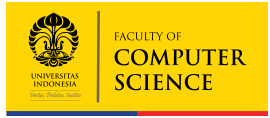
Probability, Likelihood, and Naive Bayes

Adila A. Krisnadhi*, Siti Aminah, Aruni Y. Azizah,
Dina Chahyati, Fariz Darari

Faculty of Computer Science
Universitas Indonesia

CSGE603130 - Kecerdasan Artifisial dan Sains Data Dasar
2022-11-14



Outline

Outline

Review: Probability

Consider an experiment/trial of throwing (once) two six-sided dice at-once.

- **Sample space**: set of all possible **outcomes** of a trial.
  - The sample space of the above experiment is $\Omega = \{(x, y) \mid x, y \in \{1, \ldots, 6\}\} = \{(1, 1), (1, 2), \ldots, (6, 6)\}$ with 36 poossible outcomes where for each outcome (x, y), x is the value of die 1 (number on the side facing up), while y is the value of die 2.
- An **event** is a subset of a sample space. The **event space** of Ω is the set of all events from Ω.
  - Event space of Ω is 2^{Ω}, i.e., the set of all subsets of Ω, containing in total 2^{36} possible events.
  - Event A: "die 1 lands on side 3"
    $A = \{(3, y) \mid y \in \{1, \ldots, 6\}\} = \{(3, 1), \ldots, (3, 6)\}$.
  - Event B: "value of die 2 is at least 2 and at most 4"
    $B = \{(1, 2), (1, 3), (1, 4), (2, 2), (2, 3), (2, 4), \ldots, (6, 4)\}$.
  - Event C: "two dice sum to more than 9"
    $C = \{(x, y) \mid x + y > 9, x, y \in \{1, \ldots, 6\}\} = \{(4, 6), (5, 5), (6, 4), (5, 6), (6, 5), (6, 6)\}$.

Review: Probability (contd.)

- Let A be an event from a sample space Ω. The **probability** of A, written $P(A)$, denotes
  - the degree with which we subjectively believe that the event A holds [**Bayesian interpretation**]; or
  - the fraction of times A will occur in the long run [**frequentist interpretation**].
- Axioms of probability (with respect to a sample space Ω):
  - $P(A) \geqslant 0$ for all event $A \subseteq \Omega$.
  - $P(\Omega) = 1$
  - $P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$, for **mutually exclusive** events $A_1, A_2, \dots$.
- Some consequences:
  - If $A \subseteq B$, then $P(A) \leqslant P(B)$.
  - $P(\emptyset) = 0$.
  - $0 \leqslant P(A) \leqslant 1$ for all events A.
  - The probability of A or B happening is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Here, $P(A \cap B)$ is the probability of A and B both happening, called the **joint probability** of event A and B, often written $P(A, B)$.
  - If A^c is the complement/negation of A, then $P(A^c) = 1 - P(A)$ (i.e., the probability of A not happening).

Review: Probability (contd.)

Consider the experiment of throwing two six-sided dice at-once where both dice are fair.

- Event A: “die 1 lands on side 3”
  $P(A) =$

Review: Probability (contd.)

Consider the experiment of throwing two six-sided dice at-once where both dice are fair.

- Event A: “die 1 lands on side 3”
  $P(A) =$ 6/36 = 1/6.

Review: Probability (contd.)

Consider the experiment of throwing two six-sided dice at-once where both dice are fair.

- Event A: “die 1 lands on side 3”
  $P(A) =$ 6/36 = 1/6.
- Event B: “value of die 2 is at least 2 and at most 4”
  $P(B) =$

Review: Probability (contd.)

Consider the experiment of throwing two six-sided dice at-once where both dice are fair.

- Event A: “die 1 lands on side 3”
  $P(A) =$ 6/36 = 1/6.
- Event B: “value of die 2 is at least 2 and at most 4”
  $P(B) =$ 18/36 = 1/2.

Review: Probability (contd.)

Consider the experiment of throwing two six-sided dice at-once where both dice are fair.

- Event A: “die 1 lands on side 3”
  $P(A) =$ 6/36 = 1/6.
- Event B: “value of die 2 is at least 2 and at most 4”
  $P(B) =$ 18/36 = 1/2.
- Event C: “two dice sum to more than 9”
  $P(C) =$

Review: Probability (contd.)

Consider the experiment of throwing two six-sided dice at-once where both dice are fair.

- Event A: “die 1 lands on side 3”
  $P(A) =$ 6/36 = 1/6.
- Event B: “value of die 2 is at least 2 and at most 4”
  $P(B) =$ 18/36 = 1/2.
- Event C: “two dice sum to more than 9”
  $P(C) =$ 6/36 = 1/6.

Review: Probability – Random variables

- **Random variable (RV)** represents unknown quantity of interest.
- Formally, an RV is a function from a sample space to a **target/state space**, i.e., set of quantities of interest.
  - For simplicity, the target space is often assumed to be $\mathbb{R}$ or its subset like $\mathbb{N}$, $\mathbb{Z}$, etc.
  - An RV is **discrete** if its state space is discrete. Otherwise, it's **continuous**.
- Consider the trial of throwing two six-sided dice at-once.
  - RV X_1: the value of die 1.
    $X_1((2,3)) = 2$, $X_1((5,1)) = 5$, etc.
  - RV X_2: the value of die 2.
    $X_2((2,3)) = 3$, $X_2((5,1)) = 1$, etc.
  - RV Y: the sum of both dice's values.
    $Y((4,2)) = 6$, $Y((2,6)) = 8$, etc.

Review: Probability – Random variables (contd.)

- Random variables are **not** events. Rather, assigning values to a random variable represents an event.
  - $X_1 = 3 \rightsquigarrow$ the event that "die 1's value is 3", i.e., the set $\{(3,1), \ldots, (3,6)\}$.
  - $2 \leqslant X_2 \leqslant 4 \rightsquigarrow$ the event that "die 2's value is at least 2 and at most 4", i.e., the set $\{(1,2), (1,3), (1,4), \ldots, (6,2), (6,3), (6,4)\}$
  - $Y > 9 \rightsquigarrow$ the event that "the sum of die 1's and die 2's values is more than 9", i.e., the set $\{(4,6), (5,5), (6,4), (5,6), (6,5), (6,6)\}$.
- So we can express an event in terms of some random variable and all probability laws on events still hold.
  - $P(X_1 = 3) =$

Review: Probability – Random variables (contd.)

- Random variables are **not** events. Rather, assigning values to a random variable represents an event.
  - $X_1 = 3 \rightsquigarrow$ the event that "die 1's value is 3", i.e., the set $\{(3,1), \ldots, (3,6)\}$.
  - $2 \leqslant X_2 \leqslant 4 \rightsquigarrow$ the event that "die 2's value is at least 2 and at most 4", i.e., the set $\{(1,2), (1,3), (1,4), \ldots, (6,2), (6,3), (6,4)\}$
  - $Y > 9 \rightsquigarrow$ the event that "the sum of die 1's and die 2's values is more than 9", i.e., the set $\{(4,6), (5,5), (6,4), (5,6), (6,5), (6,6)\}$.
- So we can express an event in terms of some random variable and all probability laws on events still hold.
  - $P(X_1 = 3) = 6/36 = 1/6$

Review: Probability – Random variables (contd.)

- Random variables are **not** events. Rather, assigning values to a random variable represents an event.
  - $X_1 = 3 \rightsquigarrow$ the event that "die 1's value is 3", i.e., the set $\{(3,1), \ldots, (3,6)\}$.
  - $2 \leqslant X_2 \leqslant 4 \rightsquigarrow$ the event that "die 2's value is at least 2 and at most 4", i.e., the set $\{(1,2), (1,3), (1,4), \ldots, (6,2), (6,3), (6,4)\}$
  - $Y > 9 \rightsquigarrow$ the event that "the sum of die 1's and die 2's values is more than 9", i.e., the set $\{(4,6), (5,5), (6,4), (5,6), (6,5), (6,6)\}$.
- So we can express an event in terms of some random variable and all probability laws on events still hold.
  - $P(X_1 = 3) = 6/36 = 1/6$
  - $P(2 \leqslant X_2 \leqslant 4) =$

Review: Probability – Random variables (contd.)

- Random variables are **not** events. Rather, assigning values to a random variable represents an event.
  - $X_1 = 3 \rightsquigarrow$ the event that "die 1's value is 3", i.e., the set $\{(3,1), \ldots, (3,6)\}$.
  - $2 \leqslant X_2 \leqslant 4 \rightsquigarrow$ the event that "die 2's value is at least 2 and at most 4", i.e., the set $\{(1,2), (1,3), (1,4), \ldots, (6,2), (6,3), (6,4)\}$
  - $Y > 9 \rightsquigarrow$ the event that "the sum of die 1's and die 2's values is more than 9", i.e., the set $\{(4,6), (5,5), (6,4), (5,6), (6,5), (6,6)\}$.
- So we can express an event in terms of some random variable and all probability laws on events still hold.
  - $P(X_1 = 3) = 6/36 = 1/6$
  - $P(2 \leqslant X_2 \leqslant 4) = 18/36 = 1/2$

Review: Probability – Random variables (contd.)

- Random variables are **not** events. Rather, assigning values to a random variable represents an event.
  - $X_1 = 3 \rightsquigarrow$ the event that "die 1's value is 3", i.e., the set $\{(3,1), \ldots, (3,6)\}$.
  - $2 \leqslant X_2 \leqslant 4 \rightsquigarrow$ the event that "die 2's value is at least 2 and at most 4", i.e., the set $\{(1,2), (1,3), (1,4), \ldots, (6,2), (6,3), (6,4)\}$
  - $Y > 9 \rightsquigarrow$ the event that "the sum of die 1's and die 2's values is more than 9", i.e., the set $\{(4,6), (5,5), (6,4), (5,6), (6,5), (6,6)\}$.
- So we can express an event in terms of some random variable and all probability laws on events still hold.
  - $P(X_1 = 3) = 6/36 = 1/6$
  - $P(2 \leqslant X_2 \leqslant 4) = 18/36 = 1/2$
  - $P(Y > 9) =$

Review: Probability – Random variables (contd.)

- Random variables are **not** events. Rather, assigning values to a random variable represents an event.
  - $X_1 = 3 \rightsquigarrow$ the event that "die 1's value is 3", i.e., the set $\{(3,1), \ldots, (3,6)\}$.
  - $2 \leqslant X_2 \leqslant 4 \rightsquigarrow$ the event that "die 2's value is at least 2 and at most 4", i.e., the set $\{(1,2), (1,3), (1,4), \ldots, (6,2), (6,3), (6,4)\}$
  - $Y > 9 \rightsquigarrow$ the event that "the sum of die 1's and die 2's values is more than 9", i.e., the set $\{(4,6), (5,5), (6,4), (5,6), (6,5), (6,6)\}$.
- So we can express an event in terms of some random variable and all probability laws on events still hold.
  - $P(X_1 = 3) = 6/36 = 1/6$
  - $P(2 \leqslant X_2 \leqslant 4) = 18/36 = 1/2$
  - $P(Y > 9) = 6/36 = 1/6$.

Review: Probability distribution

- Given an RV X, a **probability distribution** of X is the assignments of probability values to **all** events concerning X.
- If X is discrete, this is done by fully specifying its **probability mass function (pmf)** $P_X(a) \colon \mathbb{R} \to [0, 1]$ such that $P_X(a) = P(X = a)$.
- For RV X_1 in the previous slide:

Review: Probability distribution

- Given an RV X, a **probability distribution** of X is the assignments of probability values to **all** events concerning X.
- If X is discrete, this is done by fully specifying its **probability mass function (pmf)** $P_X(a) \colon \mathbb{R} \to [0, 1]$ such that $P_X(a) = P(X = a)$.
- For RV X_1 in the previous slide:

$$P_{X_1}(1) = 6/36 = 1/6, \quad P_{X_1}(2) = 6/36 = 1/6, \quad \ldots, \quad P_{X_1}(6) = 6/36 = 1/6$$

and $P_{X_1}(x) = 0$ if $x \notin \{1, \ldots, 6\}$

Review: Probability distribution

- Given an RV X, a **probability distribution** of X is the assignments of probability values to **all** events concerning X.
- If X is discrete, this is done by fully specifying its **probability mass function (pmf)** $P_X(a) \colon \mathbb{R} \to [0, 1]$ such that $P_X(a) = P(X = a)$.
- For RV X_1 in the previous slide:

$$P_{X_1}(1) = 6/36 = 1/6, \quad P_{X_1}(2) = 6/36 = 1/6, \quad \ldots, \quad P_{X_1}(6) = 6/36 = 1/6$$

and $P_{X_1}(x) = 0$ if $x \notin \{1, \ldots, 6\}$

- What's the probability distribution of X_2 and Y from the previous slide?

Review: Probability distribution

- Given an RV X, a **probability distribution** of X is the assignments of probability values to **all** events concerning X.
- If X is discrete, this is done by fully specifying its **probability mass function (pmf)** $P_X(a) \colon \mathbb{R} \to [0, 1]$ such that $P_X(a) = P(X = a)$.
- For RV X_1 in the previous slide:

$$P_{X_1}(1) = 6/36 = 1/6, \quad P_{X_1}(2) = 6/36 = 1/6, \quad \ldots, \quad P_{X_1}(6) = 6/36 = 1/6$$

and $P_{X_1}(x) = 0$ if $x \notin \{1, \ldots, 6\}$

- What's the probability distribution of X_2 and Y from the previous slide?
- Note: the discrete RV X_1, X_2, Y above follow **categorical distribution**.
  - The **parameters** are k (the number of categories) and $p_1, \ldots, p_k$ (probability values assigned to all categories).
  - How many parameters do the distribution of X_1, X_2, and Y have?

Review: Probability distribution

- Given an RV X, a **probability distribution** of X is the assignments of probability values to **all** events concerning X.
- If X is discrete, this is done by fully specifying its **probability mass function (pmf)** $P_X(a) \colon \mathbb{R} \to [0, 1]$ such that $P_X(a) = P(X = a)$.
- For RV X_1 in the previous slide:

$$P_{X_1}(1) = 6/36 = 1/6, \quad P_{X_1}(2) = 6/36 = 1/6, \quad \ldots, \quad P_{X_1}(6) = 6/36 = 1/6$$

and $P_{X_1}(x) = 0$ if $x \notin \{1, \ldots, 6\}$

- What's the probability distribution of X_2 and Y from the previous slide?
- Note: the discrete RV X_1, X_2, Y above follow **categorical distribution**.
  - The **parameters** are k (the number of categories) and $p_1, \ldots, p_k$ (probability values assigned to all categories).
  - How many parameters do the distribution of X_1, X_2, and Y have?
- Examples of discrete distributions: Bernoulli, categorical, binomial, discrete uniform, Poisson, geometric, etc.

Review: Probability distribution (contd.)

- If an RV X is continuous, then the corresponding distribution is called **continuous distribution**.
- Continuous distribution is specified via its **probability density function (pdf)** $f_X(x) \colon \mathbb{R} \to \mathbb{R}$ such that

$$f_X(x) \geqslant 0 \text{ for all } x \in \mathbb{R}, \qquad f_X \text{ is integrable}, \qquad \text{and } \int_0^1 f_X(x)dx = 1$$

- Probability value to an event is given by:

$$P(X \leqslant x) = \int_{-\infty}^{x} f_X(x)dx$$

- Example: a Gaussian distribution for an RV X is specified by the pdf:

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

where μ and σ are two parameters of the distribution.

Review: Probability distribution (contd.)

Important points

- Distribution = specification of probability values for all outcomes.
- Fully specifying a distribution requires fully specifying all of its parameters.

- In general, there are many more distributions than ones that can be named.
- As long as we fully specify all of its parameters, we will already obtain one particular distribution (without necessarily identifying its name).
- Some distributions can have an extremely large number of parameters.
- Training a model in machine learning corresponds to finding appropriate values of the parameters of the model.

Review: Conditional probability and product rule

$P(A|B)$: the **conditional probability** of A happening given that B happens:

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

- Intuition: "among all outcomes satisfying B, what is the proportion of them satisfies A?"
- Assumption: $P(B) > 0$. If $P(B) = 0$, then $P(A|B)$ is undefined (i.e., we cannot condition on impossible events).
- The above definition can be written as the **product rule**: $P(A,B) = P(A|B)P(B)$, which also works when $P(B) = 0$.
- The product rule can be chained:
  $P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2, A_1) \cdots P(A_n|A_1, A_2, \ldots, A_{n-1})$
- The above definition can be written in terms of random variables, yielding the so-called **joint distribution** and **conditional distribution**.

Review: Marginal probability and sum rule

- Let $B_1, \ldots, B_n$ be mutually exclusive events such that $B_1 \cup \cdots \cup B_n = \Omega$ with Ω the sample space, then the **sum rule** states:

$$\begin{aligned} P(A) = P(A, \Omega) &= P(A, B_1 \cup \cdots \cup B_n) \\ &= P(A, B_1) + \cdots + P(A, B_n) \\ &= P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n) \end{aligned}$$

  Here, $P(A)$ is also called the **marginal probability** of A, by understanding $P(A)$ as the probability of A over all possible states of B. The summing operation expressed above is also called **marginalizing** A over B.

- Writing the above definition in terms of random variables yields the so-called **marginal distribution**.

Review: Product and sum rule

Consider an experiment of throwing two fair, six-sided dice where the RV X_1 represents the value of die 1, X_2 the value of die 2, and Y is the sum of both dice's values.

$$P(X_1 = 3 \mid 2 \leqslant X_2 \leqslant 4) =$$

$$P(Y > 9 \mid 2 \leqslant X_2 \leqslant 4) =$$

Review: Product and sum rule

Consider an experiment of throwing two fair, six-sided dice where the RV X_1 represents the value of die 1, X_2 the value of die 2, and Y is the sum of both dice's values.

$$P(X_1 = 3 \mid 2 \leqslant X_2 \leqslant 4) = \frac{P(X_1 = 3, 2 \leqslant X_2 \leqslant 4)}{P(2 \leqslant X_2 \leqslant 4)} =$$

$$P(Y > 9 \mid 2 \leqslant X_2 \leqslant 4) =$$

Review: Product and sum rule

Consider an experiment of throwing two fair, six-sided dice where the RV X_1 represents the value of die 1, X_2 the value of die 2, and Y is the sum of both dice's values.

$$P(X_1 = 3 \mid 2 \leqslant X_2 \leqslant 4) = \frac{P(X_1 = 3, 2 \leqslant X_2 \leqslant 4)}{P(2 \leqslant X_2 \leqslant 4)} = \frac{3/36}{18/36} = 1/6$$

$$P(Y > 9 \mid 2 \leqslant X_2 \leqslant 4) =$$

Review: Product and sum rule

Consider an experiment of throwing two fair, six-sided dice where the RV X_1 represents the value of die 1, X_2 the value of die 2, and Y is the sum of both dice's values.

$$P(X_1 = 3 \mid 2 \leqslant X_2 \leqslant 4) = \frac{P(X_1 = 3, 2 \leqslant X_2 \leqslant 4)}{P(2 \leqslant X_2 \leqslant 4)} = \frac{3/36}{18/36} = 1/6$$

$$P(Y > 9 \mid 2 \leqslant X_2 \leqslant 4) = \frac{P(Y > 9, 2 \leqslant X_2 \leqslant 4)}{P(2 \leqslant X_2 \leqslant 4)} =$$

Review: Product and sum rule

Consider an experiment of throwing two fair, six-sided dice where the RV X_1 represents the value of die 1, X_2 the value of die 2, and Y is the sum of both dice's values.

$$P(X_1 = 3 \mid 2 \leqslant X_2 \leqslant 4) = \frac{P(X_1 = 3, 2 \leqslant X_2 \leqslant 4)}{P(2 \leqslant X_2 \leqslant 4)} = \frac{3/36}{18/36} = 1/6$$

$$P(Y > 9 \mid 2 \leqslant X_2 \leqslant 4) = \frac{P(Y > 9, 2 \leqslant X_2 \leqslant 4)}{P(2 \leqslant X_2 \leqslant 4)} = \frac{1/36}{18/36} = 1/18$$

Review: Product and sum rule

Consider an experiment of throwing two fair, six-sided dice where the RV X_1 represents the value of die 1, X_2 the value of die 2, and Y is the sum of both dice's values.

$$P(X_1 = 3 \mid 2 \leqslant X_2 \leqslant 4) = \frac{P(X_1 = 3, 2 \leqslant X_2 \leqslant 4)}{P(2 \leqslant X_2 \leqslant 4)} = \frac{3/36}{18/36} = 1/6$$

$$P(Y > 9 \mid 2 \leqslant X_2 \leqslant 4) = \frac{P(Y > 9, 2 \leqslant X_2 \leqslant 4)}{P(2 \leqslant X_2 \leqslant 4)} = \frac{1/36}{18/36} = 1/18$$

Sum rule example:

$$\begin{aligned}
P(Y > 9) &= P(Y > 9, X_2 = 1) + P(Y > 9, X_2 = 2) + P(Y > 9, X_2 = 3) + P(Y > 9, X_2 = 4) \\
&\quad + P(Y > 9, X_2 = 5) + P(Y > 9, X_2 = 6) \\
&= P(Y > 9|X_2 = 1)P(X_2 = 1) + P(Y > 9|X_2 = 2)P(X_2 = 2) \\
&\quad + P(Y > 9|X_2 = 3)P(X_2 = 3) + P(Y > 9|X_2 = 4)P(X_2 = 4) \\
&\quad + P(Y > 9|X_2 = 5)P(X_2 = 5) + P(Y > 9|X_2 = 6)P(X_2 = 6) \\
&= 0 + 0 + 0 + (1/6)(1/6) + (2/6)(1/6) + (3/6)(1/6) \\
&= 6/36 = 1/6
\end{aligned}$$

Review: Independence

- A is **(unconditionally) independent** of B, written $A \perp B$, if $P(A|B) = P(A)$, or equivalently if $P(A, B) = P(A)P(B)$.

Review: Independence

- A is **(unconditionally) independent** of B, written $A \perp B$, if $P(A|B) = P(A)$, or equivalently if $P(A, B) = P(A)P(B)$.
- Reading: "knowing that B happens doesn't change our belief about whether A happened."
- Two RVs X_1 and X_2 are independent if the events $X_1 \leqslant a$ and $X_2 \leqslant b$ are independent for every a and b.
- The events $X_1 = 3$ and $2 \leqslant X_2 \leqslant 4$ in the previous slide are independent.

Review: Independence

- A is **(unconditionally) independent** of B, written $A \perp B$, if $P(A|B) = P(A)$, or equivalently if $P(A, B) = P(A)P(B)$.
- Reading: "knowing that B happens doesn't change our belief about whether A happened."
- Two RVs X_1 and X_2 are independent if the events $X_1 \leqslant a$ and $X_2 \leqslant b$ are independent for every a and b.
- The events $X_1 = 3$ and $2 \leqslant X_2 \leqslant 4$ in the previous slide are independent.
- A set of events $\mathcal{E} = \{E_1, \ldots, E_n\}$ are **independent** if every strict subset of $\mathcal{E}$ with at least two elements are universally independent and $P(E_1, \ldots, E_n) = P(E_1)P(E_2) \cdots P(E_n)$.
- A set of events $\mathcal{E} = \{E_1, \ldots, E_n\}$ are **pairwise independent** if every pair of events from $\mathcal{E}$ are independent.

Review: Conditional independence

A is **conditionally independent** of B given another event C, written $A \perp B \mid C$, if $P(A|B,C) = P(A|C)$, or equivalently, if $P(A,B|C) = P(A|C)P(B|C)$.

Review: Conditional independence

A is **conditionally independent** of B given another event C, written $A \perp B \mid C$, if $P(A|B,C) = P(A|C)$, or equivalently, if $P(A,B|C) = P(A|C)P(B|C)$.

A box has two coins: a fair, regular coin and a fake coin with two-heads (instead of head and tail). You are to choose a coin at random and then toss it twice.

- Define the events: A = first coin toss results in a head, B = second coin toss results in a head, and C = regular coin has been selected.

Review: Conditional independence

A is **conditionally independent** of B given another event C, written $A \perp B \mid C$, if $P(A|B,C) = P(A|C)$, or equivalently, if $P(A,B|C) = P(A|C)P(B|C)$.

A box has two coins: a fair, regular coin and a fake coin with two-heads (instead of head and tail). You are to choose a coin at random and then toss it twice.

- Define the events: A = first coin toss results in a head, B = second coin toss results in a head, and C = regular coin has been selected.
- $P(C) = P(C^c) = 1/2$, $P(A|C) = P(B|C) = 1/2$, while $P(A|C^c) = P(B|C^c) = 1$

Review: Conditional independence

A is **conditionally independent** of B given another event C, written $A \perp B \mid C$, if $P(A|B,C) = P(A|C)$, or equivalently, if $P(A,B|C) = P(A|C)P(B|C)$.

A box has two coins: a fair, regular coin and a fake coin with two-heads (instead of head and tail). You are to choose a coin at random and then toss it twice.

- Define the events: A = first coin toss results in a head, B = second coin toss results in a head, and C = regular coin has been selected.
- $P(C) = P(C^c) = 1/2$, $P(A|C) = P(B|C) = 1/2$, while $P(A|C^c) = P(B|C^c) = 1$
- $P(A, B \mid C) = 1/4$, while $P(A|C) = P(B|C) = 1/2$. So, $P(A, B \mid C) = P(A|C)P(B|C)$, i.e., A and B are conditionally independent given C.

Review: Conditional independence

A is **conditionally independent** of B given another event C, written $A \perp B \mid C$, if $P(A|B,C) = P(A|C)$, or equivalently, if $P(A,B|C) = P(A|C)P(B|C)$.

A box has two coins: a fair, regular coin and a fake coin with two-heads (instead of head and tail). You are to choose a coin at random and then toss it twice.

- Define the events: A = first coin toss results in a head, B = second coin toss results in a head, and C = regular coin has been selected.
- $P(C) = P(C^c) = 1/2$, $P(A|C) = P(B|C) = 1/2$, while $P(A|C^c) = P(B|C^c) = 1$
- $P(A, B \mid C) = 1/4$, while $P(A|C) = P(B|C) = 1/2$. So, $P(A, B \mid C) = P(A|C)P(B|C)$, i.e., A and B are conditionally independent given C.
- Meanwhile, $P(A) = P(A|C)P(C) + P(A|C^c)P(C^c) = (1/2)(1/2) + (1)(1/2) = 3/4$, and $P(B) = P(B|C)P(C) + P(B|C^c)P(C^c) = (1/2)(1/2) + (1)(1/2) = 3/4$

Review: Conditional independence

A is **conditionally independent** of B given another event C, written $A \perp B \mid C$, if $P(A|B,C) = P(A|C)$, or equivalently, if $P(A,B|C) = P(A|C)P(B|C)$.

A box has two coins: a fair, regular coin and a fake coin with two-heads (instead of head and tail). You are to choose a coin at random and then toss it twice.

- Define the events: A = first coin toss results in a head, B = second coin toss results in a head, and C = regular coin has been selected.
- $P(C) = P(C^c) = 1/2$, $P(A|C) = P(B|C) = 1/2$, while $P(A|C^c) = P(B|C^c) = 1$
- $P(A, B \mid C) = 1/4$, while $P(A|C) = P(B|C) = 1/2$. So, $P(A, B \mid C) = P(A|C)P(B|C)$, i.e., A and B are conditionally independent given C.
- Meanwhile, $P(A) = P(A|C)P(C) + P(A|C^c)P(C^c) = (1/2)(1/2) + (1)(1/2) = 3/4$, and $P(B) = P(B|C)P(C) + P(B|C^c)P(C^c) = (1/2)(1/2) + (1)(1/2) = 3/4$
- Also, $P(A, B) = P(A, B \mid C)P(C) + P(A, B \mid C^c)P(C^c) = P(A|C)P(B|C)P(C) + P(A|C^c)P(B|C^c)P(C^c)$ due to conditional independence of A and B given C.
  Computing this yields $P(A, B) = 5/8 \neq P(A)P(B)$. Hence, A and B are **not** independent.

Outline

Bayes theorem

Consider two RVs X and Y (discrete, for simplicity). Product rule implies that:

$$P(X = x, Y = y) = P(X = x \mid Y = y)P(Y = y) = P(Y = y \mid X = x)$$

This yields the **Bayes rule/theorem**:

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P(Y = y)P(X = x|Y = y)}{\sum_{y'} P(Y = y')P(X = x|Y = y')}$$

- $P(Y = y)$ is **prior** probability of $Y = y$.
- $P(Y = y|X = x)$ is **posterior** probability of $Y = y$ (conditioned on the value of X).
- The denominator $\sum_{y'} P(Y = y')P(X = x|Y = y')$ is marginalizing over all possible states of Y to get the marginal probability $P(X = x)$.

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer?

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?
- The correct answer is as follows. Suppose x and y are binary random variables representing the result of the test (positive = 1, negative = 0) and actual occurrence of throat cancer (positive = 1, negative = 0), respectively.

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?
- The correct answer is as follows. Suppose x and y are binary random variables representing the result of the test (positive = 1, negative = 0) and actual occurrence of throat cancer (positive = 1, negative = 0), respectively.
  - $P(x = 1|y = 1) =$

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?
- The correct answer is as follows. Suppose x and y are binary random variables representing the result of the test (positive = 1, negative = 0) and actual occurrence of throat cancer (positive = 1, negative = 0), respectively.
  - $P(x = 1|y = 1) = 0.8$

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?
- The correct answer is as follows. Suppose x and y are binary random variables representing the result of the test (positive = 1, negative = 0) and actual occurrence of throat cancer (positive = 1, negative = 0), respectively.
  - $P(x = 1|y = 1) = 0.8$
  - $P(y = 1) =$

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?
- The correct answer is as follows. Suppose x and y are binary random variables representing the result of the test (positive = 1, negative = 0) and actual occurrence of throat cancer (positive = 1, negative = 0), respectively.
  - $P(x = 1|y = 1) = 0.8$
  - $P(y = 1) = 0.004$.
  - $P(x = 1|y = 0) =$

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?
- The correct answer is as follows. Suppose x and y are binary random variables representing the result of the test (positive = 1, negative = 0) and actual occurrence of throat cancer (positive = 1, negative = 0), respectively.
  - $P(x = 1|y = 1) = 0.8$
  - $P(y = 1) = 0.004$.
  - $P(x = 1|y = 0) = 0.1$

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?
- The correct answer is as follows. Suppose x and y are binary random variables representing the result of the test (positive = 1, negative = 0) and actual occurrence of throat cancer (positive = 1, negative = 0), respectively.
  - $P(x = 1|y = 1) = 0.8$
  - $P(y = 1) = 0.004$.
  - $P(x = 1|y = 0) = 0.1$
  - $P(y = 1|x = 1) =$

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?
- The correct answer is as follows. Suppose x and y are binary random variables representing the result of the test (positive = 1, negative = 0) and actual occurrence of throat cancer (positive = 1, negative = 0), respectively.
  - $P(x = 1|y = 1) = 0.8$
  - $P(y = 1) = 0.004$.
  - $P(x = 1|y = 0) = 0.1$
  - $P(y = 1|x = 1) = \dfrac{P(y = 1)P(x = 1|y = 1)}{P(y = 1)P(x = 1|y = 1) + P(y = 0)P(x = 1|y = 0)}$

    $=$

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?
- The correct answer is as follows. Suppose x and y are binary random variables representing the result of the test (positive = 1, negative = 0) and actual occurrence of throat cancer (positive = 1, negative = 0), respectively.
  - $P(x = 1|y = 1) = 0.8$
  - $P(y = 1) = 0.004$.
  - $P(x = 1|y = 0) = 0.1$
  - $$P(y = 1|x = 1) = \frac{P(y = 1)P(x = 1|y = 1)}{P(y = 1)P(x = 1|y = 1) + P(y = 0)P(x = 1|y = 0)}$$
    $$= \frac{(0.004)(0.8)}{(0.004)(0.8) + (0.996)(0.1)}$$
    $$=$$

Bayes theorem example

- Consider a medical test for a throat cancer. Suppose the test has sensitivity of 80%, that is, if you have cancer, the test will be positive with a probability 0.8.
- Assume probability of having throat cancer over the whole population is 0.4%.
- Moreover, the test may have a false positive/false alarm, i.e., that the test is positive despite not having throat cancer. Assume 10% false positive rate of the test.
- Suppose you test positive. Are you 80% likely to have cancer? **No.** Why?
- The correct answer is as follows. Suppose x and y are binary random variables representing the result of the test (positive = 1, negative = 0) and actual occurrence of throat cancer (positive = 1, negative = 0), respectively.
  - $P(x = 1|y = 1) = 0.8$
  - $P(y = 1) = 0.004$.
  - $P(x = 1|y = 0) = 0.1$
  - $$P(y = 1|x = 1) = \frac{P(y = 1)P(x = 1|y = 1)}{P(y = 1)P(x = 1|y = 1) + P(y = 0)P(x = 1|y = 0)}$$
    $$= \frac{(0.004)(0.8)}{(0.004)(0.8) + (0.996)(0.1)}$$
    $$= 0.031.$$

Outline

Probability and Likelihood

- Recall: probability distribution is completely specified by its parameters.
- **Probability**: the chance that a particular outcome occurs based on the values of the distribution parameters.
  - When calculating probability, we assume that the parameters are trustworthy.
  - E.g.: compute $P(\text{height} > 180 \mid \mu = 170, \sigma = 10)$.
- **Likelihood**: how well a sample/data provides support for particular values of the distribution parameter.
  - When calculating likelihood, we're trying to determine if we can trust the parameter values based on the data that we observe.
  - Given height of 5 person, say $\{190, 160, 170, 190, 200\}$, the likelihood of $\mu = 170$ and $\sigma = 10$, is given by

$$\begin{aligned} &L(\mu = 170, \sigma = 10 \mid h_1 = 190, h_2 = 160, \ldots, h_5 = 200) \\ &\qquad \propto P(h_1 = 190, h_2 = 160, \ldots, h_5 = 200 \mid \mu = 170, \sigma = 10) \end{aligned}$$

  - Likelihood can be viewed as a "score" for particular parameter values with respect to the given data.
  - Not necessarily equal, but proportional to the probability of the sample given the parameter values.

Likelihood

Let X be an RV.

- If X is discrete according to pmf $P_{\boldsymbol{\theta}}(x)$ (with parameter $\boldsymbol{\theta}$), then the **likelihood function** of $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}|X = x) = P_{\boldsymbol{\theta}}(x) = P(X = x \mid \boldsymbol{\theta})$$

  The latter is sometimes written $P(X = x; \boldsymbol{\theta})$.

- If X is continuous according to pdf $f_{\boldsymbol{\theta}}(x)$ (with parameter $\boldsymbol{\theta}$), then the likelihood of $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}|x) = f_{\boldsymbol{\theta}}(x)$$

- Likelihood is
  - **not** the posterior probability $P(\boldsymbol{\theta}|x)$,
  - **not** a probability density over the parameter $\boldsymbol{\theta}$, and
  - **not** a probability mass over the parameter $\boldsymbol{\theta}$.

  The integral of the likelihood over all parameter values may not be equal to 1.

Example

Consider a statistical model of a simple coin flip: it has a single parameter p_H expressing the probability of getting a head.

- p_H measures the coin fairness: for a perfectly fair coin, $p_H = 0.5$.

Example

Consider a statistical model of a simple coin flip: it has a single parameter p_H expressing the probability of getting a head.

- p_H measures the coin fairness: for a perfectly fair coin, $p_H = 0.5$.
- Suppose we flip the coin twice and observe HH. Assume that each coin flip is **independently and identically distributed (i.i.d)**. Probability of observing HH is

Example

Consider a statistical model of a simple coin flip: it has a single parameter p_H expressing the probability of getting a head.

- p_H measures the coin fairness: for a perfectly fair coin, $p_H = 0.5$.
- Suppose we flip the coin twice and observe HH. Assume that each coin flip is **independently and identically distributed (i.i.d)**. Probability of observing HH is $P(HH \mid p_H = 0.5) = 0.5^2 = 0.25$.

Example

Consider a statistical model of a simple coin flip: it has a single parameter p_H expressing the probability of getting a head.

- p_H measures the coin fairness: for a perfectly fair coin, $p_H = 0.5$.
- Suppose we flip the coin twice and observe HH. Assume that each coin flip is **independently and identically distributed (i.i.d)**. Probability of observing HH is $P(HH \mid p_H = 0.5) = 0.5^2 = 0.25$. Correspondingly, the likelihood that $p_H = 0.5$ given that HH is observed is:

$$L(p_H = 0.5 \mid HH) = P(HH \mid p_H = 0.5) = 0.25$$

Example

Consider a statistical model of a simple coin flip: it has a single parameter p_H expressing the probability of getting a head.

- p_H measures the coin fairness: for a perfectly fair coin, $p_H = 0.5$.
- Suppose we flip the coin twice and observe HH. Assume that each coin flip is **independently and identically distributed (i.i.d)**. Probability of observing HH is $P(HH \mid p_H = 0.5) = 0.5^2 = 0.25$. Correspondingly, the likelihood that $p_H = 0.5$ given that HH is observed is:

$$L(p_H = 0.5 \mid HH) = P(HH \mid p_H = 0.5) = 0.25$$

This is **different** from saying that $P(p_H = 0.5 \mid HH) = 0.25$ (which can only be obtained by Bayes theorem).

Example

Consider a statistical model of a simple coin flip: it has a single parameter p_H expressing the probability of getting a head.

- p_H measures the coin fairness: for a perfectly fair coin, $p_H = 0.5$.
- Suppose we flip the coin twice and observe HH. Assume that each coin flip is **independently and identically distributed (i.i.d)**. Probability of observing HH is $P(HH \mid p_H = 0.5) = 0.5^2 = 0.25$. Correspondingly, the likelihood that $p_H = 0.5$ given that HH is observed is:

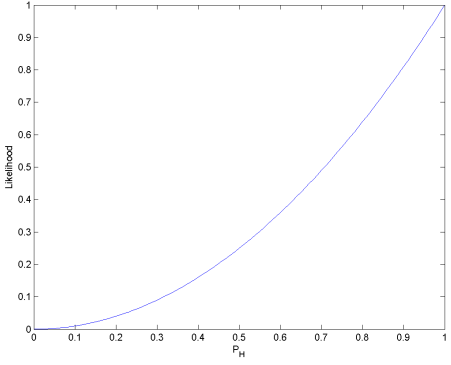
  $$L(p_H = 0.5 \mid HH) = P(HH \mid p_H = 0.5) = 0.25$$

  This is **different** from saying that $P(p_H = 0.5 \mid HH) = 0.25$ (which can only be obtained by Bayes theorem).
- Now suppose that the coin may be not fair. We can then ask what is $L(p_H = 0.5 \mid HH)$ or $L(p_H = 0.9 \mid HH)$. If the coin is indeed not fair because flipping it yields H most of the time, then the latter must be higher than the former.

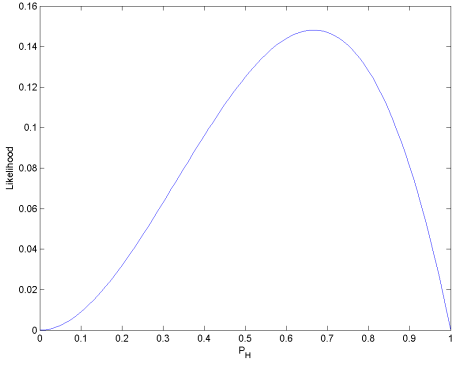
Example (contd.)

Shape of the likelihood function **changes** whenever data/observation changes.

For p_H, the probability of a coin landing heads-up (without prior knowledge of the coin's fairness), its likelihood functions depend on the samples we observe.



For sample HH, the likelihood function $L(p_H) = p_H^2$.



For sample HHT, the likelihood function $L(p_H) = p_H^2(1 - p_H)$.

Computing likelihood

- Given dataset $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of N data points, what is the likelihood of parameter $\boldsymbol{\theta}$?
  - **IID assumption**: data points in $\mathcal{D}$ are **independently** sampled from an **identical distribution**.

Computing likelihood

- Given dataset $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of N data points, what is the likelihood of parameter $\boldsymbol{\theta}$?
  - **IID assumption**: data points in $\mathcal{D}$ are **independently** sampled from an **identical distribution**.
- Suppose $p(\boldsymbol{x}_i; \boldsymbol{\theta}) = p(\boldsymbol{x}_i \mid \boldsymbol{\theta})$ is the probability mass/density of $\boldsymbol{x}_i$ under parameter $\boldsymbol{\theta}$. Then, the likelihood of $\boldsymbol{\theta}$ (using independence assumption):

Computing likelihood

- Given dataset $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of N data points, what is the likelihood of parameter $\boldsymbol{\theta}$?
  - **IID assumption**: data points in $\mathcal{D}$ are **independently** sampled from an **identical distribution**.
- Suppose $p(\boldsymbol{x}_i; \boldsymbol{\theta}) = p(\boldsymbol{x}_i \mid \boldsymbol{\theta})$ is the probability mass/density of $\boldsymbol{x}_i$ under parameter $\boldsymbol{\theta}$. Then, the likelihood of $\boldsymbol{\theta}$ (using independence assumption):

$$L(\boldsymbol{\theta}) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i \mid \boldsymbol{\theta})$$

Computing likelihood

- Given dataset $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of N data points, what is the likelihood of parameter $\boldsymbol{\theta}$?
  - **IID assumption**: data points in $\mathcal{D}$ are **independently** sampled from an **identical distribution**.
- Suppose $p(\boldsymbol{x}_i; \boldsymbol{\theta}) = p(\boldsymbol{x}_i \mid \boldsymbol{\theta})$ is the probability mass/density of $\boldsymbol{x}_i$ under parameter $\boldsymbol{\theta}$. Then, the likelihood of $\boldsymbol{\theta}$ (using independence assumption):

$$L(\boldsymbol{\theta}) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i \mid \boldsymbol{\theta})$$

- Product/multiplication of large number of small numbers can be problematic. So we often work with log-likelihood.
  - Logarithm is monotonic: $a > b$ if and only if $\log a > \log b$.
  - Maximizing/minimizing likelihood is equivalent to maximizing/minimizing log-likelihood.

$$LL(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\boldsymbol{x}_i \mid \boldsymbol{\theta})$$

Outline

Probabilistic modeling for supervised learning

- Notation: input vector $\boldsymbol{x}$, output y, (training) data $\mathcal{D}$, model parameters $\boldsymbol{\theta}$.
  - Classification: y is categorical. Regression: y is real-valued.
  - $\mathcal{D}$ consists of a number of input-output pairs $(\boldsymbol{x}, y)$, can be viewed as a sample from all possible, correct input-output pairs.
  - Aim: approximate the unknown function f^* that maps every input $\boldsymbol{x}$ to the correct output y so that incorrect mapping is avoided as much as possible not just for all input-output pairs in $\mathcal{D}$ but also other input-output pairs that have not been seen in $\mathcal{D}$.
- Modeled as the **conditional distribution** $P(y \mid \boldsymbol{x}; \boldsymbol{\theta})$, called **posterior distribution**.
  - Classification: $P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta})$ specified for all class labels c, i.e., **class posterior distribution** $P(y \mid \boldsymbol{x}; \boldsymbol{\theta})$ is discrete. It can be shown that selecting the label with the largest conditional probability leads to test error being minimized on average.
  - Regression: $P(y \mid \boldsymbol{x}; \boldsymbol{\theta})$ is continuous.
- Training the model = finding parameters $\boldsymbol{\theta}$ that best explain/fit the data $\mathcal{D}$.
  - Equivalent to **parameter estimation** problem in statistics, in particular, **point estimates** (rather than interval estimates).
  - The optimal parameter is $\hat{\boldsymbol{\theta}}$, corresponding to the **maximum likelihood estimate** given the training data:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}|\mathcal{D}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, P(\mathcal{D} \mid \boldsymbol{\theta})$$

Discriminative classifier

There are two kinds of classifiers: discriminative and generative classifiers.

Discriminative classifier

There are two kinds of classifiers: discriminative and generative classifiers.

Discriminative classifier directly models the class posterior distribution as a function that directly returns the class probability given input.

$$P(y = c|\boldsymbol{x}; \boldsymbol{\theta}) = f_c(\boldsymbol{x}; \boldsymbol{\theta})$$

where $f_c(\boldsymbol{x}; \boldsymbol{\theta})$ = probability of class label c given input $\boldsymbol{x}$ under parameter setting $\boldsymbol{\theta}$.

- Example: logistic regression models $f_c(\boldsymbol{x}; \boldsymbol{\theta})$ as the logistic function $\sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b)$.

Generative classifier

Generative classifier models class posterior distribution indirectly via Bayes theorem:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{P(y = c, \boldsymbol{x}; \boldsymbol{\theta})}{P(\boldsymbol{x}; \boldsymbol{\theta})} = \frac{P(y = c, \boldsymbol{x}; \boldsymbol{\theta})}{\sum_{c'} P(y = c', \boldsymbol{x}; \boldsymbol{\theta})} = \frac{P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta}) P(y = c; \boldsymbol{\theta})}{\sum_{c'} P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta}) P(y = c'; \boldsymbol{\theta})}$$

- $P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta})$ is the **class posterior distribution**.
- $P(y = c; \boldsymbol{\theta})$ is the **prior** over class labels.
- $P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta})$ is the **class conditional density** for class c, i.e., the distribution of the data $\boldsymbol{x}$ for that class.
- All distributions above are conditioned over values of $\boldsymbol{\theta}$, and model training/fitting is an optimization to find the optimal values for $\boldsymbol{\theta}$ (on the class prior and class conditional density).
- The classifier is called generative because it specifies a way to generate the features $\boldsymbol{x}$ for each class c by sampling from $P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta})$.
- The normalizing denominator may sometimes be ignored since it's independent of the class label c, yielding $P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) \propto P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta}) P(y = c; \boldsymbol{\theta})$
- Example: naive Bayes, Gaussian discriminant analysis.

Parametric vs. nonparametric models

- **Parametric** models:
  - Fixed-size vector $\boldsymbol{\theta}$ of parameters are estimated from a variable-sized dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n) \colon n = 1, \ldots, N\}$, but after model fitting/training, the data is thrown away.
  - Examples: linear regression, logistic regression, naive Bayes (all models described in earlier pages).

Parametric vs. nonparametric models

- **Parametric** models:
  - Fixed-size vector $\boldsymbol{\theta}$ of parameters are estimated from a variable-sized dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n) \colon n = 1, \ldots, N\}$, but after model fitting/training, the data is thrown away.
  - Examples: linear regression, logistic regression, naive Bayes (all models described in earlier pages).
- **Nonparametric** models does **not** have a fixed number of parameters
  - Effective number of parameters can grow with $|\mathcal{D}|$ (either because they keep the training data around or due to the way the training is done).

Nonparametric model examples

- K-nearest neighbor classifier:

$$P(y = c \mid \boldsymbol{x}, \mathcal{D}) = \frac{1}{K} \sum_{n \in N_K(\boldsymbol{x}, \mathcal{D})} \mathbb{I}(y_n = c)$$

  where:
  - $N_K(\boldsymbol{x}, \mathcal{D})$ contains **indices** of data points of $\mathcal{D}$ that belong to the K-closest neighbor of $\boldsymbol{x}$.
  - $\mathbb{I}(y_n = c) = 1$ if $y_n = c$ and 0 otherwise.

  We can then return this distribution or the majority labels.
- Decision trees:
  - typically not interpreted probabilistically;
  - the "parameters" are captured by the tree: a small/large dataset approximately yields a smaller/larger tree, hence decision trees are nonparametric.
- Other examples include kernel methods such as support vector machines, Gaussian process, etc.

Outline

Naive Bayes assumption

- Generative classifier where we employ the **naive Bayes assumption** for the class conditional density: “features are conditionally independent given the class label”.
- Let $\boldsymbol{x} = (x_1, \ldots, x_D)^\mathsf{T}$ be the D-dimensional input vector:

$$P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta}) = P(x_1, \ldots, x_D \mid y = c; \boldsymbol{\theta}) =$$

Naive Bayes assumption

- Generative classifier where we employ the **naive Bayes assumption** for the class conditional density: “features are conditionally independent given the class label”.
- Let $\boldsymbol{x} = (x_1, \ldots, x_D)^\mathsf{T}$ be the D-dimensional input vector:

$$P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta}) = P(x_1, \ldots, x_D \mid y = c; \boldsymbol{\theta}) = \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})$$

where $\boldsymbol{\theta}_{dc}$ are the parameters of the class conditional density for class c and feature d.

Naive Bayes classifier

- The posterior for naive Bayes classifier:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta}) P(y = c; \boldsymbol{\theta})}{\sum_{c'} P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta}) P(y = c'; \boldsymbol{\theta})} = \frac{P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})}{\sum_{c'} [P(y = c'; \boldsymbol{\theta}) P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta})]}$$

Naive Bayes classifier

- The posterior for naive Bayes classifier:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta}) P(y = c; \boldsymbol{\theta})}{\sum_{c'} P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta}) P(y = c'; \boldsymbol{\theta})} = \frac{P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})}{\sum_{c'} [P(y = c'; \boldsymbol{\theta}) P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta})]}$$

- As the denominator is independent of the class labels, the above becomes:

Naive Bayes classifier

- The posterior for naive Bayes classifier:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta}) P(y = c; \boldsymbol{\theta})}{\sum_{c'} P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta}) P(y = c'; \boldsymbol{\theta})} = \frac{P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})}{\sum_{c'} [P(y = c'; \boldsymbol{\theta}) P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta})]}$$

- As the denominator is independent of the class labels, the above becomes:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) \propto P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})$$

Naive Bayes classifier

- The posterior for naive Bayes classifier:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta}) P(y = c; \boldsymbol{\theta})}{\sum_{c'} P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta}) P(y = c'; \boldsymbol{\theta})} = \frac{P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})}{\sum_{c'} [P(y = c'; \boldsymbol{\theta}) P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta})]}$$

- As the denominator is independent of the class labels, the above becomes:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) \propto P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})$$

- Why is “proportional to” is sufficient, instead of equality, to decide the class label?

Naive Bayes classifier

- The posterior for naive Bayes classifier:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta}) P(y = c; \boldsymbol{\theta})}{\sum_{c'} P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta}) P(y = c'; \boldsymbol{\theta})} = \frac{P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})}{\sum_{c'} [P(y = c'; \boldsymbol{\theta}) P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta})]}$$

- As the denominator is independent of the class labels, the above becomes:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) \propto P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})$$

- Why is “proportional to” is sufficient, instead of equality, to decide the class label?
  - The class label for $\boldsymbol{x}$ is the c that makes $P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta})$ highest.

Naive Bayes classifier

- The posterior for naive Bayes classifier:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{P(\boldsymbol{x} \mid y = c; \boldsymbol{\theta}) P(y = c; \boldsymbol{\theta})}{\sum_{c'} P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta}) P(y = c'; \boldsymbol{\theta})} = \frac{P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})}{\sum_{c'} [P(y = c'; \boldsymbol{\theta}) P(\boldsymbol{x} \mid y = c'; \boldsymbol{\theta})]}$$

- As the denominator is independent of the class labels, the above becomes:

$$P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) \propto P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})$$

- Why is "proportional to" is sufficient, instead of equality, to decide the class label?
  - The class label for $\boldsymbol{x}$ is the c that makes $P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta})$ highest.
  - If c makes $P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta})$ highest, then c will also make $P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc})$ highest:

$$\hat{y} = \operatorname*{argmax}_c P(y = c \mid \boldsymbol{x}; \boldsymbol{\theta}) = \operatorname*{argmax}_c \left[P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right]$$

Naive Bayes classifier (contd.)

- The argmax computation in the previous slide is a product of large number of small floating point numbers (close to 0) that can be indistinguishable from 0.

Naive Bayes classifier (contd.)

- The argmax computation in the previous slide is a product of large number of small floating point numbers (close to 0) that can be indistinguishable from 0.
- As in computing likelihood, we use logarithm (usually natural base) to compute the class label $\hat{y}$:

$$\hat{y} = \underset{c}{\operatorname{argmax}} \left[P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right]$$

$$=$$

Naive Bayes classifier (contd.)

- The argmax computation in the previous slide is a product of large number of small floating point numbers (close to 0) that can be indistinguishable from 0.
- As in computing likelihood, we use logarithm (usually natural base) to compute the class label $\hat{y}$:

$$\begin{aligned}\hat{y} &= \underset{c}{\operatorname{argmax}} \left[P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right] \\ &= \underset{c}{\operatorname{argmax}} \left(\log \left[P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right] \right) \\ &= \end{aligned}$$

Naive Bayes classifier (contd.)

- The argmax computation in the previous slide is a product of large number of small floating point numbers (close to 0) that can be indistinguishable from 0.
- As in computing likelihood, we use logarithm (usually natural base) to compute the class label $\hat{y}$:

$$\begin{aligned}\hat{y} &= \underset{c}{\operatorname{argmax}} \left[P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right] \\ &= \underset{c}{\operatorname{argmax}} \left(\log \left[P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right] \right) \\ &= \underset{c}{\operatorname{argmax}} \left[\log P(y = c; \boldsymbol{\theta}) + \sum_{d=1}^{D} \log P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right]\end{aligned}$$

Naive Bayes classifier (contd.)

- The argmax computation in the previous slide is a product of large number of small floating point numbers (close to 0) that can be indistinguishable from 0.
- As in computing likelihood, we use logarithm (usually natural base) to compute the class label $\hat{y}$:

$$\begin{aligned}\hat{y} &= \underset{c}{\operatorname{argmax}} \left[P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right] \\ &= \underset{c}{\operatorname{argmax}} \left(\log \left[P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right] \right) \\ &= \underset{c}{\operatorname{argmax}} \left[\log P(y = c; \boldsymbol{\theta}) + \sum_{d=1}^{D} \log P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right]\end{aligned}$$

- Question: how do we estimate $\boldsymbol{\theta}$ for the prior $P(y = c; \boldsymbol{\theta})$ and the class conditional density $P(x_d \mid y = c; \boldsymbol{\theta}_{dc})$ for each feature x_d?

Naive Bayes classifier (contd.)

- The argmax computation in the previous slide is a product of large number of small floating point numbers (close to 0) that can be indistinguishable from 0.
- As in computing likelihood, we use logarithm (usually natural base) to compute the class label $\hat{y}$:

$$\begin{aligned}\hat{y} &= \underset{c}{\operatorname{argmax}} \left[P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right] \\ &= \underset{c}{\operatorname{argmax}} \left(\log \left[P(y = c; \boldsymbol{\theta}) \prod_{d=1}^{D} P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right] \right) \\ &= \underset{c}{\operatorname{argmax}} \left[\log P(y = c; \boldsymbol{\theta}) + \sum_{d=1}^{D} \log P(x_d \mid y = c; \boldsymbol{\theta}_{dc}) \right]\end{aligned}$$

- Question: how do we estimate $\boldsymbol{\theta}$ for the prior $P(y = c; \boldsymbol{\theta})$ and the class conditional density $P(x_d \mid y = c; \boldsymbol{\theta}_{dc})$ for each feature x_d?
  - Compute the maximum likelihood estimation of $\boldsymbol{\theta}$ for those distributions!

MLE for class prior

Suppose we have dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n) \colon n = 1, \ldots, N\}$ of N examples. The n-th example consists of a vector $\boldsymbol{x}_n = [x_{n,1}, \ldots, x_{n,d}]^\mathsf{T}$ of d feature values and its corresponding class label y_n.

Solution for MLE of the **class prior** $p(y = c; \boldsymbol{\theta})$ computed for each class c:

$$\theta_c = \frac{N_c}{N}$$

where N_c is the number of examples in $\mathcal{D}$ that belong to class c. (Why does the solution look like this?)

- In general, if we have C classes in total, then we compute C parameter values for the class prior.

MLE for class conditional density: categorical features

$P(\boldsymbol{x}_n \mid y_n = c; \boldsymbol{\theta})$ is the class conditional density where C is the total number of classes. Here, $\boldsymbol{x}_n$ consists of D features and $x_{n,i}$ is its ith feature. Also, N_c denotes the number of examples that belong to class c.

MLE for class conditional density: categorical features

$P(\boldsymbol{x}_n \mid y_n = c; \boldsymbol{\theta})$ is the class conditional density where C is the total number of classes. Here, $\boldsymbol{x}_n$ consists of D features and $x_{n,i}$ is its ith feature. Also, N_c denotes the number of examples that belong to class c.

If the ith feature is categorical with K possible categories $(1, \ldots, K)$, MLE solution:

$$\text{for each class } c\text{, for each category } k\text{, compute } \theta_{i,c,k} = \frac{N_{i,c,k}}{N_c}$$

where $N_{i,c,k}$ is the number of times that the ith feature has the value k among examples of class c.

How many parameter values correspond to a categorical feature with K categories?

MLE for class conditional density: categorical features

$P(\boldsymbol{x}_n \mid y_n = c; \boldsymbol{\theta})$ is the class conditional density where C is the total number of classes. Here, $\boldsymbol{x}_n$ consists of D features and $x_{n,i}$ is its ith feature. Also, N_c denotes the number of examples that belong to class c.

If the ith feature is categorical with K possible categories $(1, \ldots, K)$, MLE solution:

$$\text{for each class } c\text{, for each category } k\text{, compute } \theta_{i,c,k} = \frac{N_{i,c,k}}{N_c}$$

where $N_{i,c,k}$ is the number of times that the ith feature has the value k among examples of class c.

How many parameter values correspond to a categorical feature with K categories?
KC parameter values, i.e., describing $P(x_{n,i} = k \mid y_n = c) = \theta_{i,c,k} = L$ where $k = 1, \ldots, K$ and $c = 1, \ldots, C$.

MLE for class conditional density: real-valued features

If the ith feature is real-valued, we can model the class conditional density for that feature as as univariate Gaussian. MLE solution yields the sample mean and sample variance as the estimated parameter for each class c. That is, we compute **for each class** c:

$$\hat{\mu}_{i,c} = \frac{1}{N_c} \sum_{n:\ y_n=c} x_{n,i} \qquad \text{(mean of the } i\text{th feature among examples of class } c\text{)}$$

$$\hat{\sigma}^2_{i,c} = \frac{1}{N_c} \sum_{n:\ y_n=c} (x_{n,i} - \hat{\mu}_{i,c})^2 \qquad \text{(variance of the } i\text{th feature among examples of class } c\text{)}$$

MLE for class conditional density: real-valued features

If the ith feature is real-valued, we can model the class conditional density for that feature as as univariate Gaussian. MLE solution yields the sample mean and sample variance as the estimated parameter for each class c. That is, we compute **for each class** c:

$$\hat{\mu}_{i,c} = \frac{1}{N_c} \sum_{n:\, y_n = c} x_{n,i} \qquad \text{(mean of the } i\text{th feature among examples of class } c\text{)}$$

$$\hat{\sigma}^2_{i,c} = \frac{1}{N_c} \sum_{n:\, y_n = c} (x_{n,i} - \hat{\mu}_{i,c})^2 \qquad \text{(variance of the } i\text{th feature among examples of class } c\text{)}$$

Thus, for the ith feature, we have C Gaussian conditional densities.

Note: the use of N_c as denominator in $\hat{\sigma}^2_{i,c}$ yields a biased estimator of the population variance of the ith feature values.

- The expected value of the uncorrected sample variance above does **not** equal the variance of the population of the ith feature values.
- To obtain an unbiased variance estimator, we can use $N_c - 1$ as denominator (Bessel's correction).
- But, Bessel's correction does **not** yield an unbiased estimator of standard deviation.

MLE for class conditional density: real-valued features (contd.)

Given the parameters of the Gaussian conditional densities, we have $P(x_{n,i} \mid y_n = c) \sim \mathcal{N}(\hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c})$, and we can estimate the class conditional density around feature value $x_{n,i} = a$ as:

$$P(x_{n,i} \approx a \mid y_n = c; \hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c}) \approx P(a - \tfrac{\varepsilon}{2} \leqslant x_{n,i} \leqslant a + \tfrac{\varepsilon}{2} \mid y_n = c; \hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c})$$

MLE for class conditional density: real-valued features (contd.)

Given the parameters of the Gaussian conditional densities, we have $P(x_{n,i} \mid y_n = c) \sim \mathcal{N}(\hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c})$, and we can estimate the class conditional density around feature value $x_{n,i} = a$ as:

$$P(x_{n,i} \approx a \mid y_n = c; \hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c}) \approx P(a - \tfrac{\varepsilon}{2} \leqslant x_{n,i} \leqslant a + \tfrac{\varepsilon}{2} \mid y_n = c; \hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c})$$

One can use the pdf f of Gaussian distribution to compute the above (picking small ε), but it is straightforward to see that

$$P(a - \tfrac{\varepsilon}{2} \leqslant x_{n,i} \leqslant a + \tfrac{\varepsilon}{2} \mid y_n = c; \hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c}) \approx \varepsilon f_{\hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c}}(a)$$

MLE for class conditional density: real-valued features (contd.)

Given the parameters of the Gaussian conditional densities, we have $P(x_{n,i} \mid y_n = c) \sim \mathcal{N}(\hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c})$, and we can estimate the class conditional density around feature value $x_{n,i} = a$ as:

$$P(x_{n,i} \approx a \mid y_n = c; \hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c}) \approx P(a - \tfrac{\varepsilon}{2} \leqslant x_{n,i} \leqslant a + \tfrac{\varepsilon}{2} \mid y_n = c; \hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c})$$

One can use the pdf f of Gaussian distribution to compute the above (picking small ε), but it is straightforward to see that

$$P(a - \tfrac{\varepsilon}{2} \leqslant x_{n,i} \leqslant a + \tfrac{\varepsilon}{2} \mid y_n = c; \hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c}) \approx \varepsilon f_{\hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c}}(a)$$

Thus, one can simply use the pdf $f_{\hat{\mu}_{i,c}, \hat{\sigma}^2_{i,c}}(a)$ in the argmax computation of the label without altering the result.

Example

Weather	Temp.	Hum.	Wind	Play
sunny	64	65	false	yes
sunny	68	70	false	yes
cloudy	69	70	false	yes
cloudy	70	75	false	yes
cloudy	72	80	false	yes
cloudy	78	85	false	yes
rainy	76	86	true	yes
rainy	78	91	true	yes
rainy	82	90	true	yes
sunny	65	70	false	no
sunny	71	85	false	no
sunny	74	90	true	no
rainy	80	91	true	no
rainy	85	94	true	no

Decide whether to play tennis if the weather is sunny, temperature is 66, humidity is 90, and the condition is windy.

Example (contd.)

Weather	Temp.	Hum.	Wind	Play
sunny	64	65	false	yes
sunny	68	70	false	yes
cloudy	69	70	false	yes
cloudy	70	75	false	yes
cloudy	72	80	false	yes
cloudy	78	85	false	yes
rainy	76	86	true	yes
rainy	78	91	true	yes
rainy	82	90	true	yes
sunny	65	70	false	no
sunny	71	85	false	no
sunny	74	90	true	no
rainy	80	91	true	no
rainy	85	94	true	no

$$P(p = yes) = 9/14 \quad P(p = no) = 5/14$$

$$P(we = sunny \mid yes) = 2/9 \quad P(we = cloudy \mid yes) = 4/9$$
$$P(we = rainy \mid yes) = 3/9 \quad P(we = sunny \mid no) = 3/5$$
$$P(we = cloudy \mid no) = 0 \quad P(we = rainy \mid no) = 2/5$$

$$\mu_{t,yes} = 73 \quad \sigma_{t,yes} = 5.83$$
$$\mu_{t,no} = 75 \quad \sigma_{t,no} = 7.78$$
$$\mu_{h,yes} = 79 \quad \sigma_{h,yes} = 9.45$$
$$\mu_{h,no} = 86 \quad \sigma_{h,no} = 9.51$$

$$P(wi = false \mid yes) = 6/9 \quad P(wi = true \mid yes) = 3/9$$
$$P(wi = false \mid no) = 2/5 \quad P(wi = true \mid no) = 3/5$$

Example (contd.)

Weather	Temp.	Hum.	Wind	Play
sunny	64	65	false	yes
sunny	68	70	false	yes
cloudy	69	70	false	yes
cloudy	70	75	false	yes
cloudy	72	80	false	yes
cloudy	78	85	false	yes
rainy	76	86	true	yes
rainy	78	91	true	yes
rainy	82	90	true	yes
sunny	65	70	false	no
sunny	71	85	false	no
sunny	74	90	true	no
rainy	80	91	true	no
rainy	85	94	true	no

$$P(t = 66|yes) \propto f_{73,5.83}(66) = \frac{1}{5.83 \cdot \sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{66-73}{5.83})^2)$$
$$\approx 0.0333$$
$$P(t = 66|no) \propto f_{75,7.78}(66) = \frac{1}{7.78 \cdot \sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{66-75}{7.78})^2)$$
$$\approx 0.0263$$
$$P(h = 90|yes) \propto f_{79,9.45}(90) = \frac{1}{9.45 \cdot \sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{90-79}{9.45})^2)$$
$$\approx 0.0214$$
$$P(h = 90|no) \propto f_{86,9.51}(90) = \frac{1}{9.51 \cdot \sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{90-86}{9.51})^2)$$
$$\approx 0.0384$$

Example (contd.)

Weather	Temp.	Hum.	Wind	Play
sunny	64	65	false	yes
sunny	68	70	false	yes
cloudy	69	70	false	yes
cloudy	70	75	false	yes
cloudy	72	80	false	yes
cloudy	78	85	false	yes
rainy	76	86	true	yes
rainy	78	91	true	yes
rainy	82	90	true	yes
sunny	65	70	false	no
sunny	71	85	false	no
sunny	74	90	true	no
rainy	80	91	true	no
rainy	85	94	true	no

$$
\begin{aligned}
&P(p = yes \mid we = sunny, t = 66, h = 90, wi = true) \\
&\propto P(yes) \cdot P(sunny|yes) \cdot f_{73,5.83}(66) \cdot f_{79,9.45}(90) \\
&\qquad \cdot P(true|yes) \\
&= 9/14 \cdot 2/9 \cdot 0.0333 \cdot 0.0214 \cdot 3/9 \\
&= 3.393 \times 10^{-5}
\end{aligned}
$$

$$
\begin{aligned}
&P(p = no \mid we = sunny, t = 66, h = 90, wi = true) \\
&\propto P(no) \cdot P(sunny|no) \cdot f_{75,7.78}(66) \cdot f_{86,9.51}(90) \\
&\qquad \cdot P(true|no) \\
&= 5/14 \cdot 3/5 \cdot 0.0263 \cdot 0.0384 \cdot 3/5 \\
&= 1.299 \times 10^{-4}
\end{aligned}
$$

Hence, we decide that we do not play tennis.

Naive Bayes issues

- What if a feature has no example that belongs to a certain class?
  - Feature “weather = cloudy” has no example that belongs to class “no”, i.e., $P(we = cloudy|no) = 0$.
  - Class posterior for “no” for examples with feature “weather = cloudy” is

Naive Bayes issues

- What if a feature has no example that belongs to a certain class?
  - Feature “weather = cloudy” has no example that belongs to class “no”, i.e., $P(we = cloudy|no) = 0$.
  - Class posterior for “no” for examples with feature “weather = cloudy” is 0 – regardless of other feature values.

Naive Bayes issues

- What if a feature has no example that belongs to a certain class?
  - Feature “weather = cloudy” has no example that belongs to class “no”, i.e., $P(we = cloudy|no) = 0$.
  - Class posterior for “no” for examples with feature “weather = cloudy” is 0 – regardless of other feature values.
- Solution: use Laplace estimator:
  - Add one to **all counts** before normalizing, i.e., we use for categorical feature i that has K possible values:

$$\theta_{i,c,k} = \frac{1 + N_{i,c,k}}{K + N_c}$$

    where $N_{i,c,k}$ is the number of times that the ith feature has the value k among examples of class c.
  - E.g., in the previous example, we use $P(sunny|yes) = (1 + 2)/(3 + 9) = 3/12$ and $P(cloudy|no) = (1 + 0)/(3 + 5) = 1/8$.
  - This avoids zero probabilities, and usually stabilizes the overall probability estimation.
  - In some cases, one can also choose to add the counts with a value other than 1.

Naive Bayes issues

- What if a feature has no example that belongs to a certain class?
  - Feature “weather = cloudy” has no example that belongs to class “no”, i.e., $P(we = cloudy|no) = 0$.
  - Class posterior for “no” for examples with feature “weather = cloudy” is 0 – regardless of other feature values.
- Solution: use Laplace estimator:
  - Add one to **all counts** before normalizing, i.e., we use for categorical feature i that has K possible values:

    $$\theta_{i,c,k} = \frac{1 + N_{i,c,k}}{K + N_c}$$

    where $N_{i,c,k}$ is the number of times that the ith feature has the value k among examples of class c.
  - E.g., in the previous example, we use $P(sunny|yes) = (1 + 2)/(3 + 9) = 3/12$ and $P(cloudy|no) = (1 + 0)/(3 + 5) = 1/8$.
  - This avoids zero probabilities, and usually stabilizes the overall probability estimation.
  - In some cases, one can also choose to add the counts with a value other than 1.
- What if the test example contain some missing features?

Naive Bayes issues

- What if a feature has no example that belongs to a certain class?
  - Feature “weather = cloudy” has no example that belongs to class “no”, i.e., $P(we = cloudy|no) = 0$.
  - Class posterior for “no” for examples with feature “weather = cloudy” is 0 – regardless of other feature values.
- Solution: use Laplace estimator:
  - Add one to **all counts** before normalizing, i.e., we use for categorical feature i that has K possible values:

    $$\theta_{i,c,k} = \frac{1 + N_{i,c,k}}{K + N_c}$$

    where $N_{i,c,k}$ is the number of times that the ith feature has the value k among examples of class c.
  - E.g., in the previous example, we use $P(sunny|yes) = (1 + 2)/(3 + 9) = 3/12$ and $P(cloudy|no) = (1 + 0)/(3 + 5) = 1/8$.
  - This avoids zero probabilities, and usually stabilizes the overall probability estimation.
  - In some cases, one can also choose to add the counts with a value other than 1.
- What if the test example contain some missing features?
  - Then, we can simply ignore the likelihood term for the missing features when computing the class posterior without changing the result.
  - This is due to the fact that we can marginalize out the missing features in naive Bayes model.