AI1/AI&ML - Naïve Bayes

Dr Leonardo Stella

Aims of the Session

This session aims to help you:

- Describe the fundamental concepts in probability theory
- Explain Bayes’ Theorem and its application in ML
- Apply Naïve Bayes to classification for categorical and numerical independent variables

Overview

- **Fundamental concepts in Probability Theory**
- Bayes’ Theorem
- Naïve Bayes for Categorical Independent Variables
- Naïve Bayes for Numerical Independent Variables

Fundamental Concepts in Probability Theory

- **Probabilistic model**: a mathematical description of an uncertain situation. The two main elements of a probabilistic model are:
  - The **sample space** Ω, which is the set of all possible outcomes
  - The **probability law**, which assigns to a set A of possible outcomes (called an **event**) a nonnegative number $P(A)$ (called the **probability** of A)

- Every probabilistic model involves an underlying process, called the **experiment**, that produces exactly one of several possible outcomes

- A subset of the sample space Ω is called an **event**

Example: Toss of a Coin

- Consider the following experiment a single toss of a fair coin
  - The **sample space** Ω: head (H) or tail (T)
  - The **probability law**: $P(H) = 0.5$ (called the **probability** of H), $P(T) = 0.5$

Example: Toss of a Coin

- Consider the following experiment a single toss of a fair coin
  - The **sample space** Ω: head (H) or tail (T)
  - The **probability law**: $P(H) = 0.5$ (called the **probability** of H), $P(T) = 0.5$

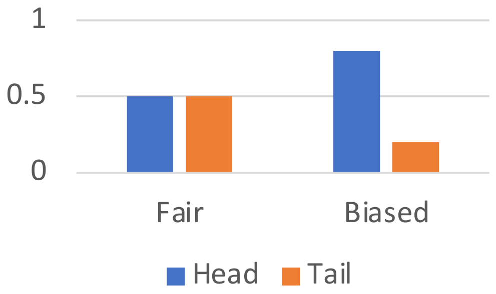
- Let us now consider the experiment consisting of 3 coin tosses. What is the probability of having exactly 2 heads? What about exactly 1 head?

Example: Toss of a Coin

- Consider the following experiment a single toss of a fair coin
  - The **sample space** Ω: head (H) or tail (T)
  - The **probability law**: $P(H) = 0.5$ (called the **probability** of H), $P(T) = 0.5$

- Let us now consider the experiment consisting of 3 coin tosses. What is the probability of having exactly 2 heads? What about exactly 1 head?

- Repeat with the biased coin: $P(H) = 0.8$



Probability Axioms

- **Nonnegativity**: $P(A) \geq 0$, for every event A

- **Additivity**: If A and B are two disjoint events, then the probability of their union satisfies: $P(A \cup B) = P(A) + P(B)$

- **Normalisation**: The probability of the entire sample space is equal to 1, namely $P(\Omega) = 1$

(Discrete) Random Variables

- Given an experiment and the corresponding sample space, a random variable maps a particular number with each outcome

- Mathematically, a random variable *X* is a real-valued function of the experimental outcome

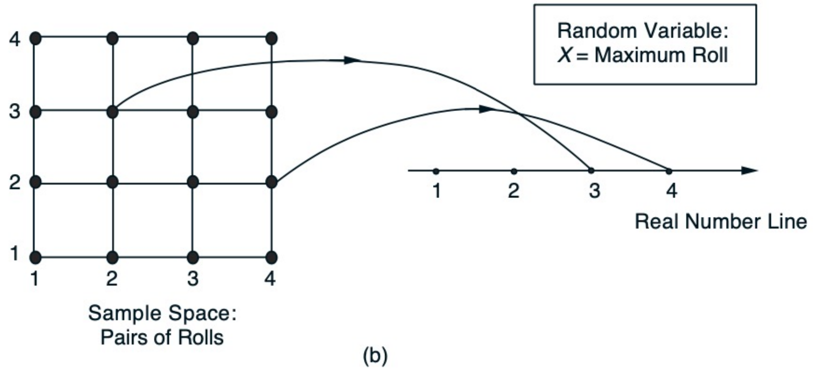


Image: taken from Introduction to Probability (Fig. 2.1 (b))

Probability Mass Function (PMF)

- The probability mass function (PMF) captures the probabilities of the values that a (discrete) random variable can take

- Let us consider the previous example:

$P(X = 1) = 1/16$

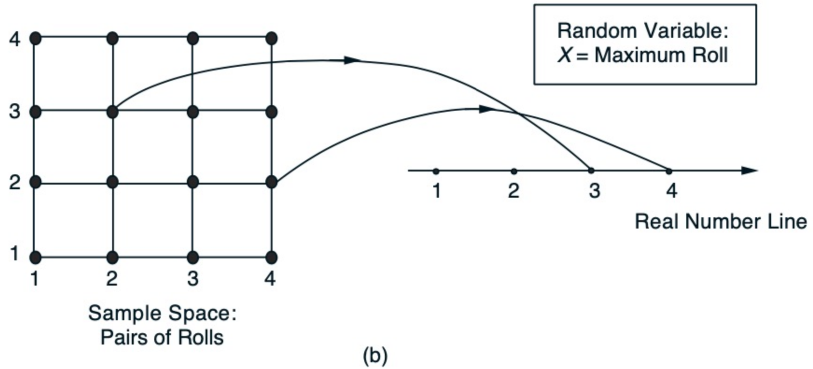


Image: taken from Introduction to Probability (Fig. 2.1 (b))

Probability Mass Function (PMF)

- The probability mass function (PMF) captures the probabilities of the values that a (discrete) random variable can take

- Let us consider the previous example:

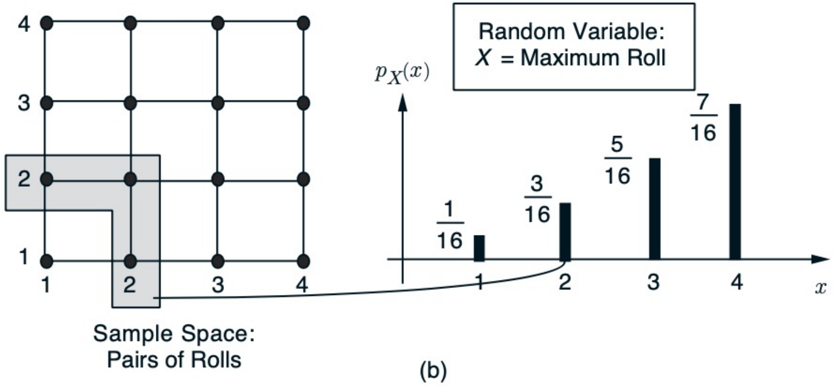
P(X = 1) = 1/16

P(X = 2) = 3/16

P(X = 3) = 5/16

P(X = 4) = 7/16

Image: taken from Introduction to Probability (Fig. 2.2 (b))



Notation

- Random variables are usually indicated with uppercase letters, e.g., X or *Temperature* or *Infection*

- The values are indicated with lowercase letters, e.g., $X \in \{true, false\}$ or $Infection \in \{low, moderate, high\}$

- Vectors are usually indicated with bold letters or a small arrow above the letter, e.g., $\boldsymbol{X}$ or $\vec{X}$

- PMF is usually indicated by the symbol $p_X(x)$

Unconditional/Conditional Probability Distributions

- An **unconditional** (or **prior**) probability distribution gives us the probabilities of all possible events without knowing anything else about the problem, e.g., the maximum value of two rolls of a 4-sided die
- $\boldsymbol{P}(X) = \{\frac{1}{15}, \frac{3}{15}, \frac{5}{15}, \frac{7}{15}\}$

- A **conditional** (or **posterior**) probability distribution gives us the probability of all possible events with some additional knowledge, e.g., the maximum value of two rolls of a 4-sided die knowing that the first roll is 3
- $\boldsymbol{P}(X \mid X_1 = 3) = \{0,0,\frac{3}{4}, \frac{1}{4}\}$

Joint Probability Distributions

- A **joint probability distribution** is the probability distribution associated to all combinations of the values of two or more random variables

- This is indicated by commas, e.g., $\boldsymbol{P}(X, Y)$ or $\boldsymbol{P}(Toothache, Cavity)$

- We can calculate the joint probability distribution by using the **product rule** as in the following:

$$\boldsymbol{P}(X, Y) = \boldsymbol{P}(X \mid Y)\, \boldsymbol{P}(Y) = \boldsymbol{P}(Y \mid X)\, \boldsymbol{P}(X)$$

Mean, Variance and Standard Deviation

- The mean (or expected value or expectation), also indicated by μ, of a random variable X with PMF $p_X(x)$ represents the centre of gravity of the PMF:

$$\mathbf{E}(X) = \sum_x x p_X(x)$$

- E.g., let us consider the random variable X, i.e., the roll of a 4-sided die. The mean is calculated as: $\mathbf{E}(X) =$ 1 * ¼ + 2 * ¼ + 3 * ¼ + 4 * ¼ = 2.5

- The variance of a random variable X provides a measure of the dispersion around the mean:

$$var(X) = \sum_x \left(x - E(X)\right)^2 p_X(x)$$

- The standard deviation is another measure of dispersion: $\sigma_X = \sqrt{var(x)}$

Continuous Random Variables

- A random variable X is called continuous if its probability law can be described in terms of a nonnegative function f_X. This function is called **probability density function** (PDF) and is the equivalent of the PMF for discrete random variables

$$P(X \in B) = \int_B f_X(x)dx$$

- Since we are dealing with continuous variables, there are an infinite number of values that X can take

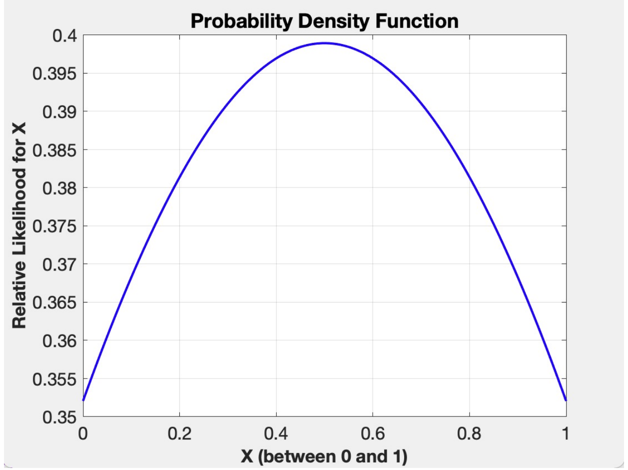
- As for the discrete case, also for continuous random variables we can have unconditional, conditional and joint probability distributions

Example: Random Number Generator

- As an example, let us consider a random number generator that returns a random value between 0 and 1: $X \in [0,1]$

- And let us model it with a Gaussian (or normal) distribution

$$P(X = a \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{(2\pi)}}\, e^{\frac{-(a-\mu)^2}{2\sigma^2}},$$



where μ is the mean and σ^2 is the variance

Also, recall that $\pi = 3.14159$ and $\mathrm{e} = 2.71828$

Overview

- Fundamental concepts in Probability Theory
- **Bayes' Theorem**
- Naïve Bayes for Categorical Independent Variables
- Naïve Bayes for Numerical Independent Variables

Bayes' Theorem

- Recall the product rule for a joint probability distribution of independent variable(s) X and dependent variable Y:

$$\boldsymbol{P}(X, Y) = \boldsymbol{P}(X \mid Y)\, \boldsymbol{P}(Y) = \boldsymbol{P}(Y \mid X)\, \boldsymbol{P}(X)$$

- By taking the second and last term from the above equation and rearranging, we get:

$$\boldsymbol{P}(X \mid Y) = \frac{\boldsymbol{P}(Y \mid X)\boldsymbol{P}(X)}{\boldsymbol{P}(Y)}$$

- The above equation is known as **Bayes' Theorem** (also Bayes' rule or Bayes' law)

ML: Probabilistic Inference

- Our ML task consists in computing the posterior probabilities for query propositions given some observed evidence: this method is **probabilistic inference**

- We use Bayes’ Theorem to make predictions about an underlying process given a knowledge base consisting of the data produced by this process

Equivalent Terminology

- Input attribute, independent variable, input variable
- Output attribute, dependent variable, output variable, label (classification)
- Predictive model, classifier (classification), or hypothesis (statistical learning)
- Learning a model, training a model, building a model
- Training examples, training data
- Example, observation, data point, instance (more frequently used for test examples)
- $P(a, b) = P(a \; and \; b) = P(a \wedge b)$

Learning Probabilities

- Consider the **training set**

Days	Sunny (X_1)	Windy (X_2)	Tennis (Y)
Day 1	yes	no	yes
Day 2	yes	no	yes
Day 3	yes	yes	yes
Day 4	no	yes	no
Day 5	no	no	no
Day 6	no	yes	no

Learning Probabilities

- Consider the **training set**

Days	Sunny (X_1)	Windy (X_2)	Tennis (Y)
Day 1	yes	no	yes
Day 2	yes	no	yes
Day 3	yes	yes	yes
Day 4	no	yes	no
Day 5	no	no	no
Day 6	no	yes	no

- Let us build the **model** for one independent variable, e.g., Windy (X_2)

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes			
Windy = no			
Total			

Learning Probabilities

- Consider the **training set**

Days	Sunny (X_1)	Windy (X_2)	Tennis (Y)
Day 1	yes	no	yes
Day 2	yes	no	yes
Day 3	yes	yes	yes
Day 4	no	yes	no
Day 5	no	no	no
Day 6	no	yes	no

- Let us build the **model** for <u>one</u> independent variable, e.g., Windy (X_2)

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1		
Windy = no			
Total			

Learning Probabilities

- Consider the **training set**

Days	Sunny (X_1)	Windy (X_2)	Tennis (Y)
Day 1	yes	no	yes
Day 2	yes	no	yes
Day 3	yes	yes	yes
Day 4	no	yes	no
Day 5	no	no	no
Day 6	no	yes	no

- Let us build the **model** for <u>one</u> independent variable, e.g., Windy (X_2)

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1		
Windy = no	2		
Total	3		

Learning Probabilities

- Consider the **training set**

Days	Sunny (X_1)	Windy (X_2)	Tennis (Y)
Day 1	yes	no	yes
Day 2	yes	no	yes
Day 3	yes	yes	yes
Day 4	no	yes	no
Day 5	no	no	no
Day 6	no	yes	no

- Let us build the **model** for <u>one</u> independent variable, e.g., Windy (X_2)

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1	2	3
Windy = no	2	1	3
Total	3	3	6

Learning Probabilities (continued)

P(Windy=yes|Tennis=yes) =

P(Windy=no|Tennis=yes) =

P(Windy=yes|Tennis=no) =

P(Windy=no|Tennis=no) =

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1	2	3
Windy = no	2	1	3
Total	3	3	6

Learning Probabilities (continued)

P(Windy=yes|Tennis=yes) = 1/3

P(Windy=no|Tennis=yes) = 2/3

P(Windy=yes|Tennis=no) = 2/3

P(Windy=no|Tennis=no) = 1/3

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1	2	3
Windy = no	2	1	3
Total	3	3	6

Learning Probabilities (continued)

$P(Windy=yes|Tennis=yes) = 1/3$

$P(Windy=no|Tennis=yes) = 2/3$

$P(Windy=yes|Tennis=no) = 2/3$

$P(Windy=no|Tennis=no) = 1/3$

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1	2	3
Windy = no	2	1	3
Total	3	3	6

$P(Windy=yes) = 3/6 = 1/2$

$P(Windy=no) = 3/6 = 1/2$

$P(Tennis=yes) = 3/6 = 1/2$

$P(Tennis=no) = 3/6 = 1/2$

Applying Bayes' Theorem

- Let us consider output class c and input value(s) a. Bayes' Theorem can be rewritten as

$$P(c \mid a) = \frac{P(a \mid c)P(c)}{P(a)}$$

- Now, given input value(s) a, we calculate the above for every class c: our prediction is the one with: $\max_{c} P(c \mid a)$

$$P(Tennis = yes \mid Windy = yes) = \frac{P(Windy = yes \mid Tennis = yes)P(Tennis = yes)}{P(Windy = yes)}$$

Applying Bayes' Theorem (continued)

$$P(Tennis = yes \mid Windy = yes) = \frac{P(Windy = yes \mid Tennis = yes)P(Tennis = yes)}{P(Windy = yes)}$$

$$= \frac{1/3 * 3/6}{3/6} = 0.33$$

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1	2	3
Windy = no	2	1	3
Total	3	3	6

Applying Bayes' Theorem (continued)

$$P(Tennis = yes \mid Windy = yes) = \frac{P(Windy = yes \mid Tennis = yes)P(Tennis = yes)}{P(Windy = yes)}$$

$$= \frac{1/3 * 3/6}{3/6} = 0.33$$

$$P(Tennis = no \mid Windy = yes) = \frac{P(Windy = yes \mid Tennis = no)P(Tennis = no)}{P(Windy = yes)}$$

$$= \frac{2/3 * 3/6}{3/6} = 0.67$$

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1	2	3
Windy = no	2	1	3
Total	3	3	6

Applying Bayes' Theorem (continued)

$$P(Tennis = yes \mid Windy = yes) = 0.33$$

$$P(Tennis = no \mid Windy = yes) = 0.67$$

$$\max_{c} P(c \mid a) = \max \{0.33, 0.67\} = 0.67$$

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1	2	3
Windy = no	2	1	3
Total	3	3	6

Normalising Factor

$$P(Tennis = yes \mid Windy = yes) = \frac{P(Windy = yes \mid Tennis = yes)P(Tennis = yes)}{P(Windy = yes)}$$

$$= \frac{1/3 * 3/6}{3/6} = 0.33$$

$$P(Tennis = no \mid Windy = yes) = \frac{P(Windy = yes \mid Tennis = no)P(Tennis = no)}{P(Windy = yes)}$$

$$= \frac{2/3 * 3/6}{3/6} = 0.67$$

- $1/P(Windy = yes)$ can be seen as a normalisation constant for the distribution: we can replace it with the constant parameter $\alpha = 1/\beta$
- $\beta = \sum_{c \in \mathcal{Y}} P(c)P(a|c)$

More than 1 Independent Variable

$$P(c|a_1, \ldots, a_n) = \frac{P(a_1,\ldots,a_n|\, c)P(c)}{\sum_{c\in\mathcal{Y}}\left(P(c)\prod_{i=1}^{n} P(a_i|c)\right)} = \alpha\, P(a_1, \ldots, a_n|\, c)P(c)$$

- P represents the probability calculated based on the frequency tables
- c represents a class
- a_i represents the value of independent variable $x_i \in \{1, \ldots, n\}$
- n is the number of independent variables
- α is the normalisation factor

Problems: Scaling and Missing Values

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

- In this example (from the book), we have 3 Boolean variables
- For a domain described by n Boolean variables, we would need an input table of size $O(2^n)$ and it would take $O(2^n)$ to process the table
- Also, it is reasonable to think that we will never see values for all possible combinations of the variables
- Naïve Bayes can be used to deal with these issues

Overview

- **Fundamental concepts in Probability Theory**
- **Bayes' Theorem**
- **Naïve Bayes for Categorical Independent Variables**
- Naïve Bayes for Numerical Independent Variables

Recall: Issues with Bayes' Theorem

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

- For increasing numbers of independent variables, all possible combinations must be considered:

$$P(c|a_1, \dots, a_n) = \alpha\, P(c)\, P(a_1, \dots, a_n|c)$$

- For a domain described by n Boolean variables, we would need an input table of size $O(2^n)$ and it would take $O(2^n)$ to process the table

Naïve Bayes: Conditional Independence

- Assumption: each input variable is **conditionally independent** of any other input variables given the output

- **Independence**: A is **independent** of B when the following equality holds (i.e., B does not alter the probability that A has occurred):
$$P(A|B) = P(A)$$

- **Conditional independence**: x_1 is **conditionally independent** of x_2 given y when the following equality holds:
$$P(x_1|x_2, y) = P(x_1, y)$$

Naïve Bayes

- **Conditional independence**: x_1 is **conditionally independent** of x_2 given y when the following equality holds:

$$P(x_1|x_2, y) = P(x_1, y)$$

$$P(c|a_1, \ldots, a_n) = \alpha\, P(c)\, P(a_1, \ldots, a_n|c)$$

Naïve Bayes

- **Conditional independence**: x_1 is **conditionally independent** of x_2 given y when the following equality holds:

$$P(x_1|x_2, y) = P(x_1, y)$$

$$P(c|a_1, \dots, a_n) = \alpha\, P(c)\, P(a_1, \dots, a_n|c)$$

$$\Downarrow$$

$$P(c|a_1, \dots, a_n) = \alpha\, P(c)\, P(a_1|c)P(a_2|c) \dots P(a_n|c)$$

Naïve Bayes

$$P(c|a_1, \dots, a_n) = \alpha\, P(c)\, P(a_1|c) P(a_2|c) \dots P(a_n|c)$$

$$\Downarrow$$

$$P(c|a_1, \dots, a_n) = \alpha\, P(c) \prod_{i=1}^{n} P(a_i|c)$$

where $\alpha = 1/\beta$ and $\beta = \sum_{c \in \mathcal{Y}} (P(c)\, \prod_{i=1}^{n} P(a_i|c))$

Example: Naïve Bayes

- Consider again the **training set**

Days	Sunny (X_1)	Windy (X_2)	Tennis (Y)
Day 1	yes	no	yes
Day 2	yes	no	yes
Day 3	yes	yes	yes
Day 4	no	yes	no
Day 5	no	no	no
Day 6	no	yes	no

- Because of conditional independence, we have a table for each variable:

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1	2	3
Windy = no	2	1	3
Total	3	3	6

Frequency Table	Tennis = yes	Tennis = no	Total
Sunny = yes	3	0	3
Sunny = no	0	3	3
Total	3	3	6

Example: Naïve Bayes (continued)

- Let us determine the predicted class for the following instance:

(Windy = no, Sunny = no, Y = ?)

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1	2	3
Windy = no	2	1	3
Total	3	3	6

Frequency Table	Tennis = yes	Tennis = no	Total
Sunny = yes	3	0	3
Sunny = no	0	3	3
Total	3	3	6

- $P(c|a_1, \dots, a_n) = \alpha\, P(c) \prod_{i=1}^{n} P(a_i|c)$
- $P(\neg T|\neg W, \neg S) = \alpha\, P(\neg T) P(\neg W|\neg T) P(\neg S|\neg T) = \alpha \frac{3}{6} * \frac{1}{3} * \frac{3}{3} = \frac{1}{6}\alpha$
- $P(T|\neg W, \neg S) = \alpha\, P(T) P(\neg W|T) P(\neg S|T) = \alpha \frac{3}{6} * \frac{2}{3} * \frac{0}{3} = 0$

Example: Naïve Bayes (continued)

- $P(\neg T|\neg W, \neg S) = \alpha\, P(\neg T) P(\neg W|\neg T) P(\neg S|\neg T) = \alpha \frac{3}{6} * \frac{1}{3} * \frac{3}{3} = \frac{1}{6}\alpha$
- $P(T|\neg W, \neg S) = \alpha\, P(T) P(\neg W|T) P(\neg S|T) = \alpha \frac{3}{6} * \frac{2}{3} * \frac{0}{3} = 0$

- $\alpha = \frac{1}{\beta} = \frac{1}{\frac{3}{6} * \frac{2}{3} * \frac{0}{3} + \frac{3}{6} * \frac{1}{3} * \frac{3}{3}} = 6$

- $P(\neg T|\neg W, \neg S) = \frac{1}{6} * 6 = 1$
- $P(T|\neg W, \neg S) = 0$

- Problem: in this example, there is no data where Tennis = yes with Sunny = no, so regardless of the value of Windy, we will get inaccuracies in doing predictions

Laplace Smoothing

- To avoid this problem, we can use Laplace smoothing by adding 1 to the frequency of all elements of our training data

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1+1	2+1	3+2
Windy = no	2+1	1+1	3+2
Total	3+2	3+2	6+4

Frequency Table	Tennis = yes	Tennis = no	Total
Sunny = yes	3+1	0+1	3+2
Sunny = no	0+1	3+1	3+2
Total	3+2	3+2	6+4

- Then we use the updated tables when calculating $P(a_i|c)$, so we do not get values with 0
- When we calculate $P(c)$, we use the original tables

Summary

Naïve Bayes Learning Algorithm

- Create frequency tables for each independent variable and the corresponding values for the frequency of an event
- Count the number of training examples of each class with each independent variable
- Apply Laplace smoothing

Naïve Bayes Model

- Consists of the frequency tables obtained from Bayes' Theorem under the conditional independence assumption (with or without Laplace smoothing)

Naïve Bayes prediction for an instance (**X=a**, Y=?)

- We use Bayes' Theorem under the conditional independence assumption

Overview

- Fundamental concepts in Probability Theory
- Bayes’ Theorem
- Naïve Bayes for Categorical Independent Variables
- **Naïve Bayes for Numerical Independent Variables**

Naïve Bayes for Numerical Independent Variables

$$P(c|a_1, \dots, a_n) = \alpha\ P(c)\ P(a_1|c)P(a_2|c) \dots P(a_n|c)$$

$$P(c|a_1, \dots, a_n) = \alpha\ P(c) \prod_{i=1}^{n} P(a_i|c)$$

where $\alpha = 1/\beta$ and $\beta = \sum_{c \in \mathcal{Y}} (P(c)\ \prod_{i=1}^{n}\ P(a_i|c))$

- We predict the class with $\max_{c} [P(c|a_1, \dots, a_n)]$

Naïve Bayes for Numerical Independent Variables

- For categorical independent variables, we can compute the probability of an event through the probability mass function associated with the training data

Frequency Table	Tennis = yes	Tennis = no	Total
Windy = yes	1+1	2+1	3+2
Windy = no	2+1	1+1	3+2
Total	3+2	3+2	6+4

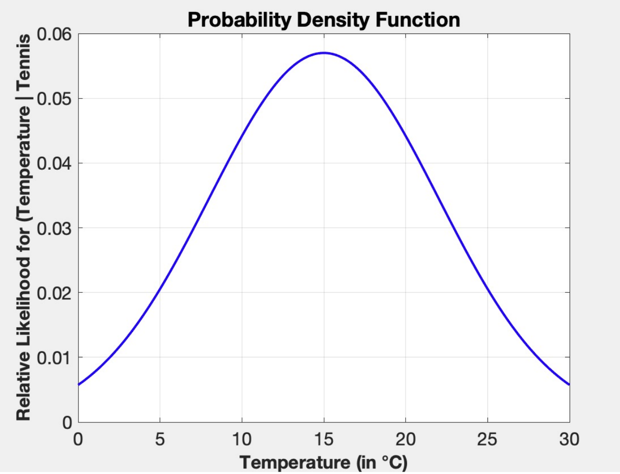
Naïve Bayes for Numerical Independent Variables

- Instead, we assume that examples are drawn from a probability distribution. We can use a Gaussian distribution as we did before

- Gaussian distribution with mean $\mu = 15$ and variance $\sigma^2 = 49$

$$P(X = a \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{(2\pi)}}\, e^{\frac{-(a-\mu)^2}{2\sigma^2}},$$

Also, recall that $\pi = 3.14159$ and $\mathrm{e} = 2.71828$



Naïve Bayes for Numerical Independent Variables

- Let us consider the training data below. We create the PDF for Tennis = yes and for Tennis = no

- So, for Tennis = yes, we calculate mean and variance

- $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{15+25+35}{3} = 25$
- $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$

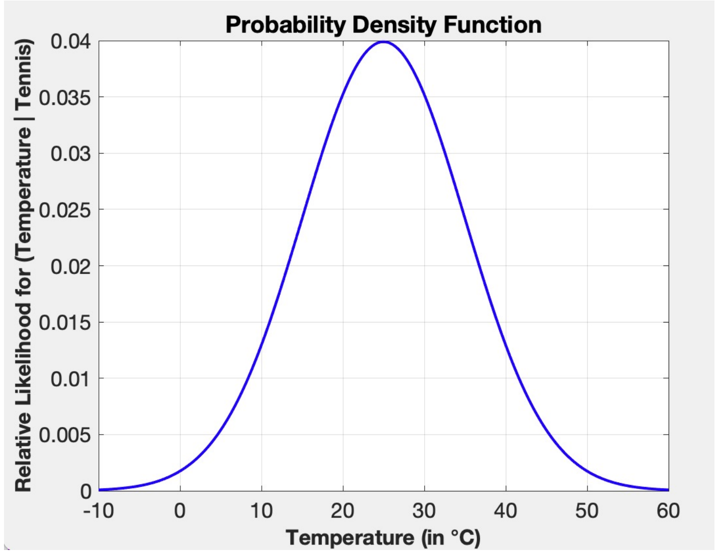
$= \frac{1}{2}\left[(15-25)^2 + (25-25)^2 + (35-25)^2\right] = 100$

Days	Sunny (X_1)	Temp. (X_2)	Tennis (Y)
Day 1	yes	15	yes
Day 2	yes	25	yes
Day 3	yes	35	yes
Day 4	no	10	no
Day 5	no	20	no
Day 6	no	5	no

Naïve Bayes for Numerical Independent Variables

- Gaussian distribution with mean $\mu = 25$ and variance $\sigma^2 = 100$

$$P(X = a \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{(2\pi)}}\, e^{\frac{-(a-\mu)^2}{2\sigma^2}}$$



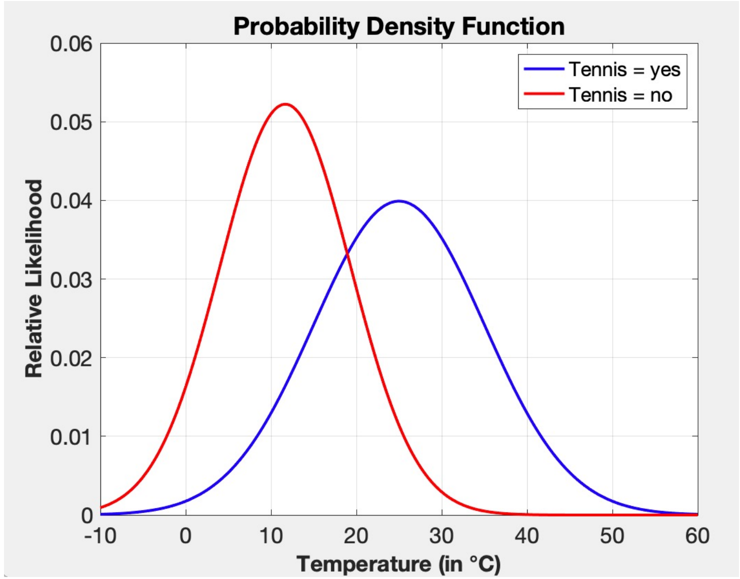
Naïve Bayes for Numerical Independent Variables

- Gaussian distribution with mean $\mu = 25$ and variance $\sigma^2 = 100$

$$P(X = a \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{(2\pi)}}\, e^{\frac{-(a-\mu)^2}{2\sigma^2}}$$

Now, if we repeat for Tennis = no

- $\mu = 11.67$
- $\sigma^2 = 58.34$



Example

- Let us build the tables for

Days	Sunny (X_1)	Temp. (X_2)	Tennis (Y)
Day 1	yes	15	yes
Day 2	yes	25	yes
Day 3	yes	35	yes
Day 4	no	10	no
Day 5	no	20	no
Day 6	no	5	no

Frequency Table	Tennis = yes	Tennis = no	Total
Sunny = yes	3+1	0+1	3+2
Sunny = no	0+1	3+1	3+2
Total	3+2	3+2	6+4

Parameter Table	Tennis = yes	Tennis = no
μ	25	11.67
σ^2	100	58.34

Example

- Now, let us use Naïve Bayes to make a prediction based on the tables:

Frequency Table	Tennis = yes	Tennis = no	Total
Sunny = yes	3+1	0+1	3+2
Sunny = no	0+1	3+1	3+2
Total	3+2	3+2	6+4

Parameter Table	Tennis = yes	Tennis = no
μ	25	11.67
σ^2	100	58.34

- $P(c|a_1, \ldots, a_n) = \alpha\, P(c) \prod_{i=1}^{n} P(a_i|c)$
- We use the frequency table for the categorical independent variables
- We use the parameter table for the numerical independent variables

Example

- Calculate P(Tennis=yes|Sunny=no,Temperature=20):

Frequency Table	Tennis = yes	Tennis = no	Total
Sunny = yes	3+1	0+1	3+2
Sunny = no	0+1	3+1	3+2
Total	3+2	3+2	6+4

Parameter Table	Tennis = yes	Tennis = no
μ	25	11.67
σ^2	100	58.34

- $P(T|\neg S, Temp = 20) = \alpha\, P(T) P(\neg S|T) P(Temp = 20|T)$

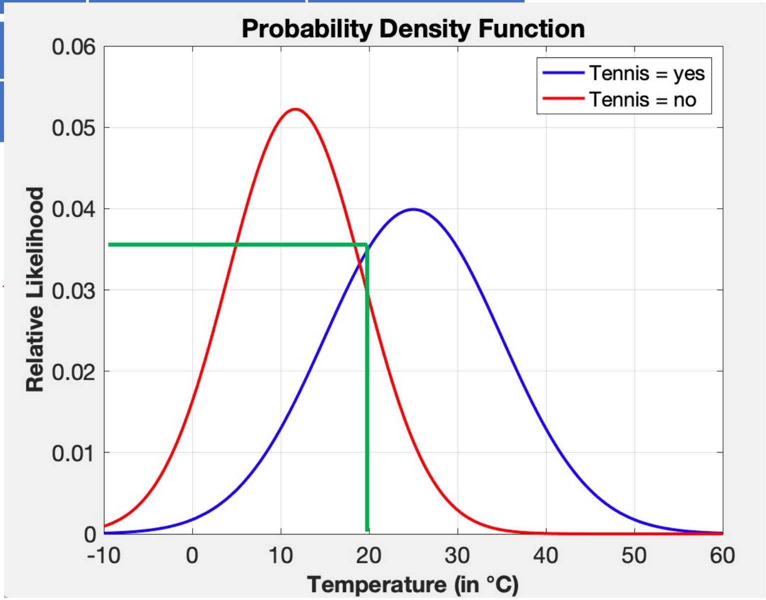
$$= \alpha \frac{3}{6} * \frac{1}{5} * P(Temp = 20|T)$$

Example

- Calculate P(Tennis=yes|Sunny=no,Temperature=20):

Frequency Table	Tennis = yes	Tennis = no	Total
Sunny = yes	3+1	0+1	3+2
Sunny = no	0+1	3+1	3+2
Total	3+2	3+2	6+4

Parameter Table	Tennis = yes	Tennis = no
μ		
σ^2		



- $P(T|\neg S, Temp = 20) = \alpha\, P(T)P(\neg S|T)$

$$= \alpha \frac{3}{6} * \frac{1}{5} * 0.035 = 0.0035\alpha$$

Example

- Calculate P(Tennis=no|Sunny=no,Temperature=20):

Frequency Table	Tennis = yes	Tennis = no	Total
Sunny = yes	3+1	0+1	3+2
Sunny = no	0+1	3+1	3+2
Total	3+2	3+2	6+4

Parameter Table	Tennis = yes	Tennis = no
μ	25	11.67
σ^2	100	58.34

- $P(\neg T|\neg S, Temp = 20) = \alpha\, P(\neg T) P(\neg S|\neg T) P(Temp = 20|\neg T)$

$$= \alpha \frac{3}{6} * \frac{4}{5} * 0.029 = 0.0116\alpha$$

Example

- Calculate P(Tennis=no|Sunny=no,Temperature=20):
- $P(T|\neg S, Temp = 20) = \alpha\ P(T)P(\neg S|T)P(Temp = 20|T)$

$$= \alpha \frac{3}{6} * \frac{1}{5} * 0.035 = 0.0035\alpha$$

- $P(\neg T|\neg S, Temp = 20) = \alpha\ P(\neg T)P(\neg S|\neg T)P(Temp = 20|\neg T)$

$$= \alpha \frac{3}{6} * \frac{4}{5} * 0.029 = 0.0116\alpha$$

- Predicted class: Tennis = no

Summary

Naïve Bayes Learning Algorithm

- Create frequency tables for each categorical independent variable and the corresponding values for the frequency of an event
- Apply Laplace smoothing
- Calculate the parameters of the PDF corresponding to each numerical independent variable

Naïve Bayes Model

- Consists of the frequency tables obtained from Bayes' Theorem under the conditional independence assumption (with or without Laplace smoothing)

Naïve Bayes prediction for an instance (**X=a**, Y=?)

- We use Bayes' Theorem under the conditional independence assumption

Pros and Cons of Naïve Bayes

Pros

- Easy to implement and fast to predict a class from training data (online learning)
- Performs well in multi-class prediction
- Good for categorical variables in general

Cons

- Data that are not observed require smoothing techniques to be applied
- For numerical variables, Gaussian distribution is assumed (strong assumption)
- Not good for regression problems

Aims of the Session

You should now be able to:

- Describe the fundamental concepts in probability theory
- Explain Bayes’ Theorem and its application in ML
- Apply Naïve Bayes to classification for categorical and numerical independent variables