Intro, Nearest Neighbors and Probability

MSc. Bui Quoc Khanh
Faculty of Information Technology
Fall 2024

Outline

1. Breaking down an ML
2. Nearest Neighbors Algorithm Review
3. Probability Review

The Machine Learning Problem

Building blocks for a machine learning problem:

- Stages
- Data
- Hyperparameters

Stages

- **Learning**: Extract information from data to make predictions.
- **Evaluation**: Check how well the algorithm/model makes prediction

Data

For supervised learning problems, we often have data of the form: n input data samples, each with d features:

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ x_{d1} & x_{d2} & x_{d3} & \dots & x_{dn} \end{bmatrix}$$

In targets, corresponding to each input sample:

$$\begin{bmatrix} y_1 & y_2 & y_3 & \dots & y_n \end{bmatrix}$$

Data

- **Training:** Used at the learning stage to extract information about the data that is relevant to the predictive task and potentially transfer that knowledge from the data such that it can be accessed to make predictions later.

- **Validation**: Used to select one out of a few possible algorithms or learned models by mimicking test time behavior. This serves as a proxy for measuring overfitting.

- **Test**: Used to evaluate algorithms' performance.

Hyperparameters

- We would like to design an algorithm or learn parameters of the model based on the training data.
- However, there are some (hyper)parameters that we, the designers of the algorithm, must determine.
  - ➢ Knobs that we tune to find a right setting for the algorithm.
  - ➢ Use validation data to choose the setting of a knob

Hyperparameters

Examples:

- Learning Rate: size of updates made to parameters
- Batch Size: amount of the data used at every step of learning
- k: number of Nearest Neighbors

Classification

- Given some data, we want to assign it to meaningful categories by learning the patterns in the training data.
- How do we store information about the learned patterns?

➢ Option 1: We don't! Like the NN algorithm, we can just look at the entire training data. This is a **Non-Parametric** classifier.

➢ Option 2: We create a model with some parameters. During the learning stage, we store information in these parameters. During evaluation, we look at the learned model only. This is a **Parametric** classifier.

Nearest Neighbors Review

Nearest Neighbors

- Let’s review the stages in the NN algorithm:

➢ Learning: None! This algorithm holds all the relevant information in the training set.

➢ Evaluation: For every test point, find the training point “close” to it and assign it the same category.

- This needs us to define a notion of “*closeness*”

Nearest Neighbors

- “Closeness” is measured as a distance between the input vectors.

  For instance, the Euclidean norm:

$$\left\| \begin{matrix} x_{11} - x_{12} \\ x_{21} - x_{22} \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ x_{d1} - x_{d2} \end{matrix} \right\|_2$$

- The NN algorithm compares these distances to determine the closest neighbor.
- Curse of Dimensionality: In higher dimensions, common distances are less meaningful.

Probability Review

Why do we care about probability?

- Uncertainty arises through:
  - ➢ Noisy measurements
  - ➢ Variability between samples
  - ➢ Finite size of data sets
- Probability provides a consistent framework for the quantification and manipulation of uncertainty.

Sample Space

- **Sample space** Ω is the set of all possible outcomes of an experiment.
- **Observations** $\omega \in \Omega$ are points in the space also called sample outcomes, realizations, or elements.
- **Events** $E \subset \Omega$ are subsets of the sample space.

Sample Space

Example: Flip a coin twice:

- **Sample space** includes all possible outcomes

  $\Omega = \{HH, HT, T\,H, T\,T\}$
- **Observation** is any single element of the sample space

  $\omega = HT \in \Omega$.
- **Event** is a subset of the sample space (eg. the event where both flips have the same outcome)

  $E = \{HH, T\,T\} \subset \Omega$

Sample Space

- The probability of an event E, P(E), satisfies three axioms:

  1: P(E) ≥ 0 for *every E*

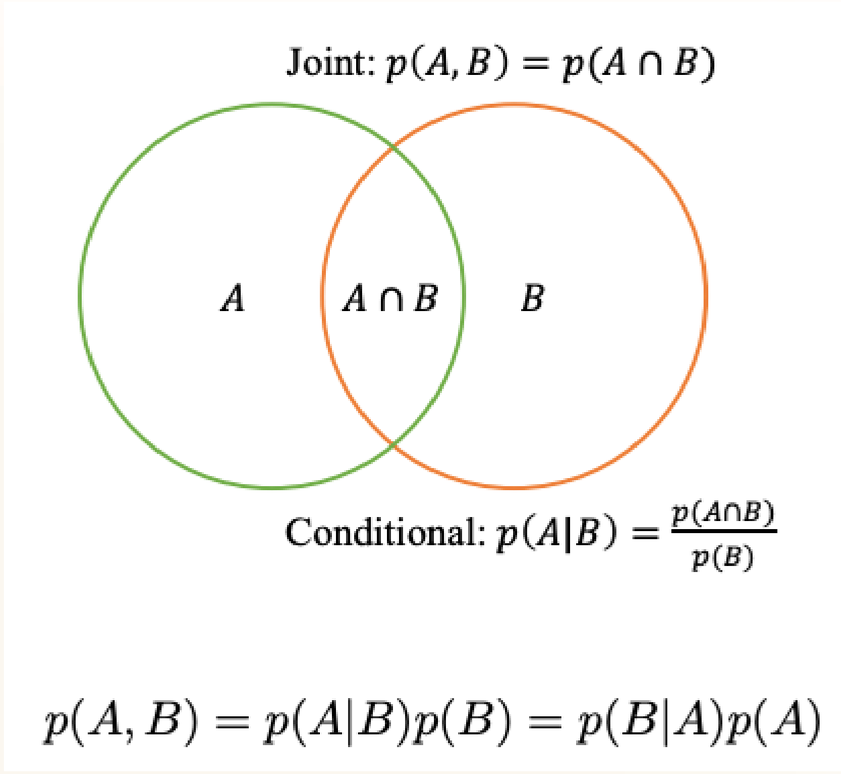
  2: P(Ω) = 1

  3: If E1, E2, . . . are disjoint then

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

Joint and Conditional Probabilities

- Joint Probability of A and B is denoted P(A, B).
- Conditional Probability of A given B is denoted P(A|B).



$$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

Conditional Example

Suppose the probability of passing the midterm is 60% and the probability of passing both the final and the midterm is 45%. What is the probability of passing the final given the student passed the midterm?

$$\begin{aligned} P(F|M) &= P(M,F)/P(M) \\ &= 0.45/0.60 \\ &= 0.75 \end{aligned}$$

Independence

Events A and B are *independent* if $P(A, B) = P(A)P(B)$.

Suppose you have 2 coins. Coin 1 always comes up Heads and Coin 2 always comes up Tails. You close your eyes, pick a coin and toss it. Then you replace it, pick again and toss again.

- **Independent:** Before seeing the result of any toss, you wonder about 2 events; A: first toss is Head, B: second toss is Head.

$$P(A, B) = 0.5 \times 0.5 = P(A)P(B)$$

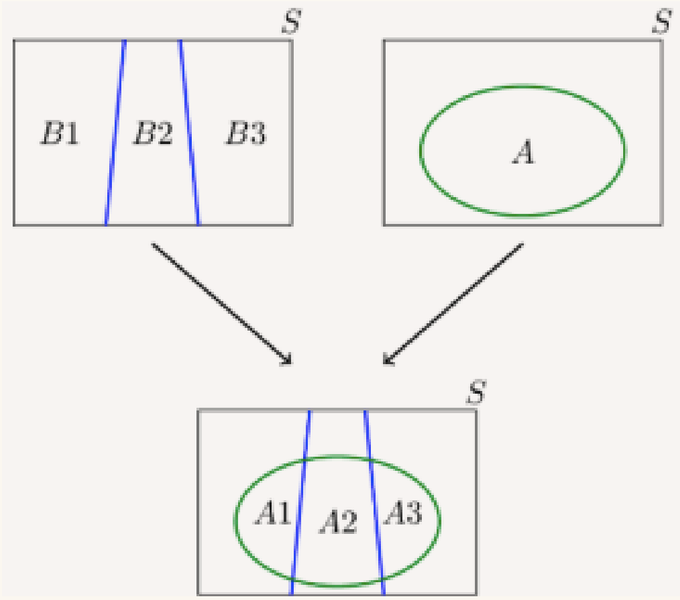
- **Not Independent:** Now you wonder about the same events A and B but you toss the same coin twice.

$$P(A, B) = 0.5 \neq P(A)P(B)$$

Marginalization and Law of Total Probability

Law of Total Probability

$$P(A) = \sum_B P(A, B) = \sum_B P(A|B)P(B)$$



Bayes' Rule

Bayes' Rule:

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

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

$$Posterior \propto Likelihood \times Prior$$

Bayes' Example

Suppose you have tested positive for a disease. What is the probability you actually have the disease?

This depends on the prior probability of the disease:

- $P(T = 1|D = 1) = 0.95$ (likelihood)
- $P(T = 1|D = 0) = 0.10$ (likelihood)
- $P(D = 1) = 0.1$ (prior)

So $P(D = 1|T = 1) =?$

Bayes' Example

P(D = 1|T = 1) =?

Solution: Use Bayes' Rule:

$$P(T=1) = P(T=1|D=1)P(D=1) + P(T=1|D=0)P(D=0)$$
$$= 0.95 * 0.1 + 0.1 * 0.90 = 0.185$$
$$P(D=1|T=1) = \frac{P(T=1|D=1)P(D=1)}{P(T=1)} = \frac{0.95 * 0.1}{P(T=1)} = 0.51$$

Random Variable

How do we connect sample spaces and events to data?

A **random variable** is a mapping which assigns a real number $X(\omega)$ to each observed outcome $\omega \in \Omega$

For example, let's flip a coin 10 times. $X(\omega)$ counts the number of Heads we observe in our sequence. If $\omega = HHT\ HT\ HHT\ HT$ then $X(\omega) = 6$. We often shorten this and refer to the random variable X.

Probability Distribution Statistics

Expectation: First Moment, μ

$$E[x] = \sum_{i=1}^{\infty} x_i p(x_i) \qquad \text{(univariate discrete r.v.)}$$

$$E[x] = \int_{-\infty}^{\infty} x p(x) dx \qquad \text{(univariate continuous r.v.)}$$

Variance: Second (central) Moment, σ^2

$$\begin{aligned} \mathrm{Var}[x] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx \\ &= E[(x-\mu)^2] \\ &= E[x^2] - E[x]^2 \end{aligned}$$

Expectations

From our example, we see that X does not have a fixed value, but rather a distribution of values it can take. It is natural to ask questions about this distribution, such as “What is the average number of heads in 10 coin tosses?”
This average value is called the expectation and denoted as E[X]. It is defined as

$$E[x] = \sum_{a \in \mathcal{A}} P[X = a] \times a$$

where A represents the set of all possible values X(w) can take

Expectation Practice

- What is the expected value of a fair die?
- Solution: X = value of roll

$$E[X] = \sum_{a \in \{1,2,3,4,5,6\}} \frac{1}{6} a$$

$$= \frac{1}{6} \sum_{a=1}^{6} a$$

$$= \frac{21}{6} = \frac{7}{2}$$

Linearity of Expectations

There are two powerful properties regarding expectations.

1. $E[X + Y] = E[X] + E[Y]$.

   This holds even if the random variables are dependent.

2. $E[cX] = cE[X]$, where c is a constant.

Note we cannot say anything in general about $E[XY]$.

Linearity of Expectation Practice 2

Suppose there are n students in class, and they each complete an assignment. We hand back assignments randomly. What is the expected number of students that receive the correct assignment? When n = 3? In general?

X = Number of students that get their assignment back

Xi = Student i gets their assignment back

Solution:

$$\begin{aligned} E[X] &= E[X_1 + X_2 + \ldots + X_n] \\ &= E[X_1] + E[X_2] + \ldots + E[X_n] \\ &= \frac{1}{n} \times n = 1 \end{aligned}$$

Variances

Knowing the expectation can only tell us so much. We have another quantity used to describe how far off we are from the expected value. It is defined as follows for a random variable X with $E[X] = \mu$:

$$Var[x] = E[(X - \mu)^2]$$

The variance can be simplified as:

$$\begin{aligned} E[(X - \mu)^2] &= E[X^2 - 2\mu X + \mu^2]. \\ &= E[X^2] - E[2\mu X] + E[\mu^2] \\ &= E[X^2] - 2\mu E[X] + E[\mu^2] \\ &= E[X^2] - \mu^2 \end{aligned}$$

Variance Properties

Constants get squared:

$$\mathrm{Var}[cX] = c^2\,\mathrm{Var}[X]$$

For independent random variables X and Y , we have

$$E[XY\,] = E[X]E[Y\,]$$

and

$$\mathrm{Var}[X + Y\,] = \mathrm{Var}[X] + \mathrm{Var}[Y\,]$$

Variance Practice

Consider a particle that starts at position 0. At each time step, the particle moves one step to the left or one step to the right with equal probability. What is the variance of the particle at time step n? X = X1 + X2 + . . . + Xn

(Solution: Each X_i is 1 or -1 with equal probability.

$$\mathrm{Var}(X_i) = 1$$

$$\mathrm{Var}(X) = \sum \mathrm{Var}(X_i) = n$$

The expected squared distance from 0 is n.)

Discrete and Continuous Random Variables

Discrete Random Variables

- Takes countably many values, e.g., number of heads
- Distribution defined by probability mass function (PMF)
- Marginalization: $p(x) = \sum_y p(x, y)$

Continuous Random Variables

- Takes uncountably many values, e.g., time to complete task
- Distribution defined by probability density function (PDF)
- Marginalization: $p(x) = \int_y p(x, y) dy$

I.I.D.

Random variables are said to be *independent and identically distributed* (i.i.d.) if they are sampled from the same probability distribution and are mutually independent. This is a common assumption for observations. For example, coin flips are assumed to be i.i.d.