Data Mining and Machine Learning

Part 4: Bayesian Learning

Melih Kandemir

University of Southern Denmark

DM566, Spring 2023

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

Can we Formalize "Confidence"?

We encountered a measure called "confidence" that should tell us how reliable a discovered association rule is.
We interpreted:

- The higher the confidence for some rule '$X \Rightarrow Y$', the more likely Y is present in transactions that contain X.
- The confidence is an estimate of the conditional probability of Y given X.

Can we formalize this interpretation?

Recommended Reading:

Mitzenmacher and Upfal [2017], Chapter 1.

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

Sample Space

The sample space Ω is the set of all (disjoint) possible outcomes of some random process.

Examples:

- *If we role a dice, we have* $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- *If we flip a coin, we have* $\Omega = \{H, T\}$.

Events

A subset $E \subseteq \Omega$ of individual outcomes of a random process can define an "event".

Examples:

- *We role a die. Every element of* $\Omega = \{1, 2, 3, 4, 5, 6\}$ *is a simple or elementary event.*
- *We could be interested in the event "The die shows an even number"* $= \{2, 4, 6\} \subseteq \Omega$.
- *We flip a coin. We could have the elementary event "head"* $\subseteq \Omega$.

A family of sets $\mathcal{F}$ represents the allowable events. Each set in $\mathcal{F}$ is a subset of Ω, i.e., $\mathcal{F} \subseteq \wp(\Omega)$.

Probability Function

Definition 1.1 (Probability Function)

A probability function is any function $\Pr : \mathcal{F} \to \mathbb{R}$ that satisfies the following conditions:

1. $\forall E : 0 \leq \Pr(E) \leq 1$;
2. $\Pr(\Omega) = 1$; and
3. for any finite or countably infinite sequence of pairwise mutually disjoint events $E_1, E_2, E_3, \ldots$:

$$\Pr\left(\bigcup_{i \geq 1} E_i\right) = \sum_{i \geq 1} \Pr(E_i).$$

Probability Space

Definition 1.2 (Probability Space)

A probability space is given by three components:

1. a sample space Ω;
2. the allowable events $\mathcal{F} \subseteq \wp(\Omega)$; and
3. a probability function $\Pr : \mathcal{F} \to \mathbb{R}$.

Event Combinations

Because events are sets, we can use standard set theory notation to express combinations of events.

- $E_1 \cap E_2$ denotes the occurrence of both, E_1 and E_2 (i.e., their co-occurrence).
- $E_1 \cup E_2$ denotes the occurrence of either E_1 or E_2 (or both).
- $E_1 \setminus E_2$ denotes the occurrence of event E_1 without E_2 occurring as well.
- $\overline{E} = \Omega \setminus E$ denotes the complementary event of E.

Event Combinations

Examples:

Suppose we roll two dice. Given events E_1 and E_2:

E_1 *the first die is a 1*

E_2 *the second die is a 1*

- $E_1 \cap E_2$: *both dice are 1*
- $E_1 \cup E_2$: *at least one of the dice lands on 1.*
- $E_1 \setminus E_2$: *the first die is a 1 and the second die is not.*

Event Combinations

Examples:

Let E be the event that by rolling a die we obtain an even number.

- *Then $\overline{E}$ is the event that we obtain an odd number.*
- *What are the events $\overline{E_1}$, $\overline{E_1 \cup E_2}$, $\overline{E_1 \cap E_2}$?*

Example: One Die

Consider the random process defined by the outcome of rolling a die:

$$\Omega_{\text{die}_1} = \{1, 2, 3, 4, 5, 6\}$$

Assuming a fair die, all sides have equal probability, thus:

$$\Pr(\{1\}) = \Pr(\{2\}) = \ldots = \Pr(\{6\}) = \frac{1}{6}$$

The probability of the event "odd number" is

$$\Pr(\{1, 3, 5\}) = \Pr(\{1\}) + \Pr(\{3\}) + \Pr(\{5\}) = \frac{1}{2}$$

Example: Two Dice

Consider the random process defined by the outcome of rolling two (fair) dice:

$$\Omega = \Omega_{\text{die}_1} \times \Omega_{\text{die}_2} = \{(i,j) | 1 \leq i,j \leq 6\}$$

Each (ordered) combination has a probability of $\frac{1}{36}$.

Example:

Probability of the event "sum $= 2$":

$$\Pr(\{(1,1)\}) = \frac{1}{36}$$

Example: Two Dice

Consider the random process defined by the outcome of rolling two (fair) dice:

$$\Omega = \Omega_{\text{die}_1} \times \Omega_{\text{die}_2} = \{(i,j) | 1 \leq i,j \leq 6\}$$

Each (ordered) combination has a probability of $\frac{1}{36}$.

Example:

Probability of the event "sum $= 3$":

$$\Pr(\{(1,2),(2,1)\}) = \Pr(\{(1,2)\}) + \Pr(\{(2,1)\}) = \frac{2}{36} = \frac{1}{18}$$

Example: Two Dice

Consider the random process defined by the outcome of rolling two (fair) dice:

$$\Omega = \Omega_{\text{die}_1} \times \Omega_{\text{die}_2} = \{(i,j) | 1 \leq i,j \leq 6\}$$

Each (ordered) combination has a probability of $\frac{1}{36}$.

Example:

Probability of the event E_1 =“sum bounded by 6”:

$$\begin{aligned} E_1 = & \{(1,1),(1,2),(1,3),(1,4),(1,5),(2,1),(2,2), \\ & (2,3),(2,4),(3,1),(3,2),(3,3),(4,1),(4,2),(5,1)\} \end{aligned}$$

$$\Pr(E_1) = \frac{15}{36}$$

Example: Two Dice

Consider the random process defined by the outcome of rolling two (fair) dice:

$$\Omega = \Omega_{\text{die}_1} \times \Omega_{\text{die}_2} = \{(i,j) | 1 \le i,j \le 6\}$$

Each (ordered) combination has a probability of $\frac{1}{36}$.

Example:

$E_2 =$ *"both dice have odd numbers"*:

$$E_2 = \{(1,1),(1,3),(1,5),(3,1),(3,3),(3,5),(5,1),(5,3),(5,5)\}$$

$$\Pr(E_2) = \frac{1}{4}$$

Example: Two Dice

Consider the random process defined by the outcome of rolling two (fair) dice:

$$\Omega = \Omega_{\text{die}_1} \times \Omega_{\text{die}_2} = \{(i,j) | 1 \leq i,j \leq 6\}$$

Each (ordered) combination has a probability of $\frac{1}{36}$.

Example:

$E_3 =$ *"sum bounded by 6 and both dice have odd numbers":*

$$\begin{aligned}\Pr(E_3) &= \Pr(E_1 \cap E_2) \\ &= \Pr(\{(1,1),(1,3),(1,5),(3,1),(3,3),(5,1)\}) \\ &= \frac{1}{6}\end{aligned}$$

Combined Probability

Lemma 1.3 (Combined Probability)

For any two events E_1 and E_2:

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$$

Proof.

$$\begin{aligned}
\Pr(E_1) &= \Pr(E_1 \setminus (E_1 \cap E_2)) + \Pr(E_1 \cap E_2) \\
\Pr(E_2) &= \Pr(E_2 \setminus (E_1 \cap E_2)) + \Pr(E_1 \cap E_2) \\
\Pr(E_1 \cup E_2) &= \Pr(E_1 \setminus (E_1 \cap E_2)) \\
&\quad + \Pr(E_2 \setminus (E_1 \cap E_2)) \\
&\quad + \Pr(E_1 \cap E_2)
\end{aligned}$$

$\square$

Union Bound

Lemma 1.4 (Union Bound)

For any finite or countably infinite sequence of events $E_1, E_2, E_3, \ldots$:

$$\Pr\left(\bigcup_{i \geq 1} E_i\right) \leq \sum_{i \geq 1} \Pr(E_i).$$

Difference from condition 3 in Definition 1.1?

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Axioms of Probability
Independence and Conditional Probability
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary

Independent Events

Definition 1.5 (Independent Events)

Two events E and F are *independent* if and only if

$$\Pr(E \cap F) = \Pr(E) \cdot \Pr(F).$$

More generally, events $E_1, E_2, \ldots, E_k$ are mutually independent if and only if

$$\forall I \subseteq [1, k] : \Pr\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \Pr(E_i).$$

Independent Events: Intuition

- If events A and B are *independent* then knowledge about event A does not change the probability of B.
- If A and B are *not independent*, then we can quantify the conditional probability of A subject to our kowledge of event B.

Example:

Probability of the event E_1 "outcome of a die roll is even": $\frac{3}{6}$.
Probability of the event E_2 "the outcome is ≤ 4": $\frac{4}{6}$.
Probability of $E_1 \cap E_2$: "the outcome is even and is ≤ 4":

$$\Pr(E_1 \cap E_2) = \frac{2}{6} = \frac{12}{36} = \frac{3}{6} \cdot \frac{4}{6} = \Pr(E_1) \cdot \Pr(E_2)$$

$\Rightarrow$ *The two events are independent.*

Independent Events: Intuition

- If events A and B are *independent* then knowledge about event A does not change the probability of B.
- If A and B are *not independent*, then we can quantify the conditional probability of A subject to our kowledge of event B.

Example:

Probability of the event E_1 "outcome of a die roll is even": $\frac{3}{6}$.
Probability of the event E_2 "the outcome is ≤ 3": $\frac{3}{6}$.
Probability of $E_1 \cap E_2$: "the outcome is even and is ≤ 3":

$$\Pr(E_1 \cap E_2) = \frac{1}{6} = \frac{6}{36} \neq \frac{9}{36} = \frac{3}{6} \cdot \frac{3}{6} = \Pr(E_1) \cdot \Pr(E_2)$$

$\Rightarrow$ *The two events are not independent.*

Conditional Probability

Definition 1.6 (Conditional Probability)

The *conditional probability* that event E occurs given that event F occurs is

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}$$

The conditional probability is well-defined only if $\Pr(F) > 0$.

Note that:

If E and F are independent and $\Pr(F) \neq 0$, we have:

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{\Pr(E)\Pr(F)}{\Pr(F)} = \Pr(E)$$

Conditional Probability: Intuition

- We look for the probability of $E \cap F$ within the sets of events defined by F.
- Because F restricts the sample space, we normalize the probabilities by dividing by $\Pr(F)$.
- If E and F are independent, information about F should not affect the probability of E.

Example:

Probability of the event E_1 "outcome of a die roll is even": $\frac{3}{6}$.
Probability of the event E_2 "the outcome is ≤ 4": $\frac{4}{6}$.
Probability of $E_1 \cap E_2$: "the outcome is even and is ≤ 4":

$$\Pr(E_1 \cap E_2) = \frac{2}{6} = \frac{12}{36} = \frac{3}{6} \cdot \frac{4}{6} = \Pr(E_1) \cdot \Pr(E_2)$$

$\Rightarrow$ *The two events are independent.*

Conditional Probability: Intuition

- We look for the probability of $E \cap F$ within the sets of events defined by F.
- Because F restricts the sample space, we normalize the probabilities by dividing by $\Pr(F)$.
- If E and F are independent, information about F should not affect the probability of E.

Example:

Probability of the event E_1 "outcome of a die roll is even": $\frac{3}{6}$.
Probability of the event E_2 "the outcome is ≤ 3": $\frac{3}{6}$.
Probability of $E_1 \cap E_2$: "the outcome is even and is ≤ 3":

$$\Pr(E_1 \cap E_2) = \frac{1}{6} = \frac{6}{36} \neq \frac{9}{36} = \frac{3}{6} \cdot \frac{3}{6} = \Pr(E_1) \cdot \Pr(E_2)$$

$\Rightarrow$ *The two events are not independent.*

The Condition Defines a Probability Space

$\Pr(X|E)$ defines a proper probability function on the sample space E (cf. Definitions 1.1 and 1.2):

$$\Pr(\emptyset|E) = \frac{\Pr(\emptyset \cap E)}{\Pr(E)} = \frac{\Pr(\emptyset)}{\Pr(E)} = 0$$

$$\Pr(E|E) = \frac{\Pr(E \cap E)}{\Pr(E)} = \frac{\Pr(E)}{\Pr(E)} = 1$$

For any two disjoint events A and B:

$$\begin{aligned}\Pr(A \cup B|E) &= \frac{\Pr((A \cup B) \cap E)}{\Pr(E)} \\ &= \frac{\Pr(A \cap E) + \Pr(B \cap E)}{\Pr(E)} \\ &= \Pr(A|E) + \Pr(B|E)\end{aligned}$$

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning

- Axioms of Probability
- Independence and Conditional Probability
  - Quality Measures for Association Rules Revisited
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

Quality Measures for Association Rules

Support: $s(X \Rightarrow Y) = s(X \cup Y)$

or in relative terms: frequency $f(X \cup Y) = \frac{s(X \cup Y)}{|\mathcal{D}|}$

Confidence: $\mathrm{conf}(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$

Lift: $\mathit{Lift}(X \Rightarrow Y) = \frac{\mathrm{conf}(X \Rightarrow Y)}{f(Y)}$

Jaccard: $\mathit{Jaccard}(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X) + s(Y) - s(X \cup Y)}$

conviction: $\mathit{conviction}(X \Rightarrow Y) = \frac{1 - f(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$

Probabilistic Interpretation: Support (Frequency)

The frequency of an itemset in the database can be seen as an empirical estimate of its probability, given the sample represented by the database:

$$\Pr(X) = \frac{s(X)}{|\mathcal{D}|}$$

Note that:

$$\Pr(X \cap Y) = \frac{s(X \cup Y)}{|\mathcal{D}|}$$

Although X and Y are sets in both cases,

- *probabilistically, $\cap$ denotes the co-occurrence of events,*
- *while for itemsets, $\cup$ denotes that both itemsets need to be present.*

Probabilistic Interpretation: Confidence

DM566
Melih Kandemir
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Quality Measures for Association Rules Revisited
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
References

The confidence is the conditional probability that a transaction contains the consequent Y given that it contains the antecedent X:

$$\begin{aligned}\mathrm{conf}(X \Rightarrow Y) &= \frac{s(X \cup Y)}{s(X)} \\ &= \frac{\Pr(X \cap Y)}{\Pr(X)} \\ &= \Pr(Y|X)\end{aligned}$$

- The confidence of a rule $X \Rightarrow Y$ is not a useful measure unless we compare it with the frequency of Y, i.e., the prior (unconditional) probability.
- If we have $\Pr(Y|X) < \Pr(Y)$ this means that in the presence of X, Y becomes less likely as it is unconditionally!
  (Not the rule, but this fact could be interesting, though!)

Probabilistic Interpretation: Lift

We can see Lift as normalization of the confidence by the prior probability of the consequent:

$$
\begin{aligned}
\mathit{Lift}(X \Rightarrow Y) &= \frac{\operatorname{conf}(X \Rightarrow Y)}{f(Y)} \\
&= \frac{\Pr(X \cap Y)}{\Pr(X)\Pr(Y)}
\end{aligned}
$$

- ratio of the observed joined probability of X and Y to the joint probability expected for statistically independent events (Definition 1.5).
- Lift is a (symmetric!) measure for the surprise of a rule.
- Values around 1: boring.
- Much smaller/larger values: interesting!

Probabilistic Interpretation: Jaccard

The Jaccard coefficient in general is a measure for the similarity between two sets:

$$
\begin{aligned}
\textit{Jaccard}(X \Rightarrow Y) &= \frac{s(X \cup Y)}{s(X) + s(Y) - s(X \cup Y)} \\
&= \frac{\Pr(X \cap Y)}{\Pr(X) + \Pr(Y) - \Pr(X \cap Y)} \\
&= \frac{\Pr(X \cap Y)}{\Pr(X \cup Y)} \qquad \text{(Lemma 1.3)}
\end{aligned}
$$

- A symmetric measure of how often both, X and Y, occur simultaneously, relative to the occurrence of both or either overall.
- Similarity of the itemsets X and Y based on their individual occurrences and their co-occurrences.

Probabilistic Interpretation: Conviction

Conviction of a rule measures the expected error: how often does X occur in a transaction where Y does not?
(How often does the rule fail?)

$$
\begin{aligned}
\textit{conviction}(X \Rightarrow Y) &= \frac{1 - f(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)} \\
&= \frac{\Pr\left(\overline{Y}\right)}{1 - \frac{\Pr(X \cap Y)}{\Pr(X)}} = \frac{\Pr(X)\Pr\left(\overline{Y}\right)}{\Pr(X) - \Pr(X \cap Y)} \\
&= \frac{\Pr(X)\Pr\left(\overline{Y}\right)}{\Pr\left(X \cap \overline{Y}\right)} = \frac{1}{\textit{Lift}\left(X \Rightarrow \overline{Y}\right)}
\end{aligned}
$$

- compares the observed joint probability of X and $\overline{Y}$ with their joint probability expected for independence
- asymmetric measure

Recommended Reading

Recommended Reading:

On the probabilistic interpretation of (even more) quality measures for association rules: Zaki and Meira Jr. [2014], Chapter 12.1.

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning

- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

The Law of Total Probability

Theorem 1.7 (The Law of Total Probability)

Let $E_1, E_2, \ldots, E_n$ be mutually disjoint events in the sample space Ω, and let $\bigcup_{i=1}^n E_i = \Omega$. Then

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B \cap E_i) = \sum_{i=1}^{n} \Pr(B|E_i) \Pr(E_i).$$

The Law of Total Probability

Proof.

Since the events $E_i (i = 1, \ldots, n)$ are disjoint and cover the entire sample space Ω, it follows that

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B \cap E_i).$$

Further, by Definition 1.6,

$$\sum_{i=1}^{n} \Pr(B \cap E_i) = \sum_{i=1}^{n} \Pr(B|E_i) \Pr(E_i).$$

$\square$

Bayes' Rule

Theorem 1.8 (Bayes' Rule)

Let $E_1, \ldots, E_n$ be mutually disjoint events, and let $\bigcup_{i=1}^n E_i = \Omega$. Then for any other event B, $\Pr(B) > 0$, $j = 1, \ldots, n$:

$$\Pr(E_j|B) = \frac{\Pr(E_j \cap B)}{\Pr(B)} \tag{1.1}$$

$$= \frac{\Pr(B|E_j)\Pr(E_j)}{\sum_{i=1}^n \Pr(B|E_i)\Pr(E_i)} \tag{1.2}$$

Proof.

From Eq. 1 to Eq. 2, we use Definition 1.6 in the numerator, and Theorem 1.7 in the denominator. □

Bayes' Rule (Simple Form)

In its simple form, we have only two events, A and B, $\Pr(A) \neq 0$, $\Pr(B) \neq 0$:

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

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

$$\Rightarrow \Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$$

$$\Rightarrow \Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

We do not require exhaustiveness of A or B here (i.e., $A \subseteq \Omega$, $B \subseteq \Omega$), since we do not apply Theorem 1.7, only Definition 1.6.

Bayes' Rule: Example 1

- We are given three coins, two of the coins are fair and the third coin is biased, showing head with probability $\frac{2}{3}$. We need to identify the biased coin.
- We flip each of the coins. The first and second coins come up with head, the third coin comes up with tail.
- What is the probability that the first coin is the biased one?
- Let E_i be the event that the i-th coin is the biased one, and let B be the event that the three coin flips came up head, head, tail.
- Prior probability: $\Pr(E_i) = \frac{1}{3}$ for $i = 1, 2, 3$.

Bayes' Rule: Example 1

$$\Pr(B|E_1) = \Pr(B|E_2) = \frac{2}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{6}$$

and

$$\Pr(B|E_3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{12}.$$

Thus, according to Bayes' rule:

$$\Pr(E_1|B) = \frac{\Pr(B|E_1)\Pr(E_1)}{\sum_{i=1}^{3} \Pr(B|E_i)\Pr(E_i)} = \frac{2}{5}.$$

The experiment increases the probability that the first coin is the biased one from $\frac{1}{3}$ to $\frac{2}{5}$.

Bayes' Rule: Example 2

- A doctor sees a patient with fever and rash.
- 80% of patient with flu, 45% of allergy patients, and 90% of infection patients have these symptoms.
- The doctor knows that 50% of the patients she sees have flu, 40% have allergy, and 10% have an infection.
- Should the doctor treat the patient for infection?

Bayesian Reasoning: The General Pattern

- There are alternative models to explain a fact.
- Each model defines a probability for the observed data.
- Which model is the best (i.e., the most likely) explanation?
- $E_1, E_2, \ldots, E_n$ are the alternative models.
- B is the observed data.
- For each model we know $\Pr(B|E_j)$ (i.e., how well the model explains the facts).

$$\Pr(E_j|B) = \frac{\Pr(E_j \cap B)}{\Pr(B)} = \frac{\Pr(B|E_j)\Pr(E_j)}{\sum_{i=1}^{n}\Pr(B|E_i)\Pr(E_i)}$$

- Difficulty: How do we know $\Pr(E_j)$?

Bayesian Approach

- Start with a *prior* model, giving some initial value to the model parameters.
- This model is then modified by incorporating new observations, to obtain a *posterior* model that captures the new information.

Example:

- A test shows that a patient has an infection.
- The test has 10% error rate.
- What is the probability that the patient has an infection?

Bayesian Approach: Example

- $A =$ "test is positive (i.e., the test says that the patient has an infection)"
- $B =$ "the patient actually has an infection"

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(A|B)\Pr(B)}{\Pr(A|B)\Pr(B) + \Pr\left(A|\overline{B}\right)\Pr\left(\overline{B}\right)}$$

- What is $\Pr(B)$?

Without any prior knowledge we set $\Pr(B) = \Pr\left(\overline{B}\right) = \frac{1}{2}$:

$$\Pr(B|A) = \frac{\frac{9}{10} \cdot \frac{1}{2}}{\frac{9}{10} \cdot \frac{1}{2} + \frac{1}{10} \cdot \frac{1}{2}} = \frac{9}{10}$$

The estimate is dominated by the reliability of the test.

Bayesian Approach: Example

DM566
Melih Kandemir
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
Summary
References

- $A =$ "test is positive (i.e., the test says that the patient has an infection)"
- $B =$ "the patient actually has an infection"

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(A|B)\Pr(B)}{\Pr(A|B)\Pr(B) + \Pr\left(A|\overline{B}\right)\Pr\left(\overline{B}\right)}$$

- What is $\Pr(B)$?

Assume that we know a priori that the probability of the patient being infected is 80%. We set $\Pr(B) = \frac{4}{5}$:

$$\Pr(B|A) = \frac{\frac{9}{10} \cdot \frac{4}{5}}{\frac{9}{10} \cdot \frac{4}{5} + \frac{1}{10} \cdot \frac{1}{5}} = \frac{36}{37} \approx 0.97$$

The *posterior* probability is sensitive to the choice of *prior* probabilities.

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
  - k Nearest Neighbor Classification Revisited
  - Bayesian Learning
  - Naïve Bayes Classifier
- Summary

Recommended Reading

Recommended Reading:

- *Mitchell [1997], Chapter 6.*
- *Zaki and Meira Jr. [2014], Chapter 18.*

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
  - k Nearest Neighbor Classification Revisited
  - Bayesian Learning
  - Naïve Bayes Classifier
- Summary

Probabilistic Classification

DM566
Melih Kandemir
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
Naïve Bayes
Summary
References

- A classifier predicts for some object x_q which class c_i it belongs to.
- Often, the prediction can be expressed as probability estimate
$$\Pr(c_i|x_q)$$
- The classifier would then decide for the most likely class:
$$h(x_q) = \arg\max_{c_i \in C} \Pr(c_i|x_q)$$
- Often, this estimate is based on an estimate of how likely the object would be, if it would belong to this or to that class:
$$\Pr(x_q|c_i)$$
How well can the object be explained if it belongs to a given class?

Proportions of Classes in the Decision Set

DM566
Melih Kandemir
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
Naïve Bayes
Summary
References

- Let $x_1, \ldots, x_k = kNN(x_q)$ be the k nearest neighbors of instance x_q, i.e., the decision set for the instance x_q.
- For a given instance x_q and classes $C = \{c_1, \ldots, c_m\}$:

$$E_j = \text{“}f(x_q) = c_j\text{”}$$

$$\Omega = \bigcup_{j=1}^{m} E_j$$

- The relative frequency of a class c_j in the decision set $kNN(x_q)$ is an empirical estimate of the probability of the event “$f(x_q) = c_j$”:

$$\Pr(E_j|x_q) = \frac{|\{x_i|x_i \in kNN(x_q) \land f(x_i) = c_j\}|}{k}$$

$$= \frac{\sum_{i=1}^{k} \delta(c_j, f(x_i))}{k} \text{ with } \delta(a,b) = \begin{cases} 1 \text{ if } a = b \\ 0 \text{ otherwise} \end{cases}$$

Proportions of Classes in the Decision Set

- We can therefore interpret the composition of the decision set as "class probability vector".
- For classes $C = \{c_1, \ldots, c_m\}$, the decision set for x_q yields a vector

$$\langle p_1, \ldots, p_m \rangle$$

where

$$\begin{aligned} p_j &= \Pr\left(\{f(x_q) = c_j\} | x_q\right) \\ &= \frac{|\{x_i | x_i \in kNN(x_q) \wedge f(x_i) = c_j\}|}{k} \\ &= \frac{\sum_{i=1}^{k} \delta(c_j, f(x_i))}{k} \end{aligned}$$

- How will the quality of the probability estimate depend on k?

Decision Rule — Weight by Distance

As discussed (Slides **??**, **??**), we can introduce a weight to the components in the decision rule:

$$h(x_q) = \arg\max_{c \in C} \sum_{i=1}^{k} w_i \delta(c, f(x_i))$$

- e.g.: $w_i = \frac{1}{\text{dist}(x_i, x_q)^2}$

$$\Pr{}_w(\{f(x_q) = c_j\} | x_q) = \frac{\sum_{i=1}^{k} w_i \delta(c_j, f(x_i))}{\sum_{i=1}^{k} w_i}$$

- How do these weights change the dependency on k?

kNN Classifier as Application of Bayes' Rule

DM566

Melih Kandemir

Bayesian Learning

Axioms of Probability

Independence and Conditional Prob.

Total Probability and Bayes' Rule

Probabilistic Learning

kNN Classification Revisited

Bayesian Learning

Naïve Bayes

Summary

References

- Let $k_i = |\{x | x \in kNN(x_q) \wedge f(x) = c_i\}|$.
- Let $n_i = |\{x | x \in O \wedge f(x) = c_i\}|$, i.e., $\Pr(c_i) = \frac{n_i}{|O|}$
- Let $V_k(x)$ be the volume of the $kNN(x)$.
- $\Pr(x|c_i) = \frac{\frac{k_i}{n_i}}{V_k(x)} = \frac{k_i}{n_i \cdot V_k(x)}$
- $\Pr(x|c_i) \cdot \Pr(c_i) = \frac{k_i}{n_i \cdot V_k(x)} \cdot \frac{n_i}{|O|} = \frac{k_i}{|O| \cdot V_k(x)}$

$$\Pr(c_i|x) = \frac{\Pr(x|c_i) \cdot \Pr(c_i)}{\sum_{j=1}^{m} \Pr(x|c_j) \cdot \Pr(c_j)}$$

$$= \frac{\frac{k_i}{|O| \cdot V_k(x)}}{\sum_{j=1}^{m} \frac{k_j}{|O| \cdot V_k(x)}} = \frac{k_i}{k}$$

$$h(x) = \arg\max_{c_i \in C} \left(\Pr(c_i|x)\right) = \arg\max_{c_i \in C} \left(\frac{k_i}{k}\right) = \arg\max_{c_i \in C} (k_i)$$

Decision Rule — Weight by Prior Class Probability

- To account for very imbalanced proportions of class sizes, we can adjust the prior probability.
- Because we *want* each class *a priori* to be equally likely, we set $\Pr(c_i) = \frac{1}{m}$.
- $\Pr(x|c_i) \cdot \Pr(c_i) = \frac{k_i}{n_i \cdot V_k(x)} \cdot \frac{1}{m} = \frac{k_i}{n_i \cdot m \cdot V_k(x)}$

$$\Pr(c_i|x) = \frac{\Pr(x|c_i) \cdot \Pr(c_i)}{\sum_{j=1}^{m} \Pr(x|c_j) \cdot \Pr(c_j)}$$

$$= \frac{\frac{k_i}{n_i \cdot m \cdot V_k(x)}}{\sum_{j=1}^{m} \frac{k_j}{n_j \cdot m \cdot V_k(x)}} = \frac{\frac{k_i}{n_i}}{\sum_{j=1}^{m} \frac{k_j}{n_j}}$$

$$h(x) = \arg\max_{c_i \in C} \left(\Pr(c_i|x)\right) = \arg\max_{c_i \in C} \left(\frac{k_i}{n_i}\right)$$

- How does the decision change with the choice of k?

Decision Rule — Weight by Prior Class Probability (Example)

- Given a training set O, $|O| = 100$, classes $C = \{A, B\}$, $n_A = 80$, $n_B = 20$
- We choose to set the prior probability $\Pr(A) = \Pr(B) = \frac{1}{2}$
- $k = 10$, classes of the $kNN(x)$ are: $\{A, A, A, A, A, A, B, B, B, B\}$, i.e., $k_A = 6$, $k_B = 4$

$$\Pr(A|x) = \frac{\Pr(x|A) \cdot \Pr(A)}{\Pr(x|A) \cdot \Pr(A) + \Pr(x|B) \cdot \Pr(B)}$$

$$= \frac{\frac{6}{80}}{\frac{6}{80} + \frac{4}{20}} = \frac{\frac{3}{40}}{\frac{3}{40} + \frac{8}{40}} = \frac{\frac{3}{40}}{\frac{11}{40}} = \frac{3}{11}$$

$$\Pr(B|x) = \frac{\frac{8}{40}}{\frac{11}{40}} = \frac{8}{11}$$

$$h(x) = \arg\max_{c_i \in C} \left(\Pr(c_i|x)\right) = \arg\max_{c_i \in C} \left(\frac{k_i}{n_i}\right)$$

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning

- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
  - k Nearest Neighbor Classification Revisited
  - Bayesian Learning
  - Naïve Bayes Classifier
- Summary

Probabilities and Learning

- The aim of machine learning (or actually of science as such) could be put as "*find the best hypothesis to explain the observations*".
- If we approach learning probabilistically, "*best*" means "*most probable*", given the data $\mathcal{D}$ plus any initial knowledge about the prior probabilities of the various hypotheses in $\mathcal{H}$.
- The prior probability $\Pr(h)$ denotes the initial probability of hypothesis h before we observe the training data.
- The prior probability could reflect any background knowledge.
- The prior probability $\Pr(\mathcal{D})$ denotes the probability of the data (observations) without any knowledge on which hypothesis holds.

Prior and Posterior Probabilities, and Bayes' Theorem

- The conditional probability $\Pr(\mathcal{D}\,|h)$ denotes the probability of the observations (likelihood of the hypothesis), given some hypothesis h (i.e., assuming, h is correct).
- The probability $\Pr(h|\,\mathcal{D})$ is called the *posterior probability*, because it reflects our confidence that hypothesis h is correct *after* we have seen the training data $\mathcal{D}$.
- Given prior probabilities $\Pr(h)$, $\Pr(\mathcal{D})$, and conditional probability $\Pr(\mathcal{D}\,|h)$, the posterior probability $\Pr(h|\,\mathcal{D})$ can be computed by Bayes' theorem (Theorem 1.8):

$$\Pr(h|\,\mathcal{D}) = \frac{\Pr(\mathcal{D}\,|h) \cdot \Pr(h)}{\Pr(\mathcal{D})}$$

- Intuitive interpretation: $\Pr(h|\,\mathcal{D})$ increases with $\Pr(\mathcal{D}\,|h)$ and with $\Pr(h)$, it decreases as $\Pr(\mathcal{D})$ increases.

Maximum Likelihood and Maximum A Posteriori Hypothesis

- A classifier shall identify the most probable hypothesis $h \in \mathcal{H}$, given the observed data.
- We call such a maximally probable hypothesis a *maximum a posteriori* (MAP) hypothesis:

$$
\begin{aligned}
h_{\text{MAP}} &\equiv \arg\max_{h \in \mathcal{H}} \Pr(h | \mathcal{D}) \\
&= \arg\max_{h \in \mathcal{H}} \frac{\Pr(\mathcal{D} | h) \cdot \Pr(h)}{\Pr(\mathcal{D})} \\
&= \arg\max_{h \in \mathcal{H}} \Pr(\mathcal{D} | h) \cdot \Pr(h)
\end{aligned}
$$

- If we assume equal prior probabilities for all hypotheses (i.e., $\Pr(h_i) = \Pr(h_j) \forall h_i, h_j \in \mathcal{H}$), MAP is given by the maximum likelihood hypothesis:

$$
h_{\text{ML}} \equiv \arg\max_{h \in \mathcal{H}} \Pr(\mathcal{D} | h)
$$

Most Probable Hypothesis vs. Most Probable Classification

- Consider some hypothesis space $\mathcal{H} = \{h_1, h_2, h_3\}$ with $\Pr(h_1|\mathcal{D}) = 0.4$, $\Pr(h_2|\mathcal{D}) = 0.3$, $\Pr(h_3|\mathcal{D}) = 0.3$.
- Obviously, h_1 is the MAP hypothesis.
- Suppose a new instance x is encountered, where

$$h_1(x) = A$$
$$h_2(x) = B$$
$$h_3(x) = B$$

- Taking all hypotheses into account, we have:

$$\Pr(A|x) = 0.4$$
$$\Pr(B|x) = 0.6$$

- The most probable classification is different from the classification generated by the MAP hypothesis.

Bayes Optimal Classification

DM566
Melih Kandemir
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
Naïve Bayes
Summary
References

- We obtain the most probable classification by combining the predictions of all hypotheses weighted by the posterior probabilities.
- For the set of classes C, for any $c_j \in C$, we have

$$\Pr(c_j|\,\mathcal{D}) = \sum_{h_i \in \mathcal{H}} \Pr(c_j|h_i)\Pr(h_i|\,\mathcal{D})$$

- The optimal classification is therefore:

$$\arg\max_{c_j \in C} \sum_{h_i \in \mathcal{H}} \Pr(c_j|h_i)\Pr(h_i|\,\mathcal{D})$$

- Any system classifying new instances according to this rule is called a "Bayes optimal classifier".

Example

With $C = \{A, B\}$ and the above example, the possible classifications of the new instance x are:

$$\begin{array}{lll} \Pr(h_1|\mathcal{D}) = 0.4 & \Pr(A|h_1) = 1 & \Pr(B|h_1) = 0 \\ \Pr(h_2|\mathcal{D}) = 0.3 & \Pr(A|h_2) = 0 & \Pr(B|h_2) = 1 \\ \Pr(h_3|\mathcal{D}) = 0.3 & \Pr(A|h_3) = 0 & \Pr(B|h_3) = 1 \end{array}$$

Therefore:

$$\sum_{h_i \in \mathcal{H}} \Pr(A|h_i) \Pr(h_i|\mathcal{D}) = 0.4$$

$$\sum_{h_i \in \mathcal{H}} \Pr(B|h_i) \Pr(h_i|\mathcal{D}) = 0.6$$

and

$$\arg\max_{c_j \in \{A,B\}} \sum_{h_i \in \mathcal{H}} \Pr(c_j|h_i) \Pr(h_i|\mathcal{D}) = B$$

Properties of the Bayes Optimal Classifier

- The predictions of the Bayes optimal classifier can correspond to the predictions of a hypothesis that is not contained in the original hypothesis space $\mathcal{H}$!
- The Bayes optimal classifier considers effectively a different hypothesis space $\mathcal{H}'$, including hypotheses that perform comparisons between linear combinations of predictions from multiple hypotheses in $\mathcal{H}$.

Note that:

The Bayes optimal learner maximizes the probability that the new instance is classified correctly, given the available data, hypothesis space, and prior probabilities over the hypotheses. Thus no other classification method using the same hypothesis space and same prior knowledge can outperform this method on average.

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
Axioms of Probability
Independence and Conditional Probability
Total Probability and Bayes' Rule
Probabilistic Learning
k Nearest Neighbor Classification Revisited
Bayesian Learning
Naïve Bayes Classifier
Summary

Estimates of Prior Probabilities

Consider a learning task to distinguish apples, oranges, and other fruits, where the objects are described by color and shape:

- 20% of the objects are apples
- 30% of the objects are oranges

} prior class probability

- 50% of the objects are round
- 40% of the objects have an orange color

} prior probability of some attribute value

Estimates of Posterior Probabilities

Posterior (conditional) probabilities model relations between attribute values and classes:

- 100% of the oranges are round: $\Pr(\text{shape=round}|\text{ORANGE})$
- 100% of the apples are round: $\Pr(\text{shape=round}|\text{APPLE})$
- 90% of the oranges have the color orange: $\Pr(\text{color=orange}|\text{ORANGE})$

Bayes Classification

For a given attribute value a_q, we can compute the posterior class probability according to Bayes' rule:

$$\Pr(c_j|a_q) = \frac{\Pr(a_q|c_j)\Pr(c_j)}{\Pr(a_q)} = \frac{\Pr(a_q|c_j)\Pr(c_j)}{\sum_{c_i \in C}\Pr(c_i)\Pr(a_q|c_i)}$$

We estimate probabilities from the training data.
For example, we have an object with color orange:

$$\begin{aligned}
&\Pr(\text{ORANGE}|\text{color=orange})\\
&\quad = \frac{\Pr(\text{color=orange}|\text{ORANGE})\Pr(\text{ORANGE})}{\Pr(\text{color=orange})}\\
&\quad = \frac{0.9 \cdot 0.3}{0.4}\\
&\quad = 0.675
\end{aligned}$$

Maximum Likelihood Classification

Given all posterior class probabilities, we predict the most likely class:

$$
\begin{aligned}
h_{\text{MAP}} &= \arg\max_{c_i \in C} \Pr(c_i | a_q) \\
&= \arg\max_{c_i \in C} \frac{\Pr(a_q | c_i) \Pr(c_i)}{\Pr(a_q)} \\
&= \arg\max_{c_i \in C} \Pr(a_q | c_i) \Pr(c_i)
\end{aligned}
$$

Discrete Attributes

DM566
Melih Kandemir
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
Naïve Bayes
Summary
References

we can count relative frequencies to estimate probabilities:

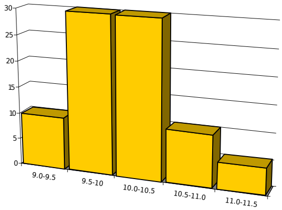
ID	shape	color	class
1	round	orange	A
2	round	green	A
3	round	yellow	A
4	square	green	A
5	oval	white	B

$$\Pr(\text{shape=round}|A) = \frac{3}{4}$$

$$\Pr(\text{color=green}|A) = \frac{2}{4}$$

$$\Pr(\text{shape=oval}|A) = \frac{0}{4}$$

Continuous Metric Attributes



$$\Pr(9.0 < \text{diameter} \leq 9.5|A) = 10\%$$
$$\Pr(9.5 < \text{diameter} \leq 10.0|A) = 30\%$$
$$\Pr(10.0 < \text{diameter} \leq 10.5|A) = 30\%$$
$$\Pr(10.5 < \text{diameter} \leq 11.0|A) = 10\%$$
$$\Pr(11.0 < \text{diameter} \leq 11.5|A) = 5\%$$

Zero Probabilities?

Note that:

- *Problem:* $\Pr($*shape=oval*$|A) = 0$ *would rule out any slight possibility of predicting an instance of class* A*.*
- *Heuristic solution: smoothing (use some artificial small minimum probability):*

$$\Pr(\text{shape} = \text{oval}|A) := \max\left\{\frac{0}{4}, \varepsilon\right\} \quad \textit{with } 0 < \varepsilon \ll 1$$

Multi-dimensional Data

DM566
Melih Kandemir
Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
Naïve Bayes
Summary
References

- So far, we considered only one attribute.
- In multi-dimensional data, we need to estimate the combined probabilities of specific attribute values:

$$\begin{aligned} h_{\text{MAP}} &= \arg\max_{c_i \in C} \Pr(c_i | a_1 \cap a_2 \cap a_3 \cap \ldots \cap a_n) \\ &= \arg\max_{c_i \in C} \frac{\Pr(a_1 \cap a_2 \cap a_3 \cap \ldots \cap a_n | c_i) \Pr(c_i)}{\Pr(a_1 \cap a_2 \cap a_3 \cap \ldots \cap a_n)} \\ &= \arg\max_{c_i \in C} \Pr(a_1 \cap a_2 \cap a_3 \cap \ldots \cap a_n | c_i) \Pr(c_i) \end{aligned}$$

Example:

ID	*shape*	*color*	*class*
1	*round*	*orange*	*A*
2	*round*	*green*	*A*
3	*round*	*yellow*	*A*
4	*square*	*green*	*A*
5	*oval*	*white*	*B*

$$\Pr(\textit{shape=round} \cap \textit{color=orange} | A) = \frac{1}{4}$$

$$\Pr(\textit{shape=round} \cap \textit{color=green} | A) = \frac{1}{4}$$

Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
Naïve Bayes
Summary
References

Problems for the Bayes Classifier in Multi-dimensional Data

Problems:

- If we have n attributes, and each can take on r different values, we have r^n different attribute combinations.
- Typically, there are not enough training instances available to reliably estimate probabilities.

Example

ID	shape	color	class
1	round	orange	A
2	round	green	A
3	round	yellow	A
4	square	green	A
5	oval	white	B

$$\Pr(\text{shape=round} \cap \text{color=orange}|A) = \frac{1}{4}$$

$$\Pr(\text{shape=round} \cap \text{color=green}|A) = \frac{1}{4}$$

$$\Pr(\text{shape=round} \cap \text{color=yellow}|A) = \frac{1}{4}$$

$$\Pr(\text{shape=round} \cap \text{color=white}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=oval} \cap \text{color=orange}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=oval} \cap \text{color=green}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=oval} \cap \text{color=yellow}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=oval} \cap \text{color=white}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=square} \cap \text{color=orange}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=square} \cap \text{color=green}|A) = \frac{1}{4}$$

$$\Pr(\text{shape=square} \cap \text{color=yellow}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=square} \cap \text{color=white}|A) = \frac{0}{4}$$

$$\Pr(\text{shape=round} \cap \text{color=orange}|B) = \frac{0}{1}$$

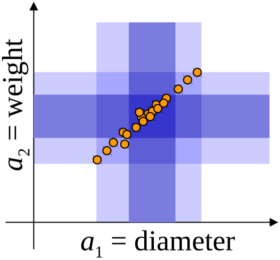
$$\vdots$$

The probability estimates are unreliable, because the sample size is too small for each instance.

The Naïve Assumption: Independence

If we assume independence among the attributes, we can estimate the combined probabilities based on Definition 1.5 ($\Pr(E \cap F) = \Pr(E) \cdot \Pr(F)$):

$$
\begin{aligned}
h_{\text{MAP}} &= \arg\max_{c_i \in C} \Pr(a_1 \cap a_2 \cap a_3 \cap \ldots \cap a_n | c_i) \Pr(c_i) \\
&= \arg\max_{c_i \in C} \prod_{j=1}^{n} \Pr(a_j | c_i) \Pr(c_i) \qquad \text{(Ass. of Indep.)}
\end{aligned}
$$



- The assumption might be wrong.
- Then we don't get the correct probabilities.
- But we *might* still get the correct maximum.
- In practice, the assumption often works despite *some* dependency among the attributes. advanced reading: Domingos

Naïve Bayes Classifier

- The "Naïve Bayes classifier" relies on the assumption of independence of attributes.
- The various $\Pr(c_i)$ and $\Pr(a_j|c_i)$ terms are estimated based on the relative frequencies of corresponding examples in the training data.
- The set of these estimates constitutes the learned hypothesis.
- The class prediction is based on these estimates according to:

$$h_{\text{naïve Bayes}} = \arg\max_{c_i \in C} \prod_{j=1}^{n} \Pr(a_j|c_i) \Pr(c_i)$$

- Because of the multiplication, the replacement of zero probabilities by some heuristic minimum probability ε (cf. slide 66) is particularly important.

Example: Should We Play Tennis Today?

DM566

Melih Kandemir

Bayesian Learning
Axioms of Probability
Independence and Conditional Prob.
Total Probability and Bayes' Rule
Probabilistic Learning
kNN Classification Revisited
Bayesian Learning
Naïve Bayes
Summary
References

ID	forecast	temperature	humidity	wind	play tennis?
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Example Prediction

Classify a new instance: $\langle \text{sunny}, \text{cool}, \text{high}, \text{strong} \rangle$

$$\begin{aligned} h_{\text{naïve Bayes}} &= \arg\max_{c_i \in \{\text{yes},\text{no}\}} \prod_{j=1}^{n} \Pr(a_j|c_i) \Pr(c_i) \\ &= \arg\max_{c_i \in \{\text{yes},\text{no}\}} \Pr(\text{sunny}|c_i) \Pr(\text{cool}|c_i) \Pr(\text{high}|c_i) \\ &\qquad \Pr(\text{strong}|c_i) \Pr(c_i) \end{aligned}$$

$$\Pr(\text{yes}) = \frac{9}{14} = 0.64 \qquad \Pr(\text{wind=strong}|\text{yes}) = \frac{3}{9} = 0.33$$

$$\Pr(\text{no}) = \frac{5}{14} = 0.36 \qquad \Pr(\text{wind=strong}|\text{no}) = \frac{3}{5} = 0.60$$

$$\vdots$$

Decision and Probability

The Naïve Bayes classifier decides by finding the class maximizing the product of probabilities:

$$\Pr(\text{sunny}|\text{yes}) \Pr(\text{cool}|\text{yes}) \Pr(\text{high}|\text{yes}) \Pr(\text{strong}|\text{yes}) \Pr(\text{yes}) = 0.0053$$
$$\Pr(\text{sunny}|\text{no}) \Pr(\text{cool}|\text{no}) \Pr(\text{high}|\text{no}) \Pr(\text{strong}|\text{no}) \Pr(\text{no}) = 0.0206$$

$$h_{\text{naïve Bayes}}\left(\langle \text{sunny}, \text{cool}, \text{high}, \text{strong} \rangle\right) = \text{no}$$

If we are interested in the conditional probability for "no", we could normalize these quantities to sum up to one:

$$\frac{0.0206}{0.0206 + 0.0053} = 0.795$$

Assumption of Independence is a Bias

- The assumption of independence can be seen as the bias inherent to the Naïve Bayes classifier.
- An unbiased probabilistic classifier is not practical due to a notorious lack of training examples.
  - In other words: in any practical scenario, it would hopelessly overfit.
- For data consisting of two classes and only 30 binary attributes, we would need more than 2 billion examples just to see each combination *once* (which is not good enough to derive reliable probability estimates).
- Relying on the bias, the classifier may have a tendency to be wrong (if the assumption does not hold).
- The bias is necessary to make generalization feasible.

Outline

Basic Probability Theory, Bayes' Rule, and Bayesian Learning
- Axioms of Probability
- Independence and Conditional Probability
- Total Probability and Bayes' Rule
- Probabilistic Learning
- Summary

Summary

You learned in this section:

- *Axioms of probability:*
  - *sample space*
  - *event*
  - *probability function*
  - *probability space*
- *independence and conditional probability*
- *probabilistic interpretation of quality measures for association rules*
- *Bayes' rule*

Summary

You learned in this section:

- *Bayesian Learning:*
  - *k nearest neighbor classifier as an application of Bayes' rule for learning*
  - *The principle of Bayesian learning:*
    - *prior and posterior probabilities*
    - *data as evidence to adapt probability estimates and to select hypotheses*
    - *MAP hypothesis*
    - *Bayesian reasoning*
  - *Bayes optimal classifier*
  - *Naïve Bayes classifier*

References I

P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *Proceedings of the 13th International Conference on Machine Learning (ICML), Bari, Italy*, pages 105–112, 1996.

T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

M. Mitzenmacher and E. Upfal. *Probability and Computing. Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2nd edition, 2017.

M. J. Zaki and W. Meira Jr. *Data Mining and Analysis. Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.