Machine Learning

Lecture 2: Probability

Hao Tang

December 2, 2022

What is a probability measure $\mathbb{P}$?

Probability measures

- Start with a set Ω.
- A subset $X \subseteq \Omega$ is called an event.
- A probability measure $\mathbb{P}$ takes a subset and returns a real value.

Probability measures

1. $\mathbb{P} : 2^{\Omega} \to \mathbb{R}$
   - 2^{Ω} is the power set, i.e., all subsets of Ω.
   - $\mathbb{P}$ is a function that takes a subset of Ω and returns a real value.
2. $0 \leq \mathbb{P}(X) \leq 1$ for any $X \subseteq \Omega$
3. $\mathbb{P}(\Omega) = 1$
4. $\mathbb{P}(X \cup Y) = \mathbb{P}(X) + \mathbb{P}(Y)$ if $X \cap Y = \emptyset$

What happens when Ω is discrete and finite?

Discrete probability distributions

- When Ω is discrete and finite, it is possible to enumerate all elements of a subset $X \subseteq \Omega$.
- For any $X \subseteq \Omega$, we can implement a probability measure $\mathbb{P}$ with another function p by letting

$$\mathbb{P}(X) = \sum_{\omega \in X} p(\omega) \tag{1}$$

- The function p is called a probability mass function or discrete probability distribution
  1. $p : \Omega \to \mathbb{R}$
  2. $0 \leq p(\omega) \leq 1$ for any $\omega \in \Omega$
  3. $\sum_{\omega \in \Omega} p(\omega) = 1$

Discrete probability distributions

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $\mathbb{P} : 2^{\Omega} \to \mathbb{R}$
  - The input to the distribution can be any subset of Ω.
  - It's valid (type-correct) to write $\mathbb{P}(\{1\})$ and $\mathbb{P}(\{1, 2\})$.
- $\mathbb{P}(\Omega) = \mathbb{P}(\{1, 2, 3, 4, 5, 6\}) = 1$
- $\mathbb{P}(\{1, 2\}) = p(1) + p(2) = 2/6$
- $\{1\}$ is an event, but 1 is not.
- $\mathbb{P}$ and p are different!

face	probability
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$

Set comprehension

- Set comprehension is a shorthand for describing sets with constraints.

  $\mathbb{P}(\omega = 3) = \mathbb{P}(\{\omega : \omega = 3\})$

  $\mathbb{P}(\omega > 3) = \mathbb{P}(\{\omega : \omega > 3\})$

  $\mathbb{P}(\omega \text{ is even}) = \mathbb{P}(\{\omega : \omega \in \{2, 4, 6\}\})$

- The variable name does not matter.

  $\mathbb{P}(\{\omega : \omega > 3\}) = \mathbb{P}(\{x : x > 3\})$

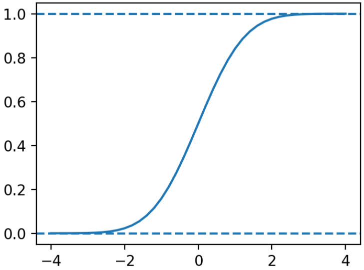
- Always ask what is random.

  $\mathbb{P}(\omega > t/\sqrt{2} + 3) = \mathbb{P}(t < \sqrt{2}(\omega - 3)) = \mathbb{P}(\{\omega : t < \sqrt{2}(\omega - 3)\})$

Continuous probability distribution

The function F is a cumulative distribution function if

1. $F : \mathbb{R} \to [0, 1]$
2. F is monotonic, i.e., $F(x) < F(y)$ if $x < y$
3. $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$



Continuous probability distribution

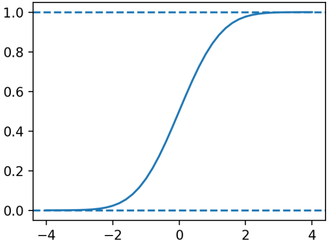
- A probability density function p is defined as $p(u) = \frac{dF}{dx}(u)$ or $F(x) = \int_{-\infty}^{x} p(u)du$.
- We can construct a probability measure $\mathbb{P}$ by letting

$$\mathbb{P}(a < X < b) = \int_{a}^{b} p(u)du = F(b) - F(a). \quad (2)$$

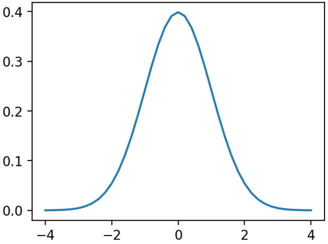
- $\Omega = \mathbb{R}$ and $\mathbb{P} : 2^{\mathbb{R}} \to \mathbb{R}$ takes a subset of $\mathbb{R}$ as input.

Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{3}$$



CDF



PDF

Sampling notation

We say that a is drawn from a Gaussian if

$$a \sim \mathcal{N}(\mu, \sigma^2). \tag{4}$$

It simply means

$$p(a) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(a-\mu)^2\right). \tag{5}$$

Expectation

- Definition

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} xp(x)dx \qquad \mathbb{E}[x] = \sum_{x \in \Omega} xp(x) \tag{6}$$

- $\mathbb{E}[x]$ is **not** a function of x, but a function of p.
- A better notation would be

$$\mathbb{E}_{x \sim p(x)}[x]. \tag{7}$$

The law of unconcious statistician (LOTUS)

- Theorem

$$\mathbb{E}_{x\sim p(x)}[f(x)] = \int_{-\infty}^{\infty} f(x)p(x)dx \qquad \mathbb{E}_{x\sim p(x)}[f(x)] = \sum_{x\in\Omega} f(x)p(x) \tag{8}$$

- The theorem needs to be formally proved.
- The $f(x)$ in $\mathbb{E}[f(x)]$ is **not** a function of x, but an expression of x.

$\mathbb{E}_{x\sim p(x)}[x^2]$

$\mathbb{E}_{x\sim p(x)}\left[(x - \mathbb{E}_{x\sim p(x)}[x])^2\right] = \mathrm{Var}[x]$

Free and bound variables

```
def p(x):
    return (1.0 / math.sqrt(2 * math.pi)
      * math.exp(-0.5 * (x - mu) * (x - mu))

mu = 0.2
p(0.5)
x = 0.3
p(x = x)
```

- Is `x` a free variable or a bound variable? When is it bound and what is it bound to?
- Is `mu` a free variable or a bound variable?

Notation hell

- When we write $p(x)$, p is **not** the name of the function, as opposed to when we write $f(x)$.
- When we have multiple distributions, the convention is to use variable names to distinguish distributions, e.g., $p(x)$, $p(y)$, and $p(z)$.
- It gets confusing when we simply write $p(a)$, and the convention is to use keyword arguments, e.g., $p(x = a)$, $p(y = a)$, and $p(z = a)$.
- Note that $p(x = a)$ does not mean $p(\{x : x = a\})$. Remember that p takes a point in Ω, not a subset of Ω.
- Sometimes people also write $p_x(a)$ to mean $p(x = a)$.

Multiple random variables

- Joint distribution $p(x, y)$
- Marginal distribution $p(x) = \int_{-\infty}^{\infty} p(x, y)dy$ or $p(x) = \sum_{y \in \Omega_Y} p(x, y)$
- Conditional distribution $p(x|y) = \dfrac{p(x, y)}{p(y)}$
- Note that these are all defined based on p not $\mathbb{P}$.

Notations again

$$p(x) = \sum_{y \in \Omega_y} p(x, y) \qquad p_x(a) = \sum_{b \in \Omega_y} p_{x,y}(a, b) \tag{9}$$

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \qquad p_{y|x}(b, a) = \frac{p_{x|y}(a, b)p_y(b)}{p_x(a)} \tag{10}$$

Bayes rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{11}$$

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y' \in \Omega_y} p(x|y')p(y')} \tag{12}$$

Independence

- We say that x and y are independent if

$$p(x, y) = p(x)p(y) \tag{13}$$

for any $x \in \Omega_x$ and $y \in \Omega_y$.

- By the definition of conditional probability,

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x)p(y)}{p(x)} = p(y). \tag{14}$$

- In other words, x and y are independent, if given x or not does not change the probability of y.

Independence and expectation

- $\mathbb{E}[cx] = c\mathbb{E}[x]$
- $\mathbb{E}[x + y] = \mathbb{E}[x] + \mathbb{E}[y]$ if x and y are independent.

  $\mathbb{E}_{x,y \sim p(x,y)}[x + y] = \mathbb{E}_{x \sim p(x)}[\mathbb{E}_{y \sim p(y)}[x + y]] = \mathbb{E}_{x \sim p(x)}[x] + \mathbb{E}_{y \sim p(y)}[y]$
- $\mathbb{E}[xy] = \mathbb{E}[x]\mathbb{E}[y]$ if x and y are independent.

Random variables

- We define events (as subsets) and probability measures (a function that maps subsets to real values).
- A probability distribution is a function that maps individual points to real values.
- For the purpose of this course, a variable is a random variable if it is associated with a probability measure.
- There is a mathematical definition, but we will not attempt to do it here.

Random variables

- If $a \sim U(0, 1)$, then a is random.
- If $a \sim \mathcal{N}(0, 1)$, then a is random.
- If $\epsilon \sim \mathcal{N}(0, 1)$, then $m + \epsilon$ is random for some real value m.
- In fact, $m + \epsilon \sim \mathcal{N}(m, 1)$.

Random variables

- If $x \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $y \sim \mathcal{N}(\mu_2, \sigma_2^2)$,

$$x + y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \tag{15}$$

- If $u_1 \sim U(0,1)$ and $u_2 \sim U(0,1)$, then

$$z_1 = \sqrt{-2\log u_1}\cos(2\pi u_2) \sim \mathcal{N}(0,1) \tag{16}$$

$$z_1 = \sqrt{-2\log u_1}\sin(2\pi u_2) \sim \mathcal{N}(0,1) \tag{17}$$

- In general, it is hard to determine the probability distribution solely based on the algebra of random variables.

Moment-generating functions

- $M_x(t) = \mathbb{E}[e^{tx}] = \int_{-\infty}^{\infty} e^{tx} p(x) dx$

$$M_x(t) = \mathbb{E}[e^{tx}] = \mathbb{E}\left[1 + \frac{t}{1!}x + \frac{t^2}{2!}x^2 + \cdots\right] \tag{18}$$

$$= 1 + \frac{t}{1!}\mathbb{E}[x] + \frac{t^2}{2!}\mathbb{E}[x^2] + \cdots \tag{19}$$

- $M_x'(0) = \mathbb{E}[x]$, $M_x''(0) = \mathbb{E}[x^2]$, ...
- If $M_x(t) = M_y(t)$, then x and y has the same probability distribution

MGF of a Gaussian

Suppose $x \sim \mathcal{N}(\mu, \sigma^2)$.

MGF of a Gaussian

Suppose $x \sim \mathcal{N}(\mu, \sigma^2)$.

$$
\begin{aligned}
\mathbb{E}[e^{tx}] &= \int e^{tx} \frac{-1}{\sqrt{2\pi\sigma^2}} e^{\frac{1}{2\sigma^2}(x-\mu)^2} dx && (20)\\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int e^{\frac{-1}{2\sigma^2}(x^2-2\mu x+\mu^2-2t\sigma^2 x))} dx && (21)\\
&= e^{\frac{1}{2\sigma^2}((\mu+t\sigma^2)^2-\mu^2)} \frac{1}{\sqrt{2\pi\sigma^2}} \int e^{\frac{-1}{2\sigma^2}(x-(\mu+t\sigma^2))^2} dx && (22)\\
&= e^{\mu t+t^2\sigma^2/2} && (23)
\end{aligned}
$$

Linear combination of Gaussians

Suppose $x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

We have $a_1 x_1 + a_2 x_2 \sim \mathcal{N}(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$.

Linear combination of Gaussians

Suppose $x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

$$\begin{aligned}
\mathbb{E}[e^{t(a_1 x_1 + a_2 x_2)}] &= \mathbb{E}[e^{t a_1 x_1}]\mathbb{E}[e^{t a_2 x_2}] && (24)\\
&= e^{t a_1 \mu_1 + t^2 a_1^2 \sigma_1^2/2} e^{t a_2 \mu_2 + t^2 a_2^2 \sigma_2^2/2} && (25)\\
&= e^{t(a_1\mu_1 + a_2\mu_2) + t^2(a_1^2\sigma_1^2 + a_2^2\sigma_2^2)/2} && (26)
\end{aligned}$$

We have $a_1 x_1 + a_2 x_2 \sim \mathcal{N}(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$.

Independence and identically distributed

- $x_1, x_2, \ldots, x_n$ are called independent and identically distributed (i.i.d.) samples if $x_1, x_2, \ldots, x_n$ are mutually independent and are drawn from the same distribution.

Maximum likelihood

- If we flip a coin 500 times and see 300 heads, how do we estimate the probability of getting a head?
- Asusme i.i.d. Bernoulli random variables $x_1, \ldots, x_n$ (with probability β to be heads).

Maximum likelihood

- If we flip a coin 500 times and see 300 heads, how do we estimate the probability of getting a head?
- Asusme i.i.d. Bernoulli random variables $x_1, \ldots, x_n$ (with probability β to be heads).
- The likelihood of β is

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2)\cdots p(x_n) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \beta^{x_i}(1-\beta)^{1-x_i} \tag{27}$$

- The maximum likelihood estimator of β is the value that maximizes the likelihood.

Maximum likelihood

$$L = \log p(x_1, \ldots, x_n) = \sum_{i=1}^{n} [x_i \log \beta + (1 - x_i) \log(1 - \beta)] \tag{28}$$

$$\operatorname*{argmax}_{\beta} \prod_{i=1}^{n} \beta^{x_i} (1 - \beta)^{1 - x_i} = \operatorname*{argmax}_{\beta} \sum_{i=1}^{n} [x_i \log \beta + (1 - x_i) \log(1 - \beta)] \tag{29}$$

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{x_i}{\beta} - \frac{(1 - x_i)}{1 - \beta} \right] = \sum_{i=1}^{n} \left[\frac{x_i - \beta}{\beta(1 - \beta)} \right] = \frac{\sum_{i=1}^{n} x_i - n\beta}{\beta(1 - \beta)} = 0 \tag{30}$$

$$\beta = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{31}$$