Probability Primer for Supervised, Unsupervised and Reinforcement Learning

Earl Wong

Two Paths

- Rigorous Development / Measure Theory Approach
  Probability: Theory and Examples, Richard Durrett

- Intuition / Heuristic Approach
  Probability Models, Sheldon Ross

- For rigorous probability proofs, we use the former.

- For practical applications, we use the latter.

- We will focus on the latter.

Preliminaries

- Sample space: S - the set of all possible outcomes of an experiment.
- Event: E - a subset of the sample space.
- Probability of Event E: P(E)

Conditions for P(E)

- 1) $0 \leq P(E) \leq 1$
- 2) $P(S) = 1$
- 3) For $E_1, E_2, E_3 \dots$ where $E_i \cap E_j = \emptyset, \quad i \neq j$

$$P(\bigcup_{n=1}^{\text{inf}} E_n) = \sum_{n=1}^{\text{inf}} P(E_n)$$

Properties

$$P(E \cup F) = P(E) + P(F) - P(E \cap F), \quad P(EF) = P(E \cap F)$$

$$P(E/F) = \frac{P(E \cap F)}{P(F)}$$

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

$$P(E \cap F) = P(E)P(F), \ for\ E\ independent\ of\ F$$

Random Variable

- Random variable: X - a real valued function defined on S

Cumulative Distribution Function (CDF)

CDF F( ) defined on random variable X:

$F(a) = P(X \leq a), \; -\infty < a < \infty$

has properties

- 1) $F(a)\ is\ a\ non-decreasing\ function\ of\ a$
- 2) $lim_{a \to \inf} F(a) = 1$
- 3) $lim_{a \to -\inf} F(a) = 0$

Example

- Experiment: Roll a pair of dice
- All outcomes S = $\{(1,1),(1,2),\ldots(1,6),(2,1),(2,2),\ldots(5,6),(6,6)\}$

$$Event\ E = \{all\ first\ dice\ values = 1\} = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}$$
$$Event\ F = \{all\ first\ dice\ values \leq 2\} = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots(2,6)\}$$

Example

$$P(E) = \frac{6}{36}$$

$$P(F) = \frac{12}{36}$$

$$P(E \cap F) = P(\{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)\} = \frac{6}{36}$$

$$P(F/E) = 1$$

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) = \frac{6}{36} + \frac{12}{36} - \frac{6}{36} = \frac{12}{36}$$

$$P(E/F) = \frac{P(E \cap F)}{P(F)} = \frac{\frac{6}{36}}{\frac{12}{36}} = \frac{1}{2}$$

$$P(E/F) = \frac{P(F/E)P(E)}{P(F)} = \frac{1 * \frac{6}{36}}{\frac{12}{36}} = \frac{1}{2}$$

$$P(E \cap F) = \frac{6}{36} \neq P(E)P(F) = \frac{6}{36}\frac{12}{36}, \textit{ E and F are not independent events}$$

Example

- Random variable: X - a real valued function defined on S
- Let the function be defined as the sum of the two dice values.

Example

$$P[X = 2] = P\{(1,1)\} = \frac{1}{36}$$

$$P[X = 3] = P\{(1,2),(2,1)\} = \frac{2}{36}$$

$$P[X = 4] = P\{(1,3),(2,2),(3,1)\} = \frac{3}{36}$$

$$\vdots$$

$$P[X = 12] = P\{(6,6)\} = \frac{1}{36}$$

$$F(b) = P[X \leq 3] = P\{(1,1),(1,2),(2,1)\} = \frac{3}{12}$$

Types of Random Variables

Discrete: $p_X(a) = P(X = a)$

- Bernoulli, Binomial, Poisson

Continuous: $P[X \in A] = \int_A f_X(x)dx$

- Gaussian / Normal

Facts

- Every random variable has a distribution.
- In the discrete case, this distribution is called the probability mass function: $p_X(x)$
- In the continuous case, this distribution is called the probability density function: $f_X(x)$
- The distribution sums (discrete) or integrates (continuous) to 1 for the input domain x.

Bernoulli & Binomial Random Variables

- Bernoulli: Success / failure

$$p_X(0) = P[X=0] = 1-p$$
$$p_X(1) = P[X=1] = p$$

- Binomial: n independent trials of Bernoulli random variable, (n,p)

$$p_X(k) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,1,2,\ldots,n$$

Poisson Random Variable

- Used to model image sensor shot noise

$$p_X(i) = P[X = i] = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0,1,2,...$$

Gaussian / Normal Random Variable

- ***Occurs naturally in nature.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad x \in \mathbb{R}$$

Expectation

- Expectation of a random variable X: E[X]
- Expectation of a function of a random variable X: E[g(X)]

Expectation

- Discrete random variable X:

$$E[X] = \sum_x x\, p_X(x) = \sum_i x_i\, P(X = x_i)$$

- Continuous random variable X:

$$E[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

Expectation for the Function of a Random Variable X

- Discrete: g(X)

$$E[g(X)] = \sum_x g(x)\, p_X(x) = \sum_i g(x_i)\, P(X = x_i)$$

- Continuous: g(X)

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

Commonly Encountered Functions of a Random Variable

- Shift and Scale of random variable X: $g(X) = aX + b$
- Variance of random variable X: $g(X) = (X - \mu)^2$

Joint Random Variables

- Previously, the CDF for random variable X:

$$F(a) = P(X \leq a), \ -\infty < a < \infty$$

- Now, the CDF for the jointly distributed random variables X and Y:

$$F(a, b) = P(X \leq a, Y \leq b), \ -\infty < a, b < \infty$$

Joint Probability Mass / Density Function

- Joint probability mass function

$$p_{X,Y}(x, y) = P[X = x, Y = y]$$

- Joint probability distribution function

$$P[X \in A, Y \in B] = \int_B \int_A f_{X,Y}(x, y) dx dy$$

Joint Probability Mass / Density Function

- Marginal (discrete)

$$p_X(x) = \sum_y p_{X,Y}(x, y)$$

$$p_Y(y) = \sum_x p_{X,Y}(x, y)$$

- Marginal (continuous)

$$f_X(x) = \int_B f_{X,Y}(x, y)dy$$

$$f_Y(y) = \int_A f_{X,Y}(x, y)dx$$

Expectation of g(X,Y)

- Discrete: g(X,Y)

$$E[g(X,Y)] = \sum_{y} \sum_{x} g(x,y) p_{X,Y}(x,y)$$

- Continuous: g(X,Y)

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dx dy$$

Commonly Encountered Functions of Random Variables

- $g(X,Y) = X + Y$
- $g(X,Y) = aX + bY$
- $E[g(X,Y)] = E[X+Y] = E[X] + E[Y]$
- $E[g(X,Y)] = E[aX+bY] = aE[X] + bE[Y]$

Independent Random Variables

- The random variables X and Y are independent if for ALL a, b:

$$P[X \leq a, Y \leq b] = P[X \leq a]P[Y \leq b]$$

- For X and Y discrete:

$$p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

- For X and Y jointly continuous:

$$f_{X,Y} = f_X(x)f_Y(y)$$

Useful Results When X and Y are Independent

- Let X and Y be independent random variables.
- $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$
- $Var(X+Y) = Var(X) + Var(Y)$
- $Cov(X,Y) = 0$

Covariance

- The covariance measures the relationship between random variables - i.e. how do the random variables “vary” with respect to one another.

- For two random variables X and Y, the covariance is defined as: E[(X-meanX)(Y-meanY)] = E[XY] - (E[X]E[Y])

- The covariance can be positive, negative or 0.

- If X and Y are independent, the covariance is 0.

- If the covariance is 0, X and Y are not necessarily independent though.

Transforming Joint Random Variables

$Let\ X_1 = X\ and\ X_2 = Y, with\ distribution\ f_{X_1,X_2}(x_1, x_2)$

$Define\ Y_1 = g_1(X_1, X_2)\ and\ Y_2 = g_2(X_1, X_2)$

$g_1\ and\ g_2\ have\ continuous\ partial\ derivatives.$

$What\ is\ the\ new\ distribution\ f_{Y_1,Y_2}(y_1, y_2)?$

Transforming Joint Random Variables

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}\left(x_1, x_2\right)\left|\begin{matrix}\frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2}\end{matrix}\right|^{-1}$$

Then, solve for x_1 *and* x_2 *in terms of* y_1 *and* y_2 *and substitute.*

$i.e.\ x_1 = h_1(y_1, y_2)$ *and* $x_2 = h_2(y_1, y_2)$

Generalization

- We have shown results for 2 random variables, X and Y.

- However, everything (previously stated) for 2 random variables, can easily be generalized to n random variables: $\{X_1, X_2, \ldots X_n\}$

Limit Theorems

- There are two main theorems that result from what we have learned so far:
- 1) Law of Large Numbers
- 2) Central Limit Theorem

Law of Large Numbers

Suppose we sample from an IID (independent, identically distributed) distribution: $X_1, X_2, \ldots X_n$

Let: $X = \dfrac{X_1 + X_2 + X_3 + \ldots + X_n}{n}$

Then: $lim_{n \to \infty} X = \mu,\ where\ E[X_i] = \mu$

Central Limit Theorem

Suppose we sample from an IID distribution with mean and variance

$$E[X_i] = \mu, \quad E[(X_i - \mu)^2] = Var(X_i) = \sigma^2$$

Then, ***

$$lim_{n \to \infty} \frac{(X_1 + X_2 + \ldots + X_n - n\mu)}{\sigma\sqrt{n}} \to N(0,1)$$

Stochastic Process / Random Process

- A stochastic process / random process is a time indexed sequence of random variables.

$$[X(t), t \in T], \; where \; X(t) \; is \; a \; collection \; of \; random \; variables \, .$$

- Stochastic processes are widely used in math, engineering, economics, physics …

Conditional Probability and Associated Conditional Probability Mass and Conditional Probability Density Function

- Previously, conditional probability: $P(E/F) = \dfrac{P(EF)}{P(F)}$

- Conditional probability mass function: $p_{X/Y}(x/y) = \dfrac{p_{X,Y}\, p(x, y)}{p_Y(y)}$

- Conditional probability density function: $f_{X/Y}(x/y) = \dfrac{f_{X,Y}\, f(x, y)}{f_Y(y)}$

Conditional Expectation

$$E[X/Y = y] = \sum_{x} x p_{X/Y}(x/y)$$

$$E[X/Y = y] = \int_{-\infty}^{\infty} x f_{X/Y}(x/y) dx$$

The Power of Conditioning

- Conditioning, allows you to introduce additional information into a problem.
- In probability space, we conditioned on events, allowing us to compute the probability for event A, given that event B had occurred.
- With random variables, conditional expectation allows us to improve the estimate of the mean of the random variable, given knowledge of it’s interaction with other random variables.
- We now show an example, where the interaction involves the total number of occurrences.

Example

- The poisson random variable counts the number of events that occur in a fixed interval of time.
- The probability mass function returns the probability of k events occurring, in a fixed interval of time.
- Suppose we observe that n total events have occurred for two independent poisson random variables X and Y. i.e. $X+Y = n$.
- Now, suppose we want to know the average number of events that occurred for the poisson random variable X, given the above information.
- How do we solve this?

Example

- Initial thought: Compute E[X], since X and Y are independent.

- Rebuttal: It is true that the random variables are independent. With no additional knowledge, we would expect the frequency of occurrence for X would be it's mean, for it's given rate parameter.

- However, we know that n events have occurred between the two random variables.

- => We need to make use of this information, to get a better estimate.

Example

- Problem formulation: $E[X \mid X + Y = n]$

- Since we want to compute the conditional expectation over X, we need the conditional probability mass function:

- $$P[X = k \mid X + Y = n] = \frac{P[X = k, X + Y = n]}{P[X + Y = n]}$$

$$P[X = k \mid X + Y = n] = \frac{P[X = k, Y = n - k]}{P[X + Y = n]}$$

Example

- Now, applying the knowledge that X and Y are independent:

$$P[X = k \,|\, X + Y = n] = \frac{P[X = k]P[Y = n - k]}{P[X + Y = n]}$$

- We can now substitute for all three terms, by applying the definition of the poisson random variables for X, Y and Z=X+Y, to obtain the conditional probability mass function.

- Finally, we apply the conditional expectation equation to the conditional probability mass function.

Conclusion

- These slides are meant to be a quick refresher for the fundamentals of probability.
- We have touched upon the highlights, focusing on the core results.
- Many ideas from deep learning, result from these simple and basic fundamentals.

Examples

- Unsupervised learning via Flow results from the idea of transforming distributions / random variables.
- Unsupervised learning via Diffusion is rooted in parameterized Markov chains (not discussed here).
- The latent vector z in supervised and unsupervised learning is based on a multivariate normal.
- Monte Carlo rollouts in reinforcement learning simulations are rooted in the law of large numbers.
- Image denoising using deep networks relies on a shot noise model rooted in the poisson distribution.

Examples

- Mutual information measures how much one random variable contains about another.

- Entropy measures the amount of uncertainty or randomness, in a random variable.

- KL Divergence measures the similarity / dis-similarity between two probability distributions.