AIML Tutorial 2: Probability

Dhruva Kashyap

February 2024

1 101

Probability Space A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ defines the formal model of probability for random experiments. The set Ω(sample space) contains a set of outcomes of interest. $\mathcal{F} \subset 2^{\Omega}$(Event space) is a σ-algebra containing Ω defining events of interest. The probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a unitary measure on the event space.

Random Variable A real-valued random variable, usually denoted with a capital letter, $X : \Omega \to \mathbb{R}$, is a function satisfying certain properties from the sample space to the real line. This is denoted by $(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$. If the range of the random variable is only on discrete points, the random variable is called a discrete random variable, and a probability mass function will always exist. If the range is continuous, the random variable is called a continuous random variable, and a density function may or may not exist.

Conditional probabilty Given two events $A, B \in \mathcal{F}$, the conditional probability of the event A given that event B has occurred, denoted by $P(A|B)$, is given by $P(A|B) = \frac{P(AB)}{P(B)}$ where AB is shorthand for the event $A \cap B$.

Excercise 1.1 (Chain rule of probability)**.** *Note that we may write $P(AB) = P(A)P(B|A) = P(B)P(A|B)$. Show that $P(A_1, ...A_n) = \prod_{i=1}^{n} P(A_i|A_1, ..., A_{i-1})$. Hint: Induction.*

Theorem 1.2 (Law of total probability)**.** *Consider two events, A and B. Let $(A_1, ..., A_n)$ form a partition of A. Then,*

$$P(B) = \sum_{n} P(A_n)P(B|A_n)$$

. For two continuous random variables X and Y with joint density p_{XY},

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y)dy = \int_{-\infty}^{\infty} p_{X|Y}(x|y)p_Y(y)dy$$

Excercise 1.3. *Prove theorem 1.2 for two events. Hint: Probability axioms.*

Theorem 1.4 (Bayes Theorem)**.** *Consider two events A and B and let $(A_1, ..., A_n)$ form a partition of A, then,*

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

The importance of Bayes theorem is in being able to compute $A|B$ using information about $B|A$.

Definition 1.5 (Independence)**.** *Two sets A and B are said to be statistically independent if $P(AB) = P(A)P(B)$.*

Excercise 1.6. *If two random variables X and Y are independent, show that $\mathbb{E}_{XY}[XY] = \mathbb{E}_X[X]\mathbb{E}_Y[Y]$ and $cov(X, Y) = 0$.*

Excercise 1.7 (Linearity of Expectation)**.** *Consider discrete random variables $X_1, ..., X_n$, which may or may not be independent. Show that, $\mathbb{E}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \mathbb{E}[X_i]$. Read up on when this is true for continuous random variables.*

Functions of random variables If X is a real-valued random variable, under certain assumptions on a function $g : \mathbb{R} \to \mathbb{R}$, $g(X)$ is also a random variable.

Excercise 1.8 (Law of the unconscious statistician(LOTUS))**.** *If $Y = g(X)$ and certain assumptions on g hold, then*

$$\mathbb{E}_Y[Y] = \mathbb{E}_X[g(X)]$$

i.e., for a discrete random variable

$$\sum_{y_j} y_j p_Y(y_j) = \sum_{x_i} g(x_i) p_X(x_i)$$

For proof of the discrete case, refer to this wiki page. The continuous case also has caveats that you may note from the same link. Under many assumptions, you may write the density for Y as $f_Y(y) = \frac{dg^{-1}}{dy}|_{y=y} f_X(g^{-1}(y))$. For multivariate random vectors, you may refer to the following wiki link for the appropriate change of variables formulation.

Note that $\mathbb{E}[X|Y]$ is a random variable, a function of the random variable Y.

Excercise 1.9 (Law of Total Expectation)**.** *Show that, for two random variables X and Y,*

$$\mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]]$$

Theorem 1.10 (Jensen's Inequality)**.** *For a convex function f,*

$$f(\mathbb{E}_X[X]) \leq \mathbb{E}_X[f(X)]$$

Excercise 1.11. *Show that $\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y|X]]$*

Excercise 1.12. *(*) If $\mathbb{E}[Y|X] = 1$, show that $var(XY) \geq var(X)$. Hint: Use the previous exercise and Jensen's inequality.*

Moment generating function The moment generating function(MGF) of a random variable X, is denoted $M_X(t)$ is defined as $M_X(t) = \mathbb{E}_X[e^{Xt}]$. The MGF function need not exist for all t and might not exist anywhere. Also related is the Characteristic function. The MGF has a few useful properties.

Excercise 1.13. *Show that for $Y = aX + b$, $M_Y(t) = e^{bt} M_X(at)$*

Excercise 1.14. *Show that $\frac{d^n M_X(t)}{dt^n}|_{t=0} = \mathbb{E}[X^n]$.*

Excercise 1.15. *Consider independent random variables $X_1, ..., X_n$. Let $Y = \sum_{i=1}^n a_i X_i$. Show that $M_Y(t) = \prod_{i=1}^n M_{X_i}(a_i(t))$.*

Theorem 1.16. *An MGF, if it exists, uniquely describes a probability distribution. That is, if for two random variables X and Y, $M_X(t) = M_Y(t) \forall t$, then $f_X(z) = f_Y(z) \forall z$.*

Refer to this stack exchange post for proof of the special case above. The general solution is beyond the scope of this course.

Excercise 1.17. *Find the moment generating function of $X \sim \mathcal{N}(\mu, \sigma^2)$.*

Excercise 1.18. *For a k-chi-squared distributed random variable $X \sim \chi^2(k)$, find the moment generating function. The density is given by*

$$f_X(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{\frac{-x}{2}}$$

Excercise 1.19. *If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then prove that $Y = aX_1 + bX_2$ is also distributed as normally for some $a, b \in \mathbb{R}$. Find the means and variances. Hint: Use the MGF. Another technique with convolution.*

2 Multiple random variables

Things become interesting when we study the interaction between multiple random variables. We start defining joint densities, correlation, covariance, etc. The real world cannot be modeled with only one random variable and requires studying multiple randomnesses. Confusingly, whether a random variable X is one-dimensional or multi-dimensional is "apparent from context."

Definition 2.1. *The covariance matrix of a vector-valued random variable X is given by*

$$\Sigma_X = \mathbb{E}_X[(X - \mathbb{E}_X[X])(X - \mathbb{E}_X[X])^T]$$

Excercise 2.2.

- *Show that the covariance matrix is always p.s.d.*
- *Show that $\Sigma_{ij} = cov(X_i, X_j)$*

The normal distribution is important and appears in a fundamental theorem. We will now define a multi-dimensional Gaussian(Normal) random variable.

2.1 Multi-dimensional Normal Random vector

We understand what it means for a one-dimensional random variable to be distributed normally. For a random vector, the definition may seem slightly different from what might have felt would be the intuitive definition. The notion of jointly Gaussian random variables is critical to defining a normally distributed random vector.

Definition 2.3. *Let $X = (X_1, ..., X_n)$ be a random vector. The random vector is said to be distributed normally if and only if the following condition holds.*

$$X \sim \mathcal{N}(\mu, \Sigma) \iff \exists A \in \mathbb{R}^{n \times m} s.t. X = Az + \mu \text{ and } z_i \sim \mathcal{N}(0, 1) \text{ i.i.d } \forall i \in [m] \text{ and } \Sigma = AA^T$$

And when Σ is positive definite, the density exists everywhere and is given by,

$$f_X(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(\frac{-1(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}\right)$$

In such a case, the random variables $(X_1, ..., X_2)$ are said to be ***jointly Gaussian****.*

Excercise 2.4.

- *If X is Gaussian, is each X_i gaussian?*
- *If X is Gaussian, are each X_i independent?*
- *On what condition on A can you say that each X_i is iid and X is Gaussian?*
- *On what condition for A can you say Σ is pd?*
-

Excercise 2.5. *Given two random variables X and Y that are jointly Gaussian with 0 mean and unit variance. Say the correlation coefficient is given by ρ. Compute $\mathbb{E}[Y^2|X]$.*

Proof. Let $Z = (X, Y)$ be a random vector. Since X and Y are jointly Gaussian, Z is normally distributed, and the properties of X and Y imply that the mean of Z is 0. for the two-dimensional case, we can easily see that

$$\Sigma_Z = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Since Z is normally distributed, we can compute the density by inverting Σ_Z

$$f_Z(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-(u^2+v^2-2\rho uv)}{2(1-\rho^2)}\right)$$

We may decompose the above formula as,

$$f_Z(u,v) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(\frac{-(v-\rho u)^2}{2(1-\rho^2)}\right)$$

Integrating with respect to v on either side and observing that the second term is the Gaussian integral with mean ρu and variance $(1-\rho^2)$

$$f_X(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right)$$

This means X is distributed as a standard normal random variable. By symmetry, this also holds for Y. Observe that X and Y are jointly Gaussian, and their marginals are also Gaussian (standard normal!), but this does not imply that they are independent.

We can also observe that

$$f_{Y|X}(v|u) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(\frac{-(v-\rho u)^2}{2(1-\rho^2)}\right)$$

which means that $Y|X = u \sim \mathcal{N}(\rho u, 1-\rho^2)$. Using $\mathbb{E}[W^2] = \text{var}(W) + \mathbb{E}[W]^2$, $\mathbb{E}[Y^2|X = u] = 1 - \rho^2 + (u\rho)^2 = 1 + (u^2-1)\rho^2$. Therefore,

$$\mathbb{E}[Y^2|X] = 1 + (X^2-1)\rho^2$$

□

One may decide to ponder on the following for the previous question.

- What does it mean when $\rho = 0$?
- What does it mean when $\rho = 1$?