AIML - CS 337

Lecture 1: Probability Theory

8 August 2022

Lecturer: Abir De *Scribe: Group 1 and Group 2*

In the first lecture of the course, we began with a review of the basics of probability theory and some discussions on "what is ML?". We discussed how ML is used to recognize patterns in data, without explicitly being told which patterns there are.

We recalled the characteristics of Supervised Learning and Unsupervised Learning, which differ in whether the training data available is labelled or not. We then began with the review of probability theory.

1 Probability

1.1 Review

- **Sample Space (S)** : In probability theory, the sample space of an experiment or random trial is defined as the set of all possible outcomes of that experiment.

$$P(S) = 1,\ P(\emptyset) = 0$$

- **Probability Distribution (p)** : Probability Distribution is a Mathematical function which outputs the probabilities of occurrence of different possible outcomes of an experiment.

$$p : S \rightarrow [0, 1]$$
$$s.t. \sum_{x \in S} p(x) = 1$$

- **Random Variable (X)**: A random variable can be defined as a numerical description of the outcome of a statistical random experiment. It is a mapping from the set of outcomes in the sample space to numbers.

- **Event(E)** : Events can be defined as the set of outcomes of an experiment. An event can be just one outcome or it can be a combination of more than one outcome from an experiment. It can be defined as a subset of the experiment's sample space.

$$E \subseteq S$$

  For example, in the dice roll experiment, the sample space is S = {1,2,3,4,5,6} ; one possible outcome is "1"; while a possible event is "The dice rolled an odd number", E = {1,3,5}.
  We often describe the events using conditions and these events can be combined using logical operations like **or**,**and**.

- **Probability of an Event** : It tells us the likelihood of happening of the event. Mathematically it is the total weight assigned to all the elements in the event by a given probability distribution.

$$P(E) = \sum_{x \in S} p(x)$$
$$P(\overline{E}) = 1 - P(E),\ \overline{E} = S \setminus E$$
$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

  - **Union bound :** This is an inequality used extensively in probability theory. It states that the probability of union of some events is less than or equal to the sum of probabilities of each individual event.

$$P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$$

  - If E_1, E_2, . . ., E_n are pairwise disjoint events, then

$$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$$

  - **Conditional Probability** It is the probability of an event happening given that another event has already occurred. For events E_1 and E_2 (s.t $P(E_2)>0$)

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

  - **Bayes' Rule** This Rule can be derived from the definition of conditional probability. It is named after Thomas Bayes(1701-1761). For events A and B (s.t P(A),P(B)>0)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

1.2 Applications of probability

1.2.1 Probability in Medical Testing

A lab test has a probability 0.95 of detecting a disease when applied to a person suffering from said disease, and a probability 0.10 of giving a false positive when applied to a non-sufferer. If 0.5% of the population are sufferers, what is the probability of a test being positive?

Solution: Let

Pos be the event that result of the test is Positive
Neg be the event that result of the test is Negative

S be the event that person is a sufferer
NS be the event that person is not a sufferer.

Given that

$$P(Pos|S) = 0.95$$
$$P(Pos|NS) = 0.10$$
$$P(S) = 0.005$$

we need to find probability of a test being positive i.e $P(Pos)$

$$P(Pos) = P(Pos \cap S) + P(Pos \cap NS)$$
$$P(Pos) = P(Pos|S) * P(S) + P(Pos|NS) * P(NS)$$
$$P(Pos) = 0.95 * 0.005 + 0.10 * 0.995$$
$$P(Pos) = 0.10425$$

0.10425 is the probability of a test being positive

1.2.2 Probability in NLP

Part Of Speech (POS) tagging is a popular Natural Language Processing (NLP) Problem.
Here Input is a set of N words and the output is POS tag for each word.

Assuming each word is independently drawn from a fixed vocabulary, find the probability that a sentence of length m contains a 'noun' given that it contains a 'verb'.

Solution: Let p_k be the probability that a word is of POS type 'k' and let A_k be the event that sentence contains POS type 'k'.

we need to find $P(A_{noun}|A_{verb})$. we already know that

$$P(A_{noun}|A_{verb}) = \frac{P(A_{noun} \cap A_{verb})}{P(A_{verb})}$$
$$P(A_{noun} \cap A_{verb}) = 1 - P(\overline{A_{noun} \cap A_{verb}})$$
$$P(A_{noun} \cap A_{verb}) = 1 - P(\overline{A_{noun}} \cup \overline{A_{verb}})$$
$$P(A_{noun} \cap A_{verb}) = 1 - (P(\overline{A_{noun}}) + P(\overline{A_{verb}}) - P(\overline{A_{noun}} \cap \overline{A_{verb}}))$$
$$P(A_{noun} \cap A_{verb}) = 1 - P(\overline{A_{noun}}) - P(\overline{A_{verb}}) + P(\overline{A_{noun}} \cap \overline{A_{verb}})$$

we now represent $P(A_k)$ in terms of p_k and length of the sentence i.e m by subtracting the probability from 1 that no POS type 'k' word is present in the sentence.

$$P(\overline{A_k}) = (1 - p_k)^m$$
$$P(A_k) = 1 - (1 - p_k)^m$$

$P(\overline{A_{noun}} \cap \overline{A_{verb}})$ is nothing but the probability of both A_{noun} and A_{verb} events not happening

$$P(\overline{A_{noun}} \cap \overline{A_{verb}}) = (1 - (p_{noun} + p_{verb}))^m$$

Now we use this to calculate the required answer

$$P(A_{noun} \cap A_{verb}) = 1 - (1 - p_{noun})^m - (1 - p_{verb})^m + (1 - (p_{noun} + p_{verb}))^m$$
$$P(A_{verb}) = 1 - (1 - p_{verb})^m$$

Final Answer:

$$P(A_{noun}|A_{verb}) = \frac{1 - (1 - p_{noun})^m - (1 - p_{verb})^m + (1 - (p_{noun} + p_{verb}))^m}{1 - (1 - p_{verb})^m}$$

2 Marginals and Independence

- Joint Distribution: It can be simply defined as the probability of two events (corresponding to two random variables) happening together. It can be extended to > 2 RVs.

$$p_{XY}(x, y) = P(X = x \ \ and \ \ Y = y)$$

- Marginal Distribution: Probability distribution of the variables contained in the subset of a collection of random variables.

$$Pr(X = x) = \sum_{y \in S_2} Pr((X, Y) = (x, y)), x \in S_1$$

- Independent Random Variable

$$X \amalg Y \Leftrightarrow P(X = x, Y = y) = P(X = x)P(Y = y) \quad \forall x, y$$

  In terms of conditional probability: $P(X|Y) = P(X)$

- Conditionally Independent Random Variable
  X and Y are conditionally independent given Z :

$$p(x, y|z) = p(x|z)p(y|z) \quad \forall x, y, z \quad (p(z) > 0)$$

In machine learning, if given a pair of dependent random variables, we often try to find a random variable Z, so that they become independent conditional to Z.

3 Expectation

For a random variable X on $\mathbb{R}$, with a probability mass function P, the expectation is defined as $\mathbf{E}[X] = \sum_x P(X = x)$. Expectation has the following properties:

- **Linearity** : $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$
- **Linearity** : $\mathbf{E}[cX] = c\mathbf{E}[X]$
- **Expectation as the minimizer of squared error** : We try to find z such that $\mathbf{E}[(X - z)^2]$ is minimized. Equating the derivative of this with respect to z to 0, we get $z = \mathbf{E}[X]$
- **Expecation of product of independent random variables** : $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$. This can be shown as:

$$\mathbf{E}[XY] = \sum_{x,y} P(X = x, Y = y)xy = \sum_x P(X = x)x \sum_y P(Y = y)y = \mathbf{E}[X]\mathbf{E}[Y]$$

Refer to the Appendix for proof of these properties.

4 Variance

Variance of a random variable is a measure of the deviation of the random variable from it's mean. The **variance** of a random variable X with E[X] = μ is defined as:

$$Var[X] = E[(X - \mu)^2] = E[X^2] - \mu^2 \tag{1}$$

Properties:

- $Var[X + \beta] = Var[X]$
- $Var[\alpha X] = \alpha^2 Var[X]$
- If $X_1, \ldots, X_n$ are pairwise independent, then $Var[\sum_i X_i] = \sum_i Var[X_i]$
- If $X_1, \ldots, X_n$ are pairwise independent, each with variance σ^2, then $Var[\sum_i X_i/n] = \frac{\sigma^2}{n}$

Refer to the Appendix for proof of these properties.

4.1 Chebyshev's Inequality

If X is a random variable with mean μ and variance σ^2, then $\forall \alpha > 0$

$$Pr[|X - \mu| \geq \alpha] \leq \frac{\sigma^2}{\alpha^2} \tag{2}$$

4.2 Covariance

It is a measure of **joint variability** of two random variables. For random variables X and Y, **covariance** is defined as:

$$Cov[X, Y] = E[(X - E(X))(Y - E(Y))] = E[XY] - E[X]E[Y] \tag{3}$$

Properties:

- $Cov[X, X] = Var[X]$
- $Cov[X + Z, Y] = Cov[X, Y] + Cov[Z, Y]$
- $Cov[\sum_i X_i, Y] = \sum_i Cov[X_i, Y]$
- $Cov[X, Y] = 0$, if X and Y are independent

5 Sum of random variables

Consider two independent random variables X, Y. Then, their sum $Z = X + Y$ is a random variable with a probabilities mass function as the convolution of the probability mass functions of X and Y.

$$P(Z = z) = \sum_x P(X = x)P(Y = z - x)$$

6 Some Important Distributions

Notation: $X \sim D$ if the random variable X takes its values according to some distribution $D : S \to [0, 1]$.

Bernoulli Random Variable
Takes values from $S = 0, 1$. A single **parameter** $q \in [0, 1]$ fully specifies a Bernoulli random variable, where $Pr[X = 1] = q$ (and $Pr[X = 0] = 1 - q$).
It is denoted by $Bern(q)$.

If $X \sim Bern(q)$ then

- $E[X] = (1 - q) \times 0 + q \times 1 = q$
- $Var[X] = q - q^2 = q(1 - q)$

Binomial Random Variable
A **binomial random variable** models how often a particular outcome occurs in a fixed number of trials of an experiment whose outcome can be modeled as a Bernoulli random variable.
It is denoted by $B(n, q)$.

Formally, if $Y = \sum_{i=1}^{n} X_i$, where $X_i \sim Bern(q)$ are i.i.d., then $Y \sim B(n, q)$. If $Y \sim B(n, q)$ then

- $P(Y = k) = \binom{n}{k} q^k (1-q)^{n-k}$
- $E[Y] = nq$
- $Var[Y] = nq(1-q)$

7 Continuous distributions

A continous random variable can be defined as a random variable which can take an infinite number of values across a range.

For a continuous random variables X, on a domain S, we define

- *Probability Distribution Function*(PDF) as a function $f : S \to \mathbb{R}_0^+$ with the probability of X taking a value inside $D \subseteq S$ being $\int_D f(x)dx$.
- *Cumulative Distribution Function*(CDF) as a function $F : \mathbb{R} \to [0, 1]$

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \int_{-\infty}^{x} f(x)dx \end{aligned}$$

Note The key difference between continous and discreet distributions lies in the the concepts of mass and density.

When the sample space S of random variable is:

- Discrete: The distribution assigns weights/probabilities to individual points
- Continuous: The distribution assigns density to all points such that the integral over the whole range is 1

7.1 Joint distribution

For two random variables X, Y on $\mathbb{R}$, the joint distribution is a function $f : \mathbb{R}^2 \to \mathbb{R}_0^+$, defined such that for all $D \subseteq \mathbb{R}^2$, we have

$$P((X, Y) \in D) = \int_D f((x, y))dxdy$$

Marginals For a joint distribution f, the marginals f_x and f_y are the probabilities of the variables X and Y taking some particular value. Hence, $f_x(t) = \int_{-\infty}^{\infty} f(t, y)dy$ and $f_y(t) = \int_{-\infty}^{\infty} f(x, t)dt$. This integral is often difficult to compute in practice, so we make use of the following trick:

$$\begin{aligned} f_x(t) &= \int_{-\infty}^{\infty} P(X = t|Y = y) f_y(y) dy \\ &= \mathbf{E}_{y \sim f_y}[P(X = t|Y = y)] \end{aligned}$$

This can be evaluated easily using the law of large numbers if it is easy to sample from the distribution f_y

Independence The variables X and Y are said to be independent if their joint distribution can be decomposed into the product of their PDFs i.e. $f(X, Y) = f_x(X)f_y(Y)$.

Conditionals The conditional density $f_x(x|Y = y)$, is the probability density of X, given that Y takes the value y. It is equal to $\frac{f(x,y)}{f_y(y)}$

8 Recovering the PDF from data

Consider a PDF $f : \mathbb{R} \to \mathbb{R}_0^+$ for a random variable X, which is unknown to us. We have a generator which allows us to generate i.i.d. samples from this PDF countably many times. We have the task of estimating what the PDF is, given the i.i.d. samples $\{x_1, x_2, x_3, \cdots x_n\}$.

We will first find the moments of the distribution using the Law of Large Numbers.

$$\mathbf{E}_f[X^k] = \lim_{n \to \infty} \frac{x_1^k + x_2^k + \cdots + x_n^k}{n}$$

Now, we find the moment generating function $M(\omega)$ of the distribution.

$$\begin{aligned} M(\omega) &= \int_{-\infty}^{\infty} e^{\iota\omega x} f(x) dx \\ &= \int_{-\infty}^{\infty} \left(1 + \iota\omega x - \frac{\omega^2 x^2}{2!} - \iota\frac{\omega^3 x^3}{3!} \cdots\right) f(x) dx \\ &= 1 + \iota\omega\mathbf{E}[X] - \frac{\omega^2\mathbf{E}[X^2]}{2!} - \iota\frac{\omega^3\mathbf{E}[X^3]}{3!} \cdots \end{aligned}$$

Once we have evaluated the moment generating function, we use the Inverse Fourier Transform to recover the PDF.

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\iota\omega x} M(\omega) d\omega$$

9 Appendix

9.1 Properties of Expectation

1. $E[X+Y] = E[X] + E[Y]$
   Proof. Let S is the sample space

$$\begin{aligned} E[X+Y] &= \Sigma_{s\in S}(X(s)+Y(s))P(s) \\ &= \Sigma(s \in S)X(s)P(s) + \Sigma_{s\in S}Y(s)P(s) \\ &= E[X] + E[Y] \end{aligned}$$

□

2. $E[(X-c)^2] \geq E[(X-\mu)^2]$ where, $\mu = E[X]$ for any constant c and random variable X
   Proof. By linearity of expectation, we have,

$$\begin{aligned} E[(X-a)^2] &= E[X^2 + a^2 - 2aX] \\ &= E[X^2] + a^2 - 2aE[X] \\ &= (E[X]-a)^2 + E[X^2] - E[X]^2 \end{aligned}$$

   As we will see later that, $Var(X) = E[X^2] - E[X]^2 = E[(X-\mu)^2]$, where $\mu = E[X]$ therefore,

$$E[(X-a)^2] \geq E[(X-\mu)^2]$$

□

3. $E[cX] = cE[X]$ for any constant c and random variable X.
   Proof.

$$\begin{aligned} E[cX] &= \Sigma_{s\in S} cX(s)P(s) \\ &= c\Sigma_{s\in S}X(s)P(s) \\ &= cE[X] \end{aligned}$$

□

4. $E[XY] = E[X]E[Y]$, If X and Y are independent
   Proof.

$$\begin{aligned} E[XY] &= \Sigma_{x,y} xyP(X=x)P(Y=y) \\ &= \Sigma_{x,y}[xP(X=x)][yP(Y=y)] \\ &= \Sigma_x xP(X=x)\ \Sigma_y yP(Y=y) \\ &= E[X]E[Y] \end{aligned}$$

□

9.2 Properties of Variance

1. $Var[X + \beta] = Var[X]$

   Proof.

$$\begin{aligned} Var[X+\beta] &= E[(X+\beta)^2] - E[(X+\beta)]^2 \\ &= E[X^2 + \beta^2 + 2X\beta] - (E[x] + \beta)^2 \\ &= E[X^2] + 2\beta E[X] + \beta^2 - E[X]^2 - \beta^2 - 2\beta E[X] \\ &= E[X^2] - E[X]^2 \\ &= Var[X] \end{aligned}$$

   □

2. $Var[\alpha X] = \alpha^2 E[X]$

   Proof.

$$\begin{aligned} Var[\alpha X] &= E[(\alpha X)^2] - E[(\alpha X)]^2 \\ &= E[\alpha^2 X^2] - (\alpha E[x])^2 \\ &= \alpha^2 E[X^2] - \alpha^2 E[X]^2 \\ &= \alpha^2 (E[X^2] - E[X]^2) \\ &= \alpha^2 Var[X] \end{aligned}$$

   □

3. $Var[X + Y] = Var[X] + Var[Y] + 2(E[XY] - E[X]E[Y])$

   Proof.

$$\begin{aligned} Var[X+Y] &= E[(X+Y)^2] - E[(X+Y)]^2 \\ &= E[X^2 + Y^2 + 2XY] - (E[X] + E[Y])^2 \\ &= E[X^2] + E[Y^2] + 2E[XY] - E[X]^2 - E[Y]^2 - 2E[X]E[Y] \\ &= (E[X^2] - E[X]^2) + (E[Y^2] - E[Y]^2) + 2(E[XY] - E[X]E[Y]) \\ &= Var(X) + Var(Y) + 2(E[XY] - E[X]E[Y]) \end{aligned}$$

   If X and Y are independent, then

$$Var[X + Y] = Var[X] + Var[Y]$$

   □

4. If $\{X_i\}_{i=1}^n$ are pairwise independent of each other
   $Var[X_1 + X_2 + \cdots + X_n] = \sum_i Var[X_i]$

9.3 Properties of Covariance

1. $Cov[X, X] = Var[X]$

   Proof.

   $$\begin{aligned} Cov[X, X] &= E[(X - E[X])(X - E[X])] \\ &= E[(X - E[X])^2] \\ &= Var(X) \end{aligned}$$

   Hence, proved. □

2. $Cov[X + Z, Y] = Cov[X, Y] + Cov[Z, Y]$

   Proof.

   $$\begin{aligned} Cov[X + Z, Y] &= E[(X + Z)Y] - E[(X + Z)]E[Y] \\ &= E[XY + ZY] - E[X + Z]E[Y] \\ &= E[XY] + E[ZY] - E[X]E[Y] - E[Z]E[Y] \\ &= (E[XY] - E[X]E[Y]) + E[ZY] - E[Z]E[Y]) \\ &= Cov[X, Y] + Cov[Z, Y] \end{aligned}$$

   □

   $Cov[\Sigma_i X_i, Y] = \Sigma_i Cov[X_i, Y]$ is a trivial extension of this property