Naïve Bayes Classifier

CI603 - Data Mining

Classification

- **Classification** is a common task in everyday life.
- Essentially it involves **dividing up** objects so that **each is assigned to one** of a number of **mutually exhaustive and exclusive categories** known as **classes**.
- The term '**mutually exhaustive** and **exclusive**' simply means that each object must be assigned to **precisely one class**.

Classification

- Many practical **decision-making tasks** can be **formulated as classification problems**, i.e. assigning people or objects to one of a number of categories, for example:
    - Customers who are *likely to buy or not buy* a particular product in a supermarket.
    - People who are at *high, medium or low risk* of acquiring a certain illness.
    - Objects on a radar display which correspond to *vehicles, people, buildings or trees*.

Classification

- Houses that are *likely to rise in value, fall in value or have an unchanged value* in 12 months' time.
- People who are at *high, medium or low risk* of a car accident in the next 12 months.
- People who are *likely to vote for each of a number* of political parties (or none).
- The *likelihood of rain* the next day for a weather forecast (*very likely, likely, unlikely, very unlikely*).

Applications Naïve Bayes

Spam Classification

Given an email, predict whether it is spam or not

Medical Diagnosis

Given a list of symptoms, predict whether a patient disease or not.

Bayesian classifiers

- **Bayesian classifiers** use the branch of Mathematics known as **probability theory** to find the most likely of the possible classifications.

- The **probability of an event** is a **number** from **0 to 1** inclusive, with **0** indicating ‘**impossible**’ and **1** indicating ‘**certain**’.

- A **probability of 0.7** implies that if we conducted a long series of trials, we would expect that the event occurs 70% of the time.

Bayesian classifiers

- Usually we are **not interested in just one event** but in a **set of alternative possible events**, which are **mutually exclusive and exhaustive**, meaning that one and only one must always occur.

- Consider the train example below, we might define four **mutually exclusive and exhaustive** events:
  - $E1$: train cancelled. $P(E1) = 0.05$
  - $E2$: train ten minutes or more late. $P(E2) = 0.1$
  - $E3$: train less than ten minutes late. $P(E3) = 0.15$
  - $E4$: train on time or early. $P(E4) = 0.7$

- They also satisfy a second important condition: the **sum of the probabilities** of a set of **mutually exclusive and exhaustive events** must always be **1**.

$$P(E1) + P(E2) + P(E3) + P(E4) = 1$$

Bayesian classifiers

- Generally we are not in a position to know the **true probability** of an event occurring.
- In practice this is often **prohibitively difficult** or **impossible to do**, especially (as in this example) the trials may potentially go on forever.
- Instead we keep **records for a sample** to estimate the four probabilities.
- The **outcome** of each trial is **recorded in one row** of a table. Each row must have **one and only one classification**.
- The **longer the series of trials** the **more reliable this estimate is likely to be**.

Bayesian classifiers

- The **training set** constitutes the **results of a sample of trials** that we can use to predict the classification of other (unclassified) instances.

  i.e.:

weekday	winter	high	heavy	????

$$P(class = on\ time) = \frac{14}{20} = 0.7$$

- We could expect to be right about 70% of the time.

day	season	wind	rain	class
weekday	spring	none	none	on time
weekday	winter	none	slight	on time
weekday	winter	none	slight	on time
weekday	winter	high	heavy	late
saturday	summer	normal	none	on time
weekday	autumn	normal	none	very late
holiday	summer	high	slight	on time
sunday	summer	normal	none	on time
weekday	winter	high	heavy	very late
weekday	summer	none	slight	on time
saturday	spring	high	heavy	cancelled
weekday	summer	high	slight	on time
saturday	winter	normal	none	late
weekday	summer	high	none	on time
weekday	winter	normal	heavy	very late
saturday	autumn	high	slight	on time
weekday	autumn	none	heavy	on time
holiday	spring	normal	slight	on time
weekday	spring	normal	none	on time
weekday	spring	normal	slight	on time

Bayesian classifiers

- What is the **probability** of the train being **on time** if we know that the season is **winter**?

day	season	wind	rain	class
weekday	spring	none	none	on time
weekday	winter	none	slight	on time
weekday	winter	none	slight	on time
weekday	winter	high	heavy	late
saturday	summer	normal	none	on time
weekday	autumn	normal	none	very late
holiday	summer	high	slight	on time
sunday	summer	normal	none	on time
weekday	winter	high	heavy	very late
weekday	summer	none	slight	on time
saturday	spring	high	heavy	cancelled
weekday	summer	high	slight	on time
saturday	winter	normal	none	late
weekday	summer	high	none	on time
weekday	winter	normal	heavy	very late
saturday	autumn	high	slight	on time
weekday	autumn	none	heavy	on time
holiday	spring	normal	slight	on time
weekday	spring	normal	none	on time
weekday	spring	normal	slight	on time

Bayesian classifiers

- What is the **probability** of the train being **on time** if we know that the season is **winter**?

$$P(class = on\ time \mid season = winter) = \frac{2}{6} = 0.33$$

- This is considerably **less than the prior probability of 0.7** and seems **intuitively reasonable**.

- The **probability of an event** occurring if we know that an **attribute** has a **particular value** (or that several variables have particular values) is called the **conditional probability** of the event occurring:

$$P(class = on\ time \mid season = winter)$$ (**posterior probability**)

Bayes’ Theorem Formula

- **Prior Probability**, in Bayesian statistical inference, is the probability of an event **before new data is collected**.
- This is the **best rational assessment of the probability of an outcome** based on the **current knowledge** before an experiment is performed.
- The prior probability of an event will be **revised** as **new data or information becomes available**, to produce a **more accurate measure** of a potential outcome (**posterior probability**).

Bayes Theorem

Naïve bayes is a probabilistic classifier. This can be applied to data mining tasks

The basis is the Bayes Theorem

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

where $P(A)$ is the prior probability of A occurring, $P(A|B)$ is the conditional probability of A given that B occurs, $P(B|A)$ is the conditional probability of B given that A occurs, $P(B)$ is the prior probability of B occurring

Naïve Bayes Classifier

- Data set contains a set of objects with features $\{x_1, x_2,\ldots, x_n\}$.
- $\{x_1, x_2,\ldots, x_n\}$ is a data record with dependent value $\{y\}$ which is the output or class.
- Example might be such that the features are: age, A-level grades, prior experience of prospective students and **y** is the classification of whether accepted to Medical School or not.

$$P(y|x_1, x_2 \ldots x_n) = \frac{P(x_1|y)P(x_2|y)..P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

$$= \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

Naïve Bayes algorithm

$$y = argmax_{yj}\, P(y_j) \prod_{i=1}^{n} P(x_i | y_j)$$

Bayesian classifiers

- The **prior probability** is estimated **before any other information is available**.

- By contrast, **posterior probability** is the **probability** that we can calculate for the **classification** after we have obtained the information that the **season** is **winter**.

- Therefore, to calculate the most likely classification for the ‘unseen’ instance we could calculate the probability of:

weekday	winter	high	heavy	????

$$P(class = on\ time \,|\, day = weekday \textbf{ and } season = winter \textbf{ and } wind = high \textbf{ and } rain = heavy)$$

Bayesian classifiers

- There are only **two instances** in the training set with that combination of attribute values, and **basing any estimates of probability on these** is **unlikely to be helpful**.

- To obtain a reliable estimate of the four classifications a more indirect approach is needed.

day	season	wind	rain	class
weekday	spring	none	none	on time
weekday	winter	none	slight	on time
weekday	winter	none	slight	on time
weekday	winter	high	heavy	late
saturday	summer	normal	none	on time
weekday	autumn	normal	none	very late
holiday	summer	high	slight	on time
sunday	summer	normal	none	on time
weekday	winter	high	heavy	very late
weekday	summer	none	slight	on time
saturday	spring	high	heavy	cancelled
weekday	summer	high	slight	on time
saturday	winter	normal	none	late
weekday	summer	high	none	on time
weekday	winter	normal	heavy	very late
saturday	autumn	high	slight	on time
weekday	autumn	none	heavy	on time
holiday	spring	normal	slight	on time
weekday	spring	normal	none	on time
weekday	spring	normal	slight	on time

Bayesian classifiers

- We could start by using **conditional probabilities** based on a **single attribute.**

$$P(class = on\ time \mid rain = heavy) = 1/5 = 0.2$$
$$P(class = late \mid rain = heavy) = 1/5 = 0.2$$
$$P(class = very\ late \mid rain = heavy) = 2/5 = 0.4$$
$$P(class = cancelled \mid rain = heavy) = 1/5 = 0.2$$

$$P(class = on\ time \mid wind = high) = 4/7 = 0.58$$
$$P(class = late \mid wind = high) = 1/7 = 0.14$$
$$P(class = very\ late \mid wind = high) = 1/7 = 0.14$$
$$P(class = cancelled \mid wind = high) = 1/7 = 0.14$$

$$P(class = on\ time \mid day = weekday) = 9/13 = 0.7$$
$$P(class = late \mid day = weekday) = 1/13 = 0.1$$
$$P(class = very\ late \mid day = weekday) = 3/13 = 0.2$$
$$P(class = cancelled \mid day = weekday) = 0/13 = 0$$

$$P(class = on\ time \mid season = winter) = 2/6 = 0.33$$
$$P(class = late \mid season = winter) = 2/6 = 0.33$$
$$P(class = very\ late \mid season = winter) = 2/6 = 0.33$$
$$P(class = cancelled \mid season = winter) = 0/6 = 0$$

Bayesian classifiers

- The **Naïve Bayes algorithm** gives us a way of combining the **prior probability** and **conditional probabilities** in a single formula, which can be used to calculate the **probability of each of the possible classifications** in turn.

- The term **Naïve** refers to the assumption that the method makes:
  - the **effect of the value of one attribute** on the probability of a given classification is **independent** of the **values of the other attributes** (in practice, that may not be the case.)

- Despite this theoretical weakness, the **Naïve Bayes method** often gives **good results** in practical use.

Bayesian classifiers

- The method uses **conditional probabilities**, but the **other way round** from before.
- Instead of the probability that the **class** is **very late** given that the **season** is **winter**:

$$P(class = \text{very late} \mid season = winter),$$

- we use the conditional probability that the **season** is **winter** given that the **class** is **very late**:

$$P(season = winter \mid class = \text{very late}).$$

Bayesian classifiers

$$P(season = winter | class = very\ late) = \frac{2}{2} = 1$$

- The number of times that **season = winter** and **class = very late** occur in the same instance, divided by the number of instances for which the **class** is **very late**.

day	season	wind	rain	class
weekday	spring	none	none	on time
weekday	winter	none	slight	on time
weekday	winter	none	slight	on time
weekday	winter	high	heavy	late
saturday	summer	normal	none	on time
weekday	autumn	normal	none	very late
holiday	summer	high	slight	on time
sunday	summer	normal	none	on time
weekday	winter	high	heavy	very late
weekday	summer	none	slight	on time
saturday	spring	high	heavy	cancelled
weekday	summer	high	slight	on time
saturday	winter	normal	none	late
weekday	summer	high	none	on time
weekday	winter	normal	heavy	very late
saturday	autumn	high	slight	on time
weekday	autumn	none	heavy	on time
holiday	spring	normal	slight	on time
weekday	spring	normal	none	on time
weekday	spring	normal	slight	on time

Naïve Bayesian Classification

- Given a set of k **mutually exclusive and exhaustive** classifications

$$c_1, c_2, \ldots, c_k,$$

which have **prior probabilities**

$P(c_1), P(c_2), \ldots, P(c_k)$, respectively, and attributes $a_1, a_2, \ldots, a_n$ which for a given instance have values $v_1, v_2, \ldots, v_n$ respectively,

the **posterior probability** of class c_i occurring for the specified instance can be shown to be proportional to

$$P(c_i) \times P(a_1 = v_1 \ and \ a_2 = v_2 \ldots \ and \ a_n = v_n | c_i)$$

Naïve Bayesian Classification

- Making the assumption that the **attributes** are **independent**, the value of this expression can be calculated using the product

$$P(c_i) \times P(a_1 = v_1 | c_i) \times P(a_2 = v_2 | c_i) \times \ldots \times P(a_n = v_n | c_i)$$

- It is often written as:

$$P(c_i) \times \Pi_{j=1}^{n} P(a_j = v_j | class = c_i)$$

- We calculate this product for each value of from to k and **choose the classification** that has the **largest value**.

Bayesian classifiers

	class = on time	class = late	class = very late	class = cancelled
day = weekday	**9/14 = 0.64**	**1/2 = 0.5**	**3/3 = 1**	**0/1 = 0**
day = saturday	2/14 = 0.14	1/2 = 0.5	0/3 = 0	1/1 = 1
day = sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
day = holiday	2/14 = 0.14	0/2 = 0	0/3 = 0	0/1 = 0
season = spring	4/14 = 0.29	0/2 = 0	0/3 = 0	1/1 = 1
season = summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
season = autumn	2/14 = 0.14	0/2 = 0	1/3 = 0.33	0/1 = 0
season = winter	**2/14 = 0.14**	**2/2 = 1**	**2/3 = 0.67**	**0/1 = 0**
wind = none	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
wind = high	**4/14 = 0.29**	**1/2 = 0.5**	**1/3 = 0.33**	**1/1 = 1**
wind = normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
rain = none	5/14 = 0.36	1/2 =0.5	1/3 = 0.33	0/1 = 0
rain = slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
rain = heavy	**1/14 = 0.07**	**1/2 = 0.5**	**2/3 = 0.67**	**1/1 = 1**
Prior Probability	**14/20 = 0.70**	**2/20 = 0.10**	**3/20 = 0.15**	**1/20 = 0.05**

day	season	wind	rain	class
weekday	spring	none	none	on time
weekday	winter	none	slight	on time
weekday	winter	none	slight	on time
weekday	winter	high	heavy	late
saturday	summer	normal	none	on time
weekday	autumn	normal	none	very late
holiday	summer	high	slight	on time
sunday	summer	normal	none	on time
weekday	winter	high	heavy	very late
weekday	summer	none	slight	on time
saturday	spring	high	heavy	cancelled
weekday	summer	high	slight	on time
saturday	winter	normal	none	late
weekday	summer	high	none	on time
weekday	winter	normal	heavy	very late
saturday	autumn	high	slight	on time
weekday	autumn	none	heavy	on time
holiday	spring	normal	slight	on time
weekday	spring	normal	none	on time
weekday	spring	normal	slight	on time

Bayesian classifiers

weekday	winter	high	heavy	????

<u>class = on time</u>

0.70 × 0.64 × 0.14 × 0.29 × 0.07 = 0.0013

<u>class = late</u>

0.10 × 0.50 × 1.00 × 0.50 × 0.50 = 0.0125

<u>**class = very late**</u>

0.15 × 1.00 × 0.67 × 0.33 × 0.67 = **0.0222**

<u>class = cancelled</u>

0.05 × 0.00 × 0.00 × 1.00 × 1.00 = 0.0000

The largest value is for **class = very late**.

	class = on time	class = late	class = very late	class = cancelled
day = weekday	**9/14 = 0.64**	**1/2 = 0.5**	**3/3 = 1**	**0/1 = 0**
day = saturday	2/14 = 0.14	1/2 = 0.5	0/3 = 0	1/1 = 1
day = sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
day = holiday	2/14 = 0.14	0/2 = 0	0/3 = 0	0/1 = 0
season = spring	4/14 = 0.29	0/2 = 0	0/3 = 0	1/1 = 1
season = summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
season = autumn	2/14 = 0.14	0/2 = 0	1/3 = 0.33	0/1 = 0
season = winter	**2/14 = 0.14**	**2/2 = 1**	**2/3 = 0.67**	**0/1 = 0**
wind = none	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
wind = high	**4/14 = 0.29**	**1/2 = 0.5**	**1/3 = 0.33**	**1/1 = 1**
wind = normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
rain = none	5/14 = 0.36	1/2 =0.5	1/3 = 0.33	0/1 = 0
rain = slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
rain = heavy	**1/14 = 0.07**	**1/2 = 0.5**	**2/3 = 0.67**	**1/1 = 1**
Prior Probability	**14/20 = 0.70**	**2/20 = 0.10**	**3/20 = 0.15**	**1/20 = 0.05**

Naïve Bayesian Classification

class = on time

$0.70 \times 0.64 \times 0.14 \times 0.29 \times 0.07 = 0.0013$

class = late

$0.10 \times 0.50 \times 1.00 \times 0.50 \times 0.50 = 0.0125$

class = very late

$0.15 \times 1.00 \times 0.67 \times 0.33 \times 0.67 = 0.0222$

class = cancelled

$0.05 \times 0.00 \times 0.00 \times 1.00 \times 1.00 = 0.0000$

- Note that the four values calculated they do **not sum to 1**.
- Each value can be **normalised** to a valid posterior probability.
- In practice, we are interested only in finding the largest value so the normalisation step is not necessary.

Naïve Bayesian Classification

- The Naïve Bayes approach is a very popular one, which often works well.
- However it has a number of potential problems, the most obvious one being that it relies on **all attributes** being **categorical**.
- A second problem is that **estimating probabilities by relative frequencies** can give a **poor estimate** if the **number of instances** with a **given attribute/ value combination** is **small**.
    - In the extreme case where it is **zero**, the **posterior probability** will inevitably be calculated as **zero**.
        - i.e. This happened for **class = cancelled** in the example.

Estimating Conditional Probabilities for Categorical Attributes

- For a categorical attribute X_i, the **conditional probability** $P(X_i = c \mid y)$ is estimated according to the fraction of training instances in class where X_i takes on a particular categorical value .

$$P(X_i = c \mid y) = \frac{n_c}{n}$$

Where n is the number of training instances belonging to class ,
out of which n_c number of instances have $X_i = c$.

Estimating Conditional Probabilities for Categorical Attributes

- i.e., in the training set:
  - Seven people have the class label **Defaulted Borrower=No**, out of which three people have **Home Owner=Yes** while the remaining four have **Home Owner=No**.

$$P(Home\ Owner = Yes \,|\, Default = No) = \frac{3}{7}$$

$$P(Marital\ Status = Single \,|\, Default = Yes) = \frac{2}{3}$$

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- The sum of conditional probabilities over all possible outcomes of X_i is equal to one.

Estimating Conditional Probabilities for Continuous Attributes

- There are **two ways** to estimate the **class-conditional probabilities** for **continuous attributes**:

  1. We can **discretise each continuous attribute** and then **replace the continuous values** with their **corresponding discrete intervals**.

     The **estimation error** of this method depends on the **discretisation strategy**, as well as the **number of discrete intervals**.

     - If the **number of intervals** is **too large**, every interval may have an **insufficient number of training instances** to provide a **reliable estimate** of $P(X_i | Y)$.

     - If the **number of intervals** is **too small**, then the discretisation process may **loose information about the true distribution** of continuous values, and thus result in **poor predictions**.

Estimating Conditional Probabilities for Continuous Attributes

2. We can assume a certain form of **probability distribution** for the **continuous variable** and **estimate** the **parameters of the distribution** using the training data.

    - i.e., **Gaussian distribution** to represent the conditional probability of continuous attributes.

    - The Gaussian distribution is characterised by two parameters, the **mean**, μ, and the **variance**, σ^2. For each class y_i, the class-conditional probability for attribute X_i is:

Naïve Bayes on Example Data

Given a Test Record: X=(Refund=No, Marital Status=Divorced, Income = 120K)

Tid	Refund	Marital Status	Taxable income	Defaulted
1	yes	single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$\prod P(X_i \mid Yes) =$
$P(Refund = No \mid Yes) \times P(Divorced \mid Yes) \times P(Income = 120K \mid Yes)$

$\prod P(X_i \mid No) =$
$P(Refund = No \mid No) \times P(Divorced \mid No) \times P(Income = 120K \mid No)$

Estimate Probabilities from Data

Tid	Refund	Marital Status	Taxable income	Defaulted
1	yes	single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

P(y) = fraction of instances of class y

–e.g., P(No) = 7/10,
P(Yes) = 3/10

For categorical attributes:

$$P(X_i = c | y) = n_c / n$$

–where $|X_i = c|$ is number of instances having attribute value $X_i = c$ and belonging to class y

–Examples:

P(Status=Married|No) = 4/7

P(Refund=Yes|Yes)=0

Estimate Probabilities from Data

For continuous attributes:

- Discretization: Partition the range into bins:

  Replace continuous value with bin value

    Attribute changed from continuous to ordinal

- Probability density estimation:

  Assume attribute follows a normal distribution

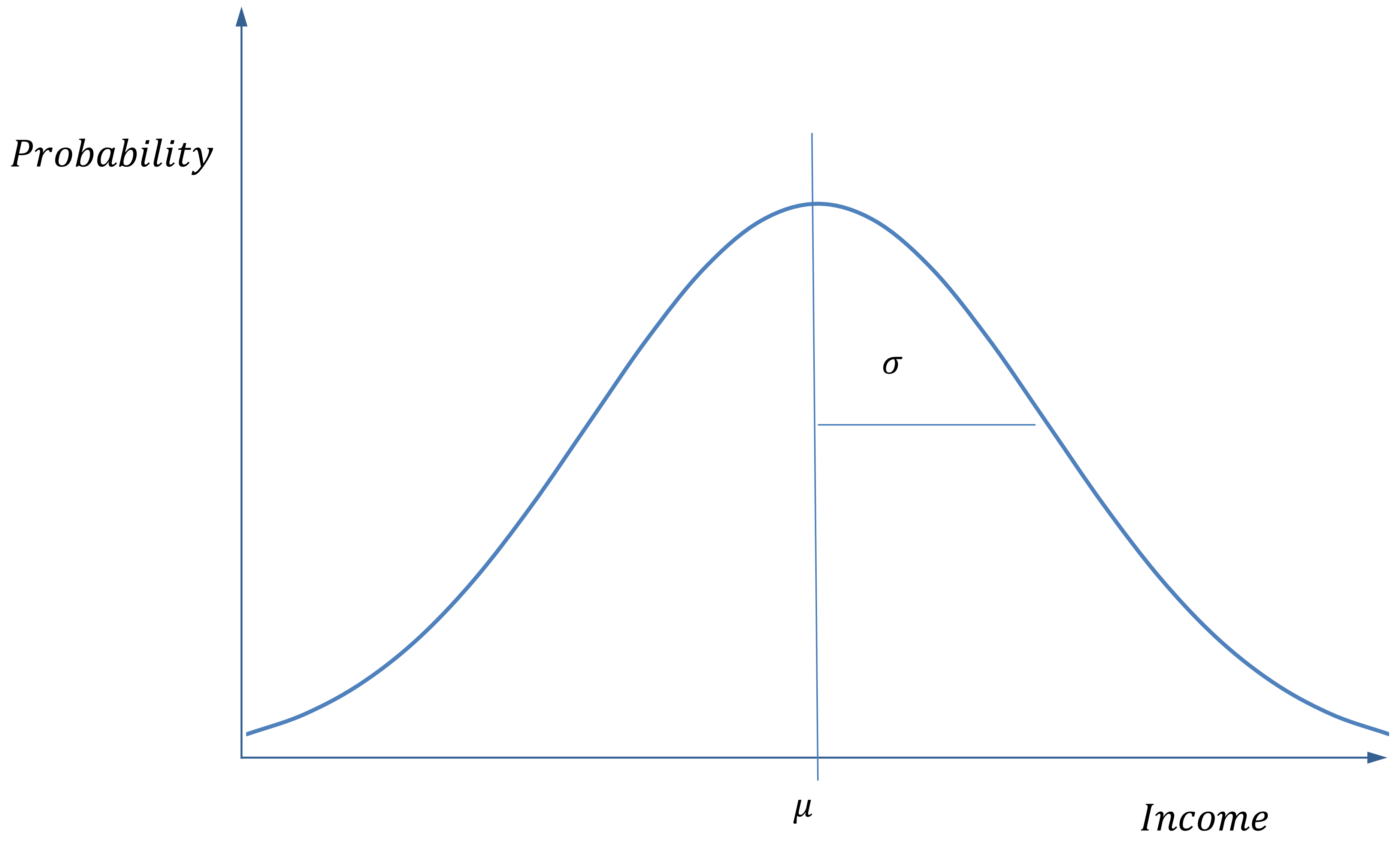
  Use data to estimate parameters of distribution

   (e.g., mean and standard deviation)

  Once probability distribution is known, use it to estimate the conditional probability $P(X_i|Y)$

Normal Distribution of the income values

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$



Estimate Probabilities from Data

Tid	Refund	Marital Status	Taxable income	Defaulted
1	yes	single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Normal distribution:

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

– One for each (X_i, Y_i) pair

For (Income, Class=No):

If Class=No

sample mean = 110

sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example working out normal distribution for class No

Tid	Refund	Marital Status	Taxable income	Defaulted
1	yes	single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Calculating the sample mean and sample variance

$$sample\ mean(\mu) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample mean $= \frac{(125+100+70+120+60+220+75)}{7} = 110$

$$sample\ variance(\sigma^2) = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}$$

Sample variance $= \frac{(125-110)^2+(100-110)^2+(70-110)^2+(120-110)^2+(60-110)^2+(220-110)^2+(75-110)^2}{7-1}$

$=2975$

Example working out normal distribution for class yes

Tid	Refund	Marital Status	Taxable income	Defaulted
1	yes	single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Calculating the sample mean and sample variance

$$sample\ mean(\mu) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample mean $=\frac{(95+85+90)}{3}$=90

$$sample\ variance(\sigma^2) = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$

Sample variance $= \frac{(95-90)^2+(85-90)^2+(90-90)^2}{3-1}$

=25

Estimating the probabilities using the normal distribution (no)

We use the sample mean and sample variance in the formula for the Normal Distribution

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

where μ = 110, σ^2 = 2975

$$P(Income = 120|No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Estimating the probabilities using the normal distribution (yes)

We use the sample mean and sample variance in the formula for the Normal Distribution

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

where μ = 90, σ^2 = 25

$$P(Income = 120 | Yes) = \frac{1}{\sqrt{2\pi}(5)} e^{-\frac{(120-90)^2}{2(25)}} = 1.2 \times 10^{-9}$$

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable income	Defaulted
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7

P(Refund = Yes | Yes) = 0/3 = 0
P(Refund = No | Yes) = 3/3 = 1

P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7

P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3= 0

Example of Naïve Bayes Classifier

Given a Test Record: $X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$

Naïve Bayes Classifier

$$y = argmaxyj\, P(y_j) \prod_{i=1}^{n} P(x_i | y_j)$$

P(No) x ∏ P(X_i | No) = P(No) x P(Refund=No | No) × P(Divorced | No) × P(Income=120K | No)

= 7/10 x 4/7 × 1/7 × 0.0072 = 0.000411

P(Yes) x ∏ P(X_i | Yes) = P(Yes) x P(Refund=No | Yes) × P(Divorced | Yes) × P(Income=120K | Yes)

= 3/10 x 1 × 1/3 × 1.2 × 10^{-9} = 1.2 × 10^{-10}

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X) $\Rightarrow$ **Class = No**

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

Tid	Refund	Marital Status	Taxable income	Defaulted
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
P(Refund = Yes | Yes) = 0/3 = 0
P(Refund = No | Yes) = 3/3 = 1
P(Marital Status = Single | No) = 2/6
P(Marital Status = Divorced | No) = 0/6 =0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3 =0
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = Yes: sample mean = 90
sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

P(X | No) = 2/6 x 0 x 0.0083 = 0

P(X | Yes) = 0 x 1/3 x 1.2 x 10^{-9} = 0

Naïve Bayes will not be able to classify X as Yes or No!

Handling Zero Conditional Probabilities

- If the **conditional probability** for **any of the attributes** is **zero**, then the **entire expression** for the class-conditional probability becomes **zero**.
- Zero conditional probabilities arise when the **number of training instances** is **small** and the **number of possible values** of an attribute is **large**.
- In such cases, it may happen that a **combination of attribute values** and **class labels** are never observed, resulting in a **zero conditional probability**.

Handling Zero Conditional Probabilities

- In a more extreme case, if the **training instances** do **not cover some combinations** of attribute values and class labels, then we may **not be able to even classify** some of the test instances.
    - If $P(Marital\ Status = Divorced \mid No) = 0$ instead of **1/7**, then a data instance with attribute set:
        - $x = (HomeOwner = Yes, MaritalStatus = Divorced, Income = 120)$ has the following class-conditional probabilities:

$$P(x \mid No) = 3/7 \times 0 \times 0.0072 = 0.$$

$$P(x \mid Yes) = 0 \times 1/3 \times 1.2 \times 10 - 9 = 0.$$

Handling Zero Conditional Probabilities

- Since both the class-conditional probabilities are 0, the naïve Bayes classifier will not be able to classify the instance.

- To address this problem, it is important to **adjust the conditional probability estimates** so that they are not as brittle as simply using fractions of training instances.

$$P(X_i = c \mid y) = \frac{n_c}{n}$$

- This can be achieved by using alternate estimates of conditional probability.

Handling Zero Conditional Probabilities

Laplace estimate: $P(Xi = c \mid y) = \dfrac{n_c + 1}{n + v}$

n: number of training instances belonging to class y

n_c: number of instances with $X_i = c$ and $Y = y$

v: total number of attribute values that X_i can take

Handling Zero Conditional Probabilities

m-estimate: $P(Xi = c \mid y) = \dfrac{n_c + mp}{n + m}$

where is the number of training instances belonging to class ,

n_c is the number of training instances with $X_i = c$ and $Y = y$,

is some initial estimate of $P(Xi = c \mid y)$ that is known a priori, and

is a hyper-parameter that indicates our confidence in using when the fraction of training instances is too brittle.

- If $n_c = 0$, both **Laplace** and **m-estimate** provide **non-zero values** of conditional probabilities.

Example with Laplace Estimates

Consider the table with Tid = 7 deleted

Tid	Refund	Marital Status	Taxable income	Defaulted
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

P(Refund = Yes | No) = (2+1)/(6+2) = 3/8
P(Refund = No | No) = (4+1)/(6+2) =5/8
P(Refund = Yes | Yes) =(0+1)/(3+2) = 1/5
P(Refund = No | Yes) = (3+1)/(3+2) = 4/5
P(Marital Status = Single | No) = (2+1)/(6+3) =1/3
P(Marital Status = Divorced | No) = (0+1)/(6+3) =1/9
P(Marital Status = Married | No) = (4+1)/(6+3) = 5/9
P(Marital Status = Single | Yes) = (2+1)/(3+3) =1/2
P(Marital Status = Divorced | Yes) = (1+1)/(3+3) =1/3
P(Marital Status = Married | Yes) = (0+1)/(3+3) =1/6
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = Yes: sample mean = 90
sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

P(X | No) = 3/8 x 1/9 x 0.0083 = 0.00035

P(X | Yes) = 1/5 x 1/3 x 1.2 x 10^{-9} = 8×10^{-11}

Naïve Bayes is now able to classify X as No!

Naïve Bayes in Sklearn

There are three types of Naïve Bayes functions available in Sklearn

GaussianNB - This classifier is employed when the predictor values are continuous and are expected to follow a Gaussian distribution.

MultinomialNB - This classifier makes use of a multinomial distribution and is often used to solve issues involving document or text classification.

BernoulliNB - When the predictors are boolean in nature and are supposed to follow the Bernoulli distribution, this classifier is utilized.

Smoothing

- Smoothing makes sense only for BernoulliNB and MultinomialNB, which have categorical features, whereas GaussianNB works with numerical features which follow a normal distribution.

- For BernoulliNB and MultinomialNB, the features often represent frequencies and can be zero. In theory, this can cause one or several of the conditional probabilities to be zero, and therefore occasionally make the posterior probability zero. It can even cause the posterior to be zero for all the classes, an inconsistency. Smoothing prevents these issues.

- In GaussianNB the conditional probability is calculated from the normal distribution so they can never be zero, thus this problem cannot happen.