PROBABILITY-BASED LEARNING

CS576 MACHINE LEARNING

Dr. Jin S. Yoo, Professor
Department of Computer Science
Purdue University Fort Wayne

Reference

- Kelleher et al., Fundamentals of ML,
  - Ch 6. Probability-based Learning
  - Appendix B. Introduction to Probability for Machine Learning
- Mitchell, Machine Learning, Ch 6

Probability-based Learning

- **Probability-based Learning** is a foundational approach in machine learning and statistics, providing a principled and coherent framework to model uncertainty, learn from data, make inferences, and make decisions under uncertainty.
- The idea behind Probability-based Learning is to use the principles of **probability theory** to make predictions and inferences about data.
- One practical example of Probability-based Learning is the **Naive Bayes classifier**, which applies **Bayes' Theorem** to model the relationship between the features and the target feature and make predictions.
- It's versatile and applicable in various domains including natural language processing, computer vision, and medical diagnosis.

Outline

- Probability-based Learning
- ☞ **Fundamentals**
  - Basics of Probability Theory
  - Bayes' Theorem
  - Bayesian Prediction
  - Conditional Independence and Factorization
- Standard Approach: The Naïve Bayes' Classifier
- Extensions and Variations
- Summary

Probabilistic Interpretation of Dataset Features

From a probability perspective,

- Each **feature** in a dataset is considered a **random variable.**
- Each row (or **example**) in a dataset represents an **experiment** that associates a value of the target feature with a set of values of descriptive features.
- The assignment of values to a set of descriptive features constitutes an **event.**
- The **sample space** for the domain related to a prediction problem includes the set of all possible combinations of assignments of values to features.

Example Dataset

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

Table A simple dataset for a MENINGITIS diagnosis with descriptive features that describe the presence or absence of three common symptoms of the disease: HEADACHE, FEVER, and VOMITING.

Probability Function

- A **probability function**, denoted as $P()$, yields the probability of a feature taking a specific value, i.e., an event occurring.
- **Example**: $P(\text{FEVER} = true)$ returns the probability of 'FEVER' being '*true*'. This probability can be calculated directly from the dataset: $P(\text{FEVER} = true)=0.4$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

Probability Function (cont.)

- Probability functions assigned to **categorical features** are termed as **probability mass functions.**
- Probability functions assigned to **continuous features** are termed as **probability density functions.**

Joint and Conditional Probability

- A **joint probability** represents the likelihood of multiple features taking on specific values simultaneously.
  - e.g., $P(\text{MENINGITIS} = true, \text{HEADACHE} = true) = 0.2$
- The joint probability **using the product rule** is

$$\boldsymbol{P(X, Y) = P(X|Y)P(Y)}$$

- A **conditional probability** indicates the likelihood of a feature taking a specific value, given the known value of another distinct feature.
  - e.g., $P(\text{MENINGITIS} = true \mid \text{HEADACHE} = true) = 0.2857$.
- The conditional probability **in terms of joint probability** is

$$\boldsymbol{P(X|Y) = \frac{P(X,Y)}{P(Y)}}$$

Product Rule and Sum Rule in Probability

- **Product rule** (also known as **multiplication rule**):
  - The probability of a conjunction of two events X and Y (joint probability of X and Y), $P(X \cap Y)$, is often expressed as:

$$\boldsymbol{P(X \cap Y) = P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)}$$

- **Sum rule**:
  - The probability of a disjunction of two events X and Y, $P(X \cup Y)$, is often expressed as:

$$\boldsymbol{P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)}$$

Chain Rule

- A **Chain Rule** states that the **probability of joint event** can be written as a product of conditional probabilities.

$$\boldsymbol{P(q[1], \ldots, q[m]) = P(q[1]) \times P(q[2] \mid q[1]) \times \ldots \times P(q[m] \mid q[m-1], \ldots, q[2], q[1])}$$

- To implement the **chain rule on a conditional probability**, each term in the expression is modified by incorporating the conditional term:

$$\boldsymbol{P(q[1], \ldots, q[m] \mid t = l) = P(q[1] \mid t = l) \times P(q[2] \mid q[1], t = l) \times \ldots \times P(q[m] \mid q[m-1], \ldots, q[3], q[2], q[1], t = l)}$$

Probability Distribution

- A **probability distribution** is a mathematical description or a data structure representing the likelihood of different possible outcomes or values that a feature can assume.
- **Example**: The probability distribution for the binary feature MENINGITIS is **P**(MENINGITIS) = <0.3, 0.7> (by convention we give the *true* probability first)
- The sum of a probability distribution must equal to **1.**
- **Notations:**
  - A probability distribution **P**()
  - A probability function *P*()

Joint Probability Distribution

- A **joint probability distribution** quantifies the probability across multiple feature assignments.
- It is represented as a multi-dimensional **matrix**, where each cell details the probability of a specific combination of feature values being assigned.
- The total of all the cell values in a joint probability distribution must equal to 1.
- **Example**: This matrix displays the join probability distribution for the four binary features (**H**EADACHE, **F**EVER, **V**OMITING, and **M**ENINGITIS)

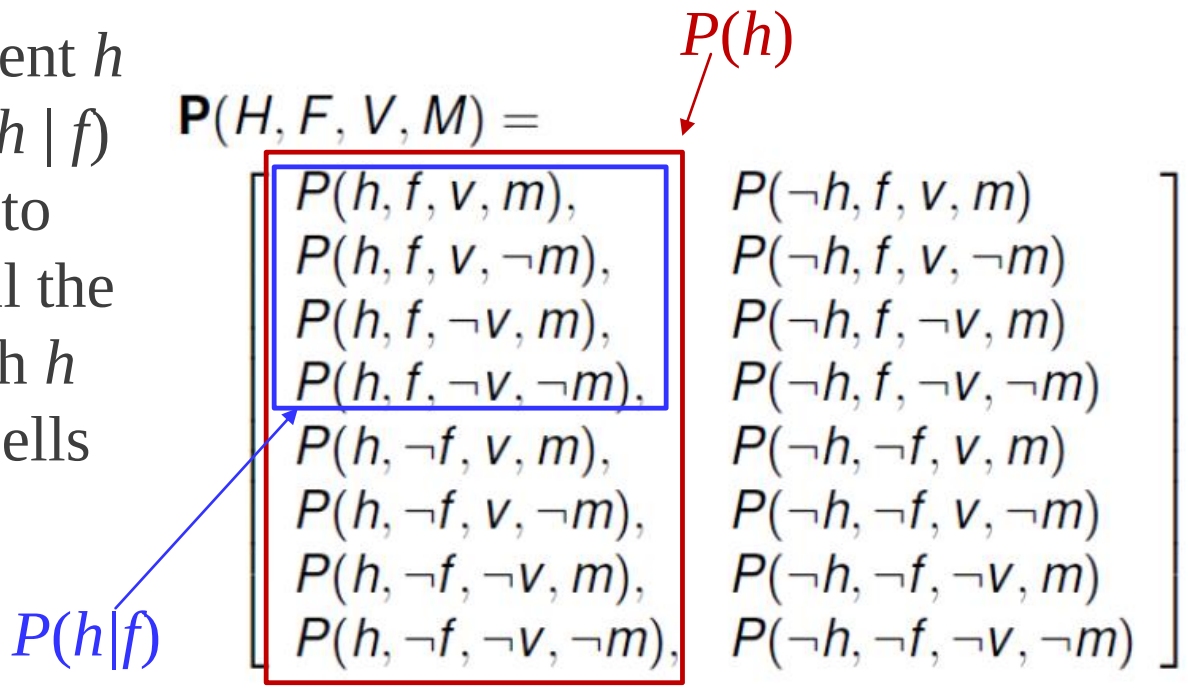
$$\mathbf{P}(H, F, V, M) = \begin{bmatrix} P(h, f, v, m), & P(\neg h, f, v, m) \\ P(h, f, v, \neg m), & P(\neg h, f, v, \neg m) \\ P(h, f, \neg v, m), & P(\neg h, f, \neg v, m) \\ P(h, f, \neg v, \neg m), & P(\neg h, f, \neg v, \neg m) \\ P(h, \neg f, v, m), & P(\neg h, \neg f, v, m) \\ P(h, \neg f, v, \neg m), & P(\neg h, \neg f, v, \neg m) \\ P(h, \neg f, \neg v, m), & P(\neg h, \neg f, \neg v, m) \\ P(h, \neg f, \neg v, \neg m), & P(\neg h, \neg f, \neg v, \neg m) \end{bmatrix}$$

Joint Probability Distribution (cont.)

- Given a join probability distribution, **the probability of any event in its domain** can be determined by summing over the cells where that event holds true.
- This method of calculating probabilities is known as **summing out**.

Examples: Summing Out

- **Example 1**: The probability of an event h, $P(h)$, in the domain specified by the joint probability distribution $\mathbf{P}(H,F,V,M)$ is simply to sum the values in the cells containing h (the cells in the first column)
- **Example 2**: The probability of an event h given an event f, $P(h \mid f)$ from $\mathbf{P}(H,F,V,M)$ is to sum the values in all the cells containing both h and f (the top four cells in the first column)



Theorem (or Law) of Total Probability

- The **Theorem of Total Probability** is a fundamental rule relating marginal probabilities to conditional probabilities.
- It provides a way to get the overall probability of an event X by considering all the different ways that it can happen, based on a partition of the sample space.
- Let $Y_1, \ldots, Y_k$ be a partition of the sample space, meaning that Y_i are **mutually exclusive** events cover with the whole space, and assume that $P(Y_i) > 0$ for each i and $\sum_{i=1}^{k} P(Y_i)=1$

$$\boldsymbol{P(X) = \sum_i P(X|Y_i)P(Y_i)} = \sum_i P(X \wedge Y_i)$$

$$= P(X|Y_1)P(Y_1) + P(X|Y_2)P(Y_2) + \cdots + P(X|Y_k)P(Y_k)$$

Example: Theorem of Total Probability

- The probability of an event h ($h$eadache=true) is

$$P(h) = \big(P(h|m) \times P(m)\big) + \big(P(h|\neg m) \times P(\neg m)\big)$$
$$= (0.6666 \times 0.3) + (0.7143 \times 0.7) = 0.7$$

,where event m represents $m$eningitis=true.

Independence

- If the occurrence of one event does not influence the probability of the other event, then the two events are **independent** of each other.
- **If two events X and Y are independent** then:

$$P(X|Y) = P(X)$$

$$P(X, Y) = P(X) \times P(Y)$$

Conditional Independence

- It's more typical to encounter situations where two or more events may be independent, provided that we know a third event has occurred.
- Two events *X* and *Y* are **conditionally independent** given a third event *Z* if the occurrence of *X* does not affect the probability of *Y* occurring, given that *Z* has occurred; mathematically, this is expressed as

$$\boldsymbol{P(X, Y|Z) = P(X|Z) \times P(Y|Z)}$$

$$\boldsymbol{P(X|Y, Z) = P(X|Z)}$$

Conditional Independence and Factorization

Withouth conditional independence

$$P(X, Y, Z|W) = P(X|W) \times P(Y|X, W) \times P(Z|Y, X, W) \times P(W)$$

With conditional independence

$$P(X, Y, Z|W) = \underbrace{P(X|W)}_{Factor1} \times \underbrace{P(Y|W)}_{Factor2} \times \underbrace{P(Z|W)}_{Factor3} \times \underbrace{P(W)}_{Factor4}$$

Outline

- Probability-based Learning
- Fundamentals
  - Basics of Probability Theory
  - ☞ **Bayes' Theorem**
  - Bayesian Prediction
  - Conditional Independence and Factorization
- Standard Approach: The Naïve Bayes' Classifier
- Extensions and Variations
- Summary

Bayes' Theorem

- **Bayes' Theorem** can be articulated as follows:

  The probability that an event has happened, given a set of evidence for it, is equal to the probability of the evidence being cause by the event, multiplied by the probability of the event itself .

  $$P(event \mid evidence) = P(evidence \mid event) \times P(event)$$

- Bayes' Theorem allows us to **determine the probability of an event using the available evidence**, **expressing it in terms of the likelihood that the event would generate the observed evidence.**

Formal Definition of Bayes' Theorem

Bayes' Theorem

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Where:

- $P(X \mid Y)$: Probability of X given Y has occurred (**Posterior**)
- $P(Y \mid X)$: Probability of Y given X has occurred (**Likelihood**)
- $P(X)$: Probability of X occurring (**Prior**)
- $P(Y)$: Probability of X occurring (**Evidence**)

- **Bayes' Theorem** defines the conditional probability of an event, X, given some evidence, Y, as the product of the inverse conditional probability, $P(Y \mid X)$ and the prior probability of the event $P(X)$.

Characteristics of Bayes' Theorem

- The division by $P(Y)$ in Bayes' Theorem, $P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$, acts as a **normalization** mechanism, ensuring that
  - $0 \leq P(X|Y) \leq 1$
  - $\sum_i P(X_i|Y) = 1.0$
- **$P(Y)$ is calculated**
  - directly from the dataset, $P(Y) = \frac{|rows\ where\ Y\ is\ the\ case|}{|\{rows\ in\ the\ data\ set|}$
  - OR using the **Theorem of Total Probability**, $P(Y) = \sum_i P(Y|X_i)P(X_i)$
- Bayes' Theorem is easily derived from the **product rule.**

$$P(X \wedge Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

$$\frac{P(X|Y)\cancel{P(Y)}}{\cancel{P(Y)}} = \frac{P(Y|X)P(X)}{P(Y)}$$

$$\Rightarrow P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Example

Example

After a yearly checkup, a doctor informs their patient that he has both bad news and good news. The bad news is that the patient has tested positive for a serious disease and that the test that the doctor has used is 99% accurate (i.e., the probability of testing positive when a patient has the disease is 0.99, as is the probability of testing negative when a patient does not have the disease). The good news, however, is that the disease is extremely rare, striking only 1 in 10,000 people.

Q. What is the actual probability that the patient has the disease? Why is the rarity of the disease good news given that the patient has tested positive for it?

Example (cont.)

$$P(d|t) = \frac{P(t|d)P(d)}{P(t)}$$

- You start with a hypothesis (**Prior**: $P(d)$)
  - The probability of having the disease as $P(d)$=0.0001
  - The probability of not having the disease as $P(\neg d) = 0.9999$
- You collect some evidence (**Likelihood**: $P(t|d)$ and **Evidence** $P(t)$)
  - The accuracy of the test as $P(t|d)$ = 0.99 and $P(t|\neg d)$=0.01
  - By Theorem of Total Probability, $P(t) = P(t|d)P(d) + P(t|\neg d)P(\neg d)$=0.0101
- You update the probability of your hypothesis based on the evidence (**Posterior**: $P(d|t)$)
  - **The probability that the patient actually has the disease** (d = yes) **based on the evidence of the test result** (t = positive):

$$P(d|t) = \frac{P(t|d)P(d)}{P(t)} = \frac{0.99 \times 0.0001}{0.0101} = 0.0098$$

➔ The probability of actually having the disease is less than 1%

Outline

- Probability-based Learning
- Fundamentals
  - Basics of Probability Theory
  - Bayes' Theorem
  - ☞ **Bayesian Prediction**
  - Conditional Independence and Factorization
- Standard Approach: The Naïve Bayes' Classifier
- Extensions and Variations
- Summary

Bayesian Approach

- **Core Foundation**: Probability-based prediction models derive much of their foundation from **Bayes' Theorem**.
- **Bayesian Approach**: This methodology provides a probabilistic perspective to inference, grounded in two **key principles**:
  - Quantities of interest are determined by underlying probability distributions
  - Optimal decision-making is derived by integrating these probabilities with observed data

Bayes' Theorem for Learning

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

- Bayes' theorem provides a way to calculate the *posterior probability* of a hypothesis $P(h|D)$ based on
  - its *prior probability* $P(h)$,
  - the probability of observing various data given the hypothesis, $P(D|h)$, and
  - the probability of observed data itself $P(D)$.

Bayes' Theorem for Learning (Cont.)

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

- $P(h)$: **Prior probability** of hypothesis h
  - It may reflect any background knowledge we have about the chance that h is a correct hypothesis
- $P(D)$: Probability that training data D will be observed
  - i.e., $P(D)$ is the probability of D given no knowledge about which hypothesis holds.
- $P(D|h)$: Probability of observing training data D given some world in which hypothesis h holds. (often called the **likelihood** of the data D given h)
- $P(h|D)$: **Posterior probability** of hypothesis h
  - We are interested in the probability that h holds given the observed training data D
  - It reflects our confidence that h holds after we have seen the training data D

Generalized Bayes' Theorem

- **To perform Bayesian predictions,** we calculate the probability of the event that a target feature ***t*** takes a specific label ***l***, given the assignment of values to a set of descriptive features derived from a query instance ***q***.

Generalized Bayes' Theorem

$$P(t = l|\mathbf{q}[1], \ldots, \mathbf{q}[m]) = \frac{P(\mathbf{q}[1], \ldots, \mathbf{q}[m]|t = l)P(t = l)}{P(\mathbf{q}[1], \ldots, \mathbf{q}[m])}$$

- $P(t = l)$, the **priori probability** of the target feature t taking the label l.
- $P(q[1], \ldots, q[m])$, the **joint probability** of the descriptive features of a query instance taking a specific set of values
- $P(q[1], \ldots, q[m]|t = l)$, the **conditional probability** of the descriptive features of a query instance taking a specific set of values given that the target feature takes the label l

How to Compute Generalized Bayes' Theorem

$$\boldsymbol{P(t = l|q[1], \ldots, q[m]) = \frac{P(q[1], \ldots, q[m]|t = l)P(t = l)}{P(q[1], \ldots, q[m])}}$$

- $P(t = l)$ is simply the relative frequency of occurrence where the target feature takes the label l in a dataset
- $P(q[1], \ldots, q[m])$ is determined by the relative frequency within a dataset where the descriptive features of an instance take the values $q[1], \ldots, q[m]$.

  Alternatively, using the Theorem of Total Probability, $P(q[1], \ldots, q[m]) = \sum_{k \in labels(t)} P(q[1], \ldots, q[m]|t = k)P(t = k)$ **or** replaced entirely with a normalization constant η
- $P(q[1], \ldots, q[m]|t = l)$ is calculated either directly from a dataset **or** using Chain Rule, $P(q[1], \ldots, q[m]|t = l) = P(q[1] \mid t = l) \times P(q[2] \mid q[1], t = l) \times \ldots \times P(q[m] \mid q[m-1], \ldots, q[3], q[2]|q[1], t = l)$

Most Probable Model

- Suppose that the probabilities of three possible hypotheses (/models), h_1, h_2 and h_3, given a training dataset, are:

  $$P(h_1|D) = .4, \quad P(h_2|D) = .3, \quad P(h_3|D) = .3$$

  , and given a new instance *x*, the three models return

  $$h_1(x) = +, h_2(x) = -, h_3(x) = -$$

- What is the most probable classification of *x*?
- We aim to determine the hypothesis, $h \in H$, from hypothesis space H, that is most probable given the observed data *D*.

Maximum A Posterior (MAP) Model

- We are interested in finding the most probable hypothesis $h \in H$ given the observed data D (or at least one of the maximally probable if there are several).
- **Maximum A Posterior (MAP) hypothesis** is the hypothesis h that is most probable given the observed data D.

$$\boldsymbol{h_{MAP}} = \underset{h \in H}{argmax} P(h|D) = \underset{h \in H}{argmax} \frac{P(D|h)P(h)}{P(D)}$$ by Bayes' Theorem

  - It considers both the likelihood of the observed data under each hypothesis $P(D|h)$ and the prior probability of the hypothesis $P(h)$
- **Maximum Likelihood (ML) hypothesis** is the hypothesis h under which the observed data D is most probable.

$$h_{ML} = \underset{h \in H}{argmax} P(D|h)$$

Perdition using Bayesian MAP Model

- Using the MAP model, we predict the label ***l*** for the target feature ***t*** that maximizes the probability, given the values assigned to a set of descriptive features from a query instance ***q***.

Bayesian MAP Prediction Model

$$\mathbb{M}_{MAP}(\mathbf{q}) = \underset{l \in levels(t)}{\text{argmax}}\ P(t = l \mid \mathbf{q}[1], \ldots, \mathbf{q}[m])$$

$$= \underset{l \in levels(t)}{\text{argmax}}\ \frac{P(\mathbf{q}[1], \ldots, \mathbf{q}[m] \mid t = l) \times P(t = l)}{P(\mathbf{q}[1], \ldots, \mathbf{q}[m])}$$

* The denominator, $P(q[1], \ldots, q[m])$, is a constant independent of the model, i.e., not dependent on the target feature

Bayesian MAP Prediction Model (without normalization)

$$\mathbb{M}_{MAP}(\mathbf{q}) = \underset{l \in levels(t)}{\text{argmax}}\ P(\mathbf{q}[1], \ldots, \mathbf{q}[m] \mid t = l) \times P(t = l)$$

Example of Bayesian Prediction

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

Table. A sample dataset for MENINGITIS diagnosis with descriptive feature that describe the presence of absence of three common symptom of the disease: HAEADCHE, FEVER and VOMITING

- A query instance:

HEADACHE	FEVER	VOMITING	MENINGITIS
true	false	true	?

- $\boldsymbol{P(M|h, \neg f, v) = ?}$
  - There are two values in the domain of the MENINGITIS feature, *true* and *false*. So, we need to consider two cases: $\boldsymbol{P(m|h, \neg f, v)}$ and $\boldsymbol{P(\neg m|h, \neg f, v)}$

Example (Cont.)

- $\boldsymbol{P(m|h,\neg f,v)}$
  - $P(m|h,\neg f,v) = \frac{P(h,\neg f,v|m)\times P(m)}{P(h,\neg f,v)} = \mathbf{0.3333}$
    - From the dataset,
      - $\boldsymbol{P(m)} = \frac{|\{\boldsymbol{d_5},\boldsymbol{d_8},\boldsymbol{d_{10}}\}|}{|\{\boldsymbol{d_1},\boldsymbol{d_2},\ldots,\boldsymbol{d_{10}}\}|} = 0.3$
      - $\boldsymbol{P(h,\neg f,v)} = \frac{|\{\boldsymbol{d_3},\boldsymbol{d_4},\boldsymbol{d_6},\boldsymbol{d_7},\boldsymbol{d_8},\boldsymbol{d_{10}}\}|}{|\{\boldsymbol{d_1},\boldsymbol{d_2},\ldots,\boldsymbol{d_{10}}\}|} = 0.6$

        or Using Chain Rule,

        $P(h,\neg f,v|m) = P(h|m)\times P(\neg f|h,m)\times P(v|\neg f,h,m)$

        $= \frac{|\{\boldsymbol{d_8},\boldsymbol{d_{10}}\}|}{|\{\boldsymbol{d_5},\boldsymbol{d_8},\boldsymbol{d_{10}}\}|} \times \frac{|\{\boldsymbol{d_8},\boldsymbol{d_{10}}\}|}{|\{\boldsymbol{d_8},\boldsymbol{d_{10}}\}|} \times \frac{|\{\boldsymbol{d_8},\boldsymbol{d_{10}}\}|}{|\{\boldsymbol{d_8},\boldsymbol{d_{10}}\}|} = 0.666$
- $\boldsymbol{P(\neg m|h,\neg f,v)}$
  - $P(\neg m|h,\neg f,v) = \frac{P(h,\neg f,v|\neg m)\times P(\neg m)}{P(h,\neg f,v)} = \mathbf{0.6667}$
    - Similarly, compute all the probabilities necessary,
- **It is twice as likely that the patient does not have meningitis compared to the likelihood that they do,** even though the patient is exhibiting symptoms such as a headache and vomiting!

Brute-Force MAP Algorithm

1. For each hypothesis h in hypothesis space H, calculate the posterior probability, $P(h|D) = \frac{P(D|h)P(h)}{P(D)}$
2. Output the hypothesis h_{MAP} with the highest posterior probability, $h_{MAP} = \underset{h \in H}{argmax} P(h|D)$

- This approach is computationally expensive due to large hypothesis space with the increase of features.

Example 2 of Bayesian Prediction

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- A query instance with HEADACHE = *true*, FEVER=*true*, and VOMITING =*false*. Does the query patient have MENINGITIS ?
- $\boldsymbol{P(M|h, f, \neg v)} = ?$
  - $P(m|h, f, \neg v)$
  - $P(\neg m|h, f, \neg v)$

Example 2 (cont.)

$$P(m \mid h, f, \neg v) = \frac{\begin{pmatrix} P(h|m) \times P(f \mid h, m) \\ \times P(\neg v \mid f, h, m) \times P(m) \end{pmatrix}}{P(h, f, \neg v)}$$

$$= \frac{0.6666 \times 0 \times 0 \times 0.3}{0.1} = 0$$

$$P(\neg m \mid h, f, \neg v) = \frac{\begin{pmatrix} P(h|\neg m) \times P(f \mid h, \neg m) \\ \times P(\neg v \mid f, h, \neg m) \times P(\neg m) \end{pmatrix}}{P(h, f, \neg v)}$$

$$= \frac{0.7143 \times 0.2 \times 1.0 \times 0.7}{0.1} = 1.0$$

Bayesian Prediction with 0

- $P(m|h,f,\neg v) = 0, \quad P(\neg m|h,f,\neg v) = 1.0$
  - $\boldsymbol{P(m|h,f,\neg v) = \frac{P(h,f,\neg v|m)\times P(m)}{P(h,f,\neg v)} = \frac{(P(h|m)\times P(f|h,m)\times P(\neg v|f,h,m))\times P(m)}{P(h,f,\neg v)} = \frac{0.6666\times 0\times 0\times 0.3}{0.1} = 0}$
  - **There is something odd about these results??**
  - The probability of a patient who has a headache and a fever having meningitis should be greater than zero!
- **Data fragmentation - A case of the curse of dimensionality**
  - In computing $P(h,f,\neg v|m)$, **the probability diminishes with addition of more conditions, by the chain rule**
  - Our dataset is not large enough. Our Bayesian model is **over-fitting** to the training data.
- The principles of **conditional independence** and **factorization** can address this limitation of the current approach.

[Reminder] Conditional Independence & Factorization

Withouth conditional independence

$$P(X, Y, Z|W) = P(X|W) \times P(Y|X, W) \times P(Z|Y, X, W) \times P(W)$$

With conditional independence

$$P(X, Y, Z|W) = \underbrace{P(X|W)}_{Factor1} \times \underbrace{P(Y|W)}_{Factor2} \times \underbrace{P(Z|W)}_{Factor3} \times \underbrace{P(W)}_{Factor4}$$

Conditional Independence and Chain Rule

- If the event $t = l$ causes the events $q[1], \ldots, q[m]$, and these events are **conditionally independent of one another** given $t = l$, then the **chain rule definition** for $P(q[1], \ldots, q[m] | t = l)$ can be simplified as follows:

$$P(q[1], \ldots, q[m] | t = l)$$
$$= P(q[1] | t = l) \times P(q[2] | t = l) \times \ldots \times P(q[m] | t = l)$$
$$= \prod_{i=1}^{m} \boldsymbol{P(q[i] | t = l)}$$

Bayesian Prediction with Conditional Independence

- Using conditional independence and chain rule, we can simplify the computations in Bayes' Theorem.

$$P(t = l|q[1], \dots, q[m]) = \frac{\prod_{i=1}^{m} \boldsymbol{P(q[i]|t=l)} \times P(t=l)}{P(q[1],\dots,q[m])}$$

Without normalization

$$P(t = l|q[1], \dots, q[m]) = \prod_{i=1}^{m} \boldsymbol{P(q[i]|t = l)} \times P(t = l)$$

- This is based on the assumption that there is **conditional independence among the descriptive features**, given the label *l* of the target feature.

Example

- The join probability distribution for the four binary features (**H**eadache, **F**ever, **V**omiting, and **M**eningitis)

$$\mathbf{P}(H,F,V,M) = \begin{bmatrix} P(h,f,v,m), & P(\neg h,f,v,m) \\ P(h,f,v,\neg m), & P(\neg h,f,v,\neg m) \\ P(h,f,\neg v,m), & P(\neg h,f,\neg v,m) \\ P(h,f,\neg v,\neg m), & P(\neg h,f,\neg v,\neg m) \\ P(h,\neg f,v,m), & P(\neg h,\neg f,v,m) \\ P(h,\neg f,v,\neg m), & P(\neg h,\neg f,v,\neg m) \\ P(h,\neg f,\neg v,m), & P(\neg h,\neg f,\neg v,m) \\ P(h,\neg f,\neg v,\neg m), & P(\neg h,\neg f,\neg v,\neg m) \end{bmatrix}$$

Example (Cont.)

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- Assuming the three descriptive features are **conditionally independent** of each other, given target feature value,
  we only need to store four factors for the joint probability distribution because

$$Factor_1 : < P(m) = 0.3 >$$
$$Factor_2 : < P(h|m) = 0.6666, P(h|\neg m) = 0.7413 >$$
$$Factor_3 : < P(f|m) = 0.3333, P(f|\neg m) = 0.4286 >$$
$$Factor_4 : < P(v|m) = 0.6666, P(v|\neg m) = 0.5714 >$$

$$P(H, F, V, M) = P(M) \times P(H|M) \times P(F|M) \times P(V|M)$$

- The Bayesian classifier is a very simple form.

Bayesian Inference with Factor Values

- $Factor_1 : < P(m) = 0.3 >$
  $Factor_2 : < P(h|m) = 0.6666, P(h|\neg m) = 0.7413 >$
  $Factor_3 : < P(f|m) = 0.3333, P(f|\neg m) = 0.4286 >$
  $Factor_4 : < P(v|m) = 0.6666, P(v|\neg m) = 0.5714 >$

- A query instance with HEADACHE = *true*, FEVER=*true*, and VOMITING =*true*. Does the patient have MENINGITIS?

  Bayes' Theorem: $P(t = l|q[1], \ldots, q[m]) = \prod_{i=1}^{m} P(q[i]|t = l) \times P(t = l)$

  - $P(m|h, f, v) = P(h|m) \times P(f|m) \times P(v|m) \times P(m)$

    $= 0.6666 \times 0.3333 \times 0.6666 \times 0.3 = 0.0443112$

  - $P(\neg m|h, f, v) = P(h|\neg m) \times P(f|\neg m) \times P(v|\neg m) \times P(\neg m)$

    $= 0.7413 \times 0.4286 \times 0.5714 \times 0.7 = 0.127082$

  - The MAP model prediction would be MENINGITIS =***false*** !!

Outline

- Probability-based Learning
- Fundamentals
  - Basics of Probability Theory
  - Bayes' Theorem
  - Bayesian Prediction
  - Conditional Independence and Factorization
- ☞ **Standard Approach: The Naïve Bayes' Classifier**
- Extensions and Variations
- Summary

Naïve Bayes Classifier

- A **naïve Bayes model** outputs a **MAP** (Maximum A Posterior) prediction, computing the posterior probabilities for the levels of the target feature **under the assumption of conditional independence** between the descriptive features in an instance, given a level of the target feature.

- More formally, the **naïve Bayes model** is defined as follows:

Naive Bayes' Classifier

$$\mathbb{M}(\mathbf{q}) = \underset{l \in levels(t)}{\operatorname{argmax}} \left(\prod_{i=1}^{m} P(\mathbf{q}[i] \mid t = l) \right) \times P(t = l)$$

, where **q** is a query instance.

Naïve Bayes' Classifier

- Naïve Byes domain
  - The domain representation in Naïve Byes specifies a conditional probability for each possible value within the domain of a descriptive feature, corresponding to each label in the domain of the target
- Naïve Bayes' classifier is **simple to train!**

1. Calculate the priors for each of the target levels (class labels)
2. Calculate the conditional probabilities for each descriptive feature given each target level.

$$\mathbb{M}(\mathbf{q}) = \underset{l \in levels(t)}{\operatorname{argmax}} \left(\prod_{i=1}^{m} P(\mathbf{q}[i] \mid t = l) \right) \times P(t = l)$$

Example

- Build a model that predicts whether loan applications are fraudulent or genuine.

ID	CREDIT HISTORY	GUARANTOR/ COAPPLICANT	ACCOMMODATION	FRAUD
1	current	none	own	true
2	paid	none	own	false
3	paid	none	own	false
4	paid	guarantor	rent	true
5	arrears	none	own	false
6	arrears	none	own	true
7	current	none	own	false
8	arrears	none	own	false
9	current	none	rent	false
10	none	none	own	true
11	current	coapplicant	own	false
12	current	none	own	true
13	current	none	rent	true
14	paid	none	own	false
15	arrears	none	own	false
16	current	none	own	false
17	arrears	coapplicant	rent	false
18	arrears	none	free	false
19	arrears	none	own	false
20	paid	none	own	false

Table. A dataset from a loan application fraud detection domain. FRAUD is the binary target feature.

Example (cont.)

- **To train a Naïve Bayes model**, we need to compute:
  - the prior probabilities for each level within the domain of the target feature, and
  - the conditional probability for each value within the domain of every descriptive feature, conditioned on each level that the target feature can take.
- This dataset shows two levels in the target feature domain, four levels in the CREDIT HISTORY domain, three in the GUARANTOR/COAPPLICANT domain, and three in the ACCOMMODATION domain.
- So, after computing $2+(2\times4)+(2\times3)+(2\times3)=22$ probabilities, the naive Bayes model is ready.

Probabilities needed by the Naïve Bayes' Prediction

$P(fr)$	=	0.3
$P(\text{CH} = none \mid fr)$	=	0.1666
$P(\text{CH} = paid \mid fr)$	=	0.1666
$P(\text{CH} = current \mid fr)$	=	0.5
$P(\text{CH} = arrears \mid fr)$	=	0.1666
$P(\text{GC} = none \mid fr)$	=	0.8334
$P(\text{GC} = guarantor \mid fr)$	=	0.1666
$P(\text{GC} = coapplicant \mid fr)$	=	0
$P(\text{ACC} = own \mid fr)$	=	0.6666
$P(\text{ACC} = rent \mid fr)$	=	0.3333
$P(\text{ACC} = free \mid fr)$	=	0

$P(\neg fr)$	=	0.7
$P(\text{CH} = none \mid \neg fr)$	=	0
$P(\text{CH} = paid \mid \neg fr)$	=	0.2857
$P(\text{CH} = current \mid \neg fr)$	=	0.2857
$P(\text{CH} = arrears \mid \neg fr)$	=	0.4286
$P(\text{GC} = none \mid \neg fr)$	=	0.8571
$P(\text{GC} = guarantor \mid \neg fr)$	=	0
$P(\text{GC} = coapplicant \mid \neg fr)$	=	0.1429
$P(\text{ACC} = own \mid \neg fr)$	=	0.7857
$P(\text{ACC} = rent \mid \neg fr)$	=	0.1429
$P(\text{ACC} = free \mid \neg fr)$	=	0.0714

Notation key: FR = FRAUD, CH = CREDIT HISTORY, GC = GUARANTOR/COAPPLICANT, ACC =ACCOMMODATION.

Prediction using the Naïve Bayes model

- A query instance is

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMODATION	FRAUDULENT
paid	none	rent	?

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.0139$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.0245$$

The prediction to Fraudulent is **'false'.**

$P(fr)$	=	0.3	$P(\neg fr)$	=	0.7
$P(\text{CH} = none \mid fr)$	=	0.1666	$P(\text{CH} = none \mid \neg fr)$	=	0
$P(\text{CH} = paid \mid fr)$	=	0.1666	$P(\text{CH} = paid \mid \neg fr)$	=	0.2857
$P(\text{CH} = current \mid fr)$	=	0.5	$P(\text{CH} = current \mid \neg fr)$	=	0.2857
$P(\text{CH} = arrears \mid fr)$	=	0.1666	$P(\text{CH} = arrears \mid \neg fr)$	=	0.4286
$P(\text{GC} = none \mid fr)$	=	0.8334	$P(\text{GC} = none \mid \neg fr)$	=	0.8571
$P(\text{GC} = guarantor \mid fr)$	=	0.1666	$P(\text{GC} = guarantor \mid \neg fr)$	=	0
$P(\text{GC} = coapplicant \mid fr)$	=	0	$P(\text{GC} = coapplicant \mid \neg fr)$	=	0.1429
$P(\text{ACC} = own \mid fr)$	=	0.6666	$P(\text{ACC} = own \mid \neg fr)$	=	0.7857
$P(\text{ACC} = rent \mid fr)$	=	0.3333	$P(\text{ACC} = rent \mid \neg fr)$	=	0.1429
$P(\text{ACC} = free \mid fr)$	=	0	$P(\text{ACC} = free \mid \neg fr)$	=	0.0714

Table: The Naïve Bayes Model - Probabilities needed by a Naïve Bayes Prediction

Outline

- Probability-based Learning
- Fundamentals
- Standard Approach: The Naïve Bayes' Classifier
- ☞ **Extensions and Variations**
  - Smoothing
  - Handling Continuous Features
- Summary

Motivation Example

Table: A dataset from a loan application fraud detection domain.

ID	Credit History	Guarantor/ CoApplicant	Accomodation	Fraud
1	current	none	own	true
2	paid	none	own	false
3	paid	none	own	false
4	paid	guarantor	rent	true
5	arrears	none	own	false
6	arrears	none	own	true
7	current	none	own	false
8	arrears	none	own	false
9	current	none	rent	false
10	none	none	own	true
11	current	coapplicant	own	false
12	current	none	own	true
13	current	none	rent	true
14	paid	none	own	false
15	arrears	none	own	false
16	current	none	own	false
17	arrears	coapplicant	rent	false
18	arrears	none	free	false
19	arrears	none	own	false
20	paid	none	own	false

Credit History	Guarantor/CoApplicant	Accommodation	Fraudulent
paid	guarantor	free	?

Example (Cont.)

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMMODATION	FRAUDULENT
paid	guarantor	free	?

- When there are no instances in the training data that match a specific combination of target feature and descriptive feature values, the conditional probability is equal to zero.

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = paid \mid fr)$	$=$	0.1666	$P(CH = paid \mid \neg fr)$	$=$	0.2857
$P(GC = guarantor \mid fr)$	$=$	0.1666	$P(GC = guarantor \mid \neg fr)$	$=$	0
$P(ACC = free \mid fr)$	$=$	0	$P(ACC = free \mid \neg fr)$	$=$	0.0714

Table. The relevant probabilities

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.0$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.0$$

This Bayesian model is unable to return a prediction for this query.

Smoothing in Naïve Bayes Model

- In the Naive Bayes model, **smoothing** is used to handle situations where a given class and feature value do not occur together in the training data, leading to a conditional probability of zero.
- Smoothing assigns a small probability to the unobserved occurrences of feature values in each class.
- A commonly used technique for smoothing in the context of the Naive Bayes model is **Laplace smoothing** (or **add-one smoothing**).

Laplace Smoothing

- The formula with **Laplace smoothing** for conditional probabilities is defined as

Laplacian Smoothing (conditional probabilities)

$$P(f = v|t) = \frac{count(f = v|t) + k}{count(f|t) + (k \times |Domain(f)|)}$$

  - $count(f = v|\ t)$ is how often the event $f = v$ occurs in the subset of rows in the dataset where the target level is t,
  - $count(f\ |\ t)$ is how often the feature, f, took any value in the subset of rows in the dataset where the target level is t,
  - $|Domain(f\ )|$ is the number of values in the domain of the feature, and
  - k is a predetermined parameter. **Larger values of k mean that more smoothing occurs**—that is more probability mass is taken from the larger probabilities and given to the small probabilities. Typically k takes small values such as 1, 2, or 3.

Example

$$P(f = v|t) = \frac{count(f = v|t) + k}{count(f|t) + (k \times |Domain(f)|)}$$

Raw Probabilities	$P(GC = none\|\neg fr)$	=	0.8571
	$P(GC = guarantor\|\neg fr)$	=	0
	$P(GC = coapplicant\|\neg fr)$	=	0.1429
Smoothing Parameters	k	=	3
	$count(GC\|\neg fr)$	=	14
	$count(GC = none\|\neg fr)$	=	12
	$count(GC = guarantor\|\neg fr)$	=	0
	$count(GC = coapplicant\|\neg fr)$	=	2
	$\|Domain(GC)\|$	=	3
Smoothed Probabilities	$P(GC = none\|\neg fr) = \frac{12+3}{14+(3\times3)}$	=	0.6522
	$P(GC = guarantor\|\neg fr) = \frac{0+3}{14+(3\times3)}$	=	0.1304
	$P(GC = coapplicant\|\neg fr) = \frac{2+3}{14+(3\times3)}$	=	0.2174

Table: Smoothing the posterior probabilities for the GUARANTOR/COAPPLICANT feature conditioned on FRAUDULENT being False.

Example: Prediction after Smoothing

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMMODATION	FRAUDULENT
paid	guarantor	free	?

Table. The relevant smoothed probabilities

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(ACC = Free \mid fr)$	$=$	0.2	$P(ACC = Free \mid \neg fr)$	$=$	0.1739

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[m] \mid fr)\right) \times P(fr) = 0.0036$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[m] \mid \neg fr)\right) \times P(\neg fr) = 0.0043$$

Outline

- Probability-based Learning
- Fundamentals
- Standard Approach: The Naïve Bayes' Classifier
- Extensions and Variations
  - Smoothing
  - ☞ **Handling Continuous Features**
- Summary

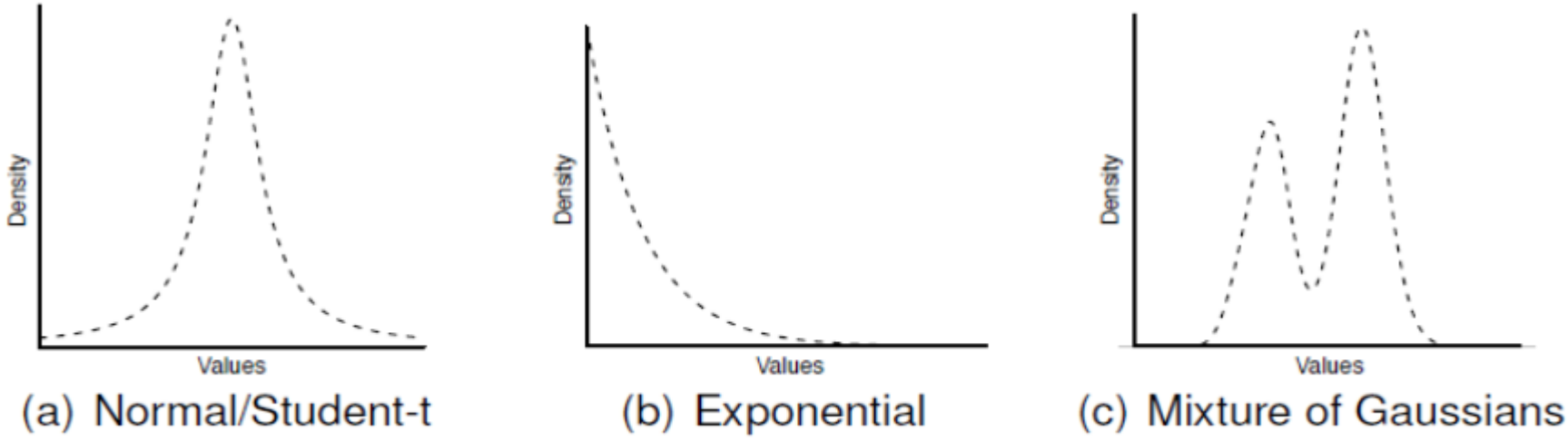
Continuous Features – Motivation

- A continuous feature can have an infinite number of values in its domain
- The relative frequency of any particular value for a continuous feature will be indistinguishable from zero given a large dataset.

  ➔ The problem of zero probabilities.
- The way to solve the problem of zero probabilities is to think in terms of how the probability of a continuous feature taking a value is distributed across the range of values that a continuous feature can take.

Probability Density Functions

- A **probability density function (PDF)** represents the probability distribution of a continuous feature using a mathematical function.
- There are a large number of standard, well-defined probability distributions such as **normal**, **exponential**, and **mixture of Gaussians** distributions —that are commonly used in probabilistic prediction models.



(a) Normal/Student-t (b) Exponential (c) Mixture of Gaussians

Probability Density Functions

- A PDF defines a density curve and the shape of the curve is determined by:
  - the statistical distribution that is used to define the PDF, and
  - the values of the statistical distribution parameters
- All standard PDFs have *parameters* that alter the shape of the density curve defining that distribution
- In order to use a PDF to represent the probability of a continuous feature taking different values, we need to choose these parameters to fit the characteristics of the data.

Some Standard Probability Distribution

Parameters **Probability Density Function**

Normal

$x \in \mathbb{R}$

$\mu \in \mathbb{R}$

$\sigma \in \mathbb{R}_{>0}$

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Student-t

$x \in \mathbb{R}$

$\phi \in \mathbb{R}$

$\rho \in \mathbb{R}_{>0}$

$\kappa \in \mathbb{R}_{>0}$

$z = \frac{x - \phi}{\rho}$

$$\tau(x, \phi, \rho, \kappa) = \frac{\Gamma(\frac{\kappa+1}{2})}{\Gamma(\frac{\kappa}{2}) \times \sqrt{\pi\kappa} \times \rho} \times \left(1 + \left(\frac{1}{\kappa} \times z^2\right)\right)^{-\frac{\kappa+1}{2}}$$

Exponential

$x \in \mathbb{R}$

$\lambda \in \mathbb{R}_{>0}$

$$E(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Mixture of n Gaussians

$x \in \mathbb{R}$

$\{\mu_1, \ldots, \mu_n | \mu_i \in \mathbb{R}\}$

$\{\sigma_1, \ldots, \sigma_n | \sigma_i \in \mathbb{R}_{>0}\}$

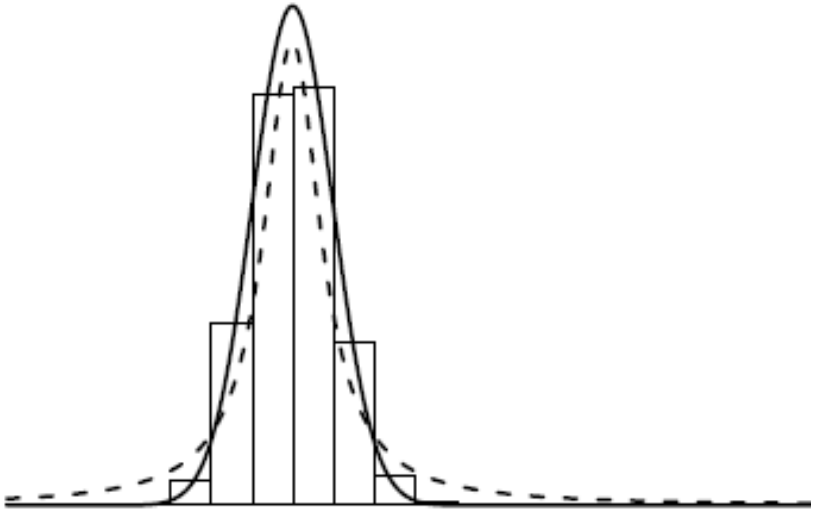
$\{\omega_1, \ldots, \omega_n | \omega_i \in \mathbb{R}_{>0}\}$

$\sum_{i=1}^{n} \omega_i = 0$

$$N(x, \mu_1, \sigma_1, \omega_1, \ldots, \mu_n, \sigma_n, \omega_n) = \sum_{i=1}^{n} \frac{\omega_i}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

Probability from PDF

- A PDF is an abstraction over a density histogram and consequently PDF represents probabilities in terms of area under the curve.



- To use a PDF to calculate a probability
  1. Decide on the interval you wish to calculate the probability for
  2. Calculate the area under the density curve for that interval

- Rather than exact probabilities, we only need to calculate the relative likelihood of a continuous feature taking a value given different labels of a target function. ➔ The height of the density curve defined by an appropriately defined PDF at a particular feature value

Example

Table: The dataset from the loan application fraud detection domain with a new continuous descriptive features added: ACCOUNT BALANCE

ID	Credit History	Guarantor/ CoApplicant	Accommodation	Account Balance	Fraud
1	current	none	own	56.75	true
2	current	none	own	1,800.11	false
3	current	none	own	1,341.03	false
4	paid	guarantor	rent	749.50	true
5	arrears	none	own	1,150.00	false
6	arrears	none	own	928.30	true
7	current	none	own	250.90	false
8	arrears	none	own	806.15	false
9	current	none	rent	1,209.02	false
10	none	none	own	405.72	true
11	current	coapplicant	own	550.00	false
12	current	none	free	223.89	true
13	current	none	rent	103.23	true
14	paid	none	own	758.22	false
15	arrears	none	own	430.79	false
16	current	none	own	675.11	false
17	arrears	coapplicant	rent	1,657.20	false
18	arrears	none	free	1,405.18	false
19	arrears	none	own	760.51	false
20	current	none	own	985.41	false

Example: PDFs for ACCOUNT BALANCE

- We need to define two PDFs for the new ACCOUNT BALANCE (AB) feature with each PDF conditioned on a different level in the domain of the target:
  - $P(AB = x|fr) = \boldsymbol{PDF_1(AB = c|fr)}$
  - $P(AB = x|\neg fr) = \boldsymbol{PDF_2(AB = c|\neg fr)}$

    NOTE: These two PDFS do not have to be defined using the same distribution.

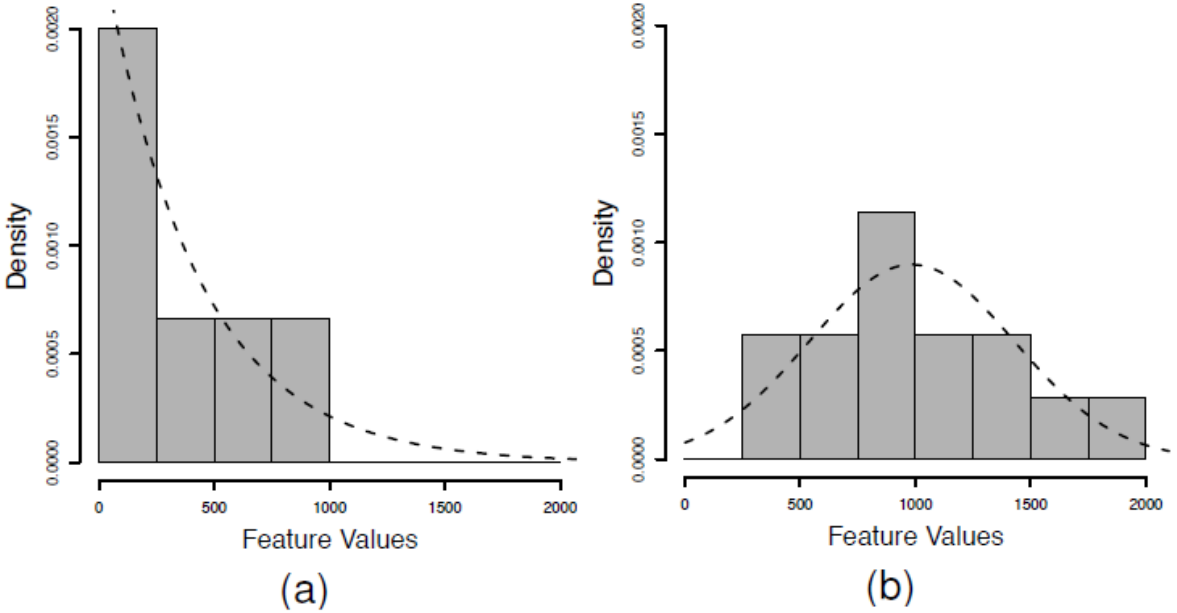


Figure: Histograms, using a bin size of 250 units, and density curves for the ACCOUNT BALANCE feature: **(a)** the fraudulent instances overlaid with a fitted **exponential distribution**; **(b)** the non-fraudulent instances overlaid with a fitted **normal distribution**.

Example (Cont.)

- Once we have selected the distributions the next step is to fit the distributions to the data.
  - To fit the **exponential distribution** we simply compute the sample mean, $\bar{x}$, of the ACCOUNT BALANCE feature in the set of instances where FRAUDULENT='True' and set the λ parameter equal to one divided by $\bar{x}$.
  - To fit the **normal distribution** to the set of instances where FRAUDULENT='False' we simply compute the sample mean and sample standard deviation, *s*, for the ACCOUNT BALANCE feature for this set of instances and set the parameters of the normal distribution to these values.
- Most data analytics packages and programming APIs provide functions that implement methods to fit a specified distribution to a given datasets

Probabilities for the Naïve Bayes Model

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = none \mid fr)$	$=$	0.2222	$P(CH = none \mid \neg fr)$	$=$	0.1154
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(CH = current \mid fr)$	$=$	0.3333	$P(CH = current \mid \neg fr)$	$=$	0.2692
$P(CH = arrears \mid fr)$	$=$	0.2222	$P(CH = arrears \mid \neg fr)$	$=$	0.3462
$P(GC = none \mid fr)$	$=$	0.5333	$P(GC = none \mid \neg fr)$	$=$	0.6522
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(GC = coapplicant \mid fr)$	$=$	0.2	$P(GC = coapplicant \mid \neg fr)$	$=$	0.2174
$P(ACC = own \mid fr)$	$=$	0.4667	$P(ACC = own \mid \neg fr)$	$=$	0.6087
$P(ACC = rent \mid fr)$	$=$	0.3333	$P(ACC = rent \mid \neg fr)$	$=$	0.2174
$P(ACC = free \mid fr)$	$=$	0.2	$P(ACC = free \mid \neg fr)$	$=$	0.1739
$P(AB = x \mid fr)$			$P(AB = x \mid \neg fr)$		
	$\approx$	$E\begin{pmatrix} x, \\ \lambda = 0.0024 \end{pmatrix}$		$\approx$	$N\begin{pmatrix} x, \\ \mu = 984.26, \\ \sigma = 460.94 \end{pmatrix}$

Table: The Laplace smoothed (with k = 3) probabilities needed by a naive Bayes prediction model

Prediction

- A query instance:

Credit History	Guarantor/ CoApplicant	Accomodation	Account Balance	Fraudulent
paid	guarantor	free	759.07	?

- Probabilities need for Naive Bayes prediction

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(ACC = free \mid fr)$	$=$	0.2	$P(ACC = free \mid \neg fr)$	$=$	0.1739
$P(AB = 759.07 \mid fr)$ $\approx E\begin{pmatrix} 759.07, \\ \lambda = 0.0024 \end{pmatrix}$	$=$	0.00039	$P(AB = 759.07 \mid \neg fr)$ $\approx N\begin{pmatrix} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{pmatrix}$	$=$	0.00077

- The Naive Bayes model :

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.0000014$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.0000033$$

➔ Fraud = ‘FALSE’

Alternative Technique for Continuous Features

- A commonly used alternative to representing a continuous feature using a probability density function is to convert the feature into a categorical feature using **binning**
- Binning techniques
  - Equal-width binning
  - **Equal-frequency binning**
- Example: a continuous feature, LOAN AMOUNT

ID	LOAN AMOUNT	BINNED LOAN AMOUNT	FRAUD
15	500	bin1	false
19	500	bin1	false
1	900	bin1	true
10	9,500	bin1	true
12	9,850	bin1	true
4	10,000	bin2	true
17	15,450	bin2	false
16	16,000	bin2	false
11	16,750	bin2	false
8	18,500	bin2	false

ID	LOAN AMOUNT	BINNED LOAN AMOUNT	FRAUD
9	20,000	bin3	false
7	25,000	bin3	false
5	32,000	bin3	false
20	35,000	bin3	false
3	48,000	bin3	false
18	50,000	bin4	false
14	65,000	bin4	false
13	95,500	bin4	true
2	150,000	bin4	false
6	250,000	bin4	true

Alternative Technique (cont.)

- Once we have discretized the data we need to record the raw continuous feature threshold between the bins so that we can use these for query feature values.
  - E.g.,

LOAN AMOUNT

Bin Thresholds		
	Bin1	$\leq 9,925$
$9,925 <$	Bin2	$\leq 19,250$
$19,225 <$	Bin3	$\leq 49,000$
$49,000 <$	Bin4	

Probabilities needed for the Naïve Bayes Model

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = none \mid fr)$	$=$	0.2222	$P(CH = none \mid \neg fr)$	$=$	0.1154
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(CH = current \mid fr)$	$=$	0.3333	$P(CH = current \mid \neg fr)$	$=$	0.2692
$P(CH = arrears \mid fr)$	$=$	0.2222	$P(CH = arrears \mid \neg fr)$	$=$	0.3462
$P(GC = none \mid fr)$	$=$	0.5333	$P(GC = none \mid \neg fr)$	$=$	0.6522
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(GC = coapplicant \mid fr)$	$=$	0.2	$P(GC = coapplicant \mid \neg fr)$	$=$	0.2174
$P(ACC = own \mid fr)$	$=$	0.4667	$P(ACC = own \mid \neg fr)$	$=$	0.6087
$P(ACC = rent \mid fr)$	$=$	0.3333	$P(ACC = rent \mid \neg fr)$	$=$	0.2174
$P(ACC = free \mid fr)$	$=$	0.2	$P(ACC = free \mid \neg fr)$	$=$	0.1739
$P(AB = x \mid fr) \approx E\begin{pmatrix} x, \\ \lambda = 0.0024 \end{pmatrix}$			$P(AB = x \mid \neg fr) \approx N\begin{pmatrix} x, \\ \mu = 984.26, \\ \sigma = 460.94 \end{pmatrix}$		
$P(BLA = bin1 \mid fr)$	$=$	0.3333	$P(BLA = bin1 \mid \neg fr)$	$=$	0.1923
$P(BLA = bin2 \mid fr)$	$=$	0.2222	$P(BLA = bin2 \mid \neg fr)$	$=$	0.2692
$P(BLA = bin3 \mid fr)$	$=$	0.1667	$P(BLA = bin3 \mid \neg fr)$	$=$	0.3077
$P(BLA = bin4 \mid fr)$	$=$	0.2778	$P(BLA = bin4 \mid \neg fr)$	$=$	0.2308

Prediction Example

Credit History	Guarantor/ CoApplicant	Accomodation	Account Balance	Loan Amount	Fraudulent
paid	guarantor	free	759.07	8,000	?

$$P(fr) = 0.3 \qquad P(\neg fr) = 0.7$$

$$P(CH = paid|fr) = 0.2222 \qquad P(CH = paid|\neg fr) = 0.2692$$

$$P(GC = guarantor|fr) = 0.2667 \qquad P(GC = guarantor|\neg fr) = 0.1304$$

$$P(ACC = free|fr) = 0.2 \qquad P(ACC = free|\neg fr) = 0.1739$$

$$P(AB = 759.07|fr) \approx E\begin{pmatrix} 759.07, \\ \lambda = 0.0024 \end{pmatrix} = 0.00039$$

$$P(AB = 759.07|\neg fr) \approx N\begin{pmatrix} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{pmatrix} = 0.00077$$

$$P(BLA = bin1|fr) = 0.3333 \qquad P(BLA = bin1|\neg fr) = 0.1923$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.000000462$$

$$\left(\prod_{k=1}^{n} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.000000633$$

Outline

- Probability-based Learning
- Fundamentals
- Standard Approach: The Naïve Bayes' Classifier
- Extensions and Variations
- ☞ **Summary**

Summary of Naïve Bayes' Classifier

- A Naive Bayes' classifier naively assumes that each of the descriptive features in a domain is conditionally independent of all of the other descriptive features, given the state of the target feature.
- This assumption, although often wrong, enables the Naïve Bayes' model to maximally factorize the representation that it uses the domain.
- Surprisingly, given the naivety and strength of the assumption it depends upon, a Naive Bayes' model often performs reasonably well.