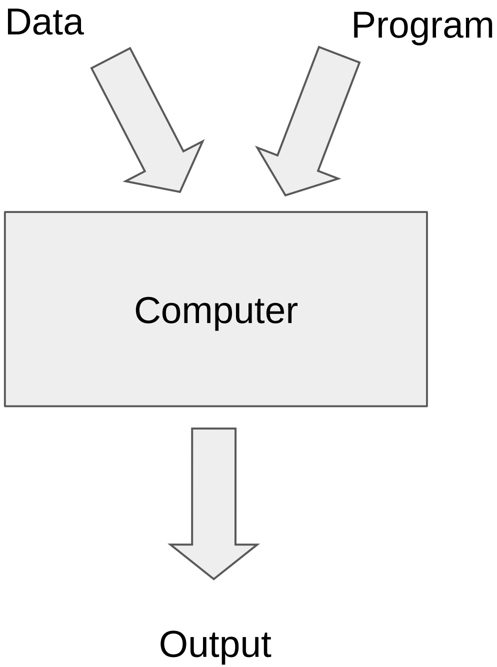
Introduction to Machine Learning

What is learning?

The ability to improve one’s behaviour with experience.

How machine learning is different from traditional programming?



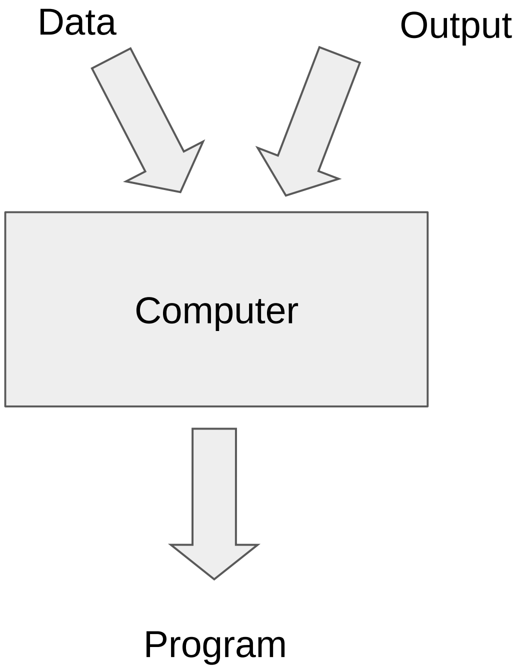
Data

Program

Computer

Output

Traditional Programming



Data

Output

Computer

Program

Machine Learning

Definitions

Arthur Samuel (1959):

Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed.

Tom Mitchell (1998):

Well-posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

What ML does?

Machine Learning explores algorithms that learn from data or build models from data that perform some tasks.

The tasks can be:

- Making predictions
- Classifications
- Clustering
- Decision Making
- Solving tasks/problems

Machine Learning Process

Machine Learning Process

1. Choose the training experience/data.
2. Choose the target function (that is to be learnt).
3. Choose the target class of the function.
4. Choose a learning algorithm to infer the target function.

Types of Machine Learning

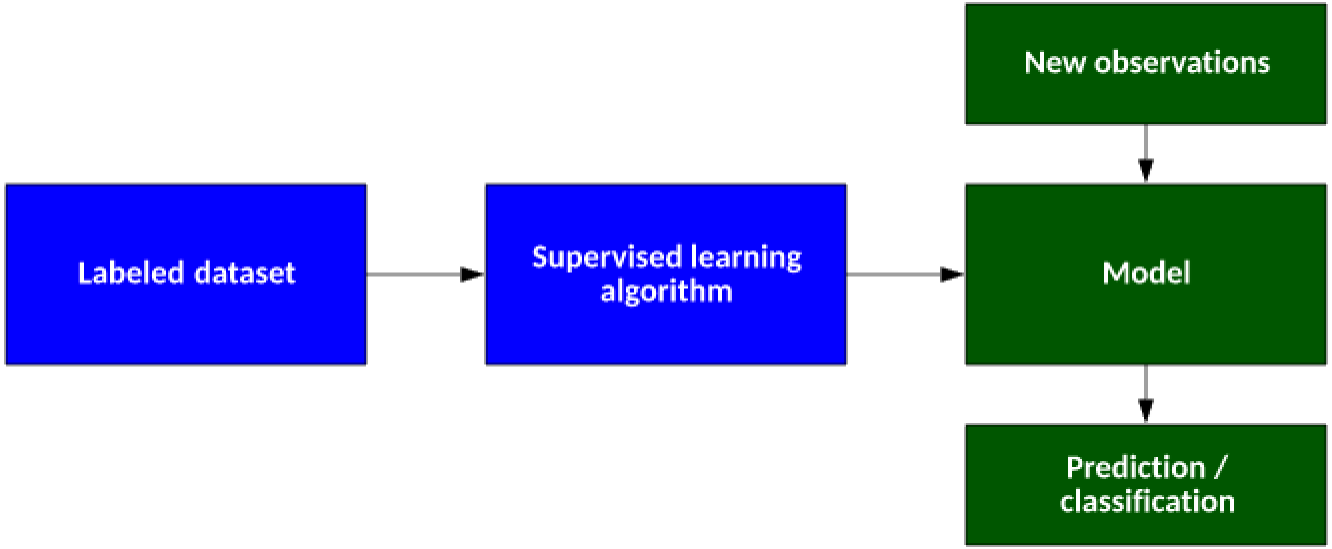
Type of Machine Learning

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

Supervised Learning

- Supervised learning involves learning from a training set of labelled data.
- Every point in the training is an input-output pair, where the input maps to an output.
- The learning problem consists of inferring the function that maps between the input and the output, such that the learned function can be used to predict the output from future input.
- It is called “supervised” because of the presence of the outcome variable to guide the learning process.
- Supervised learning problems are further categorised into **Regression** and **Classification** problems.

Supervised Learning



Labeled dataset

Supervised learning algorithm

New observations

Model

Prediction / classification

Applications of Supervised Learning

- Prediction
  - Stock prices
  - House prices
  - Weather Forecasting
  - Sales Volume Forecasting
- Classification
  - Disease identification
  - Sentiment Analysis
  - Spam Mail Detection
  - Handwritten Digit Recognition

Supervised Learning Algorithms

- Linear Regression
- Logistic Regression
- Naive Bayes Classifier
- Decision Trees
- Neural Networks
- Support Vector Machines
- K-nearest Neighbour

Unsupervised Learning

- It is a type of machine learning in which the algorithm is not provided with any pre-assigned labels or scores for the training data.
- Unsupervised learning algorithms must first self-discover any naturally occurring patterns in that training data set.
- Common examples include clustering, and principal component analysis,
- We observe only the features and have no measurements of the outcome.
- The task of learner is to describe how the data are organized or clustered.

Advantages and Disadvantages of Unsupervised Learning

- A minimal workload to prepare and audit the training set.
- Greater freedom to identify and exploit previously undetected patterns that may not have been noticed by the "experts".
- The cost of unsupervised techniques requiring a greater amount of training data and converging more slowly to acceptable performance.
- Increased computational and storage requirements during the exploratory process,
- Potentially greater susceptibility to artifacts or anomalies in the training data that might be obviously irrelevant or recognized as erroneous by a human, but are assigned undue importance by the unsupervised learning algorithm.

Applications of Unsupervised Learning

- Clustering
  - Customer Segmentation
  - Grouping products in a supermarket
- Visualization
- Dimensionality reduction
- Finding association rules
  - Customer that buy item X will buy item Y too.
- Anomaly detection
  - Fraudulent card transaction
  - Malware detection
  - Identification of human errors during data entry

Unsupervised Learning Algorithms

- K-Means Clustering
- Expectation Maximization
- Principal Component Analysis
- Hierarchical Clustering

Basic Terminology

- The inputs are often called the ***predictors*** or more classically the ***independent variables***.

  In the pattern recognition literature the term ***features*** is preferred.

- The outputs are called the ***responses***, or classically the ***dependent variables***.
- Such type of problems (learnings) is called inductive learning problems because we identify a function by inducting on data.

Types of Variables

- Quantitative Variables
  - Variables whose values exist on a continuous scale
  - Examples: temperature, salary, pressure, sales, price etc.
- Qualitative Variables
  - Variables that have values from a discrete set of values.
  - Also referred as categorical or discrete variables.
  - Example: Spam/Not-spam, Malignant/Benign etc.
- Ordered Categorical Variables
  - There is an ordering between the values, but no metric notion is appropriate (the difference between medium and small need not be the same as that between large and medium).
  - Example: Small, Medium and Large

Train/Test/Validation Set

Training Set: Set of examples/data that is used to train or build the model (find parameters).

Testing Set: Set of examples/data that is used to estimate the model’s performance i.e. how well the model fits the data

Validation Set: Set of data/examples used to tune the parameters of a classifier. It is not required if no fine-tuning of hyperparameters is required.

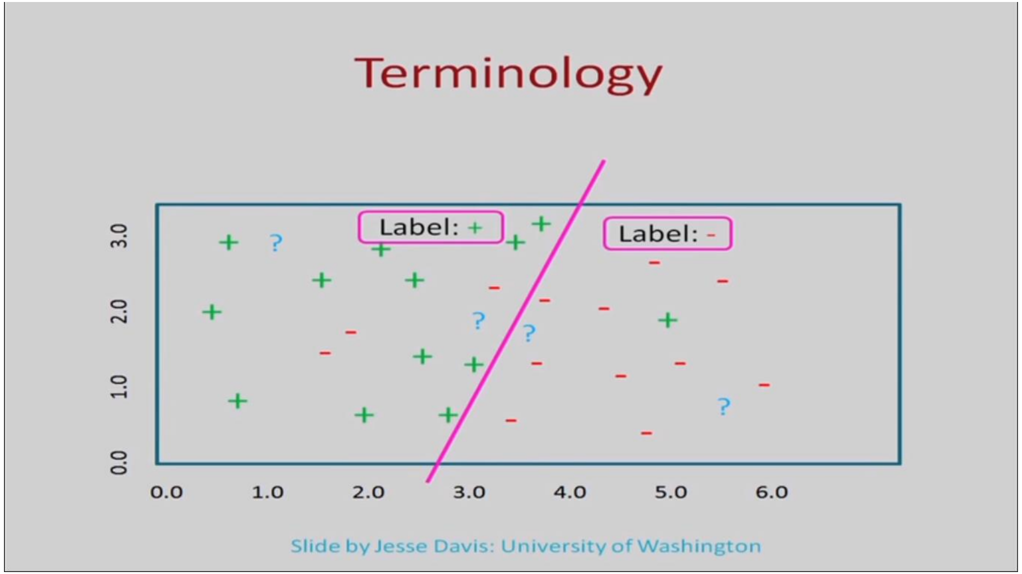
Using validation and test sets will increase the generalizing capability of the model on new unseen data.

Hypothesis Space and Inductive Bias

Hypothesis

- A hypothesis is a function that best describes the target in supervised machine learning.
- Represented by h_1, h_2, h_3 etc.

Machine learning involves finding a model (hypothesis) that best explains the training data.



Terminology

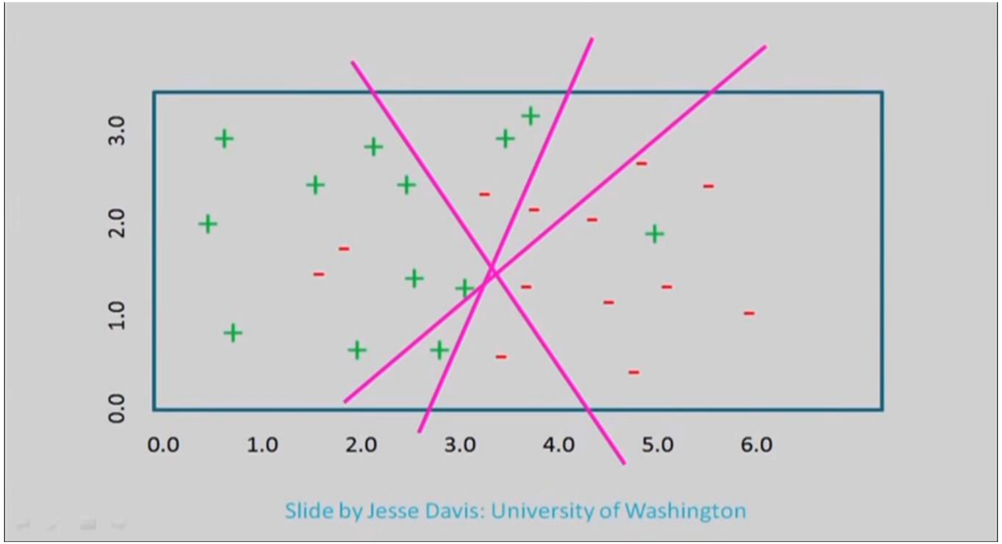
Slide by Jesse Davis: University of Washington

Hypothesis Space

Hypothesis space is a set of valid hypothesis, i.e. all possible functions.

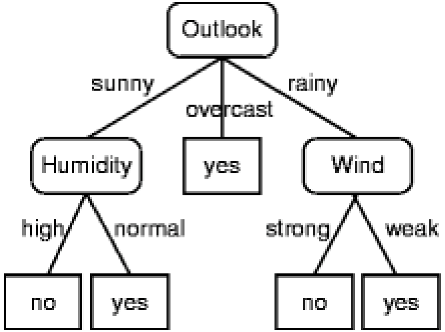
It is typically defined by a Hypothesis Language, possibly in conjunction with a Language Bias.

Represented by symbol H.



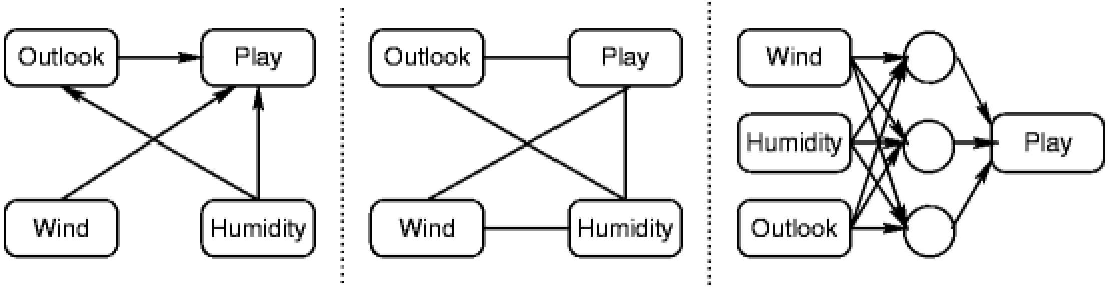
Slide by Jesse Davis: University of Washington

Hypothesis Language



IF Outlook=sunny AND Humidity=high THEN Play=no
IF Outlook=sunny AND Humidity=normal THEN Play=yes
IF Outlook=overcast THEN Play=yes
IF Outlook=rainy AND Wind=strong THEN Play=no
IF Outlook=rainy AND Wind=weak THEN Play=yes

Decision Tree and Rule Set



Bayesian Network, Markov Network, Neural Network

Example: Hypothesis Space

Let’s take the features which are boolean i.e. X1, X2, X3, X4 are 4 features which are boolean. Thus, X1 can take either 1(=T) or 0(=F). Similarly, X2,X3,X4 can take either a 0 or 1 as shown below.

X1	**X2**	**X3**	**X4**	**Output Class**
1	0	1	0	POSITIVE
0	0	0	1	NEGATIVE
1	1	1	1	POSITIVE

Thus, there are 2^4 = 16 possible instances.

How Many Boolean Functions Are Possible?

The number of functions is the number of possible subsets of the 16 instances. So, the possible number of subsets are (2^{16}).

This can be generalised to N boolean features. If there are N boolean features then the number of possible instances is 2^{n} and the number of possible functions will be $2^{(2^N)}$.

Thus, it can be inferred that the hypothesis space is gigantic as the number of features increases and it is not possible to look at every hypothesis individually in order to select the best hypothesis.

So, one puts restrictions in the Hypothesis Space to consider only specific Hypothesis Space. These restrictions are also referred as Bias.

Inductive Bias

- The inductive bias (also known as learning bias) of a learning algorithm is the set of assumptions that the learner uses to predict outputs.
- Bias is of two types, constraints and preferences.
  - Constraints or Restrictions limit the hypothesis space.
  - Preferences impose ordering on the hypothesis space.
- Examples:

  1. Instead of considering all Boolean formulas, we are going to consider only conjunctive Boolean formulas, this can be an example of Constraints.

  2. Giving preference that out of all possible polynomials, I will prefer polynomials of lower degree. This is an example of Preferences.

Bayesian Learning Methods

Probability Basics

Let D1 and D2 be two random variable

D1: value on die 1
D2: value on die 2

P(D1+D2<=5)=10/36
P(D1=2|D1+D2<=5)=3/10

P(D1=2 ∩ D1+D2<=5)/P(D1+D2<=5)
=(3/36)/(10/36)= 3/10

Hence, P(A|B)=P(A ∩ B)/P(B) given P(B)!=0

D2

D1	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

- Independent Events: Two events are said to be independent if

$$P(A|B)=P(A) \text{ and } P(B|A)=P(B)$$

A: Getting value 6 on die1

B: Getting value 3 on die2

$P(D1=6|D2=3)=P(D1=6)$

$P(D2=3|D1=6)=P(D2=3)$

- Mutually Exclusive Events: Two events A and B are said to be mutually exclusive if P(A|B)=P(B|A)=0 i.e. P(A ∩ B)=0

  P(D1=3|D1=6)=0

Product Rule: P(A ∩ B)=P(A|B).P(B)=P(B|A).P(A)

Sum Rule=P(A U B)=P(A)+P(B)-P(A ∩ B)

Theorm of total probability: If A_1……..A_N are mutually exclusive events with $\sum_{i=1}^{N} P(A_i) = 1$ then

P(B)=$\sum_{i=1}^{N} P(B|A_i) \cdot P(A_i)$

Bayes Theorem

Bayes' theorem, named after 18th-century British mathematician ***Thomas Bayes***, is a mathematical formula for determining ***conditional probability***.

Conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring.

Bayes theorem is the cornerstone of Bayesian learning methods as it provides a way to calculate the posterior probability.

Bayes' theorem relies on incorporating ***prior probability*** distributions in order to generate ***posterior probabilities***.

Bayes Theorem

Bayes' theorem gives the probability of an event based on new information that is, or may be related, to that event.

For instance, say a single card is drawn from a complete deck of 52 cards. The probability that the card is a king is 4/52, which equals 1/13 or approximately 7.69%.

Now, suppose it is revealed that the selected card is a face card. The probability the selected card is a king, given it is a face card, is 4/12, or approximately 33.3%, as there are 12 face cards in a deck.

Formula for Bayes' Theorem

$$P\left(A|B\right)=\frac{P\left(A\bigcap B\right)}{P\left(B\right)}=\frac{P\left(A\right)\cdot P\left(B|A\right)}{P\left(B\right)}$$

where:

$P\left(A\right)=$ The probability of A occurring

$P\left(B\right)=$ The probability of B occurring

$P\left(A|B\right)=$ The probability of A given B

$P\left(B|A\right)=$ The probability of B given A

$P\left(A\bigcap B\right))=$ The probability of both A and B occurring

Example: Bayes Theorem

In a clinic for liver disease, past data tells you that 10% of patients entering the clinic have liver disease. Five percent of the clinic’s patients are alcoholics. Also among the patients diagnosed with liver disease, 7% are alcoholics. Find out a patient’s probability of having liver disease given that he is an alcoholic.

P(alcoholic) = 0.05

P(liver-disease) = 0.1

P(alcoholic | liver-disease) = 0.07

P(liver-disease | alcoholic) = ?

Answer

$$P(liverdisease \mid alcoholic) = \frac{P(alcoholic|liverdisease)P(liverdisease)}{P(alcoholic)}$$

$$= \frac{(0.07)\cdot(0.1)}{0.05}$$

$$= 0.14$$

Thus, if the patient is an alcoholic, their chances of having liver disease is 0.14 (14%).

Practice Problem

Imagine there is a drug test that is 98% accurate, meaning 98% of the time it shows a true positive result for someone using the drug and 98% of the time it shows a true negative result for nonusers of the drug. It is known that 0.5% of people use the drug. If a person selected at random tests positive for the drug, determine the probability the person is actually a user of the drug.

P(drug) = 0.005 (Prior Probability)

P(+/drug) = 0.98 P(-/no-drug) = 0.98

P(drug/+) = ?

Bayes theorem in Machine Learning

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.

Prior Probability is the probability of an event before new data is collected.

In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.

Notations

P(*h*) denotes the prior probability of h and may reflect any background knowledge we have about the chance that *h* is a correct hypothesis. It is independent of data D.

P(D) denotes the prior probability that training data D will be observed. Also, called as Evidence

P(D|*h*) denotes the probability of observing data D given some world in which hypothesis *h* holds. Also called as Likelihood.

P(h|D) denotes the posterior probability of *h*, it reflects our confidence that *h* holds after we have seen the training data D.

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

Maximum A Posteriori (MAP) Hypothesis

A learning problem considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis $h \in H$ given the observed data D.

Any such maximally probable hypothesis is called a ***maximum a posteriori (MAP) hypothesis***.

MAP hypothesis is determined by using Bayes theorem to calculate the posterior probability of each candidate hypothesis.

MAP Hypothesis

$$
\begin{aligned}
h_{MAP} &\equiv \underset{h \in H}{\operatorname{argmax}}\, P(h|D) \\
&= \underset{h \in H}{\operatorname{argmax}}\, \frac{P(D|h)\,P(h)}{P(D)} \\
&= \underset{h \in H}{\operatorname{argmax}}\, P(D|h)\,P(h)
\end{aligned}
$$

Notice, in final step the denominator P(D) has been dropped because it is constant independent of *h*.

Example: MAP Hypothesis

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 have cancer.

The available data is from a particular laboratory test with two possible outcomes: + (positive) and - (negative). We have prior knowledge that over the entire population of people only. 0.008 have this disease. Furthermore, the lab test is only an imperfect indicator of the disease. The test returns a correct positive result in only 98% of cases in which the disease is actually present and a correct negative result in only 97% of cases in which disease is not present. In other cases, the test returns the opposite result. Suppose we observe a new patient for whom the lab test returns a positive result. Should we diagnose the patient as having cancer or not?

Solution

Possible Hypotheses:

h_1: Person has cancer given his test is positive

h_2: Person does not have cancer given his test is positive.

Hypothesis having highest probability will be selected.

To find: P(cancer | +) =? And P(~cancer | +) = ? and select the one that has higher probability.

Solution (contd.)

P(cancer) = 0.008 P(~cancer) = 1 - 0.008 = 0.992

P(+|cancer) = 0.98 P(-|cancer) = 1 - 0.98 = 0.02

P(+|~cancer) = 0.03 P(-|~cancer) = 1 - 0.03 = 0.97

$$P(cancer \mid +) = \frac{P(+|cancer) \cdot P(cancer)}{P(+)}$$

$$= \frac{(0.98) \cdot (0.008)}{P(+)}$$

$$= \frac{0.0078}{P(+)}$$

$$P({\sim}cancer \mid +) = \frac{P(+|{\sim}cancer) \cdot P({\sim}cancer)}{P(+)}$$

$$= \frac{(0.03) \cdot (0.992)}{P(+)}$$

$$= \frac{0.0298}{P(+)}$$

Answer

As P(~cancer | +) > P(cancer | +), thus h_{MAP} = ~cancer , that patient does not have cancer.

Maximum Likelihood Hypothesis

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).

In such case, we can further simplify the equation by only considering the term P(D|*h*) to find the most probable hypothesis. As P(D|*h*) is known as likelihood thus any hypothesis that maximises P(D|*h*) is called as **Maximum Likelihood *h*$_{ML}$ Hypothesis.**

$$\begin{aligned} h_{MAP} &\equiv \underset{h \in H}{\operatorname{argmax}}\, P(h|D) \\ &= \underset{h \in H}{\operatorname{argmax}}\, \frac{P(D|h)\,P(h)}{P(D)} \\ &= \underset{h \in H}{\operatorname{argmax}}\, P(D|h)\,P(h) \end{aligned}$$

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

Bayes Optimal Classifier

MAP Hypothesis is used to know “what is the most probable hypothesis given the training data?”

However, we are mostly interested in knowing:

> “What is the most probable **classification** of the new instance given the training data?”

Bayes Optimal Classifier is an improvement over MAP Hypothesis.

Bayes Optimal Classifier is also known as Bayes Optimal Learner.

It is a probabilistic model that makes the most probable prediction for a new example.

It is called Bayes Optimal because no other classification method using the same hypothesis space and prior knowledge can outperform it on average.

Intuition

Consider a hypothesis space containing three hypotheses h_1, h_2 and h_3.

Suppose that posterior probabilities of these hypotheses given the training data are as follows:

$P(h_1|D) = 0.4$ $P(h_2|D) = 0.3$ $P(h_3|D) = 0.3$

What is the MAP Hypothesis?

h1 is the MAP Hypothesis.

Intuition (contd.)

Suppose a new instance (example) x is encountered, which is classified positive(+) by h_1 but negative(-) by h_2 and h_3.

Taking all hypotheses into account, the probability of x is positive is 0.4 (from h_1) and the probability that it is negative is 0.6 (0.3 + 0.3, from h_2 and h_3 combined).

The most probable classification (negative) in this case is different from classification generated by MAP Hypothesis.

The most probable classification of the new instance is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities.

If the possible classification of the new example can take on any value v_j from some set V, then the probability $P(v_j|D)$ that the correct classification for the new instance is v_j, is just

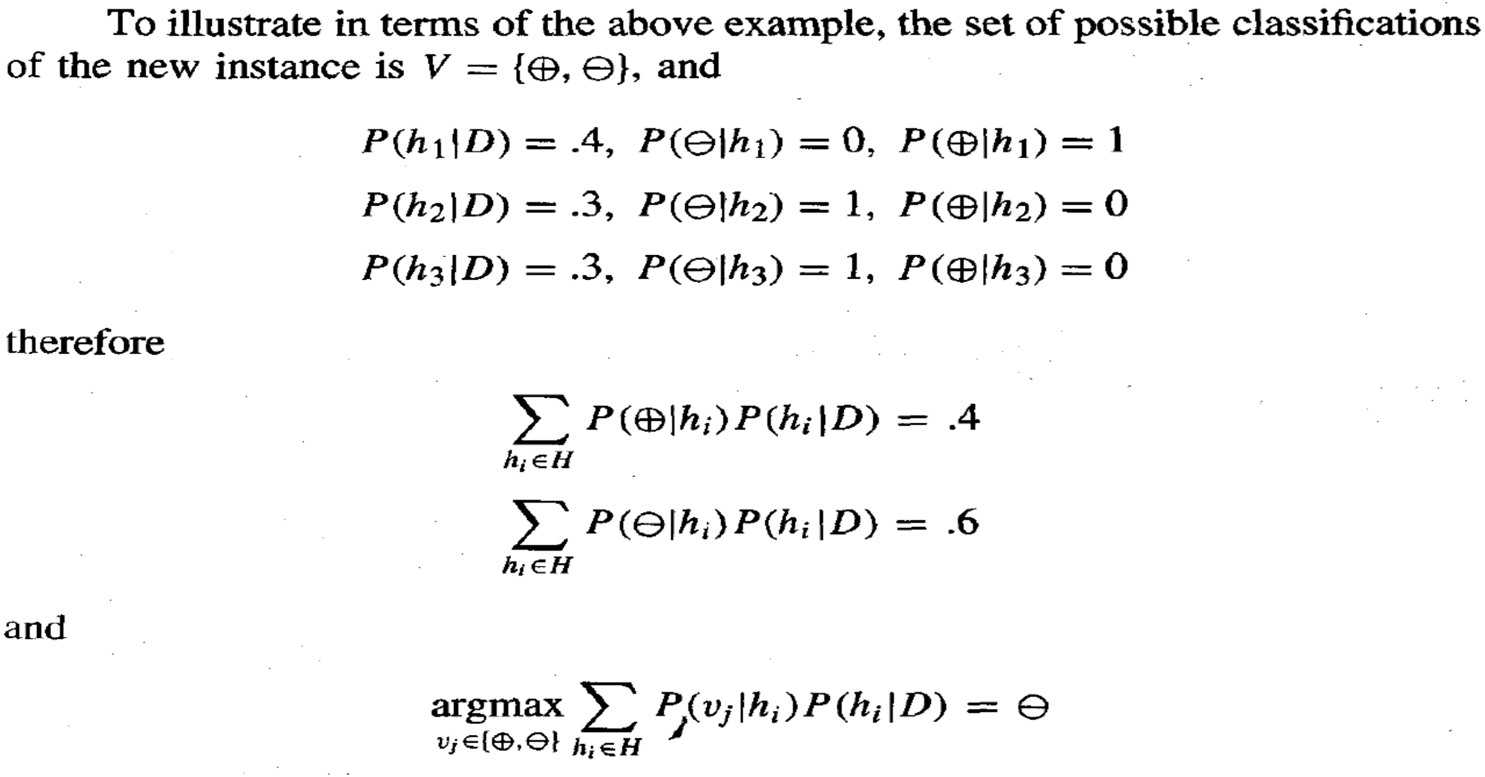
$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

The optimal classification of the new instance is the value v_j, for which $P(v_j|D)$ is maximum.

Bayes optimal classification:

$$\underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Example: Bayes Optimal Classifier



To illustrate in terms of the above example, the set of possible classifications of the new instance is $V = \{\oplus, \ominus\}$, and

$$P(h_1|D) = .4, \;\; P(\ominus|h_1) = 0, \;\; P(\oplus|h_1) = 1$$

$$P(h_2|D) = .3, \;\; P(\ominus|h_2) = 1, \;\; P(\oplus|h_2) = 0$$

$$P(h_3|D) = .3, \;\; P(\ominus|h_3) = 1, \;\; P(\oplus|h_3) = 0$$

therefore

$$\sum_{h_i \in H} P(\oplus|h_i) P(h_i|D) = .4$$

$$\sum_{h_i \in H} P(\ominus|h_i) P(h_i|D) = .6$$

and

$$\underset{v_j \in \{\oplus, \ominus\}}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i) P(h_i|D) = \ominus$$

Example of Bayes Optimal Classification

Let there be 5 hypotheses h_1 through h_5.

$P(h_i \mid D)$	$P(F \mid h_i)$	$P(L \mid h_i)$	$P(R \mid h_i)$
0.4	1	0	0
0.2	0	1	0
0.1	0	0	1
0.1	0	1	0
0.2	0	1	0

Then, the MAP hypothesis suggests the robot should go forward (F).

What does the Bayes optimal procedure suggest?

Example of Bayes Optimal Classification

$$\sum_{h_i \in H} P(F \mid h_i) P(h_i \mid D) = 0.4$$

$$\sum_{h_i \in H} P(L \mid h_i) P(h_i \mid D) = 0.2 + 0.1 + 0.2 = 0.5$$

$$\sum_{h_i \in H} P(R \mid h_i) P(h_i \mid D) = 0.1$$

Thus, Bayes optimal recommends the robot turn left.

Strength and Weakness of Bayes Optimal Classifier

No other classification method using same hypothesis space and same prior knowledge can outperform this method on average. That’s why it is called as “Optimal Classifier”.

This method maximizes the probability that the new instance is classified correctly, given the available data, hypothesis space and prior probabilities over the hypothesis.

In practice, the Bayes Optimal Classifier is computationally expensive as it computes the posterior probability of every hypothesis in H and then combines the predictions of each hypothesis to classify each new instance.

Gibbs Algorithm and Naive Bayes Classifier are two simplified approaches over Bayes Optimal Classifier.

Bayes Error

Although the Bayes Optimal Classifier makes optimal predictions, it is not perfect given the uncertainty in the training data and incomplete coverage of the problem domain and hypothesis space. As such, the model will make errors. These errors are often referred to as **Bayes errors**.

Because the Bayes classifier is optimal, the Bayes error is the **minimum** possible error that can be made.

Bayes Error: The minimum possible error that can be made when making predictions.

The Bayes classifier produces the lowest possible test error rate, called the **Bayes error rate**.

Naïve Bayes Classifier

Naïve Bayes Classifier

It is a popular Bayesian Learning Method.

Its performance has been shown to be comparable to that of neural networks and decision tree learning.

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 any value from some finite set V.

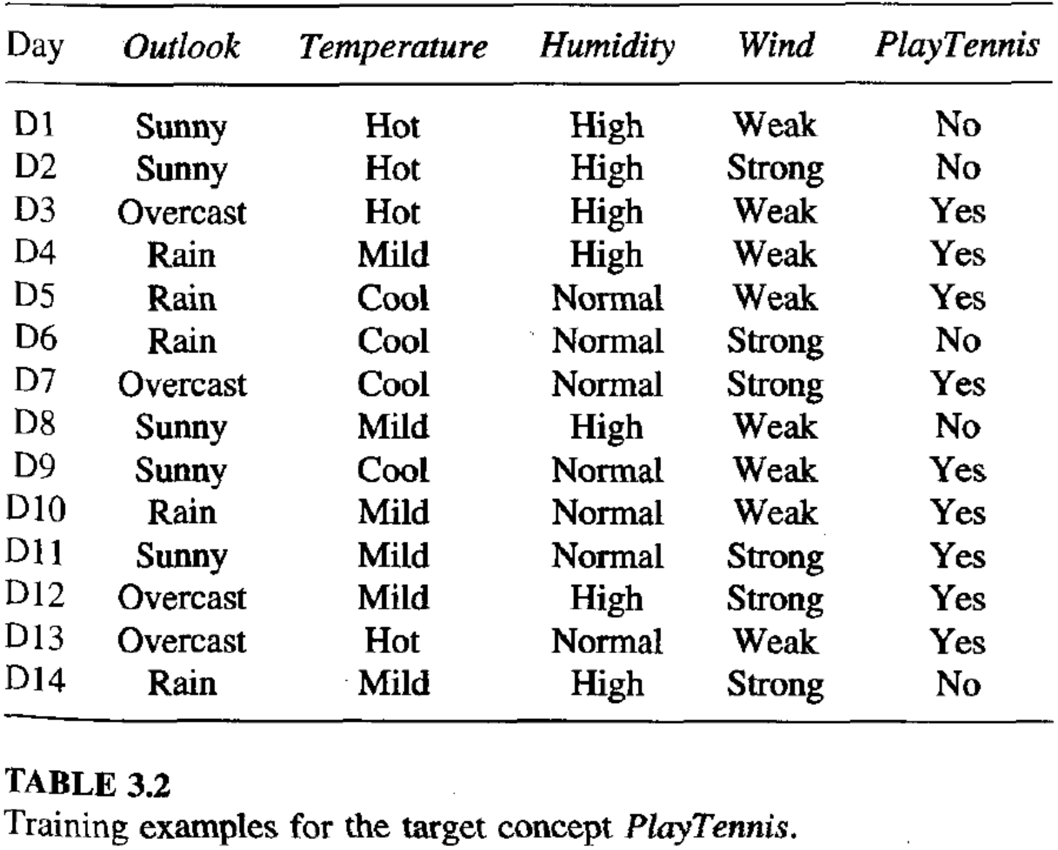
It is based on the simplifying assumption that the attribute values are conditionally independent given the target value. Hence called as Naive Classifier.

Example

Target Concept: PlayTennis

Attributes:

- Outlook
- Temperature
- Humidity
- 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

TABLE 3.2
Training examples for the target concept *PlayTennis*.

Example

Using NBC and training data, classify the given new instance:

Conjunction of attribute values

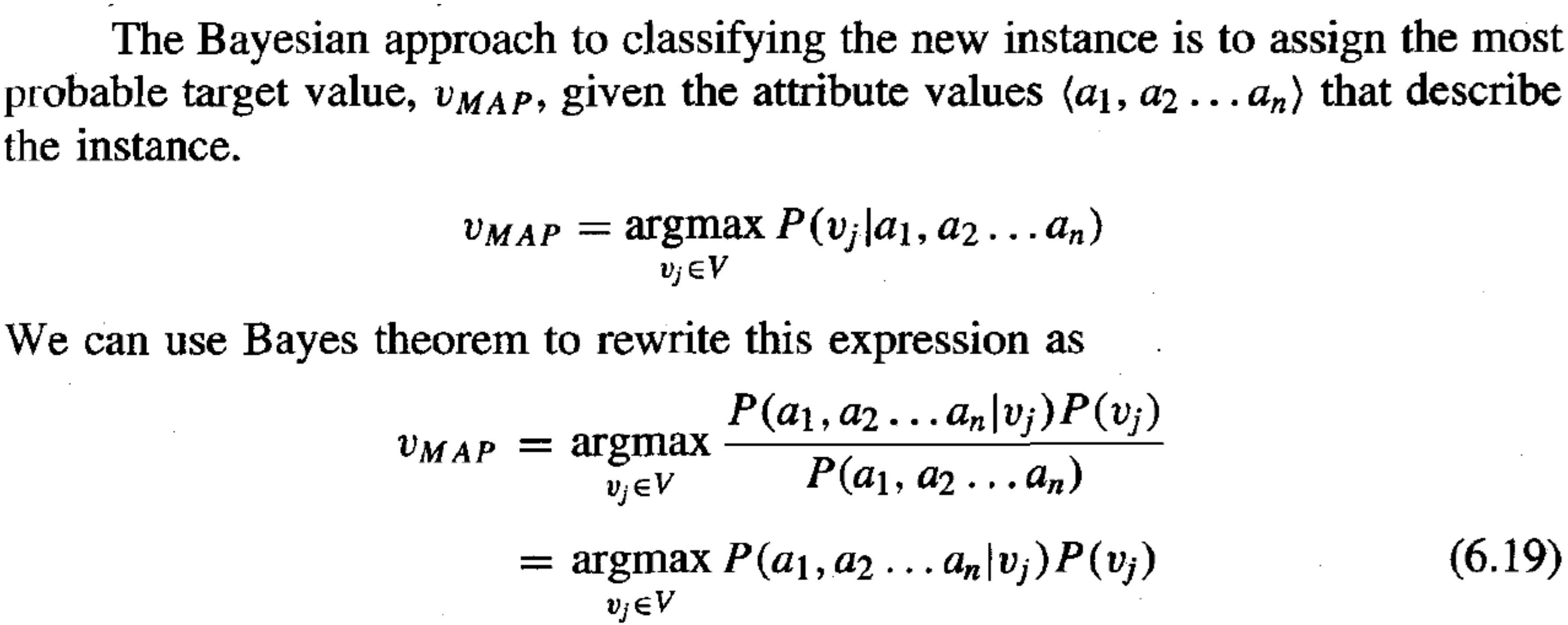
<Outlook = sunny, Temp = cool, Humidity = high, Wind = strong>

Predict PlayTennis = Yes or No?

Target Function f(x)

V = {YES, NO}

Naïve Bayes Classifier



The Bayesian approach to classifying the new instance is to assign the most probable target value, v_{MAP}, given the attribute values $\langle a_1, a_2 \dots a_n \rangle$ that describe the instance.

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

We can use Bayes theorem to rewrite this expression as

$$\begin{aligned} v_{MAP} &= \operatorname*{argmax}_{v_j \in V} \frac{P(a_1, a_2 \dots a_n | v_j) P(v_j)}{P(a_1, a_2 \dots a_n)} \\ &= \operatorname*{argmax}_{v_j \in V} P(a_1, a_2 \dots a_n | v_j) P(v_j) \qquad (6.19) \end{aligned}$$

Naïve Bayes Classifier

$P(v_j)$ can be easily calculated from the training data.

However, estimating different $P(a_1,a_2…a_n|v_j)$ terms is not feasible unless we have a very very large set of training data.

From the naive bayes assumption, the probability of observing the conjunction $a_1,a_2…a_n$ is just the product of the probabilities of the individual attributes:

$$P(a_1, a_2 ... a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

Naïve Bayes Classifier

Substituting this into Equation (6.19)

Naive Bayes Classifier:

$$v_{NB} = arg \max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

where v_{NB} denotes the target value output by the Naive Bayes Classifier.

Example (contd.)

Using the training data provided, given a novel instance as:

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

Predict the value of target concept “PlayTennis” labelled as “yes” or “no” for this new instance. Using the NBC equation

$$v_{NB} = arg \max_{v_j \in \{yes,no\}} P(v_j) \prod_i P(a_i \mid v_j)$$

$$= arg \max_{v_j \in \{yes,no\}} P(outlook = sunny \mid v_j) \cdot P(temp = cool \mid v_j)$$

$$\cdot P(humidity = high \mid v_j) \cdot P(wind = strong \mid v_j)$$

<table>
<tr><td colspan="2"><b>P(PlayTennis = yes) = 9/14 = 0.64</b></td><td colspan="2"><b>P(PlayTennis = no) = 5/14 = 0.36</b></td></tr>
<tr><td><b>P(sunny|yes) = 2/9 = 0.22</b></td><td><b>P(cool|yes) = 3/9 = 0.33</b></td><td><b>P(high|yes) = 3/9 = 0.33</b></td><td><b>P(strong|yes) = 3/9 =0.33</b></td></tr>
<tr><td><b>P(sunny|no) = ⅗ = 0.6</b></td><td><b>P(cool|no) =⅕ = 0.2</b></td><td><b>P(high|no) = ⅘=0.8</b></td><td><b>P(strong|no) =⅗ = 0.6</b></td></tr>
<tr><td colspan="4"><b>P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes) = (0.64)*(0.22)*(0.33)*(0.33)*(0.33) = 0.005</b></td></tr>
<tr><td colspan="4"><b>P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no) = (0.36)*(0.6)*(0.2)*(0.8)*(0.6) = 0.0207</b></td></tr>
</table>

<table>
<tr><td colspan="2">P(PlayTennis = yes) = 9/14 = 0.64</td><td colspan="2">P(PlayTennis = no) = 5/14 = 0.36</td></tr>
<tr><td>P(sunny|yes) = 1/9=0.11</td><td>P(cool|yes) = 3/9 = 0.33</td><td>P(high|yes) = 3/9 = 0.33</td><td>P(strong|yes) = 3/9 = 0.33</td></tr>
<tr><td>P(sunny|no) = ⅗ = 0.6</td><td>P(cool|no) = ⅕ = 0.2</td><td>P(high|no) = ⅘ = 0.8</td><td>P(strong|no) = ⅗ = 0.6</td></tr>
<tr><td colspan="4">P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes) = (0.64)*(0.11)*(0.33)*(0.33)*(0.33) = 0.0026</td></tr>
<tr><td colspan="4">P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no) = (0.36)*(0.6)*(0.2)*(0.8)*(0.6) = 0.0207</td></tr>
</table>

The probability for PlayTennis = No is higher than PlayTennis=Yes so the output class is “No”.

Outlook = overcast, temp=hot, humidity=normal and wind=weak

Find out the output class?

Gaussian Naive Bayes Classifier

Naive Bayes Classifier can be extended to real-valued attributes, most commonly by assuming a Gaussian distribution.

This extension of naive Bayes is called Gaussian Naive Bayes.

With real-valued inputs, we can calculate the mean and standard deviation of input values (x) for each class to summarize the distribution.

Gaussian Probability Density Function

Probabilities of new instance x values are calculated using the Gaussian Probability Density Function (PDF).

Gaussian PDF is calculated as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

f(x)	=	probability density function
σ	=	standard deviation
μ	=	mean

Gaussian Naive Bayes Example

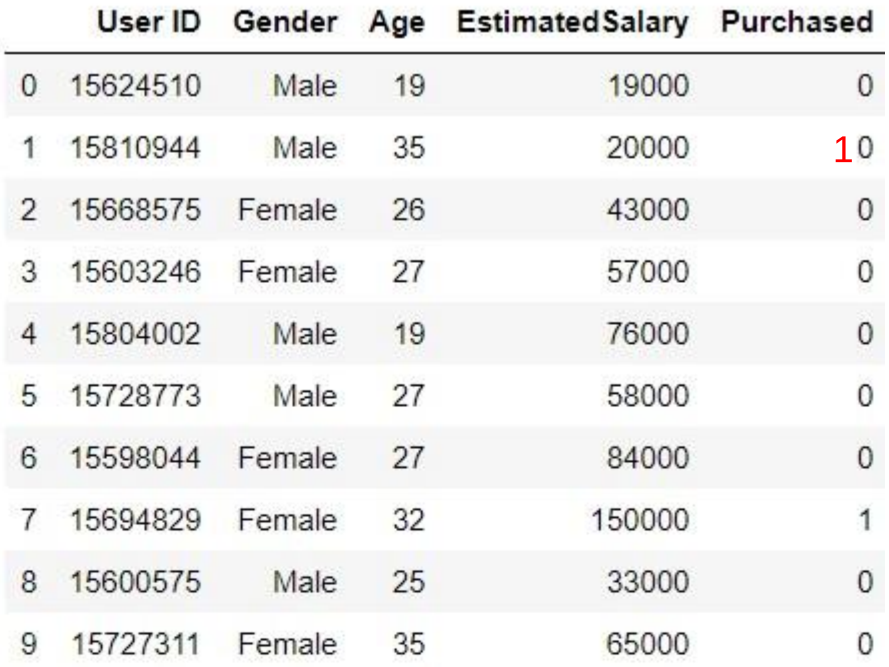
Given this data,

Predict whether a person with

age = 24, gender=female

And salary=50,000 will purchase

the car or not?



	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	1 0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
5	15728773	Male	27	58000	0
6	15598044	Female	27	84000	0
7	15694829	Female	32	150000	1
8	15600575	Male	25	33000	0
9	15727311	Female	35	65000	0

	Case1	y=no				
	age	square(age-agemean)		salary	square(salary-salarymean)	
	19	43.890625		19000	1251390625	
	26	0.140625		43000	129390625	
	27	1.890625		57000	6890625	
	19	43.890625		76000	467640625	
	27	1.890625		58000	13140625	
	27	1.890625		84000	877640625	
	25	0.390625		33000	456890625	
	35	87.890625		65000	112890625	
mean	25.625	25.98214286		54375	473696428.6	
s.d		5.097268176			21764.56819	
	Case2	y=yes		Case2	y=yes	
	age	square(age-agemean)		salary	square(salary-salarymean)	
	35	2.25		20000	4225000000	
	32	2.25		150000	4225000000	
mean	33.5	4.5		85000	8450000000	
s.d		2.121320344			91923.88155	

Calculating Probabilities

P(yes) = 0.2

P(no) = 0.8

P(female|yes) = 0.5

P(female|no) = 0.5

$\text{Mean}_{\text{age|yes}}$ = 33.5 $\text{Std.Dev}_{\text{age|yes}}$ = 2.12

$\text{Mean}_{\text{age|no}}$ = 25.62 $\text{Std.Dev}_{\text{age|no}}$ = 5.09

$\text{Mean}_{\text{salary|yes}}$ = 85000 $\text{Std.Dev}_{\text{salary|yes}}$ = 91923.88

$\text{Mean}_{\text{salary|no}}$ = 54375 $\text{Std.Dev}_{\text{salary|no}}$ = 21764.57

Solution

Let, person1 = {age=24,gender=female,salary=50,000}

P(yes|person1) =? P(no|person1)=?

P(yes|person1) = P(yes) * P(age=24|yes) * P(gender=female|yes) * P(salary=50000|yes)

P(no|person1) = P(no) * P(age=24|no) * P(gender=female|no) * P(salary=50000|no)

Calculating Probabilities (contd.)

P(yes|person1) = P(yes) * P(age=24|yes) * P(gender=female|yes) * P(salary=50000|yes)

= 0.2 * P(age=24|yes) * 0.5 * P(salary=50000|yes)

= 0.2 * 8.304151e-06 * 0.5 * 4.036471e-06

= 3.351946e-12

P(no|person1) = P(no) * P(age=24|no) * P(gender=female|no) * P(salary=50000|no)

= 0.8 * P(age=24|no) * 0.5 * P(salary=50000|no)

= 0.8 * 0.07438809 * 0.5 * 1.796328e-05

= 5.345016e-07

Result

$$P(no|person1) > P(yes|person1)$$

Thus, from given data it can be predicted that a person with age=24, gender = female and salary = 50000 will **NOT** purchase the car.

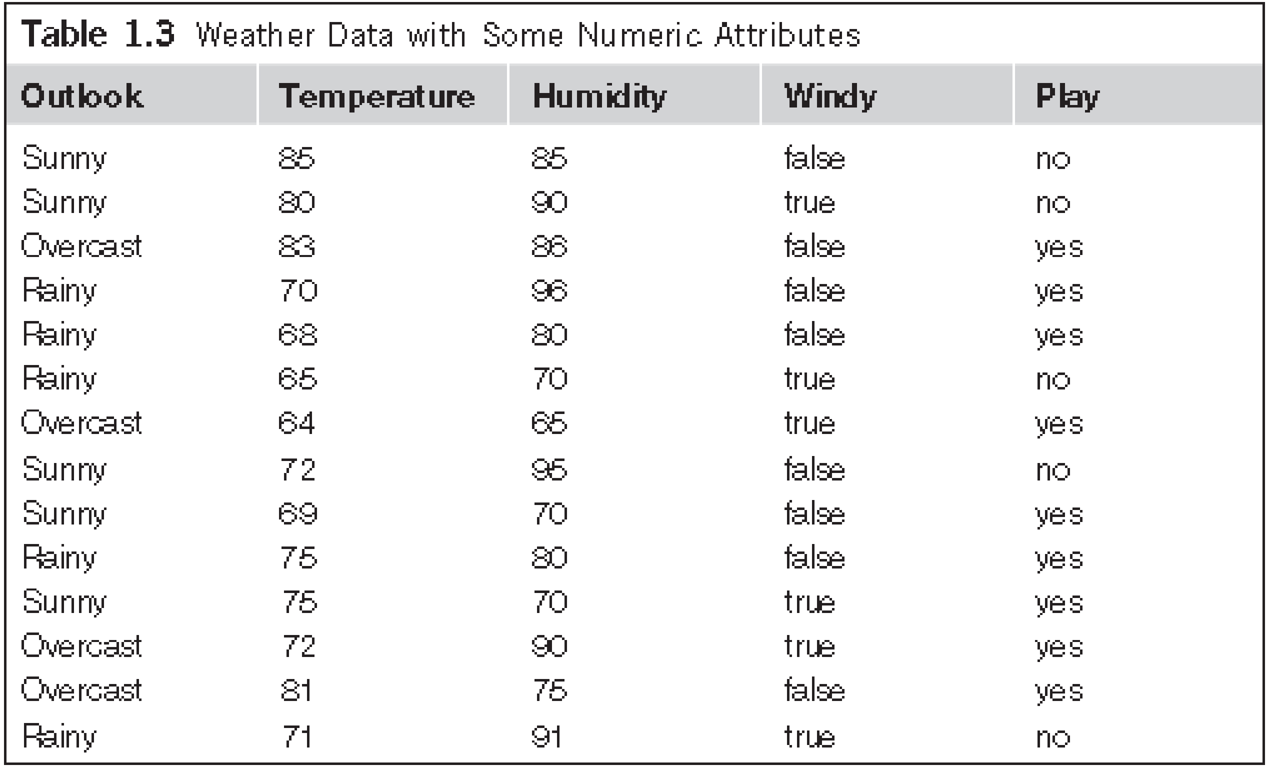


Table 1.3 Weather Data with Some Numeric Attributes

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

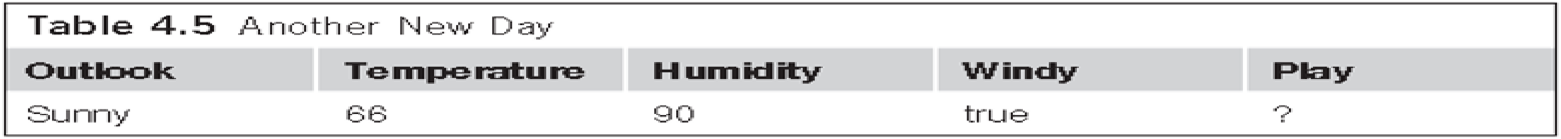


Table 4.5 Another New Day

Outlook	Temperature	Humidity	Windy	Play
Sunny	66	90	true	?

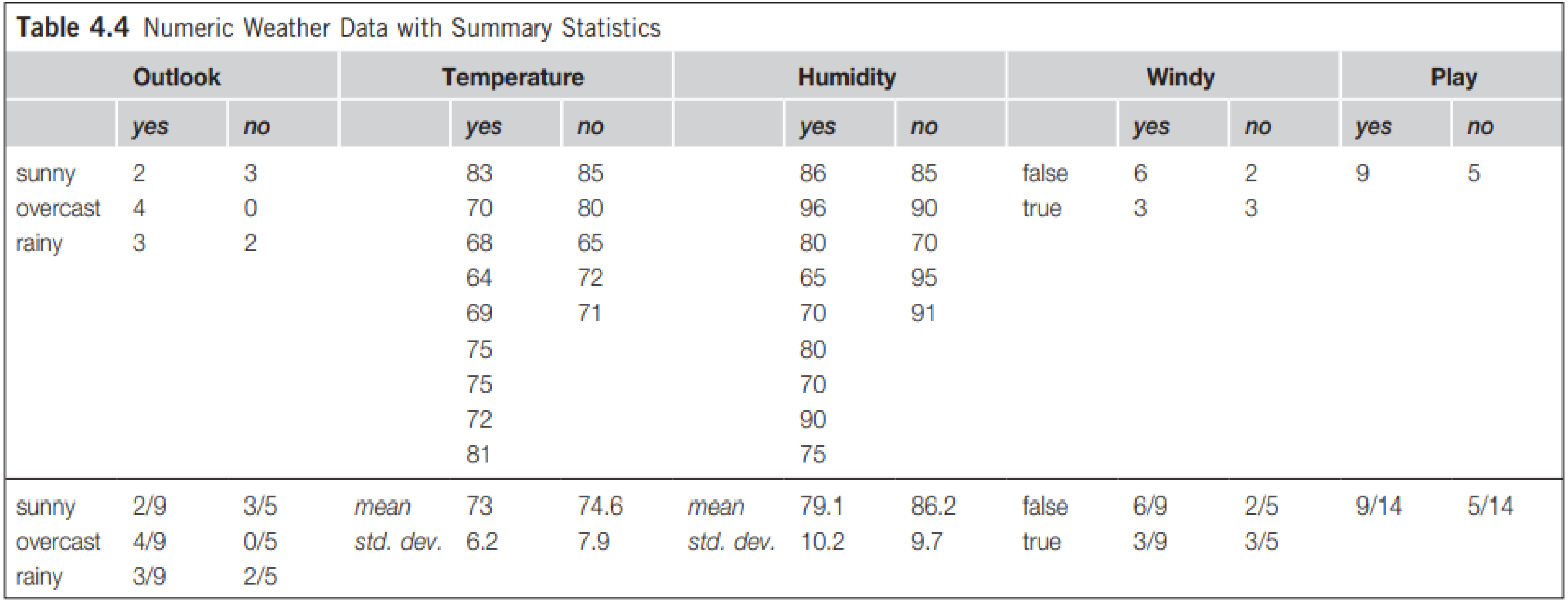


Table 4.4 Numeric Weather Data with Summary Statistics

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	*std. dev.*	6.2	7.9	*std. dev.*	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

Using these probabilities for the new day in Table 4.5 yields

$$\text{Likelihood of } \mathit{yes} = 2/9 \times 0.0340 \times 0.0221 \times 3/9 \times 9/14 = 0.000036$$

$$\text{Likelihood of } \mathit{no} = 3/5 \times 0.0279 \times 0.0381 \times 3/5 \times 5/14 = 0.000137$$

which leads to probabilities

$$\text{Probability of } \mathit{yes} = \frac{0.000036}{0.000036 + 0.000137} = 20.8\%$$

$$\text{Probability of } \mathit{no} = \frac{0.000137}{0.000036 + 0.000137} = 79.2\%$$

These figures are very close to the probabilities calculated earlier for the new day in Table 4.3 because the *temperature* and *humidity* values of 66 and 90 yield similar probabilities to the *cool* and *high* values used before.

Zero Probability Problem

- x = <Outlook=overcast, Temperature=66, Humidity=90, Windy=True>

likelihood for play=yes
P(x/yes) * P(yes) = P(overcast/yes) * P(Temperature=66/yes) * P(Humidity=90/yes) * P(True/yes) * P(yes)
likelihood for play=no
P(x/no) * P(no) = P(overcast/no) * P(Temperature=66/no) * P(Humidity=90/no) * P(True/no) * P(no)

The new values needed to calculate the above equations are:

- P(overcast/yes) = 4/9 and P(overcast/no) = 0/5 = 0

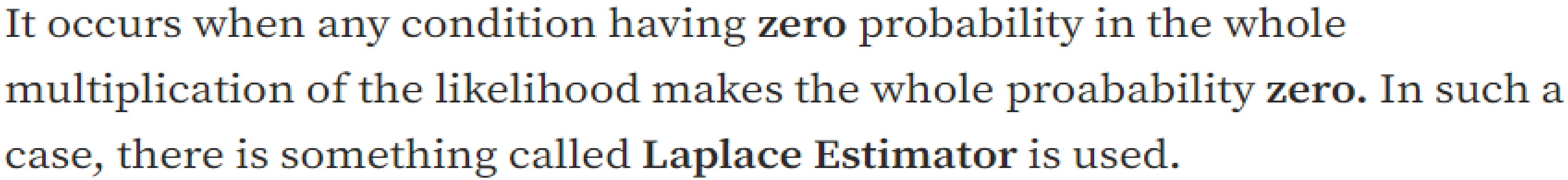
$P(x/\text{yes}) * P(\text{yes}) = (2/9) * 0.034 * 0.0221 * (3/9) * (9/14) = 0.000036$

$P(x/\text{no}) * P(\text{no}) = 0 * 0.0279 * 0.0381 * (3/5) * (5/14) = 0$

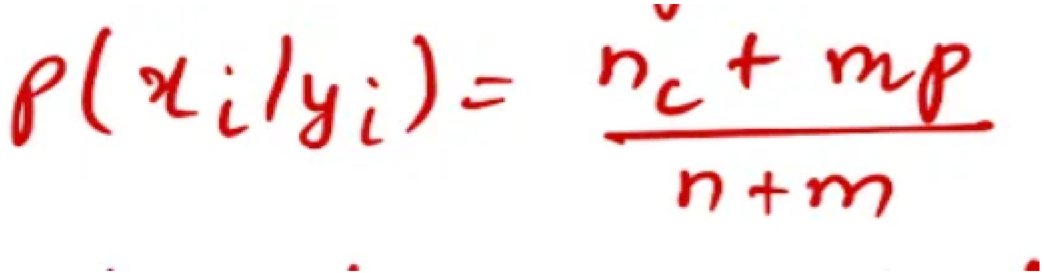
0.000036 > 0

Classification — YES

Solution-Laplace Estimator



It occurs when any condition having **zero** probability in the whole multiplication of the likelihood makes the whole proabability **zero.** In such a case, there is something called **Laplace Estimator** is used.



$$p(x_i|y_i) = \frac{n_c + mp}{n+m}$$

Image 3

where,

n_c = number of instances where xi = x and yi = y,

n = number of instances where yi = y,

p = prior estimate, example: when assuming a uniform distribution of attribute values p=1/*m*, with *m* defining the number of different (unique) attribute values.

m = number of unique values for that attribute.

So, if a uniform distribution is assumed the formula in **Image 3** modifies to the following:

$$P(x_i|y_i) = \frac{n_c+1}{n+m}$$

Applying Laplace Estimator

For **Outlook = overcast,** the new probability becomes
P(overcast/yes) = (4 + 3 * (1/3)) / (9 + 3)= 5/12
Where,

nc = 4, since 4 instances where Outlook = overcast & play = yes,
n = 9, since total instances where play = yes,
m = 3, since the attribute Outlook has 3 unique values (sunny, overcast, rainy),
p = 1/*m* = 1/3, since the uniform distribution is assumed

Similarly,

P(overcast/no) = (0 + 3 * (1/3)) / (5 + 3)= 1/8
where,
nc = 0, since 0 instances where Outlook = overcast & play = no,
n = 5, since total instances where play = no,
m = 3, since the attribute Outlook has 3 unique values (sunny, overcast, rainy),
p = 1/*m* = 1/3, since the uniform distribution is assumed

- Note: While applying Laplace Estimator, ensure that you apply it to all the ordinal attributes. You can’t just apply is to the attribute where the Zero frequency problem is occurring.

Since, the other ordinal attribute in our instance to classify is the attribute Windy, we need to apply Laplace Estimator t here as well. After applying the modified probabilities are:

For Windy = True, the new probability becomes
P(True/yes) = (3 + 2 * (1/2)) / (9 + 2) = 4/11
where,
nc = 3, since 3 instances where Windy = True & play = yes,
n = 9, since total instances where play = yes,
m = 2, since the attribute Windy has 2 unique values (True, False),
p = 1/m = 1/2, since the uniform distribution is assumed

Similarly,

P(True/no) = (3 + 2* (1/2)) / (5 + 2)= 4/7

where,

nc = 3, since 3 instances where Windy = True & play = no,

n = 5, since total instances where play = no,

m = 2, since the attribute Windy has 2 unique values (True, False),

p = 1/m = 1/2, since the uniform distribution is assumed

P(*x*/yes) * P(yes) = (5/12) * 0.034 * 0.0221 * (4/11) * (9/14) = 0.0000731

P(*x*/no) * P(no) = 1/8 * 0.0279* 0.0381* (4/7) * (5/14) = 0.0000271

0.0000731 > 0.0000271

Classification — YES

Even though the classification did not change but now we have a better scientific reasoning behind our conclusion.

Possible Exam Questions

How is Bayes optimal classifier different from MAP Hypothesis?

References

1. Machine Learning, Tom Mitchell, McGraw Hill, 1997.
2. https://machinelearningmastery.com/bayes-optimal-classifier/