Chapter 9

Evidence and Probabilities

증거　　　　　　　　확률

↓

classification

Evidence-Based Classification

- So far we have examined several methods for classification
  – Now we examine a ***different*** way of classification

- Evidence
  – The things that really happened (i.e., the feature values of a data instance)
  – We can think each feature value as evidence for or against a target value

Chills	Runny nose	Headache	Fever	Flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
…	…	…	…	…

연관성 $\propto$ strength

- If we know the ***strength*** of the evidence given by each feature, we can combine them ***probabilistically*** to classify the instance
  – We obtain the strength of each piece of evidence from the training data

(Ex) Online Targeted Advertising (1/6)

- Consider ***targeting online advertisements*** to consumers
  - Based on what webpages they have visited in the past

- Let's consider ***display advertising***
  - The ads that appear on the top, sides, and bottom of pages

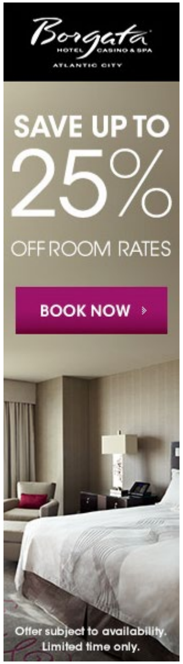
- The characteristics of display advertising
  - It is different from search advertising 쉽다
    - i.e., the ads that appear with the search results
  - The users have not typed in any search keywords
  - Therefore, we need to infer whether the users would be interested in a particular advertisement ***based on their feature values***

추론

(evidence)

(Ex) Online Targeted Advertising (2/6)

- Let's define our ad targeting problem more precisely
- Assume that we are working for ***a very large content provider***
  - Has a wide variety of content
  - Sees many online consumers
  - Can show advertisements to these consumers
  - (ex) Yahoo!, Facebook
- For simplicity, assume we have ***one advertising campaign***
  - For which we would like to target some subset of the online customers that visit our sites
  - This campaign is for the upscale hotel chain, ***Luxhote***
  - The goal of Luxhote is for people to book rooms



(Ex) Online Targeted Advertising (3/6)

- To obtain the training data, we have run this campaign in the past, selecting online consumers ***randomly***

- We now want to run a ***targeted*** campaign
  - Hopefully getting more bookings per dollar spent on ad impressions

- How would you define our ad targeting problem?
  - What will be an instance?
  - What will be the target variable?
  - What will be the features?
  - How will we get the training data?
  - What classification model will we use?

(Ex) Online Targeted Advertising (4/6)

- Now we define our ad targeting problem as follows:
  - **Instance**
    - A consumer 고객 data
  - **Target variable**
    - Did/will the consumer book a Luxhote room within one week after having seen the Luxhote advertisement?
  - **Features** → evidence
    - A key question (will be discussed in the next slide)
  - **Training data**
    - We will have a binary value for the target variable for each consumer (Y/N)
  - **Classification model**
    ↳ evidence-base classifier
    - We will use the ***Naive Bayes classifier*** to estimate the probability that a consumer will book a room after having seen an ad

(Ex) Online Targeted Advertising (5/6)

- A key question feature들의 value가 evidence기 때문
  - What will be the ***features*** we will use to describe the consumers?
    - Such that we might be able to differentiate those that are more or less likely to be good customers for Luxhote

- For this example, we will use the following features:
  - ***The set of content pieces*** that a consumer has viewed (or liked) previously
    - (ex) Jessie = {www.sookmyung.ac.kr, “Avengers: Endgame”, …}
    - Recorded via browser cookies or some other mechanism

- We will use our historical data to estimate both the direction and strength of each piece of evidence (i.e., each content piece)
  - We will then combine it to estimate the likelihood of class membership

(Ex) Online Targeted Advertising (6/6)

- There are many other problems that fit the mold of our example framework
  - Each instance is described by ***a set of pieces of evidence***
  - We combine the strength of each piece of evidence to classify the instance

- Example: ***spam detection*** → data = email
  - Each email is a collection of words
  - Each word provides some evidence for or against if the email is a spam
  - We combine the evidence probabilistically to classify the email
  - Problem definition
    - **Instance**: an email message
    - **Target classes**: *spam* or *not-spam*
    - ***Features***: the words (and symbols) in the email message

핵심단어

Combining Evidence Probabilistically (1/2)

- We want to know the ***probability*** of a consumer booking a room after being shown an ad

- We will represent the probability of an event C as $p(C)$
  - (ex) $p(\text{“}A\ consumer\ books\ a\ room\text{”}) = 0.0001$
    - We would expect about 1 in 10,000 consumers to book rooms

- Now, we are interested in $\boldsymbol{p(C \mid E)}$
  - $p(C \mid E)$: the conditional probability of C given some evidence E
    - $E = \{e_1, e_2, \ldots, e_k\}$: the set of websites visited by a consumer
    - C: an event that the consumer books a room
  - We would expect that $p(C \mid E)$ would be different for different E
    - i.e., consumers who visited different websites will show different behaviors

Combining Evidence Probabilistically (2/2)

- How can we infer $p(C \mid E)$?
  - We would like to use our ***training data*** to infer $p(C \mid E)$
    - (ex) the labeled data from our randomly targeted campaign

- However, this introduces a ***key*** problem
  - For any particular $E = \{e_1, e_2, \ldots, e_k\}$, there may ***not*** be enough cases with exactly the same evidence in the training data
    - (ex) what is the chance that in our training data there are consumers with ***exactly*** the same visiting patterns? → it is infinitesimal (극히 희박)

↳ yes라고 분류된 사람과 똑같은 사이트를 방문한 사람이 얼마나 있겠는가?

- Therefore, we will consider the different pieces of evidence (i.e., e_1, e_2, …, e_k) ***separately***
  - Then combine evidence

Joint Probability (1/2)

결합확률

- Notations
  - A, B: two events
  - $p(A)$, $p(B)$: the probability that A and B occur, respectively
  - $p(AB)$: the probability that ***both*** A and B occur → ***joint probability***

독립

- If events A and B are ***independent***, then $p(AB) = p(A){\cdot}p(B)$
  - Knowing about A or B tells you nothing about the likelihood of the other

서로 영향을 미치지 X

- Example: rolling a ***fair*** die
  - Let A = "roll #1 shows a six" and B = "roll #2 shows a six"
  - Then, $p(A) = 1/6$ and $p(B) = 1/6$
  - Even if we ***know*** that A occurs, still $p(B) = 1/6$
  - In this case, the probability of the joint event is $p(AB) = p(A){\cdot}p(B) = 1/36$

Joint Probability (2/2)

- However, if events A and B are ***not* independent,** then

$$p(AB) = p(A) \cdot p(B \mid A)$$

- Example: rolling a ***trick* die**
  - Suppose we have six trick dice
    모든 face가 똑같은 숫자로 된 주사위 6개
  - Each trick dice has one of the numbers from 1 to 6 on ***all*** faces
  - We pull a die at random and then roll it twice
  - Let A = "roll #1 shows a six" and B = "roll #2 shows a six"
  - Then, $p(A) = 1/6$ and $p(B) = 1/6$
  - However, if we ***know*** that A occurs, $p(B \mid A) = 1.0 \neq p(B)$
    - Since if the first roll was a six, then the second roll is guaranteed to be a six
  - Thus, in this case, $p(AB) = p(A) \cdot p(B \mid A) = 1/6 \cdot 1 = 1/6$

Bayes' Rule (1/2)

베이즈정리

- In $p(AB) = p(A) \cdot p(B \mid A)$, the order of A and B is arbitrary

$$p(AB) = p(A) \cdot p(B \mid A) = p(B) \cdot p(A \mid B)$$

$$p(B \mid A) = \frac{p(A \mid B) \cdot p(B)}{p(A)}$$

- Now we rename A and B with E and H, respectively
  - H: some hypothesis that we want to assess the likelihood of

    가설
  - E: some evidence that we have observed

어떤 E가 벌어졌을때
가설 (event)가 일어날확률

$$p(H \mid E) = \frac{p(E \mid H) \cdot p(H)}{p(E)}$$

Bayes' Rule (2/2)

- This is the famous Bayes' Rule

$$p(H \mid E) = \frac{p(E \mid H) \cdot p(H)}{p(E)}$$

  - Named after the Reverend Thomas Bayes who derived a special case of the rule back in the 18th century

- Meaning
  - We can compute the probability of our hypothesis H given some evidence E by ***instead*** looking at the probability of E given H
    - As well as the probability of H and E
  - Simply speaking, ***we can obtain $p(H \mid E)$ from $p(E \mid H)$, $p(H)$, and $p(E)$***

Importance of Bayes' Rule (1/2)

$$p(H \mid E) = \frac{p(E \mid H) \cdot p(H)}{p(E)}$$

- $p(E \mid H)$, $p(H)$, and $p(E)$ may be ***easier*** to determine than $p(H \mid E)$

- Example: medical diagnosis
  - Assume you're a doctor and a patient arrives with red spots
  - You guess that the patient has measles
  - You want to determine $p(H \mid E)$, where $H =$ "measles" and $E =$ "red spots"
  - However, it is likely impossible to directly estimate $p(H \mid E)$
    - Because we would need to think through all the different cases a person might exhibit red spots and what proportion of them would be measles

Importance of Bayes' Rule (2/2)

$$P(\text{홍역} \mid \text{반점}) = \frac{P(\text{반점} \mid \text{홍역}) \cdot P(\text{홍역})}{P(\text{반점})}$$

(전체인구)

$$p(H \mid E) = \frac{p(E \mid H) \cdot p(H)}{p(E)}$$

- However, consider instead estimating $p(E \mid H)$, $p(H)$ and $p(E)$
  - $p(E \mid H)$: the probability that one has red spots given that one has measles
    - An expert may well know this or be able to estimate it relatively accurately
  - $p(H)$: the probability that someone has measles
    - Simply the prevalence of measles in the population
  - $p(E)$: the probability that someone has red spots
    - Simply the prevalence of red spots in the population
- Thus, Bayes' Rule has made estimating $p(H \mid E)$ much ***easier***
  - $p(E \mid H)$, $p(H)$, and $p(E)$ are much easier to estimate than $p(H \mid E)$ is

Applying Bayes' Rule to Classification (1/3)

- A large portion of data science is based on ***Bayesian*** methods
  - They have at their core reasoning based on Bayes' Rule

- Bayes' Rule for ***classification***

evidence들을 보고 classify할 확률

$$p(C = c \mid E) = \frac{p(E \mid C = c) \cdot p(C = c)}{p(E)}$$

C가 c라고 classify

  - $\boxed{C = c}$: the event that the value of the target variable is c
    - (ex) $C =$ "YES" or $C =$ "NO"
  - E: the evidence (i.e., the vector of feature values)
  - $p(C = c \mid E)$: the probability that $C = c$ ***given*** the evidence E
    - This is the quantity we would like to estimate
    - Called the ***posterior*** probability

사후확률

Applying Bayes' Rule to Classification (2/3)

- Bayes' Rule for ***classification*** (cont'd)

구하기 쉬운 사전확률을 이용해
구하기 어려운 사후확률을 구하는 베이즈정리

$$p(C = c \mid E) = \frac{p(E \mid C = c) \cdot p(C = c)}{p(E)}$$

사전확률

  - $p(C = c)$: the "prior" probability of the class c : 이미 알고있음
    - The percentage of all examples that are of class c
  - $p(E \mid C = c)$: the likelihood of seeing the evidence E when the class $C = c$
    - The percentage of examples of class c that have feature vector E
  - $p(E)$: the likelihood of E (i.e., how common is E among all examples?)
    - The percentage occurrence of E among all examples

✓ $p(C = c)$, $p(E \mid C = c)$, and $p(E)$ can be calculated easily from training data

Applying Bayes' Rule to Classification (3/3)

사후확률

- The posterior probability $p(C = c \mid E)$
  - Can be used directly as an estimate of class probability
  - Can be used as a score to rank instances
  - Or, we could choose the maximum $p(C = c \mid E)$ across the different values c as the classification
    - (ex) $p(C = \text{"Yes"} \mid E) = 0.7$, $p(C = \text{"No"} \mid E) = 0.3$ → We determine $C = \text{"Yes"}$

- Classification example: spam detection
  - $w_1, \ldots, w_n$: the words in an email
  - The probability that the email is spam is:

$$p(c \mid E) = \frac{p(E \mid c) \cdot p(c)}{p(E)}$$

$$p(\text{Spam} \mid w_1, \ldots, w_n) = \frac{p(w_1, \ldots, w_n \mid \text{Spam})\, p(\text{Spam})}{p(w_1, \ldots, w_n)}$$

Major Difficulty in Computing $p(C = c \mid \mathbf{E})$

$$p(C = c \mid E) = \frac{p(E \mid C = c) \cdot p(C = c)}{p(E)}$$

- Let $E = \langle e_1, e_2, \ldots, e_k \rangle$
  - E is a possibly large, specific collections of conditions
  - Then, $p(E \mid C = c) = p(e_1, e_2, \ldots, e_k \mid C = c)$

evidence랑 모두 들어맞기 어려움 ↓ 대표성↓

- However, it is difficult to measure $\boldsymbol{p(e_1, e_2, \ldots, e_k \mid C = c)}$
  - ★ We may ***never*** see a specific example in the training data that exactly matches a given $E = \langle e_1, e_2, \ldots, e_k \rangle$
  - Even if we do, it may be ***unlikely*** we'll see enough of them to estimate a probability with any confidence

⇒ naive Bayes' rule 이용 (다음시간)

Conditional Independence (1/2)

조건부 + 독립

- The conditional probability that ***both*** A and B occur given C

$$p(AB \mid C) = p(A \mid C) \cdot p(B \mid AC) = p(B \mid C) \cdot p(A \mid BC)$$

  – A and B are ***not independent*** when C occurs

- However, if A and B are ***conditionally independent*** given C, then

A,B가 독립, 조건부(C) 확률

$$p(AB \mid C) = p(A \mid C) \cdot p(B \mid C) = p(B \mid C) \cdot p(A \mid C)$$

  – A and B are ***independent*** when C occurs

✓ The second equation is much ***easier*** to compute probabilities from the data

Conditional Independence (2/2)

Bayes' Rule의 부족한점 보완

- Thus, Bayesian methods for data science deal with this issue by making a $\boxed{\textit{strong assumption}}$ of conditional independence

$$p(E \mid c) = p(e_1 \wedge e_2 \wedge \ldots \wedge e_k \mid c)$$
$$= p(e_1 \mid c) \cdot p(e_2 \mid c) \cdot \ldots \cdot p(e_k \mid c)$$

  - We assume that each e_i is ***independent*** of every other e_j given the class c
  - For simplicity of presentation, we replace $C = c$ simply by c
  - Each $p(e_i \mid c)$ can be computed ***directly*** from the data
    - Now we don't need to look for an entire matching feature vector
    - Simply count up the proportion of the time that we see individual feature e_i in the instances of class c
    - There are likely to be relatively many occurrences of e_i

(simple)

Naive Bayes Classifier (1/2)

베이즈정리 + E의 각 요소가 모두 조건부독립이라는 가정

- Combining the previous equation with Bayes Rule, we get the ***Naive Bayes equation*** as follows:

$$p(c \mid E) = \frac{\boldsymbol{p(E \mid c)} \cdot p(c)}{p(E)}$$

$$= \frac{\boldsymbol{p(e_1 \mid c)} \cdot \boldsymbol{p(e_2 \mid c)} \cdot \boldsymbol{\ldots} \cdot \boldsymbol{p(e_k \mid c)} \cdot p(c)}{p(E)}$$

- The Naive Bayes classifier
  - Using the Native Bayes equation, estimate the probability that the example belongs to ***each*** class
  - Report the class with ***highest*** probability

ex. yes 30%, no 50%. (각각 계산해야함. yes+no ≠ 100%.)

→ "no"

Naive Bayes Classifier (2/2)

- If we are interested only in classification, we can just look to see which ***numerator*** is larger 상대적으로 비교. yes(20%), no(10%)
  - Because $p(E)$ is the ***same*** for all classes →yes. 퍼센트는 중요하지 않음

$$p(c \mid E) = \frac{p(e_1 \mid c) \cdot \ p(e_2 \mid c) \ \cdot \ \dots \ \cdot \ p(e_k \mid c) \cdot p(c)}{\boldsymbol{p(E)}}$$

$$\approx p(e_1 \mid c) \cdot \ p(e_2 \mid c) \ \cdot \ \dots \ \cdot \ p(e_k \mid c) \cdot p(c)$$

분모를 없애버림. 어차피 같고 비교하는데에는 불필요하다.

- Thus, in practice, we compute ***only*** the numerator
  - Because the denominator $p(E)$ is effectively constant for all classes

Advantages of Naive Bayes (1/2)

- It is a very ***simple*** classifier f 보고자하는 E는 모두 고려하면서도
  - Yet it still takes all the feature evidence into account

- It is very ***efficient*** in terms of storage space and execution time
  - **Training**: consists only of storing $p(c)$ and $p(e_i \mid c)$ for each c and e_i
    - $p(c)$: we count the proportions of examples of class c among all examples
    - $p(e_i \mid c)$: we count the proportion of examples in class c for which e_i appears
  - **Classification**: requires only simple multiplications of them

C 그럼에도

- In spite of its simplicity and the strict independence assumption, it performs ***surprisingly well*** on many real-world tasks
  - Because the violation of the independence assumption tends not to hurt classification performance
  - What if two pieces of evidence are actually NOT independent and we treat them as being independent? ➔double counting of the evidence
    - However, double counting will not tend to hurt us (i.e., probability will be simply overestimated) B이기 이전에 A에 포함된 것 (사실상 count x2) ↳ 조금 더 크게 측정되지만 상대적인 비교가 중요하므로
    - E.g.) P(AB) = P(A) x P(B|A) vs P(AB) = P(A) P(B)

0.3

Advantages of Naive Bayes (2/2)

- However, due to the independence assumption, the probability estimates themselves should ***not*** be considered accurate 확률값 자체 X
  – Thus, practitioners use Naïve Bayes regularly for ***ranking***, where the actual values of the probabilities are not relevant
  상대적 비교가 중요

- It is naturally an "***incremental learner***" 새로운 data가 들어왔을 때 전체를 다시 돌릴 필요 X
  – i.e., we can update the model one training example at a time
  – It does not need to reprocess all past training examples when new training data become available
  – Especially advantageous in applications where we would like to update the model whenever new labeled data become available 실시간 update
  – (ex) Creating a personalized spam email classifier
    - New labeled data become available when the user clicks the "spam" button in her browser

Example (1): Weather Forecast (1/7)

- Suppose we have the following weather dataset:

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

- How can we predict the class of the following instance?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Cool	High	True	?

Example (1): Weather Forecast (2/7)

- First, we compute all necessary probabilities as follows:

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

개수

  - $p(c)$: the prior probability of the class c
    - $\boxed{p(\text{Yes}) = 9/14,\ \ p(\text{No}) = 5/14}$

$p(C|E) \approx p(e_1|C) \cdot p(e_2|C) \cdot \cdots p(e_k|C) \cdot p(C)$

Example (1): Weather Forecast (3/7)

- First, we compute all necessary probabilities as follows:

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

  – $p(e_i \mid c)$ for "Outlook" feature
    - $p(\text{Sunny} \mid \text{Yes}) = 3/9$, $p(\text{Overcast} \mid \text{Yes}) = 4/9$, $p(\text{Rainy} \mid \text{Yes}) = 2/9$
    - $p(\text{Sunny} \mid \text{No}) = 2/5$, $p(\text{Overcast} \mid \text{No}) = 0/5$, $p(\text{Rainy} \mid \text{No}) = 3/5$

Example (1): Weather Forecast (4/7)

- First, we compute all necessary probabilities as follows:

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

  - $p(e_i \mid c)$ for "Temp" feature
    - $p(\text{Hot} \mid \text{Yes}) = 2/9,\ p(\text{Mild} \mid \text{Yes}) = 4/9,\ p(\text{Cool} \mid \text{Yes}) = 3/9$
    - $p(\text{Hot} \mid \text{No}) = 2/5,\ p(\text{Mild} \mid \text{No}) = 2/5,\ p(\text{Cool} \mid \text{No}) = 1/5$

Example (1): Weather Forecast (5/7)

- First, we compute all necessary probabilities as follows:

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

– $p(e_i \mid c)$ for "Humidity" feature

- $p(\text{High} \mid \text{Yes}) = 3/9,\ \ p(\text{Normal} \mid \text{Yes}) = 6/9$
- $p(\text{High} \mid \text{No}) = 4/5,\ \ p(\text{Normal} \mid \text{No}) = 1/5$

Example (1): Weather Forecast (6/7)

- First, we compute all necessary probabilities as follows:

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

– $p(e_i \mid c)$ for "Windy" feature

- $p(\text{True} \mid \text{Yes}) = 3/9$, $p(\text{False} \mid \text{Yes}) = 6/9$
- $p(\text{True} \mid \text{No}) = 2/5$, $p(\text{False} \mid \text{No}) = 3/5$

Example (1): Weather Forecast (7/7)

- Finally, we compute $p(c \mid E)$ for each class c (i.e., Yes, No)

$$p(\text{Yes} \mid \text{Rainy}, \text{Cool}, \text{High}, \text{True})$$
$$\approx p(\text{Rainy} \mid \text{Yes}) \cdot p(\text{Cool} \mid \text{Yes}) \cdot p(\text{High} \mid \text{Yes}) \cdot p(\text{True} \mid \text{Yes}) \cdot p(\text{Yes})$$
$$= 2/9 \cdot 3/9 \cdot 3/9 \cdot 3/9 \cdot 9/14 = 0.00529$$

$$p(\text{No} \mid \text{Rainy}, \text{Cool}, \text{High}, \text{True})$$
$$\approx p(\text{Rainy} \mid \text{No}) \cdot p(\text{Cool} \mid \text{No}) \cdot p(\text{High} \mid \text{No}) \cdot p(\text{True} \mid \text{No}) \cdot p(\text{No})$$
$$= 3/5 \cdot 1/5 \cdot 4/5 \cdot 3/5 \cdot 5/14 = 0.02057$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Cool	High	True	?

- Because $p(\text{Yes} \mid \text{Rainy, Cool, High, True}) < p(\text{No} \mid \text{Rainy, Cool, High, True})$, we determine that **Play Golf = No**

Example (2): Patient Diagnosis (1/7)

- Suppose we have a patient dataset as follows:

Chills	Runny nose	Headache	Fever	Flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

- How can we predict the class of the following instance?

Chills	Runny nose	Headache	Fever	Flu?
Y	N	Mild	N	?

Example (2): Patient Diagnosis (2/7)

- First, we compute all necessary probabilities as follows:

Chills	Runny nose	Headache	Fever	Flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

  - $p(c)$: the prior probability of the class c
    - $p(\text{Flu} = \text{Y}) = 5/8,\ p(\text{Flu} = \text{N}) = 3/8$

Example (2): Patient Diagnosis (3/7)

- First, we compute all necessary probabilities as follows:

Chills	Runny nose	Headache	Fever	Flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

– $p(e_i \mid c)$ for "Chills" feature
  - $p(\text{Chills} = \text{Y} \mid \text{Flu} = \text{Y}) = 3/5,\ p(\text{Chills} = \text{N} \mid \text{Flu} = \text{Y}) = 2/5$
  - $p(\text{Chills} = \text{Y} \mid \text{Flu} = \text{N}) = 1/3,\ p(\text{Chills} = \text{N} \mid \text{Flu} = \text{N}) = 2/3$

Example (2): Patient Diagnosis (4/7)

- First, we compute all necessary probabilities as follows:

Chills	Runny nose	Headache	Fever	Flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

– $p(e_i \mid c)$ for "Runny nose" feature

- $p(\text{Runny nose} = \text{Y} \mid \text{Flu} = \text{Y}) = 4/5$, $p(\text{Runny nose} = \text{N} \mid \text{Flu} = \text{Y}) = 1/5$
- $p(\text{Runny nose} = \text{Y} \mid \text{Flu} = \text{N}) = 1/3$, $p(\text{Runny nose} = \text{N} \mid \text{Flu} = \text{N}) = 2/3$

Example (2): Patient Diagnosis (5/7)

- First, we compute all necessary probabilities as follows:

Chills	Runny nose	Headache	Fever	Flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

– $p(e_i \mid c)$ for "Headache" feature

- $p(\text{Headache} = \text{No} \mid \text{Flu} = \text{Y}) = 1/5$, $p(\text{Headache} = \text{Mild} \mid \text{Flu} = \text{Y}) = 2/5$, $p(\text{Headache} = \text{Strong} \mid \text{Flu} = \text{Y}) = 2/5$
- $p(\text{Headache} = \text{No} \mid \text{Flu} = \text{N}) = 1/3$, $p(\text{Headache} = \text{Mild} \mid \text{Flu} = \text{N}) = 1/3$, $p(\text{Headache} = \text{Strong} \mid \text{Flu} = \text{N}) = 1/3$

Example (2): Patient Diagnosis (6/7)

- First, we compute all necessary probabilities as follows:

Chills	Runny nose	Headache	Fever	Flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

  - $p(e_i \mid c)$ for "Fever" feature
    - $p(\text{Fever} = \text{Y} \mid \text{Flu} = \text{Y}) = 4/5,\ \ p(\text{Fever} = \text{N} \mid \text{Flu} = \text{Y}) = 1/5$
    - $p(\text{Fever} = \text{Y} \mid \text{Flu} = \text{N}) = 1/3,\ \ p(\text{Fever} = \text{N} \mid \text{Flu} = \text{N}) = 2/3$

Example (2): Patient Diagnosis (7/7)

- Finally, we compute $p(c \mid E)$ for each class c (i.e., Y, N)

$$p(\text{Flu} = \text{Y} \mid \text{Chills} = \text{Y}, \text{Runny nose} = \text{N}, \text{Headache} = \text{Mild}, \text{Fever} = \text{N})$$
$$\approx p(\text{Chills} = \text{Y} \mid \text{Flu} = \text{Y}) \cdot p(\text{Runny nose} = \text{N} \mid \text{Flu} = \text{Y}) \cdot$$
$$p(\text{Headache} = \text{Mild} \mid \text{Flu} = \text{Y}) \cdot p(\text{Fever} = \text{N} \mid \text{Flu} = \text{Y}) \cdot p(\text{Flu} = \text{Y})$$
$$= 3/5 \cdot 1/5 \cdot 2/5 \cdot 1/5 \cdot 5/8 = 0.006$$

$$p(\text{Flu} = \text{N} \mid \text{Chills} = \text{Y}, \text{Runny nose} = \text{N}, \text{Headache} = \text{Mild}, \text{Fever} = \text{N})$$
$$\approx p(\text{Chills} = \text{Y} \mid \text{Flu} = \text{N}) \cdot p(\text{Runny nose} = \text{N} \mid \text{Flu} = \text{N}) \cdot$$
$$p(\text{Headache} = \text{Mild} \mid \text{Flu} = \text{N}) \cdot p(\text{Fever} = \text{N} \mid \text{Flu} = \text{N}) \cdot p(\text{Flu} = \text{N})$$
$$= 1/3 \cdot 2/3 \cdot 1/3 \cdot 2/3 \cdot 3/8 \approx 0.0185$$

Chills	Runny nose	Headache	Fever	Flu?
Y	N	Mild	N	?

- Because $p(\text{Flu} = \text{Y} \mid E) < p(\text{Flu} = \text{N} \mid E)$, we determine that **Flu = N**

Example (3): Spam Detection (1/9)

- Suppose we have a spam dataset as follows:

Email content	Spam?
"send us your password"	spam
"send us your review"	ham (not spam)
"review your password"	ham
"review us"	spam
"send your password"	spam
"send us your account"	spam

- How can we predict the class of the following instance?

Email content	Spam?
"review your account"	?

Example (3): Spam Detection (2/9)

- First, we compute all necessary probabilities as follows:

Email content	Spam?
“send us your password”	spam
“send us your review”	ham
“review your password”	ham
“review us”	spam
“send your password”	spam
“send us your account”	spam

– $p(c)$: the prior probability of the class c
  - $p(\text{spam}) = 4/6$, $p(\text{ham}) = 2/6$

$p(c|E) \approx p(e_1|c) \cdot p(e_2|c) \cdot \dots \cdot p(e_k|c) \cdot p(c)$

Example (3): Spam Detection (3/9)

- First, we compute all necessary probabilities as follows:

Email content	Spam?
“send us your password”	spam
“send us your review”	ham
“review your password”	ham
“review us”	spam
“send your password”	spam
“send us your account”	spam

각각의 단어가 evidence가 됨

– $p(e_i \mid c)$ for the word “password”
  - $p(\text{password} \mid \text{spam}) = 2/4$, $p(\neg\text{password} \mid \text{spam}) = 2/4$
  - $p(\text{password} \mid \text{ham}) = 1/2$, $p(\neg\text{password} \mid \text{ham}) = 1/2$

(not)

Example (3): Spam Detection (4/9)

- First, we compute all necessary probabilities as follows:

Email content	Spam?
“send us your password”	spam
“send us your review”	ham
“review your password”	ham
“review us”	spam
“send your password”	spam
“send us your account”	spam

– $p(e_i \mid c)$ for the word “review”

  - $p(\text{review} \mid \text{spam}) = 1/4$, $p(\neg\text{review} \mid \text{spam}) = 3/4$
  - $p(\text{review} \mid \text{ham}) = 2/2$, $p(\neg\text{review} \mid \text{ham}) = 0/2$

Example (3): Spam Detection (5/9)

- First, we compute all necessary probabilities as follows:

Email content	Spam?
“send us your password”	spam
“send us your review”	ham
“review your password”	ham
“review us”	spam
“send your password”	spam
“send us your account”	spam

– $p(e_i \mid c)$ for the word “send”

- $p(\text{send} \mid \text{spam}) = 3/4$, $p(\neg\text{send} \mid \text{spam}) = 1/4$
- $p(\text{send} \mid \text{ham}) = 1/2$, $p(\neg\text{send} \mid \text{ham}) = 1/2$

Example (3): Spam Detection (6/9)

- First, we compute all necessary probabilities as follows:

Email content	Spam?
“send us your password”	spam
“send us your review”	ham
“review your password”	ham
“review us”	spam
“send your password”	spam
“send us your account”	spam

– $p(e_i \mid c)$ for the word “us”

- $p(\mathrm{us} \mid \mathrm{spam}) = 3/4$, $p(\neg\mathrm{us} \mid \mathrm{spam}) = 1/4$
- $p(\mathrm{us} \mid \mathrm{ham}) = 1/2$, $p(\neg\mathrm{us} \mid \mathrm{ham}) = 1/2$

Example (3): Spam Detection (7/9)

- First, we compute all necessary probabilities as follows:

Email content	Spam?
"send us your password"	spam
"send us your review"	ham
"review your password"	ham
"review us"	spam
"send your password"	spam
"send us your account"	spam

– $p(e_i \mid c)$ for the word "your"

- $p(\text{your} \mid \text{spam}) = 3/4$, $p(\neg\text{your} \mid \text{spam}) = 1/4$
- $p(\text{your} \mid \text{ham}) = 2/2$, $p(\neg\text{your} \mid \text{ham}) = 0/2$

Example (3): Spam Detection (8/9)

- First, we compute all necessary probabilities as follows:

Email content	Spam?
"send us your password"	spam
"send us your review"	ham
"review your password"	ham
"review us"	spam
"send your password"	spam
"send us your account"	spam

– $p(e_i \mid c)$ for the word "account"
  - $p(\text{account} \mid \text{spam}) = 1/4$, $p(\neg\text{account} \mid \text{spam}) = 3/4$
  - $p(\text{account} \mid \text{ham}) = 0/2$, $p(\neg\text{account} \mid \text{ham}) = 2/2$

Example (3): Spam Detection (9/9)

- Finally, we compute $p(c \mid E)$ for each class c (i.e., spam, ham)

$$
\begin{aligned}
&p(\text{spam} \mid \neg\text{password}, \text{review}, \neg\text{send}, \neg\text{us}, \text{your}, \text{account}) \\
&\quad \approx p(\neg\text{password} \mid \text{spam}) \cdot p(\text{review} \mid \text{spam}) \cdot p(\neg\text{send} \mid \text{spam}) \cdot \\
&\quad\quad p(\neg\text{us} \mid \text{spam}) \cdot p(\text{your} \mid \text{spam}) \cdot p(\text{account} \mid \text{spam}) \cdot p(\text{spam}) \\
&\quad = 2/4 \cdot 1/4 \cdot 1/4 \cdot 1/4 \cdot 3/4 \cdot 1/4 \cdot 4/6 \approx 0.000976
\end{aligned}
$$

$$
\begin{aligned}
&p(\text{ham} \mid \neg\text{password}, \text{review}, \neg\text{send}, \neg\text{us}, \text{your}, \text{account}) \\
&\quad \approx p(\neg\text{password} \mid \text{ham}) \cdot p(\text{review} \mid \text{ham}) \cdot p(\neg\text{send} \mid \text{ham}) \cdot \\
&\quad\quad p(\neg\text{us} \mid \text{ham}) \cdot p(\text{your} \mid \text{ham}) \cdot p(\text{account} \mid \text{ham}) \cdot p(\text{ham}) \\
&\quad = 1/2 \cdot 2/2 \cdot 1/2 \cdot 1/2 \cdot 2/2 \cdot 0/2 \cdot 2/6 = 0
\end{aligned}
$$

Email content	Spam?
"review your account"	?

- Because $p(\text{spam} \mid E) > p(\text{ham} \mid E)$, we determine that the email is **spam**

A Model of Evidence "Lift"

$\frac{\text{positive}}{\text{Sampling(targeting)}}$

증거향상도

- Recall the notion of ***lift*** as a metric for evaluating a classifier
  - i.e., measures how much more prevalent the positive class is in the selected subpopulation over the prevalence in the population as a whole

- We can consider Naive Bayes as a product of ***evidence lifts***

$$p(c \mid E) = \frac{p(e_1 \mid c) \cdot p(e_2 \mid c) \cdot \ldots \cdot p(e_k \mid c) \cdot p(c)}{p(E)}$$

$$= \frac{p(e_1 \mid c) \cdot p(e_2 \mid c) \cdot \ldots \cdot p(e_k \mid c) \cdot p(c)}{p(e_1) \cdot p(e_2) \cdot \ldots \cdot p(e_k)}$$

  - Note that we further assume full feature independence
    - i.e., $p(E) = p(e_1, e_2, \ldots, e_k) = p(e_1) \cdot p(e_2) \cdot \ldots \cdot p(e_k)$

Probability as a Product of Evidence Lifts

- The terms in the previous equation can be rearranged to yield:

$$p(c \mid E) = \frac{p(e_1 \mid c) \cdot p(e_2 \mid c) \cdot \ldots \cdot p(e_k \mid c) \cdot p(c)}{p(e_1) \cdot p(e_2) \cdot \ldots \cdot p(e_k)}$$

$$= p(c) \cdot \frac{p(e_1 \mid c)}{p(e_1)} \cdot \frac{p(e_2 \mid c)}{p(e_2)} \cdot \cdots \cdot \frac{p(e_k \mid c)}{p(e_k)}$$

$$= p(c) \cdot \mathrm{lift}_c(e_1) \cdot \mathrm{lift}_c(e_2) \cdot \cdots \cdot \mathrm{lift}_c(e_k)$$

class c의 evidence lift
: 해당클래스로 classify할때 각각의 항이
얼만큼 lift시켜주는지 (도움을주는지)

  – where $\mathrm{lift}_c(e_i)$ is defined as:

$$\mathrm{lift}_c(e_i) = \frac{p(e_i \mid c)}{p(e_i)}$$

  – FYI, definition of lift

$$Lift = \frac{The\ percentage\ of\ positive\ instances\ targeted}{The\ percentage\ of\ instances\ targeted}$$

Probability as a Product of Evidence Lifts

- How these evidence lifts apply a new example $E = <e_1, e_2, \ldots, e_k>$?

$$p(c \mid E) = p(c) \cdot \frac{p(e_1 \mid c)}{p(e_1)} \cdot \frac{p(e_2 \mid c)}{p(e_2)} \cdot \cdots \cdot \frac{p(e_k \mid c)}{p(e_k)}$$

$$= p(c) \cdot \mathrm{lift}_c(e_1) \cdot \mathrm{lift}_c(e_2) \cdot \cdots \cdot \mathrm{lift}_c(e_k)$$

- Interpretation

"yes"라고 분류될 사전확률

  - We start at the prior probability, $p(c)$, and go through our example E
  - Each piece of evidence, e_i, raises or lowers the probability of the class
    - By a factor equal to $\mathrm{lift}_c(e_i)$
  - If $\mathrm{lift}_c(e_i) > 1$, then the probability is increased
  - If $\mathrm{lift}_c(e_i) < 1$, then the probability is diminished

(Ex) Evidence Lifts from Facebook "Likes"

- Let's examine some evidence lifts from real data

- Researchers recently published a paper showing striking results
  - Michal Kosinski et al., "Private traits and attributes are predictable from digital records of human behavior," *National Academy of Sciences*, 2013.

    ex. facebook의 Like

- What people "Like" on Facebook is quite predictive of all manner of traits that usually are not directly apparent:
  - How they score on intelligence tests
  - How they score on psychometric tests
  - Whether they are (openly) gay
  - Whether they drink alcohol or smoke
  - Their religion and political views

(Ex) Evidence Lifts from Facebook "Likes"

- What are the Likes that give strong evidence lifts for "high IQ"?
  - Some Facebook page "Likes" that give the highest evidence lifts

Like	Lift	Like	Lift
Lord Of The Rings	1.69	Wikileaks	1.59
One Manga	1.57	Beethoven	1.52
Science	1.49	NPR	1.48
Psychology	1.46	*Spirited Away*	1.45
The Big Bang Theory	1.43	Running	1.41
Paulo Coelho	1.41	Roger Federer	1.40
The Daily Show	1.40	*Star Trek (Movie)*	1.39
Lost	1.39	Philosophy	1.38
Lie to Me	1.37	*The Onion*	1.37
How I Met Your Mother	1.35	*The Colbert Report*	1.35
Doctor Who	1.34	*Star Trek*	1.32
Howl's Moving Castle	1.31	Sheldon Cooper	1.30
Tron	1.28	*Fight Club*	1.26
Angry Birds	1.25	*Inception*	1.25
The Godfather	1.23	*Weeds*	1.22

lift > 1
→ boost

칼텍...

The probability of a high-IQ person liking "Sheldon Cooper" is 30% higher than the probability in the general population

(Ex) Evidence Lifts from Facebook "Likes"

- Taking a sample of the Facebook population, if we define our target variable as the binary variable IQ > 130, about 14% of the sample is positive → base rate
- The independence assumption made, we can calculate the probability that someone has very high IQ based on the things they Like.
  - If I Like nothing, the my estimated probability of IQ > 130 is just the base rate the population (i.e., 14%)
  - If on Facebook I had Liked "Sheldon Cooper", then my estimated probability would increase by 30% to 0.14 x 1.3 = 18%
  - If I have three Likes (Sheldon Cooper, Star Trek, and the Lord of the Rings), then my estimated probability of IQ > 130 increases to 0.14 x 1.3 x 1.39 x 1.69 = 43%

Summary (1/2)

- Modeling techniques in prior chapters
  - Ask "What is the best way to ***distinguish*** (segment) target values?"
  - Classification trees, linear classifiers, etc.
  - These are termed ***discriminative*** methods
    - i.e., they try directly to discriminate different targets

- A new family of methods introduced in this chapter
  - Asks "How do different target segments ***generate*** feature values?"
  - They attempt to model how the data was generated
  - When faced with a new example to be classified, they use the models to answer the question: "Which class most likely generate this example?"
  - Thus, this approach is called ***generative***

Summary (2/2)

- Bayesian methods
  - A family of popular generative methods, which depend on ***Bayes' Rule***
- Naive Bayes classifier
  - A particularly common and simple but very useful Bayesian method
  - It is "naive" in the sense that it simply assumes conditional independence
  - Because of its simplicity, it is very fast and efficient
  - Furthermore, in spite of its naïveté, it is surprisingly effective
  - Thus, in data science, it is a common "baseline" method
- Evidence lifts
  - Used to examine large number of possible pieces of evidence for or against a conclusion

ex. IQ가 120이다 라는 결론