]$

$$P(\cup E_n) = \sum_n P(E_n) = 1$$

Example: Try to estimate the probability of a customer buying a pair of female shoes (**FS**) with:

- {No information}
- {Gender}
- {Gender, past purchase }

Distributions of 100 transactions

Gender	**FS Past purchase**	**FS**	**Other**
Male	False	0	30
Male	True	1	19
Female	False	8	28
Female	True	1	13

Uncertainty decreases as the data increases

Random Variables

Remember sample spaces!
A random variable is a real-valued function defined on a sample space:

$$X : \text{Sample space} \longrightarrow R^1$$

It is a variable since it takes different values: For each trial, it assumes a different value.
Example: Sum of the values on two fair dices is a random variable. X takes the integer values between [2, 12]

- $P(X = 2) = P(\{1, 1\}) = 1/36$
- $P(X = 12) = P(\{6, 6\}) = 1/36$

Random Variables

Example: Coin tossing experiment. $P(H) = p$. Define

$N =$ the number of flips required till the first appearance of a head

Then

$$
\begin{aligned}
P(N = 1) &= p \\
P(N = 2) &= (1 - p)p \\
P(N = 3) &= (1 - p)^2 p \\
&\vdots \\
P(N = k) &= (1 - p)^{(k-1)} p
\end{aligned}
$$

Cumulative distribution function

Random variables are completely characterized by its cumulative distribution function:

$$F_X(x) = P(X <= x)$$

i F is non-decreasing

ii $F(-\infty) = P(X <= -\infty) = 0$

iii $F(-\infty) = P(X >= \infty) = 1$

Example: CDF for Bernoulli r.v. with $P(H) = p$

$$f(x) = \begin{cases} 0 & : x < 0 \\ (1-p) & : x \in [0,1) \\ 1 & : x \geq 1 \end{cases}$$

Discrete Random Variables

X is discrete $\Leftrightarrow$ if X takes a countably finite number of values
Assume that the values X can take are in the set $\{x_1, x_2, \cdots, x_n\}$.
The *probability mass function* of X is defined as

$$p(x) \equiv P(X = x)$$

Note that

$$\begin{aligned}\sum_i p(x_i) &= \sum_i P(X = x_i) = 1 \\ F_X(a) &= \sum_{x_i <= a} p(x_i)\end{aligned}$$

Bernoulli (binary outcome)

Experiments with 2 outcomes: Event happens (positive event), Event does not happen (negative non-event).
Define the Bernoulli r.v. as follows:

$$p(x) = \begin{cases} X = 1 & \text{: if event happens} \\ X = 0 & \text{: if event does not happen} \end{cases}$$

- $P(event) = P(X = 1) = p$
- $P(non - event) = P(X = 0) = 1 - p$

where p is *event probability*
Examples:

- Credit Risk: {default, no-default}
- E-commerce: {purchase, no-purchase}
- Healthcare: {disease, no-disease}
- Schroedinger's cat: {(dead, alive}

Binomial r.v. (sums of Bernoullis)

Number of events in a sequence of **identical** and **independent** Bernoullis

- $S_n = \sum_{k=1}^{n} X_k = X_1 + X_2 + ... + X_n$
- S_n takes values from 0 (no event) to n (all event)
- No event prob.: $P(S_n = 0) = (1 - p)^n$
- $P(S_n = k) = (nk)p^k(1 - p)^(n - k)$
- All event prob.: $P(S_n = 0) = (1 - p)^n$

Note: Statistics is (mostly) about sums and limits-of-sums of random variables

Poisson random variable

A random variable taking non-negative integer values with the following mass function is a Poisson r.v.:

$$P(X = i) = p(i) = e^{(-\lambda)} \frac{\lambda^i}{i!}$$

Poisson random variable is defined to model count of events (so called arrival processes). The parameter λ corresponds to the density of events. Examples:

- Number of customers entering the branch since the morning
- Number of accidents in the highway each day
- Number of days since the default event
- Number of years left till death

Continuous random variables

- The range of X is not finite but (potentially) infinite
- NO probability mass function (i.e. P(X = x))
- One can only talk about $P(X) \in [x, x + \delta x]$ where delta δx is infinitely small

$f_X(x) = P(X) \in [x, x + \delta x]$ is called the **density function**.
Remember CDF $F_X(x) = P(X \leq x)$: $F_X(x)$ and $f_X(x)$ carries the same information: PDF is the derivative of CDF

$$\frac{dF_X(x)}{dx} = f_x(x)$$

Examples:

- Income of a customer
- Lifetime of a product
- Life expectancy of a person
- Number of sales per product/store/total

Uniform random variable

$$f_X(x) = \begin{cases} 1 & : x \in [0,1] \\ 0 & : \textit{elsewhere} \end{cases}$$

- Most important cont. RV From a **computational** perspective: All other interesting RVs could be derived from it
- let X be uniform. Define Y as follows:

$$f(x) = \begin{cases} Y = 1 & : X \leq p \\ 0 & : \textit{otherwise} \end{cases}$$

Then Y is Bernoulli

Normal (Gaussian) random variables:

The single most important (continuous) random variable encountered in nature (due to Central Limit Theorem)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp^{-\frac{(x-\mu)^2}{\sigma^2}}$$

- μ is the location parameter, σ is the dispersion parameter
- Used to model quantities that arise from the sum of many independent events
- Appears in central limit theorem: Sums of random variables converge to normal r.v.s

Expectation (average) of a random variable

Expected value of a random variable is defined as

- Discrete case: $E[X] = \sum_{x_i} x_i p(x_i) = \sum_{x_i} x_i P(X = x_i)$
- Continuous case: $E[X] = \int_{-\infty}^{\infty} x f_X(x)$

Expectation operator is linear

$$E[aX + bY] = aEX + bE[Y]$$

- Conceptually it refers to the central tendency(average) of X
- If you want to summarize a random variable with a single number, $E[X]$ is your number

Expected values of important random variables:

- $E[\text{Bernoulli}(p)] = p$
- $E[\text{Poission}(\lambda)] = \lambda$
- $E[\text{Uniform}[a, b] = 0.5 \times (a + b)$
- $E[\text{Normal}(\mu, \sigma^2)] = \mu$

Expectation of a function of a random variable

Most of the time, one is interested in the expectation of **a function** of the random variable

- Discrete case:
  $E[g(X)] = \sum_{x_i} g(x_i)p(x_i) = \sum_{x_i} g(x_i)P(X = x_i)$
- Continuous case: $E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)$

Example: $X \sim \text{Bernoulli}(p)$.

$$
\begin{aligned}
E[X^2] &= \sum_{x_i} x_i^2 P(X = x_i) = 0^2 P(X = 0) + 1^2 P(X = 1) \\
&= 0 \times (1 - p) + 1 \times p = p
\end{aligned}
$$

Expectation of a function of a random variable: cont'd

Example: Variance of an r.v. is defined as
$Var(X) = E[(X - E[X])^2]$

- $E[X]$: central value
- $Var(X)$: deviations (distance, dispersion) from the central value

If $X \sim \mathrm{N}(\mu, \sigma^2)$.

$$Var(X) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma}} \exp^{-\frac{(x-\mu)^2}{\sigma^2}} dx = \sigma^2$$

Jointly distributed random variables

Joint analysis of ≥ 2 RVs together. Why important?

- ML algorithms analyzes many RVs at once: many outputs, many inputs
- An ML algorithm **in essence** tries to **learn** joint density from **finite** samples

Both **discrete**: $X \in x_1, x_2, \cdots, x_m$ and $Y \in y_1, y_2, \cdots, y_n$. joint PMF is defined as

$$p(x_i, y_j) = P(X = x_i, Y = y_j) i = 1, 2, ...m, j = 1, 2, ...n$$

Both **continuous**:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$
$$f_{XY} = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}$$

Independence of Random Variables

X and Y are independent if and only if one of the following holds

$$\begin{aligned} P(X <= a, Y <= b) &= P(X <= a) \times P(Y <= b) \text{ for all a, b} \\ F_{XY}(x,y) &= F_X(x) \times F_Y(y) \text{ distributions separable} \\ f_{XY}(x,y) &= f_X(x) \times f_Y(y) \text{ densities separable} \\ E[g(X)h(Y)] &= E[g(X)] \times E[h(Y)] \ \forall g, h \end{aligned}$$

Insight: Knowing X does NOT tell you any information about Y and vice versa

Variance, Covariance, Correlation

- $Cov(X, Y) \equiv E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- If X, Y are independent $Cov(X, Y) = 0$
- $Var(X) = Cov(X, X)$
- $Corr(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$ where $\sigma_X = \sqrt{Var(X}$

  If $Abs(Cov(X, Y) > 0)$, X, Y are correlated: Knowing one of them tells you information about the other

  if $Cov(X, Y)$ is large and negative, it means that while X increases away from its mean, Y decreases away from its mean

Conditional Probability and Conditional Expectations

- Measure Theory(MT) is a branch of Mathematics: Generalizes the notions of length, area, and volume
- Probability Theory = MT + Conditionality
- Conditional Probabilities and Expectations: The machinery of ML computations
- The most important concepts Probability for ML.

Remember events (subsets of a probability sample space).
Conditional probability of E given F is defined as

$$P(E|F) \equiv P(EF)/P(F)$$

$P(E|F)$ meaning: the probability of event E given that the event F occurred

Example:

- E: purchase of woman shoes
- M: $\{\text{gender} = \text{male}\}$
- F: $\{\text{gender} = \text{female}\}$

$$P(E|F) = \frac{P(EF)}{P(F)},\ P(E|M) = \frac{P(EF)}{P(M)}$$

If $P(F) \geq P(M) \Rightarrow P(E|F) \geq P(E|M)$

- $P(\text{Event})$: (The marginal) probability of event with no condition present
- $P(\text{Event}|\mathit{Condition})$: (The conditional) probability of event in the presence of a condition

Consider the probabilites $P(death)$, $P(death|young)$, $P(death|old)$

$$P(death|old) \geq P(death) \geq P(death|young)$$

Conditional Expectations: Discrete Case

Conditional probability mass function:

$$
\begin{aligned}
P_{X|Y}(x,y) &\equiv \frac{p_{XY}(x,y)}{p_Y(y)} \\
p_X(x) &= \text{marginal dist of } X \\
p_Y(y) &= \text{marginal dist of } Y \\
p_{XY}(x,y) &= \text{joint dist of } X, Y \\
p_{X|Y}(x,y) &= \text{cond. dist of } X \text{ given } Y
\end{aligned}
$$

Conditional expectation now is defined as

$$E[X|Y=y_j] = \sum_{x_i} x_i P(X = X_i | Y = y_j)$$

Note: Conditional expectation is a random variable and is a function of Y

Conditional Expectations: Discrete Case

Conditional probability mass function:

$$
\begin{aligned}
P_{X|Y}(x,y) &\equiv \frac{p_{XY}(x,y)}{p_Y(y)} \\
p_X(x) &= \text{marginal dist of } X \\
p_Y(y) &= \text{marginal dist of } Y \\
p_{XY}(x,y) &= \text{joint dist of } X, Y \\
p_{X|Y}(x,y) &= \text{cond. dist of } X \text{ given } Y
\end{aligned}
$$

Conditional expectation now is defined as

$$E[X|Y = y_j] = \sum_{x_i} x_i P(X = X_i|Y = y_j)$$

Note: Conditional expectation is a random variable and is a function of Y

Bayes Rule:

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

$$\begin{aligned} P(\text{model}|\text{data}) &= P(\text{data}|\text{model})P(\text{model})/P(\text{data}) \\ P(\text{output}|\text{input}) &= P(\text{input}|\text{output})P(\text{output})/P(\text{input}) \\ P(\text{model}|\text{evidence}) &= P(\text{evidence}|\text{model})P(\text{model})/P(\text{evidence}) \end{aligned}$$

Note: Think about the minorities and the prejudices!
Example:

$$P(\text{Race}|\text{Crime}) = \frac{P(\text{Crime}|\text{Race})P(\text{Race})}{P(\text{Crime})}$$

READ: Chapter-14 from **Thinking Fast and Slow** from Daniel Kahneman