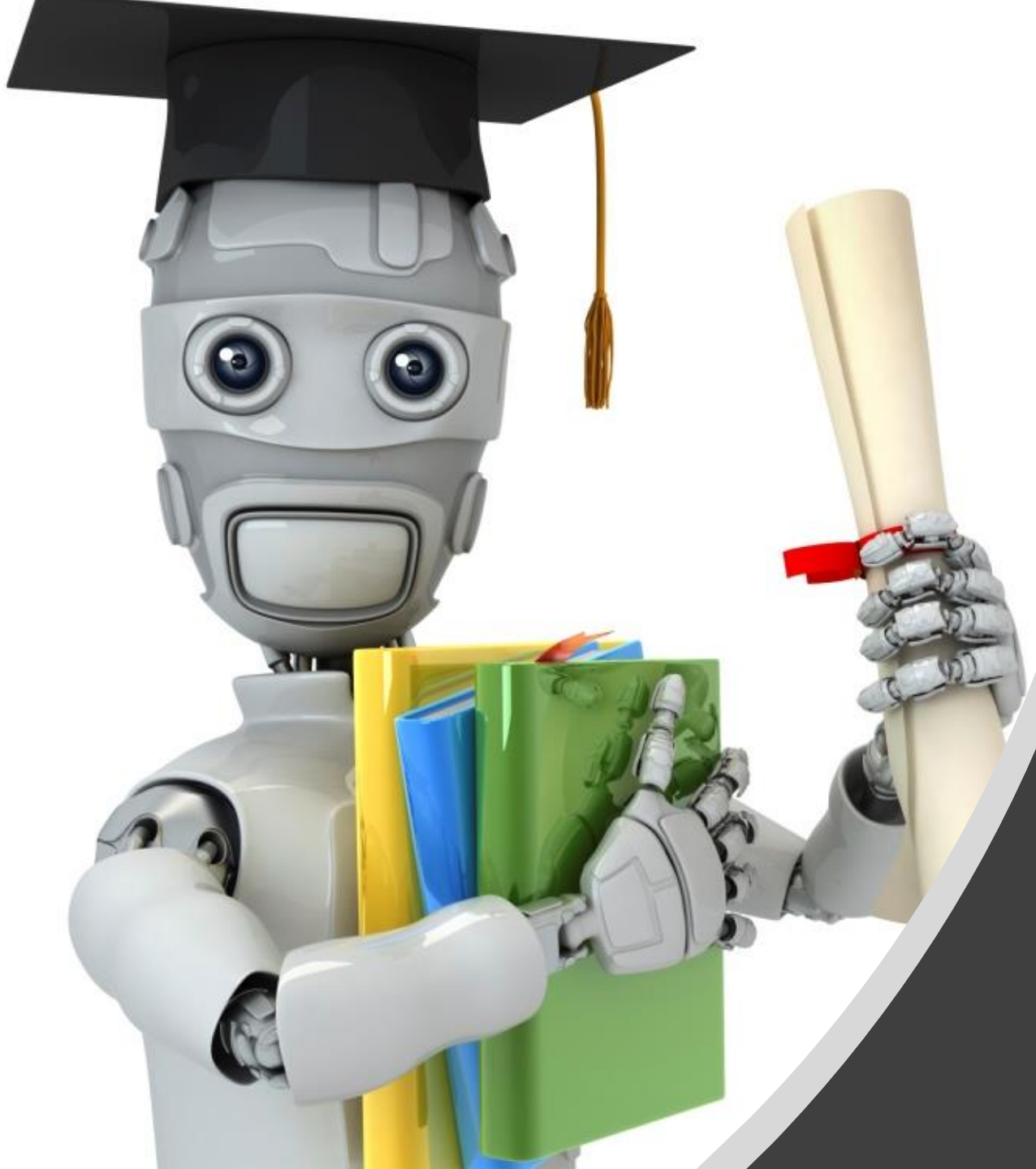
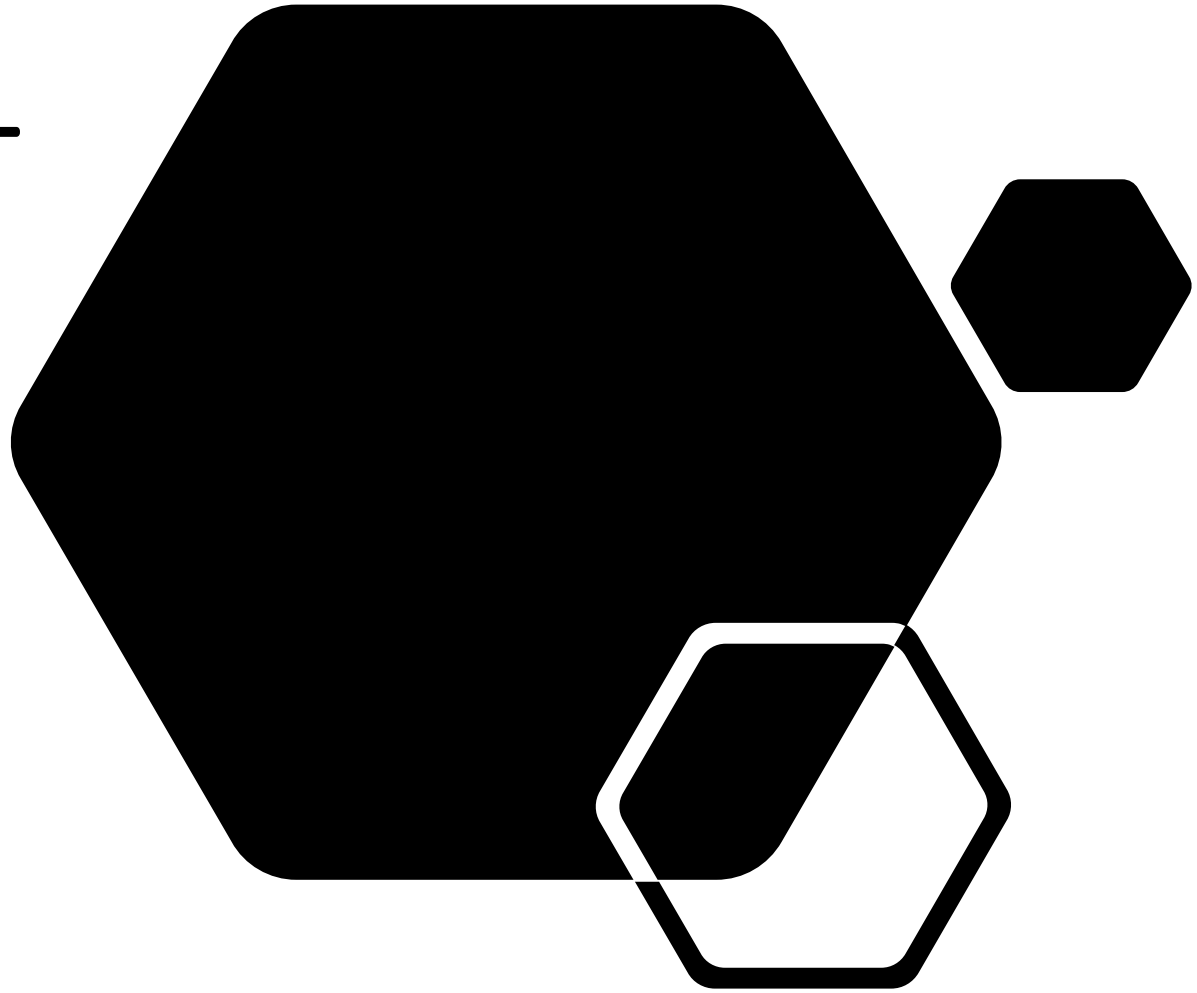


Machine Learning

Prepared By:

Dr. Sara Sweidan

Probabilistic Learning – Classification Using Naive Bayes



To Know

- The technique describes the probability of events and how probabilities should be revised in the light of additional information. These principles formed the foundation for what are now known as **Bayesian methods**.
- It suffices to say that a probability is a number between 0 and 1 (that is, between 0 percent and 100 percent), which captures the chance that an event will occur in the light of the available evidence. The lower the probability, the less likely the event is to occur. A probability of 0 indicates that the event will not occur, while a probability of 1 indicates that the event will occur with 100 percent certainty.

To Know

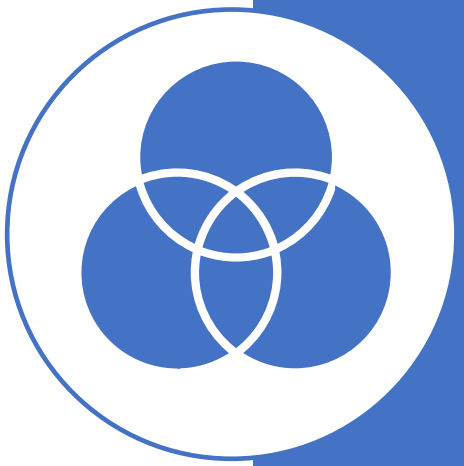
- Classifiers based on Bayesian methods utilize training data to calculate an observed probability of each outcome based on the evidence provided by feature values. When the classifier is later applied to unlabeled data, it uses the observed probabilities to predict the most likely class for the new features.
- Bayesian classifiers have been used for:
- • Text classification, such as junk e-mail (spam) filtering.
- • Intrusion or anomaly detection in computer networks.
- • Diagnosing medical conditions given a set of observed symptoms.

Basic concepts of Bayesian methods

- The probability of an event is estimated from the observed data by dividing the number of trials in which the event occurred by the total number of trials.
- To denote these probabilities, we use notation in the form *P(A)*, which signifies the probability of event *A*. For example,

  P(rain) = 0.30 and *P(spam) = 0.20*.

- The probability of all the possible outcomes of a trial must always sum to 1 because a trial always results in some outcome happening.
- Because an event cannot simultaneously happen and not happen, an event is always mutually exclusive and exhaustive with its **complement**, or the event comprising of the outcomes in which the event of interest does not happen.



The Naive Bayes algorithm

- The **Naive Bayes** algorithm describes a simple method to apply Bayes' theorem to classification problems. Although it is not the only machine learning method that utilizes Bayesian methods, it is the most common one.

Strengths	Weaknesses
• Simple, fast, and very effective. • Does well with noisy and missing data. • Requires relatively few examples for training, but also works well with very large numbers of examples. • Easy to obtain the estimated probability for a prediction.	• Relies on an often-faulty assumption of equally important and independent features. • Not ideal for datasets with many numeric features. • Estimated probabilities are less reliable than the predicted classes.

Strengths

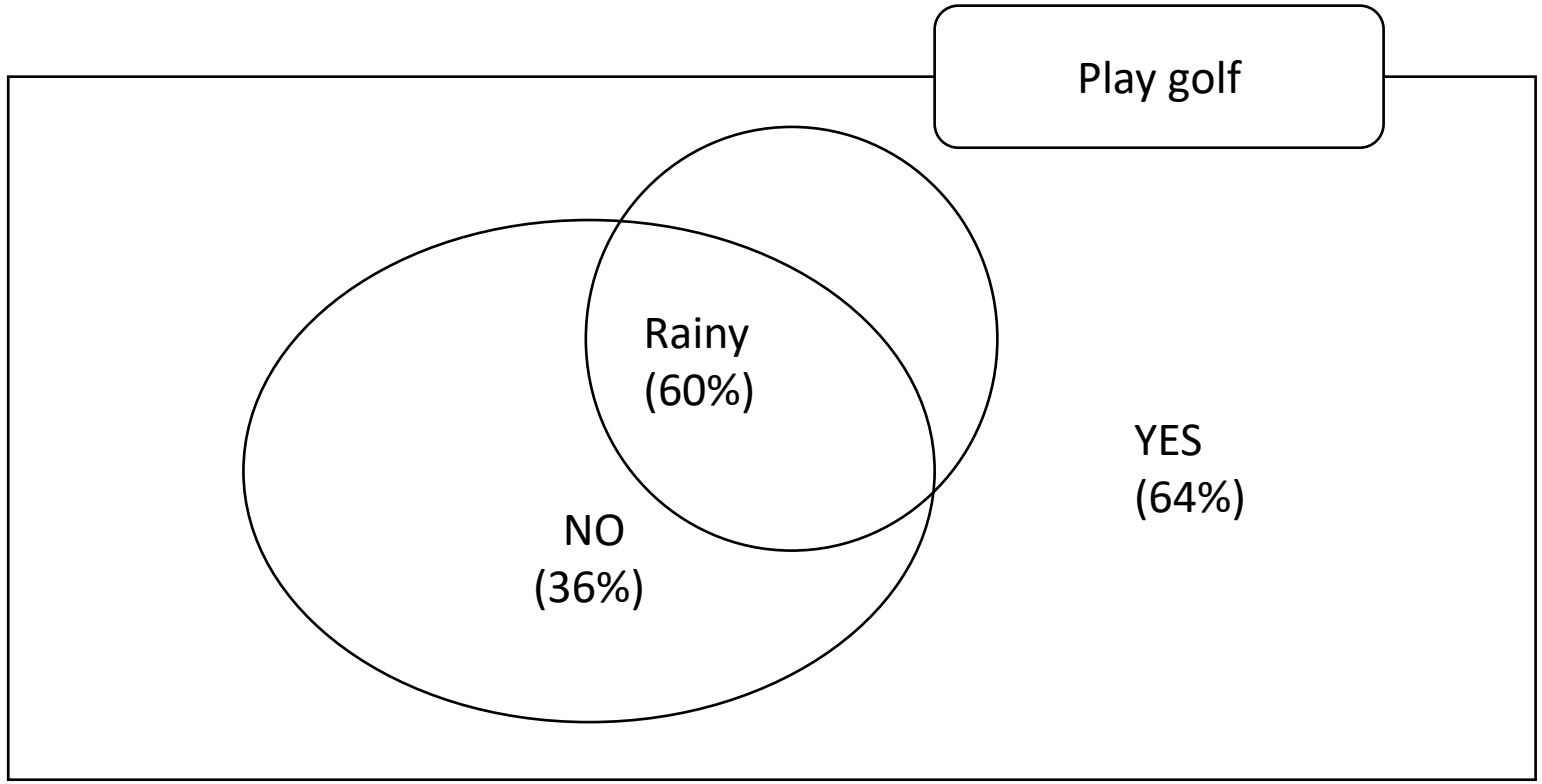
- *Simplicity* - Naive Bayes is very simple to understand and implement. It relies on just Bayes' theorem and conditional independence assumptions between features. This makes it fast to train and easy to interpret.
- *Speed* - Naive Bayes is very fast for training and prediction due to its simplicity. It can handle very large datasets efficiently. This makes it useful for real-time predictions.
- *Effectiveness* - Despite its simplicity, Naive Bayes often performs surprisingly well and can outdo more sophisticated classifiers. The conditional independence assumption rarely holds true in real data, but the algorithm is robust enough to create good models still.
- *Data Efficiency* - Naive Bayes can be trained with a small amount of training data. Since it only estimates a few parameters, it doesn't need huge datasets to estimate probabilities well. This makes it especially useful when data is limited.
- *Handles Missing Data* - Naive Bayes gracefully handles missing data during prediction by ignoring the missing fields. This is useful for real-world data which often has missing values.
- *Probability Estimates* - Naive Bayes provides a probability estimate for each prediction, which is useful for ranking predictions and dealing with uncertainty.

Weaknesses

- *Strong feature independence assumption* - The biggest limitation is that Naive Bayes assumes all features are independent, which is often not true in real data. This independence assumption greatly simplifies computation but can degrade performance if strong feature dependencies exist.
- *Not suitable for continuous data* - Naive Bayes performs poorly if the dataset has many continuous, numerical features. This is because it assumes Gaussian distributions on the features which may not fit real data well.
- *Probability estimates can be extreme* - Since Naive Bayes multiplies probabilities, the estimated probabilities can easily become very small or large for particular combinations of feature values. This can impact the reliability of probability estimates.
- *Overfitting on small training sets* - While Naive Bayes works well with small data, it can easily overfit if the training set is too small. Its simplicity means it has high variance and can model noisy patterns found in tiny training sets.
- *Limited model complexity* - Naive Bayes has a simple linear model structure. So it may not be able to learn more complex non-linear decision boundaries found in many real datasets.


Day
Discount
Free Delivery
Weekdays
weekend
HOLIDAY
Yes
No
Yes
No
Probabilities
Buy
Don't Buy

Understanding joint probability



Understanding joint probability (Probabilistic model)

Naive Bayes is a conditional probability model, it assigns probabilities

$$p(\ C_k\ |x_1, \ldots\ldots, x_n)$$

for each of the *K* possible outcomes or *classes* C_k given a problem instance to be classified, represented by a vector $x = (x_1, \ldots\ldots, x_n)$ encoding some *n* features (independent variables).

Understanding joint probability (Probabilistic model)

The problem with the above formulation is that if the number of features *n* is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. The model must therefore be reformulated to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as:

$$p(C_k \mid x) = \frac{p(x|C_k)p(C_k)}{P(x)} \quad \rightarrow \quad Posterior = \frac{prior\ X\ likelihood}{evidence}$$

Understanding joint probability (Probabilistic model)

In practice, there is interest only in the numerator of that fraction because the denominator does not depend on C and the values of the features X_i are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

- ***P(play Golf= yes)= 9/14 = 0.64***
- ***P(play Golf = no) = 5/14 = 0.36***

outlook	Yes	No
Rainy	2/9	3/5
Overcast	4/9	0/5
sunny	3/9	2/5

temp	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
cool	3/9	1/5

humidity	Yes	No
High	3/9	4/5
normal	6/9	1/5

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

windy	Yes	No
False	6/9	2/5
true	3/9	3/5

Computing conditional probability with bayes' theorem

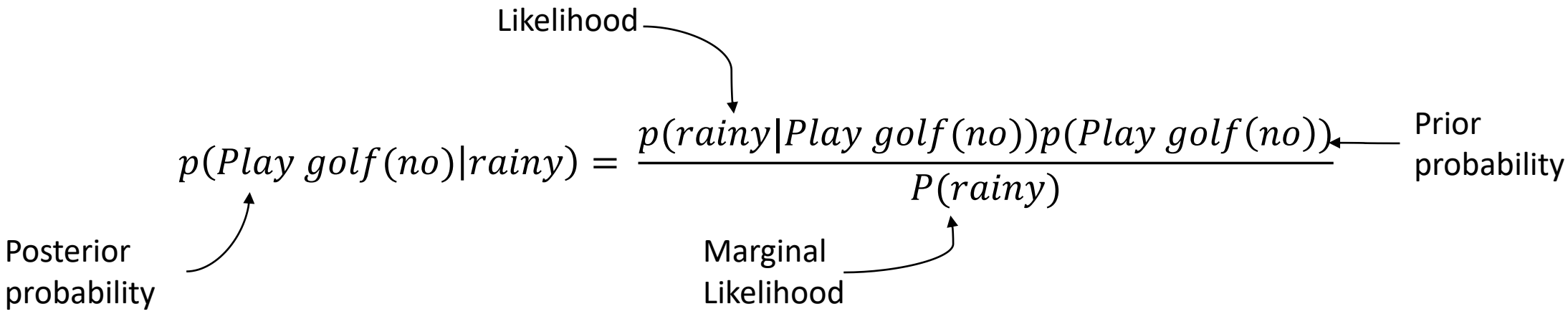
- The relationships between dependent events can be described using **Bayes' theorem**, as shown in the following formula:

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

- The notation *P(A|B)* is read as the probability of event *A*, given that event *B* occurred. This is known as **conditional probability**, since the probability of *A* is dependent (that is, conditional) on what happened with event *B*.

Computing conditional probability with bayes' theorem

- Bayes' theorem tells us that our estimate of *P(A|B)* should be based on *P(A ∩ B)*, a measure of how often *A* and *B* are observed to occur together, and *P(B)*, a measure of how often *B* is observed to occur in general.



frequency	Rainy		
	Yes	no	total
Play golf (yes)	2	7	9
Play golf (no)	3	2	5
Total	5	9	14

likelihood	Rainy		
	Yes	no	total
Play golf (yes)	2/9 0.22	7/9 0.78	9/14 0.64
Play golf (no)	3/5 0.6	2/5 0.4	5/14 0.36
Total	5/14	9/14	14

$$p(Play\ golf(no)|rainy)$$

$$= \frac{p(rainy|Play\ golf(no))p(Play\ golf(no))}{p(rainy|Play\ golf(no))p(Play\ golf\ (no)) + p(rainy|\neg Play\ golf(no)p(\neg Play\ golf(no))}$$

$$p(Play\ golf(no)|rainy) = \frac{0.6 * 0.36}{0.6*0.36+0.22*0.64} = 0.60$$

likelihood	Outlook w1			Temp w2			Humidity w3		Windy w4		
	Rainy w11	Overcast W12	Sunny w13	Hot w21	Mild w22	Cool w23	High w31	Normal w32	False w41	True w42	total
Play golf (yes)	2/9	4/9	3/9	2/9	4/9	3/9	3/9	6/9	6/9	3/9	9/14
Play golf (no)	3/5	0/5	2/5	2/5	2/5	1/5	4/5	1/5	2/5	3/5	5/14
Total	5/14	4/14	5/14	4/14	6/14	4/14	7/14	7/14	8/14	6/14	14

As a new case is received, we need to calculate the posterior probability to determine whether they are more likely to play golf or not, given the likelihood of the available features.
For example, suppose the weather as follows: rainy, cool, high, and windy.
Using Bayes' theorem, we can define the problem as shown in the following formula.
It captures the probability of playing golf, given that *outlook* = *rainy*, *temp* = *cool*, *humidity* = *high*, and *windy* = *true*.

$$p(Play\ golf(no)|w11 \cap w23 \cap w31 \cap w42) = \frac{p(w11 \cap w23 \cap w31 \cap w42|Play\ golf(no))p(Play\ golf(no))}{P(w11 \cap w23 \cap w31 \cap w42)}$$

$$p(Play\ golf(yes)|w11 \cap w23 \cap w31 \cap w42) = \frac{p(w11 \cap w23 \cap w31 \cap w42|Play\ golf(yes))p(Play\ golf(yes))}{P(w11 \cap w23 \cap w31 \cap w42)}$$

likelihood	Outlook w1			Temp w2			Humidity w3		Windy w4		
	Rainy	overcast	sunny	hot	mild	Cool	High	Normal	False	True	total
Play golf (yes)	2/9	4/9	3/9	2/9	4/9	3/9	3/9	6/9	6/9	3/9	9/14
Play golf (no)	3/5	0/5	2/5	2/5	2/5	1/5	4/5	1/5	2/5	3/5	5/14
Total	5/14	4/14	5/14	4/14	6/14	4/14	7/14	7/14	8/14	6/14	14

Where states that *P(A ∩ B) = P(A) * P(B)*.

$$p(Play\ golf(no)|w11 \cap w23 \cap w31 \cap w42) = (3/5 * 1/5 * 4/5 * 3/5 * 5/14) = 0.020$$

$$p(Play\ golf(yes)|w11 \cap w23 \cap w31 \cap w42) = (2/9 * 3/9 * 3/9 * 3/9 * 9/14) = 0.003$$

The probability of not playing golf = 0.020/(0.020+0.003)=0.87

The probability of playing golf = 0.003/ (0.020+0.003)= 0.13

Find a probability of dangerous Fire when there is Smoke

Let's say:

•the probability of dangerous fires are rare (1%)

•but smoke is fairly common (10%) due to barbecues,

•and 90% of dangerous fires make smoke

Can you find the probability of dangerous Fire when there is Smoke?

Answer

We can then discover the **probability of dangerous Fire when there is Smoke** using the Bayes' Theorem:

$$P(Fire \mid Smoke) = P(Fire) * P(Smoke \mid Fire) / P(Smoke)$$

$$= 0.01 * 0.90 / (0.1*0.99)+(0.01*0.90)$$

$$\approx 0.09 \ (9\%)$$

- P(fire)= 0.01
- P(not fire)= 0.99
- P(fire) p(bar)

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

- P(Smoke | Fire)= 0.90
- P(Smoke | bar) = 0.10

P(Fire | Smoke) = P(Fire) * P(Smoke | Fire) / P(Smoke)

- = 0.01 * 0.90 / (0.01*0.90 + 0.10* 0.99)

- After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease and that the test is 99% accurate (i.e., the probability of testing positive when you do have the disease is 0.99, as is the probability of testing negative when you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people of your age. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

- Answer:

$P(\text{test}|\text{disease}) = 0.99$

$P(\neg\text{test}|\neg\text{disease}) = 0.99$

$P(\text{disease}) = 0.0001$

$$P(\text{disease}|\text{test}) = \frac{P(\text{test}|\text{disease})P(\text{disease})}{P(\text{test}|\text{disease})P(\text{disease})+P(\text{test}|\neg\text{disease})P(\neg\text{disease})}$$

$$= \frac{0.99\times0.0001}{0.99\times0.0001+0.01\times0.9999} = 0.009804$$

- **How do we perform Bayesian classification when some features are missing?**

◯ We assuming the missing values as the mean of all values.

◯ We ignore the missing features.

● We integrate the posteriors probabilities over the missing features.

◯ Drop the features completely.

- **True or False: In a naive Bayes algorithm, when an attribute value in the testing record has no example in the training set, then the entire posterior probability will be zero.**

◯ True ◯ false

Which of the following statement is TRUE about the Bayes classifier?

◯ Bayes classifier works on the Bayes theorem of probability.

◯ Bayes classifier is an unsupervised learning algorithm.

◯ Bayes classifier is also known as maximum a priori classifier.

◯ It assumes the independence between the independent variables or features.


THANK
YOU!