Machine Learning

Supervised Learning-II

Naïve Bayes classifier

Topics

- **Introduction**
- **Bayes Theorem**
- **Naive Bayes Classifier**

Introduction

- Bayes theorem is given by an English statistician, philosopher, and Presbyterian minister named **Mr. Thomas Bayes** in 17th century.
- **Bayes theorem** (**Bayes rule or Bayes Law**) describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

  (For example, if the risk of developing health problems is known to increase with age)
- **Bayes theorem** is also widely used in Machine Learning where we need to predict classes **precisely and accurately.**

Introduction

- An important concept of Bayes theorem named **Bayesian method** is used to calculate **conditional probability** in Machine Learning application that includes classification tasks.
- **Bayesian method** is used to calculate the probability of occurring one event while other one already occurred.
- It is a best method to relate the **condition probability** and **marginal probability (unconditional probability).**
- A simplified version of Bayes theorem known as **Naïve Bayes classification** is used to reduce computation time and average cost of the classification

Introduction

- **Bayes theorem is one of the most popular machine learning concepts that helps to calculate the probability of occurring one event with uncertain knowledge while other one has already occurred.**

Introduction

Prerequisites for Bayes Theorem

- **An experiment** is defined as the **planned operation carried out under controlled condition** such as tossing a coin, drawing a card and rolling a dice.
- **Sample space:** During an experiment what we get as a result is called as possible outcomes and the set of all possible outcome of an event is known as **sample space.**

  if we are rolling a dice, sample space will be: S1 = {1, 2, 3, 4, 5, 6}

  toss a coin : S2 = {Head, Tail}

Introduction

Prerequisites for Bayes Theorem

- **Event** is defined as **subset of sample space** in an experiment. Further, it is also called as set of **outcomes.**

  Assume in our experiment of rolling a dice, there are two event A and B such that;

  A = Event when an even number is obtained = {2, 4, 6}

  B = Event when a number is greater than 4 = {5, 6}

  Probability of the event A P(A) = 3/6

  Probability of the event A P(B) = 2/6

- **Union of event A and B:** $A \cup B = \{2, 4, 5, 6\}$

  P($A \cup B$ **)** = 4/6

- **Intersection of event A and B:** $A \cap B = \{6\}$, $P(A \cap B) = 1/6$

Introduction

Prerequisites for Bayes Theorem

- **Disjoint Event:** If the intersection of the event A and B is an empty set or null then such events are known as **disjoint event** or **mutually exclusive events.**
- A **random variable** takes on some random values and each value having some probability.
- **Exhaustive event** : Two events A and B are said to be exhaustive if either A or B definitely occur at a time and both are **mutually exclusive** for e.g., while tossing a coin, either it will be a **Head or may be a Tail.**

(A set of events are called exhaustive events if at least one of them necessarily occurs whenever the experiment is performed. Also, the union of all these events constitutes the sample space of that experiment.**)**

Introduction

Prerequisites for Bayes Theorem

- **Independent Event:** Two events are said to be **independent** when occurrence of one event **does not affect the occurrence of another event**. In simple words we can say that the probability of outcome of both events does not depends one another.

Mathematically, two events A and B are said to be independent if:

$$P(A \cap B) = P(AB) = P(A)*P(B)$$

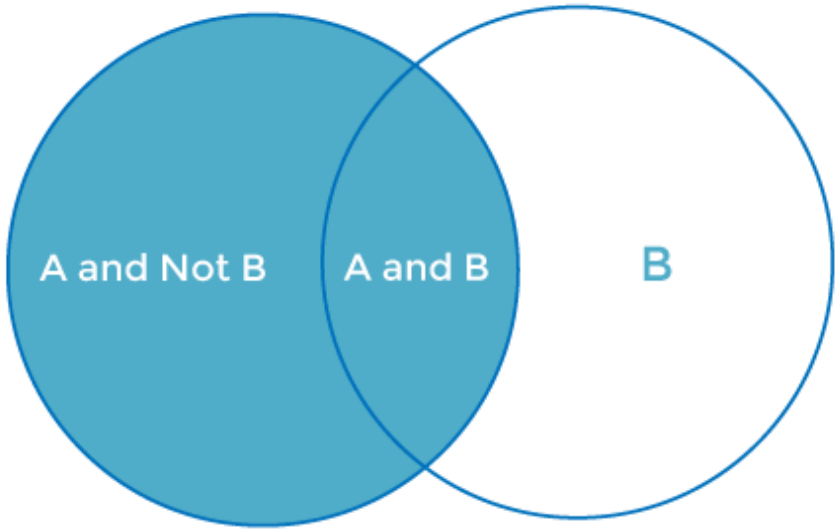
Introduction

Prerequisites for Bayes Theorem

- **Marginal Probability:** Marginal probability is defined as the probability of an event A occurring independent of any other event B.

$$P(A) = P(A|B)*P(B) + P(A|\sim B)*P(\sim B)$$

Here ~B represents the event that B does not occur.



BAYES THEOREM

- In machine learning we are often interested in determining the **best hypothesis** from some **space H**, given the observed training data **D**.
- That is, we demand the most **probable hypothesis**, given the **data D** plus any initial knowledge about the **prior probabilities of the various hypotheses in H.**
- More precisely, Bayes theorem provides a way to calculate the probability of a hypothesis based on its **prior probability**, the probabilities of **observing various data given the hypothesis**, and **the observed data itself**.

BAYES THEOREM

- Let **P(h)** to denote the **initial probability** that hypothesis **h** holds, before we have observed the training data.
- **P(h)** is often called the **prior probability** of h and may reflect any **background knowledge** we have about the chance that **h is a correct hypothesis.**
- If **we have no such prior knowledge,** then we might simply **assign the same prior probability** to each candidate hypothesis.

BAYES THEOREM

- Let **P(D)** to denote the **prior probability** that training data D will be observed (i.e., the probability of D given no knowledge about which hypothesis holds). **(Marginal Probability)**
- **P(D|h)** to denote the probability of observing data D given some world in which hypothesis h holds. (**Likelihood probability**)
- In machine learning problems we are interested in the probability **P (h|D), that h holds given the observed training data D.**

BAYES THEOREM

- **P(h|D)** is called the **posterior probability of h**, because it reflects our confidence that **h** holds after we have seen the training data D.
- Notice the posterior probability P(h|D) reflects the **influence of the training data D**, in contrast to the prior probability P(h) , which is independent of D. (before data is seen)

BAYES THEOREM

- Bayes theorem provides a way to calculate the posterior probability **P(hlD),** from the **prior probability P(h)**, together with **P(D) and P(D(h).**

Bayes theorem:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

BAYES THEOREM

Note :

- **P(hlD)** increases with **P(h)** and with **P(D|h)** according to Bayes theorem.
- **P(hlD)** decreases as **P(D)** increases. (the more probable it is that D will be observed independent of h, the less evidence D provides in support of h)

BAYES THEOREM

Example : Mrs. Jane, wants to takes a test to determine if she has diabetes. Let's say that the overall probability having diabetes is **5%;** this would be our **prior probability.** However, if she obtains **a positive result from her test**, the prior probability is updated to account for this additional information, and it then becomes our **posterior probability.** This example can be represented with the following equation, **using Bayes' Theorem:**

$$\boldsymbol{P(diabetes \mid +test) = \frac{P(+test|diabetes)P(diabetes)}{P(+test)}}$$

Where

$$\boldsymbol{P(+test) = P(+test|diabetes)P(diabetes) + P(+test|no\ diabetes)P(no\ diabetes)}$$

(this remains constant independent of diabetes)

BAYES THEOREM

We are interested in finding the most probable hypothesis $h \in H$ given the observed data D (or at least one of the maximally probable if there are several). Any such maximally probable hypothesis is called a **maximum a posteriori (MAP) hypothesis**. We can determine the MAP hypotheses by using Bayes theorem to calculate the posterior probability of each **candidate hypothesis.**

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h)$$

we dropped the term ***P(D)*** because it is a constant independent of ***h***

BAYES THEOREM

- In some cases, we will assume that every hypothesis in H is equally probable a priori ($P(h_i) = P(h_j)\ for\ all\ h_i and\ h_j\ in\ H$).
- Thus, In this case we c need only consider the term $P(D|h)$ to find the most probable hypothesis.
- $P(D|h)$ is often called the **likelihood** of the data D given h, and any hypothesis that maximizes $P(D|h)$ is called a maximum likelihood (ML) hypothesis, h_{ML}

$$h_{ML} \equiv \underset{h \in H}{\operatorname{argmax}}\ P(D|h)$$

- **We consider cases where H is a hypothesis space containing possible target functions and the data D are training examples.**

BAYES THEOREM – An Illustration

- Consider a medical diagnosis problem in which there are two alternative hypotheses:
  1. **that the patient has a particular form of cancer.**
  2. **that the patient does not.**
- The available data is from a particular laboratory test with two possible outcomes: $\oplus$ (positive) and $\ominus$ (negative).
- We have prior knowledge that over the entire population of people only **.008** have this disease.

BAYES THEOREM – An Illustration

- The test returns **a correct positive** result in only **98%** of the cases in which the disease is actually present and **a correct negative** result in only **97%** of the cases in which the disease is not present.
- In other cases, the test returns **the opposite result.** The above situation can be summarized by the following probabilities:

$$P(cancer) = .008, \qquad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98, \qquad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03, \qquad P(\ominus|\neg cancer) = .97$$

BAYES THEOREM – An Illustration

- Suppose we now observe a new patient for whom the lab test returns a **positive result**.
- **Should we diagnose the patient as having cancer or not?** The **maximum a posteriori hypothesis can be found using the Equation:**

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h)$$

BAYES THEOREM – An Illustration

- Thus,

$$\boldsymbol{P(cancer|\oplus) = P(\oplus|cancer)P(cancer) = (.98).008 = .0078}$$

$$\boldsymbol{P(\neg cancer|\oplus) = P(\oplus|\neg cancer)P(\neg cancer) = (.03).992 = .0298}$$

Thus, $h_{MAP} = \neg cancer.$

BAYES THEOREM – An Illustration

- The exact posterior probabilities can also be determined by normalizing the above quantities so that they sum to 1

$$\boldsymbol{P(cancer|\oplus) = \frac{P(\oplus|cancer)P(cancer)}{P(\oplus)} = 0.0078/(0.0078 + 0.0298) = .21}$$

$$\boldsymbol{P(\neg cancer|\oplus) = \frac{P(\oplus|\neg cancer)P(\neg cancer)}{P(\oplus)} = 0.0298/(0.0298 + 0.0078) = .79}$$

So that

$$\boldsymbol{P(cancer|\oplus) + P(\neg cancer|\oplus) = 1}$$

BAYES THEOREM – An Illustration

Note

- Notice that while the posterior probability of cancer is significantly higher than its prior probability, the most probable hypothesis is still that the patient does not have cancer.

- This example illustrates that he result of Bayesian inference depends strongly **on the prior probabilities**, which must be available in order to apply the method directly.

BAYES THEOREM

1. Each observed training example **can incrementally decrease or increase the estimated probability that a hypothesis is correct**. This provides a more flexible approach to learning than algorithms that completely eliminate a hypothesis if it is found to be inconsistent with any single exampl

BAYES THEOREM

2. Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.

   In Bayesian learning, prior knowledge is provided by asserting

   - a prior probability for each candidate hypothesis, P(h)
   - a probability distribution over observed data for each possible hypothesis. $(P(D|h))$

BAYES THEOREM

3. Bayesian methods can accommodate hypotheses that make probabilistic predictions
4. New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.
5. Even in cases where Bayesian methods prove computationally intractable, they can provide a standard of optimal decision making against which other practical methods can be measured.

BAYES THEOREM

Practical difficulty in applying Bayesian methods

1. One practical difficulty in applying Bayesian methods is that they **typically require initial knowledge of many probabilities.** When these probabilities are not known in advance they are often **estimated based on background knowledge, previously available data, and assumptions** about the form of the underlying distributions.

BAYES THEOREM

Practical difficulty in applying Bayesian methods

2. A second practical difficulty is the significant computational cost required to determine the Bayes optimal hypothesis in the general case. In certain specialized situations, this computational cost can be significantly reduced.

NAIVE BAYES CLASSIFIER

- The naive Bayes classifier applies to learning tasks where each instance x is described **by a conjunction of attribute values** and where the **target function f (x) can take on a value from some finite set V.**
- A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values $(a_{1,}a_{2,}, a_{3,}\ldots a_{n})$.
- The learner is asked to predict the target value, or classification, for this new instance.

NAIVE BAYES CLASSIFIER

- The Bayesian approach to classifying the new instance is to assign the **most probable target value**, V_{MAP} , given the attribute values $(a_1, a_2, a_3, \ldots a_n)$ that describe the instance

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2 \ldots a_n)$$

- We can use Bayes theorem to rewrite this expression as

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

NAIVE BAYES CLASSIFIER

- The naive Bayes classifier is based on the assumption that the **attribute values are conditionally independent** given the target value.
- Means, the assumption is that given the target value of the instance, the probability of observing the conjunction $(a_{1,}a_{2,},a_{3,}\ldots a_n)$ is just the product of the probabilities for the individual attributes:

$$P(a_1, a_2 \ldots a_n|v_j) = \prod_i P(a_i|v_j)$$

- Substituting this into Equation in the previous eqn we get :

$$\mathbf{V_{NB} = \underset{v_j \in V}{argmax}\, P(v_j) \prod_i P(a_i|v_j)}$$

An Illustrative Example

- Let us apply the naive Bayes classifier to a concept learning problem i.e., classifying days according to whether someone will play tennis. The below table provides a set of 14 training examples of the target value **PlayTennis**, where each day is described by the attributes **Outlook, Temperature, Humidity, and Wind**

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

An Illustrative Example

- Use the naive Bayes classifier and the training data from this table to classify the following novel instance:

< Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong >

- Our task is to predict the target value **(yes or no)** of the target concept **PlayTennis** for this new instance

$$V_{NB} = \underset{v_j \in \{yes, no\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)$$

$$V_{NB} = \underset{v_j \in \{yes, no\}}{\operatorname{argmax}} P(v_j)\, P(\text{Outlook}=\text{sunny}|v_j)\, P(\text{Temperature}=\text{cool}|v_j)\, P(\text{Humidity}=\text{high}|v_j)\, P(\text{Wind}=\text{strong}|v_j)$$

An Illustrative Example

- The probabilities of the **different target values** can easily be estimated based on their frequencies over the 14 training examples
  - P(PlayTennis = yes) = 9/14 = 0.64
  - P(PlayTennis = no) = 5/14 = 0.36

- Similarly, estimate the conditional probabilities.

$$P(sunny|yes) = \frac{2}{9} \qquad P(sunny|no) = \frac{3}{5}$$

$$P(cool|yes) = \frac{3}{9} \qquad P(cool|no) = \frac{1}{5}$$

$$P(high|yes) = \frac{3}{9} \qquad P(high|no) = \frac{4}{5}$$

$$P(strong|yes) = \frac{3}{9} \qquad P(strong|no) = \frac{3}{5}$$

An Illustrative Example

- Calculate V_{NB} according to the Equation

$$P(yes)\ P(sunny|yes)\ P(cool|yes)\ P(high|yes)\ P(strong|yes) = .0053$$

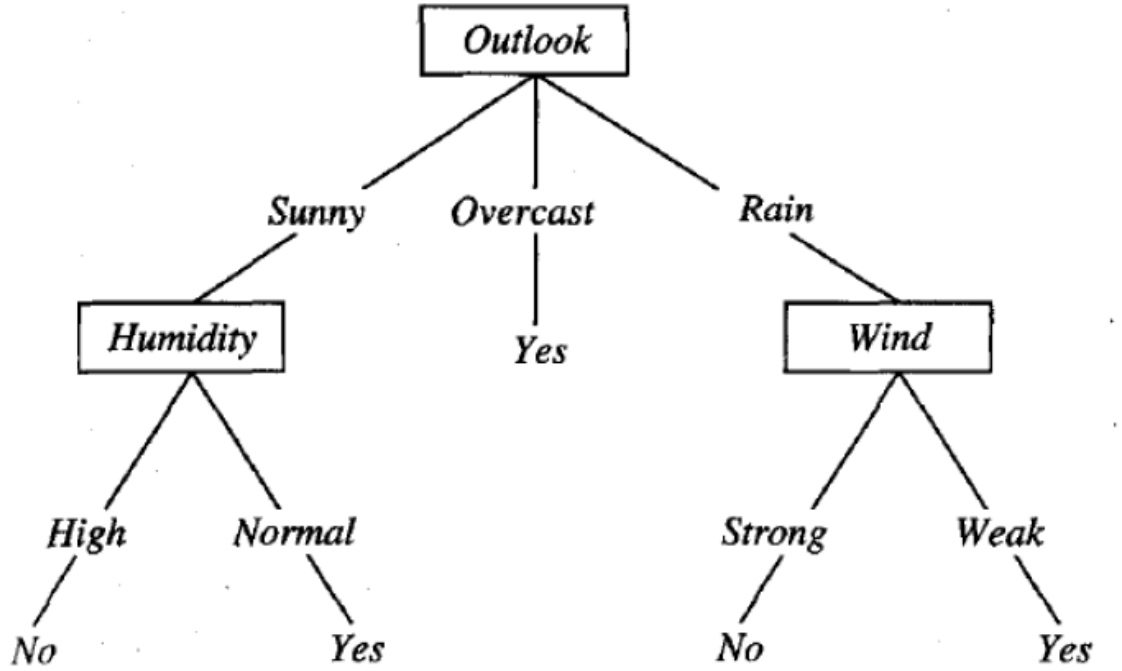
$$P(no)\ P(sunny|no)\ P(cool|no)\ P(high|no)\ P(strong|no) = .0206$$

- Thus, the naive Bayes classifier assigns the target value **PlayTennis = no** to this new instance, based on the probability estimates learned from the training data.

An Illustrative Example

Using the decision tree, we get the same class.

< Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong >



NAIVE BAYES - Handling Continuous Data

NAIVE BAYES - Handling Continuous Data

- Most real-world data are continuous (such as age, weight, price etc) . Relative frequency is then impractical.
- For example, it is easy to establish that the probability of a student being male is P (male) = 0.7. However, the probability that this student's body weight is 184.5 pounds is extremely small (as there are infinite weight values)

Discretizing Continuous Attributes

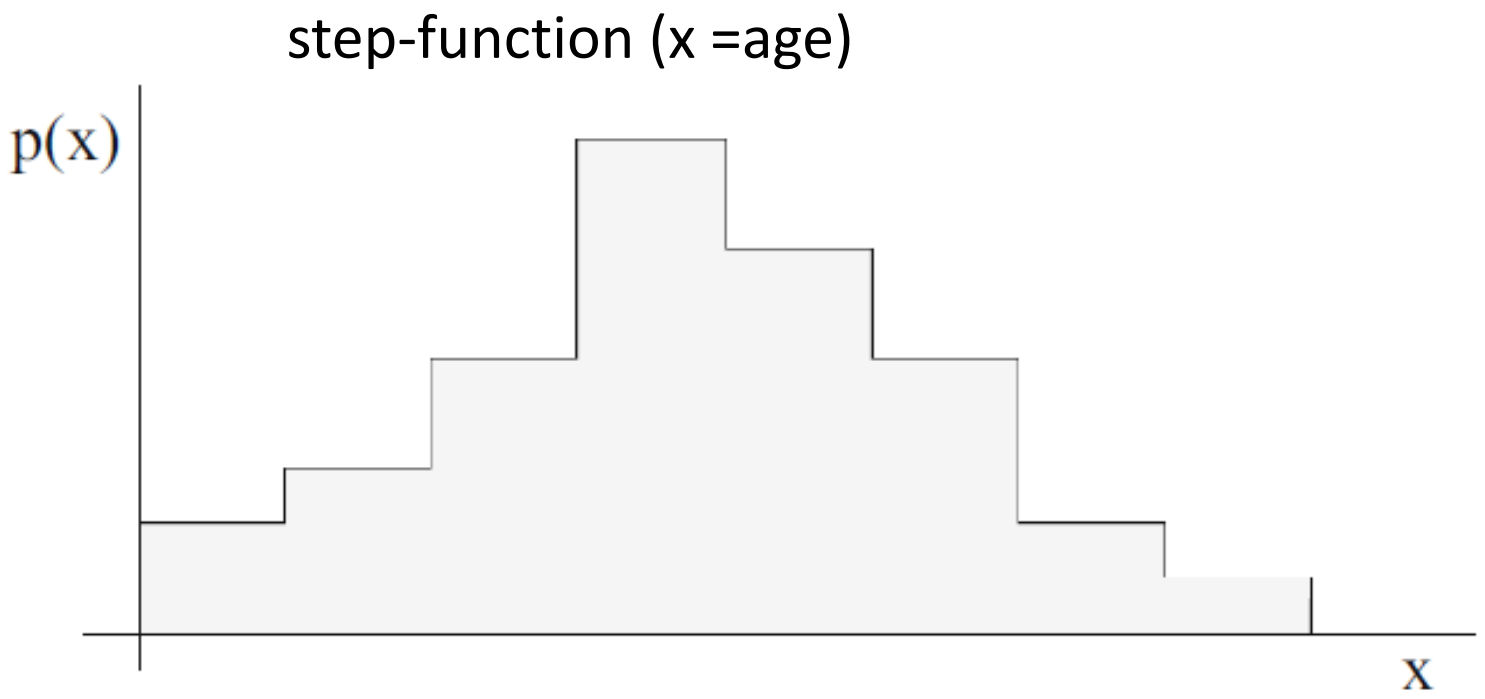
- One simplest approach for discretizing continuous attributes is to replace them with **Boolean attribute.**
- For instance, replacing **age attribute** with the Boolean attribute true (old) **if age > 60** and **false** (Young) otherwise. Thus, relative frequency can easily ne used to **probability of a person being old** , for example.
- But in this case, we see that **information gets lost**; If he/she is old, then we no longer know how old he is? Nor do we know whether one old person is older than another old person.

Discretizing Continuous Attributes

- This loss will be mitigated if we divide the original domain into several intervals, say **[0-10] ,[11-18],…,[91-100].**
- If N is number of training examples and N_i is the number of values in the i^{th} interval, then the $P\left(person\ age\ is\ in\ i^{th}\ interval \right) = \frac{N_i}{N}$
- Thus, we have a mechanism to estimate the probability not of a **concrete value of age,** but rather of this value falling into the given interval.

Discretizing Continuous Attributes

- Suppose that the age attribute values are such that most of the age values are concentrated **near the average age**. So in this case we reach the situation which is depicted below .

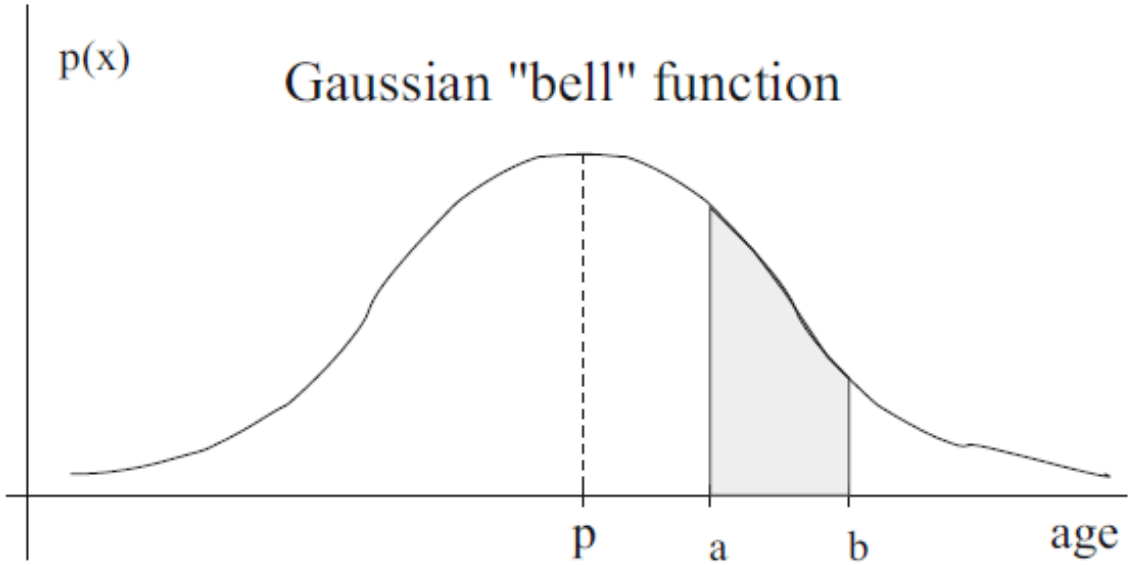


Discretizing Continuous Attributes

- We may fine-tune the step function by dividing the original domain into shorter—and thus more numerous—intervals, provided that the number of age values in each interval is sufficient for reliable probability estimates.
- If the training set is **infinitely large**, we can, theoretically speaking, keep reducing the lengths of the intervals until they become **infinitesimally** small. The resulting function we get is known as **Probability Density Function** $(\mathbf{pdf})$ **(normal/Gaussian bell function)**

Discretizing Continuous Attributes

- A high value of $p(x)$ indicates that there are many examples with age close to x;
  conversely, a low value of $p(x)$ tells us that age in the vicinity of x is rare.



- **The *pdf* at *x* is denoted by a lowercase letter, *p*(x).**
- Now the probability of $age \in [a, b]$ is the relative size of the area below the corresponding section of the pdf. (see the shaded region)

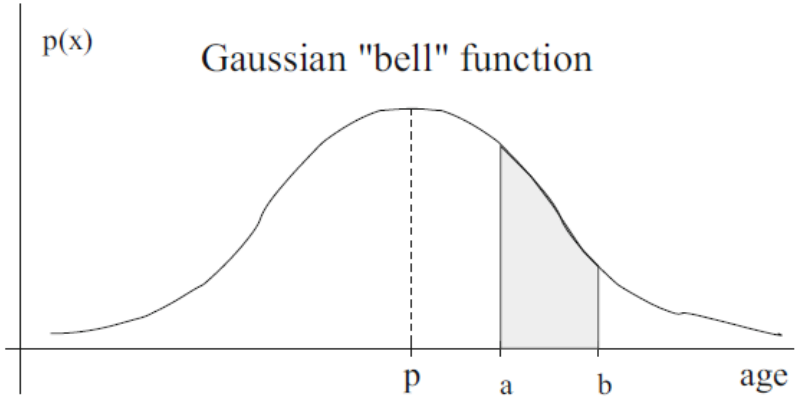
The Gaussian function

- The gaussian function is defined by the following formula where e is the base of the natural logarithm

$$p(x) = k \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where

$$k = \frac{1}{\sqrt{2\pi\sigma^2}}$$



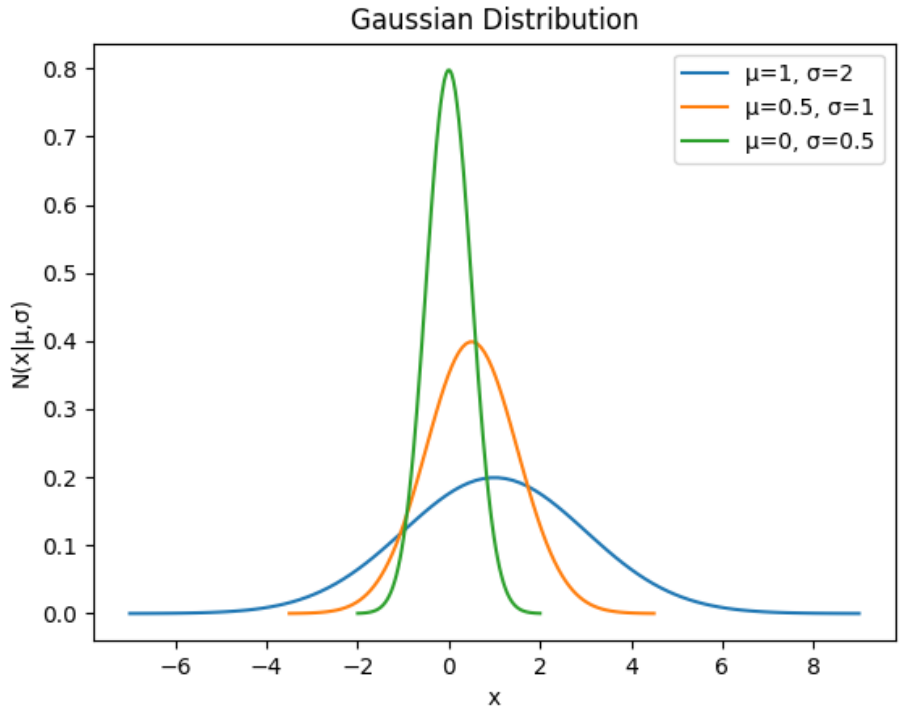
The Gaussian function

Note :

- Note that the greater the difference between x and μ, the greater the exponent's numerator, and thus the smaller the value of $p(x)$ because the exponent is negative.
- Greater the variance σ^2 the greater the width of the curve. A smaller variance defines a narrower bell curve.
- The task for the coefficient k is to make the area under the bell function equal to 1 as required by the theory of probability.

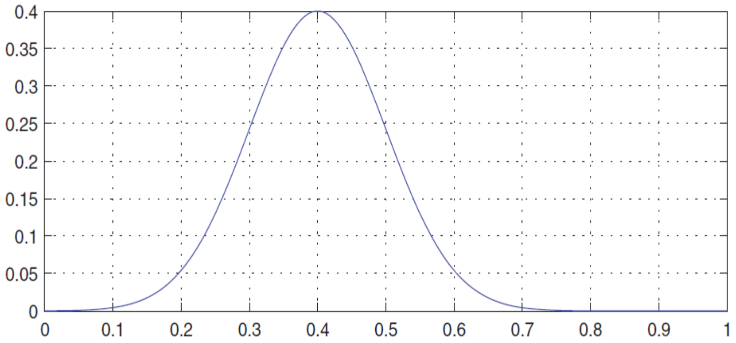
The Gaussian function with multiple peaks

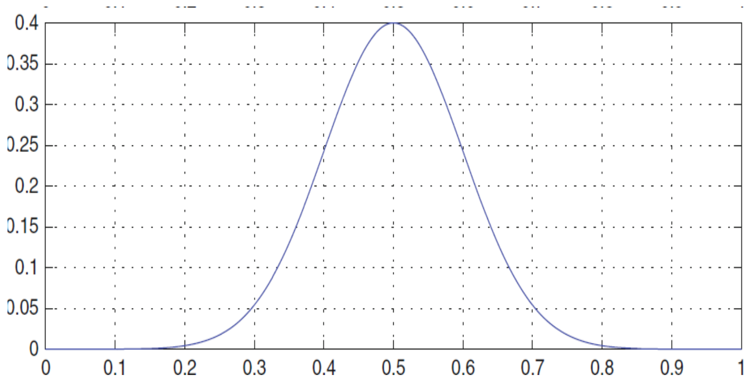
- Suppose that we have the attribute body-weight, which includes body weights of children , men and women. Then the pdf (gaussian function) will observe three peaks: one for the kids, one for men and the other for the women similar to one shown below

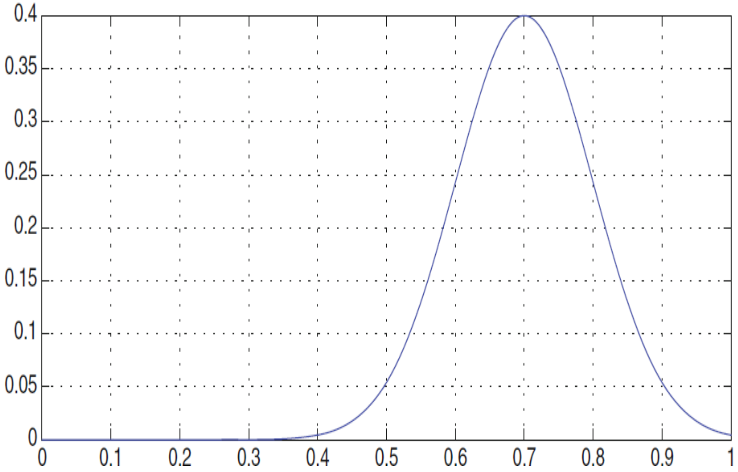


The Gaussian function with multiple peaks

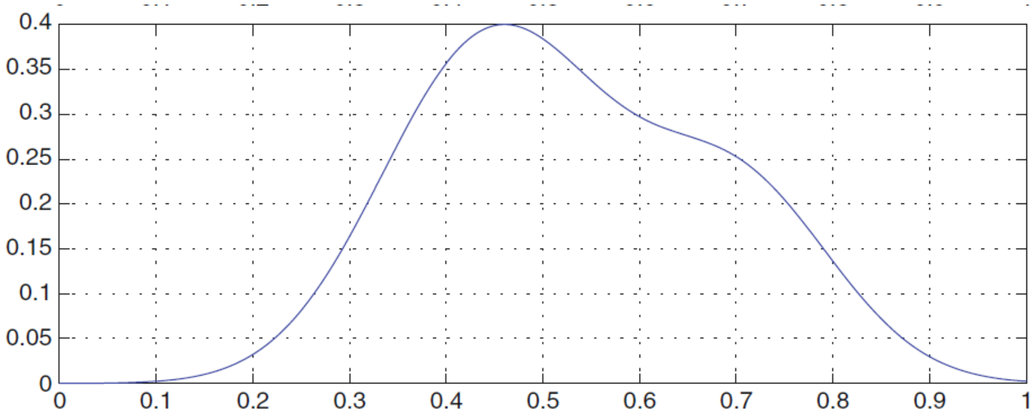
- In this case, we might create a separate gaussian for each source, and then combine by superimposing the bell functions on each other.







Three bell functions are combined into a single bell function



The Gaussian function with multiple peaks

- Suppose that there are m bell functions, then the pdf of the combined m bell functions ids given by :

$$p(x) = k \cdot \sum_{i=1}^{m} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}$$

where

$$k = \frac{1}{m\sigma\sqrt{2\pi}}$$

And $\sigma^2 = 1$ (when i^{th} attribute value is μ_i on its own)

Bayes Formula for Continuous Attributes

- We only replace the conditional probability $P(x|c_i)$ with $p_{c_i}(x)$ and $P(x)$ with $p(x)$.

$$P(c_i \mid x) = \frac{p_{c_i}(x) \cdot P(c_i)}{p(x)}$$

- $P(c_i)$ is estimated by the relative frequency of c_i in the training set,
- $p(x)$ is the **pdf created from all training examples**
- $p_{c_i}(x)$ the **pdf** created from those training examples that belong to $\boldsymbol{c_i}$**.**

Bayes Formula for Continuous Attributes

- **Naive Bayes makes** the assumption that all attributes are mutually independent.
- Suppose we encounter an example described by

$$x = (x_1, x_2, \ldots, x_n,) \quad \text{(mutually independent attributes)}$$

Then

$$p_{c_j}(\mathbf{x}) = \prod_{i=1}^{n} p_{c_j}(x_i)$$

Bayes Formula for Continuous Attributes

Example :Suppose we are given a training set that consists of the following six examples, $ex_1, \ldots, ex_6$, each described by three continuous attributes, at_1, at_2, at_3. Using Naive Bayes find the most probable class of $x = (9, 2.6, 3.3)$.

Example	at_1	at_2	at_3	Class
ex_1	3.2	2.1	2.1	pos
ex_2	5.2	6.1	7.5	pos
ex_3	8.5	1.3	0.5	pos
ex_4	2.3	5.4	2.45	neg
ex_5	6.2	3.1	4.4	neg
ex_6	1.3	6.0	3.35	neg

Bayes Formula for Continuous Attributes

Solution : Given $x = (9, 2.6, 3.3)$, our task is to find

$P(c_i | x) = \prod_{i=1}^{n} p_{c_i}(x)\ P(c_i)$ for each class c_i and select the maximum.

$\prod_{i=1}^{3} p_{pos}(x) = p_{pos}(at_1).p_{pos}(at_2).p_{pos}(at_3)$

$\prod_{i=1}^{3} p_{neg}(x) = p_{neg}(at_1).p_{neg}(at_2).p_{neg}(at_3)$

$P(pos) = P(neg) = \frac{3}{6} = \frac{1}{2}$ (so we can ignore this)

Bayes Formula for Continuous Attributes

The terms on the right-hand sides is calculated as follows. Here $m = 3$ and $\sigma^2 = 1$.

$$p_{pos}(at_1 = 9) = \frac{1}{3\sqrt{2\pi}}\left[e^{-0.5(9-3.2)^2} + e^{-0.5(9-5.2)^2} + e^{-0.5(9-8.5)^2}\right]$$

$$= 0.0561$$

$$p_{pos}(at_2 = 2.6) = \frac{1}{3\sqrt{2\pi}}\left[e^{-0.5(2.6-2.1)^2} + e^{-0.5(2.6-6.1)^2} + e^{-0.5(2.6-1.3)^2}\right]$$

$$= 0\,.\,0835$$

$$p_{pos}(at_2 = 3.3) = \frac{1}{3\sqrt{2\pi}}\left[e^{-0.5(3.3-2.1)^2} + e^{-0.5(3.3-7.5)^2} + e^{-0.5(3.3-0.5)^2}\right]$$

$$= 0.0322$$

Bayes Formula for Continuous Attributes

The terms on the right-hand sides is calculated as follows. Here $m = 3$ and $\sigma^2 = 1$.

$$p_{neg}(at_1 = 9) = \frac{1}{3\sqrt{2\pi}}\left[e^{-0.5(9-2.3)^2} + e^{-0.5(9-6.2)^2} + e^{-0.5(9-1.3)^2}\right]$$

$$= 0.0023$$

$$p_{neg}(at_2 = 2.6) = \frac{1}{3\sqrt{2\pi}}\left[e^{-0.5(2.6-5.4)^2} + e^{-0.5(2.6-3.1)^2} + e^{-0.5(2.6-6.0)^2}\right]$$

$$= 0.0575$$

$$p_{neg}(at_2 = 3.3)$$

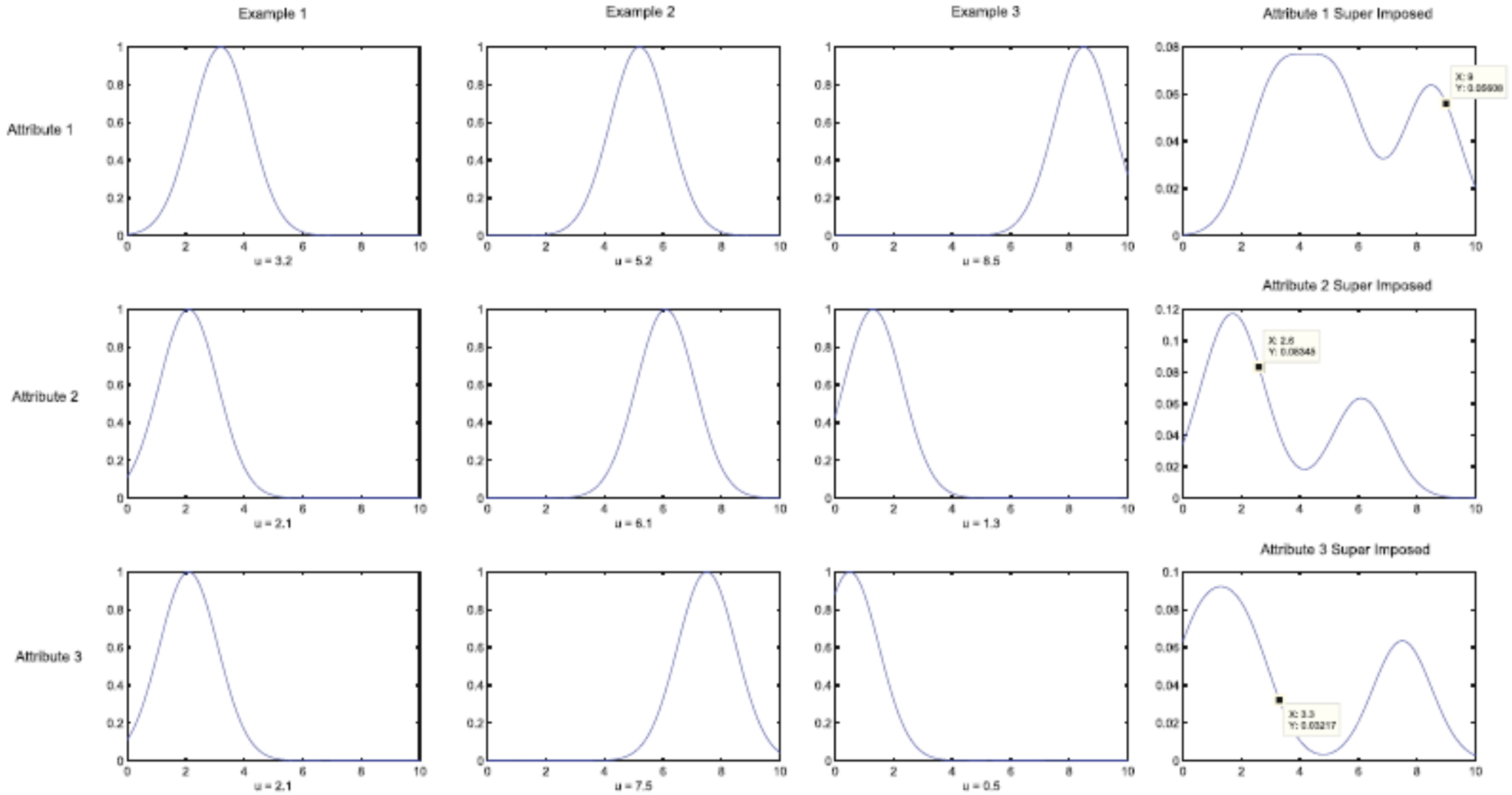
$$= \frac{1}{3\sqrt{2\pi}}\left[e^{-0.5(3.3-2.45)^2} + e^{-0.5(3.3-4.4)^2} + e^{-0.5(3.3-3.35)^2}\right] = 0.1423$$

Bayes Formula for Continuous Attributes

$$p_{\mathrm{pos}}(\mathbf{x}) = 0.0561 \times 0.0835 \times 0.0322 = 0.00015$$
$$p_{\mathrm{neg}}(\mathbf{x}) = 0.0023 \times 0.0575 \times 0.1423 = 0.00001$$

Observing that $p_{pos}(\mathbf{x}) > p_{neg}(\mathbf{x})$, we label $\mathbf{x}$ with the class $\mathtt{pos}$.

Bayes Formula for Continuous Attributes



Bayes Formula for Continuous Attributes