Machine learning

Overview of probability theory

Hamid Beigy

Sharif University of Technology

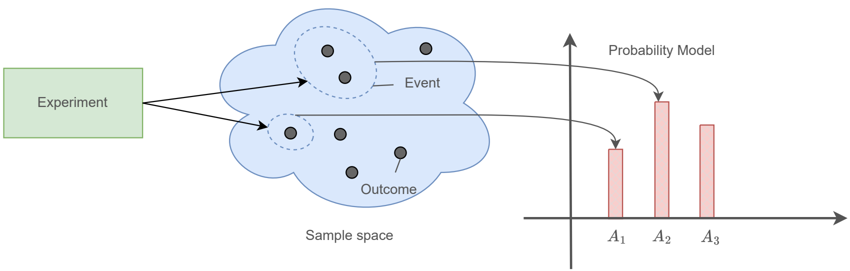
February 13, 2023

Table of contents

Probability

Probability

1. Probability theory is the study of uncertainty.
2. Elements of probability
   - Sample space Ω is the set of all the outcomes of a random experiment.
   - Event space $\mathcal{F}$ is a set whose elements $A \in \mathcal{F}$ (called events) are subsets of Ω.
   - Probability measure is a function $P : \mathcal{F} \mapsto [0, 1]$.



Definition (Probability measure)

A probability measure on the sample space Ω is a function, denoted P, from subsets of Ω to the real numbers $\mathbb{R}$, such that the following hold:

- $P(A) \geq 0$, for all $A \in \mathcal{F}$.
- $P(\Omega) = 1$.
- If $A_1, A_2, \ldots$ are disjoint events (i.e.,$A_i \cap A_j = \emptyset$ whenever $i \neq j$),then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

Example (Tossing two coins)

In tossing two coins, we have

- The sample space equals to $\Omega = \{HH, HT, TT, TH\}$
- An event space $\mathcal{F}$ that only one head is a subset of Ω such as $\mathcal{F} = \{TH, HT\}$
- The probabilities are $P(TH) = \frac{1}{4}$ and $P(HT) = \frac{3}{4}$

1. If $A \subseteq B \implies P(A) \leq P(B)$.
2. $P(A \cap B) \leq \min(P(A), P(B))$.
3. $P(A \cup B) \leq P(A) + P(B)$. This property is called union bound.
4. $P(\Omega \setminus A) = 1 - P(A)$.
5. If $A_1, A_2, \ldots, A_k$ are disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$,then

$$\sum_{i=1}^{k} P(A_i) = 1$$

   This property is called law of total probability.

Conditional probability and independence

1. Let B be an event with $P(B) \geq 0$. The conditional probability of any event A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

   In other words, $P(A \mid B)$ is the probability measure of the event A after observing the occurrence of event B.

2. Two events are called independent if and only if

$$P(A \cap B) = P(A)P(B),$$

   or equivalently, $P(A \mid B) = P(A)$.
   Therefore, independence is equivalent to saying that observing B does not have any effect on the probability of A.

Classical definition (Laplace, 1814). $P(A)$ is the number of outcomes that are favorable to A divided by the total number of outcomes.

$$P(A) = \frac{N_A}{N}$$

where N mutually exclusive equally likely outcomes, N_A of which result in the occurrence of A.

Frequentist definition. $P(A)$ is the relative frequency of occurrence of A in infinite number of trials as

$$P(A) = \lim_{N\to\infty} \frac{N_A}{N}$$

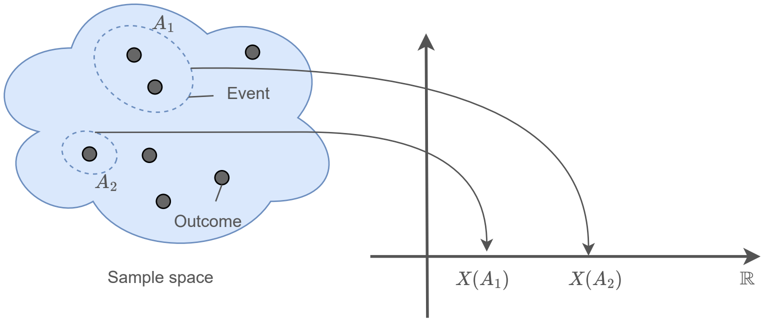
Bayesian definition (de Finetti, 1930s). $P(A)$ is a degree of belief.

Example (Bayesian vs. Frequentist)

1. We have a coin with unknown probability θ of coming up heads.
2. We must determine this probability as accurately as possible using experimentation.
3. Experimentat is to repeatedly tossing the coin.
4. Let us denote two possible outcomes of a single toss by 1 (Heads) and 0 (Tails).
5. If we toss the coin m times, then we can record the outcomes as $x_1, \ldots, x_m$, where each $x_i \in \{0, 1\}$ and $P[x_i = 1] = \theta$ independently of all other x_i's.
6. What would be a reasonable estimate of θ?
7. In Frequentist view, by Law of Large Numbers, in a long sequence of independent coin tosses, the relative frequency of heads will eventually approach the true value of θ with high probability. Hence,
$$\hat{\theta} = \frac{1}{m} \sum_i x_i$$
8. In Bayesian view, θ is a random variable and has a distribution.

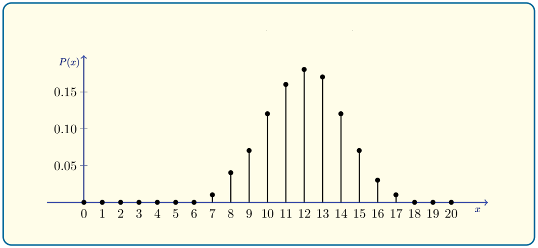
Random variables

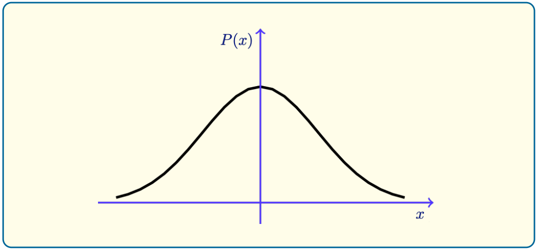
1. Consider an experiment with 10 coin flips, and we want to know the number of coins that come up heads.
2. Here, the elements of the sample space Ω are 10-length sequences of heads and tails.
3. We usually do not care about the probability of any particular sequence of heads and tails.
4. Instead we usually care about real-valued functions of outcomes, such as
   - the number of heads that appear among our 10 tosses, or
   - the length of the longest run of tails.
5. These functions, under some technical conditions, are known as random variables.
6. More formally, a random variable X is a function $X : \Omega \to \mathbb{R}$.
7. We denote random variables using upper case letters $X(\omega)$ or more simply X, where ω is an event.



8. We will denote the value that a random variable X may take on using lower case letter x.

1. A random variable can be discrete or continuous.

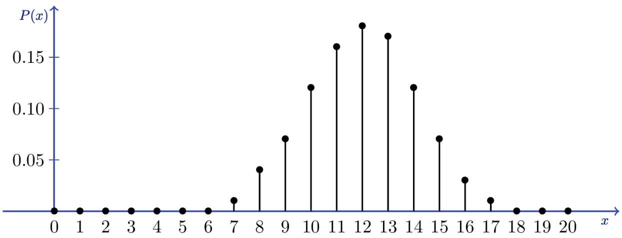




2. A random variable is associated with a probability mass function or probability distribution.

1. For a discrete random variable X, $p(x)$ denotes the probability that $p(X = x)$.
2. $p(x)$ is called the probability mass function (PMF).
3. This function has the following properties:

$$p(x) \geq 0$$
$$p(x) \leq 1$$
$$\sum_{x} p(x) = 1$$

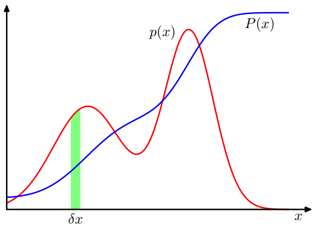


Continuous random variables

1. For a continuous random variable X, a probability $p(X = x)$ is meaningless.
2. Instead we use $p(x)$ to denote the probability density function (PDF).

$$p(x) \geq 0$$
$$\int_x p(x) = 1$$

3. Probability that a continuous random variable $X \in (x, x + \delta x)$ is $p(x)\delta x$ as $\delta x \to 0$.



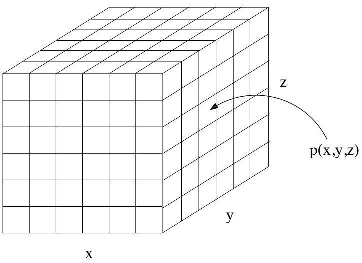
4. Probability that $X \in (-\infty, z)$ is given by cumulative distribution function (CDF) $P(z)$

$$P(z) = p(X \leq z) = \int_{-\infty}^{z} p(x)dx$$
$$p(x) = z \left| \frac{dP(z)}{dz} \right|_{z=x}$$

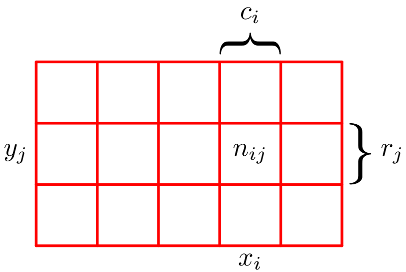
Joint probability

1. Two or more random variables may interact.
2. Thus, the probability of one taking on a certain value depends on which value(s) the others are taking.
3. We write this as

$$p(x, y) = P(X = x, Y = y).$$



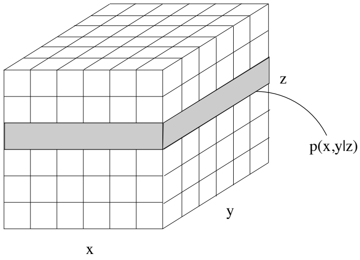
1. Let n_{ij} be the number of times events x_i and y_j simultaneously occur.



2. Let $N = \sum_i \sum_j n_{ij}$.
3. Joint probability is
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}.$$
4. Let $c_i = \sum_j n_{ij}$, and $r_j = \sum_i n_{ij}$.
5. The probability of X irrespective of Y is
$$p(X = x_i) = \frac{c_i}{N}.$$

1. If we know that some event has occurred, it changes our belief about the probability of other events.
2. This is like taking a **slice** through the joint table.
3. We write this as

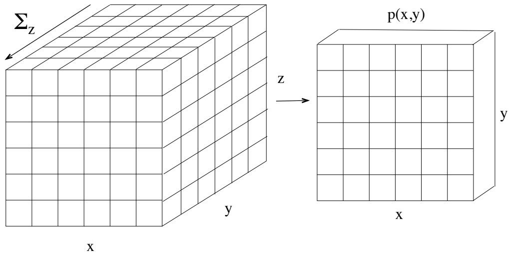
$$p(x|y) = \frac{p(x, y)}{p(y)}.$$



1. We can sum out part of a joint distribution to get the marginal distribution of a subset of variables:

$$p(x) = \sum_y p(x, y)$$

2. This is like adding slices of the table together.

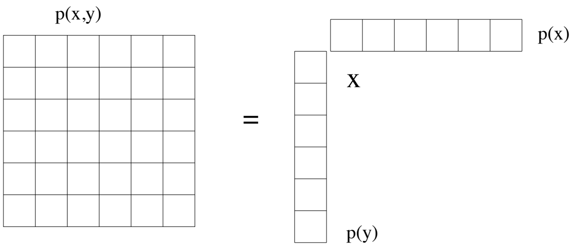


3. Another equivalent definition

$$p(x) = \sum_y p(x|y)p(y)$$

1. Two variables are independent iff their joint factors:

$$p(x, y) = p(x)p(y)$$



2. Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:

$$p(x, y|z) = p(x|z)p(y|z) \qquad \forall z$$

1. Expectation, expected value, or mean of a random variable X, denoted by $\mathbb{E}[X]$, is the average value of X in a large number of experiments.

$$\mathbb{E}[X] = \sum_{x} xp(x)$$
$$\mathbb{E}[X] = \int_{x} xp(x)dx$$

2. The definition of Expectation also applies to functions of random variables (e.g., $\mathbb{E}[f(x)]$)
3. Linearity of expectation

$$\mathbb{E}[\alpha f(x) + \beta g(x)] = \alpha\,\mathbb{E}[f(x)] + \beta\,\mathbb{E}[g(x)]$$

Variance and and Covariance

1. Variance (σ^2) measures how much X varies around the expected value and is defined as.

$$Var(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mu^2$$

2. Standard deviation is defined as

$$std[X] = \sqrt{Var[X]} = \sigma$$

3. Covariance of two random variables X and Y indicates the relationship between two random variables X and Y.

$$Cov(X, Y) = \mathop{\mathbb{E}}_{X,Y}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top\right]$$

Probability distributions

Common probability distributions

We will use these probability distributions extensively to model data as well as parameters

- Some discrete distributions and what they can model:
  1. Bernoulli : Binary numbers, e.g., outcome (head/tail, 0/1) of a coin toss
  2. Binomial : Bounded non-negative integers, e.g., the number of heads in n coin tosses
  3. Multinomial : One of $K(> 2)$ possibilities, e.g., outcome of a dice roll
  4. Poisson : Non-negative integers, e.g., the number of words in a document
- Some continuous distributions and what they can model:
  1. Uniform: Numbers defined over a fixed range
  2. Beta: Numbers between 0 and 1, e.g., probability of head for a biased coin
  3. Gamma: Positive unbounded real numbers
  4. Dirichlet : Vectors that sum of 1 (fraction of data points in different clusters)
  5. Gaussian: Real-valued numbers or real-valued vectors

Exponential family

1. For (continuous or discrete) random variable x

$$p(x|\theta) = \frac{1}{Z(\theta)} h(x) exp\left[\theta^\top \phi(x)\right]$$
$$= h(x) exp\left[\theta^\top \phi(x) - A(\theta)\right]$$

where

$$Z(\theta) = \int_x h(x) exp\left[\theta^\top \phi(x)\right] dx$$
$$A(\theta) = \log Z(\theta)$$

is an exponential family distribution with natural parameter θ.

- $\phi(x)$ is called a vector of sufficient statistics.
- $Z(\theta)$ is called the partition function.
- $A(\theta)$ is called the log partition function.
- $h(x)$ is the a scaling constant, often 1.

Probability distributions

Discrete distributions

Bernoulli distribution

1. For a binary random variable $x \in \{0, 1\}$ with $p(x = 1) = \pi$, like a coin-toss outcome

$$\begin{aligned} Ber(x|\pi) &= \pi^x (1-\pi)^{1-x} \\ &= \exp\left\{\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right\} \end{aligned}$$

2. The expected value and the variance of X are equal to

$$\begin{aligned} \mathbb{E}[X] &= \pi \\ Var(X) &= \pi(1-\pi) \end{aligned}$$

3. The Bernoulli for $x \in \{0, 1\}$ can be written in exponential family form as follows:

$$\begin{aligned} Ber(x|\pi) &= \pi^x (1-\pi)^{1-x} \\ &= \exp[x \log \pi + (1-x) \log(1-\pi)] \\ &= \exp\left[\theta^\top \phi(x)\right] \end{aligned}$$

where

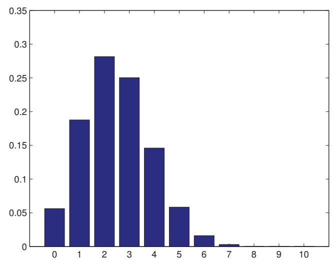
$$\begin{aligned} \phi(x) &= [\mathbb{I}[x=0], \mathbb{I}[x=1]] \\ \theta &= [\log \pi, \log(1-\pi)] \end{aligned}$$

Binomial distribution

1. Suppose we toss a coin n times.
2. Let $x \in \{0, 1, \ldots, n\}$ be the number of heads.
3. If probability of heads is π, x has a binomial distribution, written as

$$\text{Bin}(k|n, \pi) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

4. Binomial distribution for $n = 10$ and $\pi = 0.25$



5. The expected value and the variance of x are equal to

$$\mathbb{E}[x] = n\pi$$
$$Var(x) = n\pi(1-\pi)$$

1. For a categorical random variable taking K values, let π_k be the probability of k^{th} value.
2. Using a binary vector $(x_1, \ldots, x_K)$ where $x_k = 1$ iff the variable takes on its k^{th} value.
3. Now we can write,

$$Cat(x|\pi) = \prod_{k=1}^{K} \pi_k^{x_k} = \exp\left[\sum_{k=1}^{K} x_k \log \pi_k\right]$$

4. Suppose n such trials are made where outcome k occurred n_k times with $\sum_{k=1}^{K} n_k = n$.
5. The joint distribution of $n_1, n_2, \ldots, n_K$ is multinomial

$$P(n_1, n_2, \ldots, n_K) = n! \prod_{i=1}^{K} \frac{\pi_i^{n_i}}{n_i!}$$

Homework (Representing a exponential family)

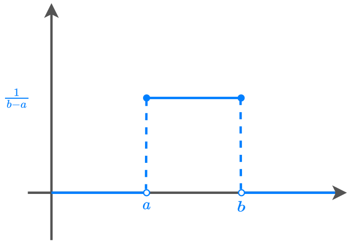
Represent the multinoulli distribution as a special case of exponential family.

Probability distributions

Continuous distributions

1. Models a continuous random variable X distributed uniformly over a finite interval $[a, b]$.

$$U(X; a, b) = \frac{1}{b - a}$$



2. The expected value and the variance of X are equal to

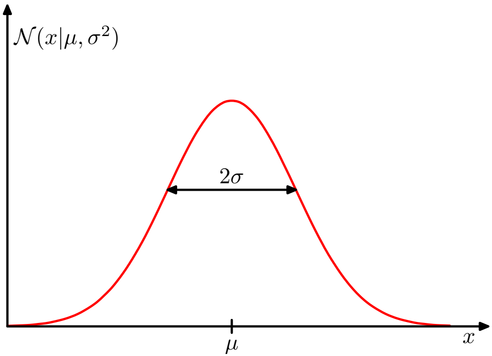
$$\mathbb{E}[X] = \frac{b + a}{2}$$

$$Var(X) = \frac{(b - a)^2}{12}$$

Normal (Gaussian) distribution

1. For 1-dimensional normal or Gaussian distributed variable X with mean μ and variance σ^2

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



2. The expected value and the variance of X are equal to

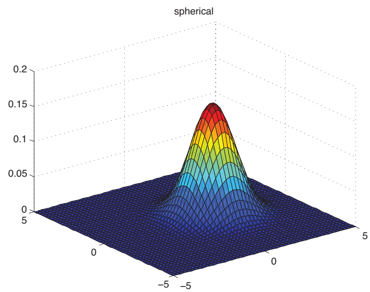
$$\mathbb{E}[X] = \mu$$
$$Var(X) = \sigma^2$$

3. Precision (inverse variance): $\beta = \frac{1}{\sigma^2}$

Multivariate Gaussian distribution

1. Distribution over a multivariate random variables vector $x \in \mathbb{R}^D$ of real numbers
2. Defined by a mean vector $\mu \in \mathbb{R}^D$ and a $D \times D$ covariance matrix Σ

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}$$



3. The covariance matrix Σ must be symmetric and positive definite
   3.1 All eigenvalues are positive
   3.2 $z^\top \Sigma z > 0$ for any real vector z.
4. Often we parameterize a multivariate Gaussian using the precision matrix $\Lambda = \Sigma^{-1}$.

Bayes theorem

Bayes theorem

1. Bayes theorem

$$p(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum_Y p(X|Y)p(Y)}$$

   1.1 Prior of Y ($p(Y)$) : We have this information before observing anything about Y.
   1.2 Posterior of Y ($p(Y|X)$) : This is the distribution of Y after observing X.
   1.3 Likelihood of X ($p(X|Y)$) : This is the conditional probability that an event Y has the associated observation X.
   1.4 Evidence ($p(X)$) : This is the marginal probability that an observation X is seen.

2. In other words

$$posterior = \frac{prior \times likelihood}{evidence}.$$

Prior distribution

1. What does the shape of a prior tell us?
   It tells us your belief as to how the underlying parameter should be distributed.
2. Which prior should we choose?
   2.1 Based on your preference

       You know from historical data that the parameter should behave in certain ways.

   2.2 Based on physics

       The parameter has a physical interpretation, so you need to abide by the physical laws.

   2.3 Choose a prior that is computationally friendlier.

       This is the topic of the conjugate prior, which is a prior that does not change the form of the posterior distribution.

Maximum a posteriori estimation

1. In many learning scenarios, the learner considers some set $\mathcal{Y}$ and is interested in finding the most probable $Y \in \mathcal{Y}$ given observed data X.
2. This is called maximum a posteriori estimation (MAP) and can be estimated using Bayes theorem.

$$
\begin{aligned}
Y_{MAP} &= \underset{Y \in \mathcal{Y}}{argmax} \; p(Y|X) \\
&= \underset{Y \in \mathcal{Y}}{argmax} \; \frac{P(X|Y)P(Y)}{P(X)} \\
&= \underset{Y \in \mathcal{Y}}{argmax} \; P(X|Y)P(Y)
\end{aligned}
$$

3. $P(X)$ is dropped because it is constant and independent of Y.

$$
\begin{aligned}
Y_{MAP} &= \underset{Y \in \mathcal{Y}}{argmax} \; P(X|Y)P(Y) \\
&= \underset{Y \in \mathcal{Y}}{argmax} \; \{\log P(X|Y) + \log P(Y)\} \\
&= \underset{Y \in \mathcal{Y}}{argmin} \; \{-\log P(X|Y) - \log P(Y)\}
\end{aligned}
$$

Maximum likelihood estimation

1. In some cases, we will assume that every $Y \in \mathcal{Y}$ is equally probable.
2. This is called maximum likelihood estimation.

$$\begin{aligned} Y_{ML} &= \underset{Y \in \mathcal{Y}}{argmax} \; P(X|Y) \\ &= \underset{Y \in \mathcal{Y}}{argmax} \; \log P(X|Y) \\ &= \underset{Y \in \mathcal{Y}}{argmin} \; \{-\log P(X|Y)\} \end{aligned}$$

3. Let $x_1, x_2, \dots, x_N$ be random samples drawn from $p(X, Y)$.
4. Assuming statistical independence between the different samples,we can form $p(X|Y)$ as

$$p(X|Y) = p(x_1, x_2, \dots, x_N|Y) = \prod_{n=1}^{N} p(x_n|Y)$$

5. This method estimates Y so that $p(X|Y)$ takes its maximum value.

$$Y_{ML} = \underset{Y \in \mathcal{Y}}{argmax} \; \prod_{n=1}^{N} p(x_n|Y)$$

Maximum likelihood estimation(cont.)

1. A necessary condition that Y_{ML} must satisfy in order to be a maximum is the gradient of the likelihood function with respect to Y to be zero.

$$\frac{\partial \prod_{n=1}^{N} p(x_n|Y)}{\partial Y} = 0$$

2. Because of the monotonicity of the logarithmic function, we define the log likelihood function as

$$L(Y) = \ln \prod_{n=1}^{N} p(x_n|Y)$$

3. Equivalently, we have

$$\begin{aligned} \frac{\partial L(Y)}{\partial Y} &= \sum_{n=1}^{N} \frac{\partial \ln p(x_n|Y)}{\partial Y} \\ &= \sum_{n=1}^{N} \frac{1}{p(x_n|Y)} \frac{\partial p(x_n|Y)}{\partial Y} = 0 \end{aligned}$$

Readings

1. Chapter 2 of Pattern Recognition and Machine Learning Book (Bishop 2006).
2. Chapter 2 of Machine Learning: A probabilistic perspective (Murphy 2012).
3. Chapter 1 of Probabilistic Machine Learning: An introduction (Murphy 2022).

- Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.
- Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.
- — (2022). *Probabilistic Machine Learning: An introduction*. The MIT Press.

Questions?