diese Objekte klassifiziert. Als label stehen z.B. alle Worte des bekannten Vokabulars zur Verfügung oder alle isolierten Phoneme der Sprache.

### 5.1.1 Verwendete Symbole

**V** : Fettgedruckte Symbole in Formeln stellen Vektoren dar
K : Die Anzahl der Klassen bzw. Gruppen, die wir unterscheiden
D : Die Anzahl der Dimensionen des Merkmalsraumes
$\Omega_k$ : Bezeichnet die k-te Klasse
$P(\mathbf{x}|\Omega_i)$ : Kurzschreibweise für $P(\mathbf{x}|\mathbf{x} \in \Omega_i)$ [1]
$N_k$ : Anzahl der Mischverteilungskomponenten der k-ten Klasse
$\theta$ : Parametervektor einer parametrisierten Verteilungsfunktion
$C_k$ : Anzahl der Mischverteilungskomponenten der k-ten Klasse ($\Omega_k$)

### 5.1.2 Mathematische Vorbereitungen

Hier werden nur einige der wichtigsten Begriffe genannt und kurz erläutert. Dies ist bestenfalls eine Auffrischung von Kenntnissen des Lesers. Für eine umfangreichere Beschreibung verweisen wir auf die Anhänge zu diesem Buch und, falls darüber hinaus notwendig, auf weiterführende Lehrbücher.

$P(X, Y)$ **Verbundwahrscheinlichkeit**:
Die Verbundwahrscheinlichkeit ist die Wahrscheinlichkeit, dass das Ereignis X *und* das Ereignis Y gemeinsam eintreten. Es gilt:

$$P(X, Y) = P(X|Y) \cdot P(Y) = P(Y|X) \cdot P(X) \tag{5.1}$$

**Bayes-Formel**:

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} = \frac{P(Y|X) \cdot P(X)}{\sum_{x_i \in X} P(Y|x_i) P(x_i)} \tag{5.2}$$

$P(X|Y)$ **Bedingte Wahrscheinlichkeit**:
Wahrscheinlichkeit, dass das Ereignis X auftritt, wenn das Ereignis Y bereits eingetreten ist. Meist verwendet als $P(\mathbf{x}|\Omega_i)$, der Wahrscheinlichkeit, dass ein Merkmalsvektor $\mathbf{x}$ von der Klasse $\Omega_i$ erzeugt wird.

Es gibt drei beachtenswerte Sonderfälle bei bedingten Wahrscheinlichkeiten:

1. Das Ereignis Y führt automatisch zum Ereignis X, d.h. wir wissen genau, dass X sicher ist, wenn Y eingetreten ist. Mathematisch geschrieben bedeutet das $P(X|Y) = 1$ und mit Gl. (5.1) folgt

$$P(X, Y) = P(Y) \text{ und } P(Y|X) = \frac{P(Y)}{P(X)} \tag{5.3}$$

[1] Die Schreibweise $\mathbf{x} \in \Omega_i$ ist mathematisch eigentlich nicht korrekt, da $\Omega_i$ hier *nicht* als Menge aufgefasst wird. Gemeint ist hier: „Der Vektor $\mathbf{x}$ wurde von der Klasse $\Omega_i$ erzeugt."

2. Das Ereignis Y führt automatisch zum Ereignis X und umgekehrt, d.h. $P(X|Y) = 1$ und $P(Y|X) = 1$, dann folgt sofort $P(X) = P(Y)$, da die Ereignisse X und Y dann immer gemeinsam auftreten und damit identisch sind.

3. Die Ereignisse X und Y sind statistisch unabhängig, d.h. das Wissen, dass das Ereignis Y eingetreten ist, macht das Ereignis X nicht wahrscheinlicher und umgekehrt. D.h. $P(A|B) = P(A)$ und $P(B|A) = P(B)$. Unter Verwendung von Gl. (5.1) ist

$$P(A, B) = P(A) \cdot P(B). \tag{5.4}$$

Ebenso gilt in diesem Falle:

$$P(A, B|X) = P(A|X) \cdot P(B|X). \tag{5.5}$$

$P(Y) = \sum_{x_i \in X} P(Y|x_i)P(x_i) = \sum_{x_i \in X} P(Y, x_i)$ **Marginalisierung**:
Das Ereignis Y wird bedingt auf das Eintreten eines anderen Ereignisses X. Wird über *alle* Ausgänge von X summiert, so ändert sich nichts. Diese Technik wird im Folgenden häufig angewandt. Die Umkehrung dieser Gleichung, also die Summierung, liefert höher akkumulierte Wahrscheinlichkeiten, bei denen eine Bedingung aufsummiert wird. Es entstehen dabei die so genannten **Randwahrscheinlichkeiten**.

Zu diesen Begriffen bringen wir nun ein Beispiel:

**Beispiel 5.2 (Berechnung verschiedener Wahrscheinlichkeiten):**
Bei einer Person zwischen 20 und 60 Jahren bezeichne X das Geschlecht (weiblich: w, männlich: m), Y bezeichnet, ob die Person erwerbstätig ist (erwerbstätig: e, nicht erwerbstätig: ne).
Die *Verbundwahrscheinlichkeit* gibt vollständige Information über die Personengruppe: Es sei

$$\begin{aligned} P(X = m, Y = e) = 0.4 \quad &; \quad P(X = w, Y = e) = 0.3 \\ P(X = m, Y = ne) = 0.1 \quad &; \quad P(X = w, Y = ne) = 0.2. \end{aligned}$$

Durch Umkehrung der Marginalisierung lassen sich daraus die *Randwahrscheinlichkeiten* berechnen:

$$\begin{aligned} P(X - m) &= P(X - m, Y - e) + P(X - m, Y - ne) - 0.4 + 0.1 - 0.5 \\ P(X = w) &= P(X = w, Y - e) + P(X = w, Y = ne) = 0.3 + 0.2 = 0.5 \\ P(Y = e) &= P(X = m, Y = e) + P(X = w, Y = e) = 0.4 + 0.3 = 0.7 \\ P(Y = ne) &= P(X = m, Y = ne) + P(X = w, Y = ne) = 0.1 + 0.2 = 0.3 \end{aligned}$$

Daraus entnehmen wir *bedingte Wahrscheinlichkeiten*, z.B.

$$\begin{aligned} P(X = m|Y = e) &= P(X = m, Y = e)/P(Y = e) = 0.4/0.7 = 0.57 \\ P(X = m|Y = ne) &= P(X = m, Y = ne)/P(Y = ne) = 0.1/0.3 = 0.33. \end{aligned}$$

Mit Hilfe der *Bayes-Formel* können wir nun *posteriori- Wahrscheinlichkeiten* berechnen, z.B.:

$$P(Y = e|X = m) = P(X = m, Y = e)/P(X = m) = 0.4/0.5 = 0.8$$
$$P(Y = e|X = w) = P(X = w, Y = e)/P(X = w) = 0.3/0.5 = 0.6$$

Aus den letzten Ergebnissen können wir sehen:

- wissen wir, dass eine Person erwerbstätig ist, so können wir schwerlich ihr Geschlecht daraus abschätzen, denn die Wahrscheinlichkeiten sind sehr ähnlich (m: 0.57, w: 0.43).
- wissen wir, dass eine Person nicht erwerbstätig ist, so können wir ihr Geschlecht besser daraus abschätzen, da die Wahrscheinlichkeiten sich deutlicher unterscheiden (m: 0.33, w: 0.66).
- am besten gelingt die Abschätzung im posteriori-Fall: für einen Mann können wir Erwerbstätigkeit mit Wahrscheinlichkeit 0.8 vorhersagen, für eine Frau immerhin noch mit 0.6. □

### A-priori, a-posteriori und klassenbedingte Wahrscheinlichkeiten

In der Bayes'schen Theorie ist es notwendig, für jegliche Berechnungen oder Modellierungen Wahrscheinlichkeiten vorzugeben. Sie werden A-Priori Wahrscheinlichkeiten genannt. Diese müssen empirisch ermittelt werden. Modellieren wir z.B. Spracherzeugung, so müssen wir Annahmen über diesen Prozess in Form von Wahrscheinlichkeiten in unserem Modell vorraussetzen. Als A-Posteriori Wahrscheinlichkeiten bezeichnet man die Wahrscheinlichkeiten, die keine empirischen Annahmen sind.

Abb. 5.1 zeigt uns 2 Beispiele. Wir gehen dabei von dem unten näher erläuterten statistischen Erzeugungsmodell aus, das hier zunächst nur kurz angerissen werden soll. Angenommen, wir möchten 2 verschiedene Laute unterscheiden und haben daher zwei Klassen $\Omega_1$ und $\Omega_2$. Wir gehen davon aus, dass der Prozess der Spracherzeugung wie folgt abläuft:

Es wird zuerst per Zufall einer der beiden Laute und damit eine der beiden Klassen ausgewählt und dann, einer klassenabhängigen Verteilung folgend, ein zufälliger Merkmalsvektor $\mathbf{x}$ erzeugt. (Das bedeutet jede Klasse besitzt eine eigene Verteilung, nach der sie ihre Merkmalsvektoren $\mathbf{x}$ produziert). Diese beiden bedingten Klassenverteilungen $P(\mathbf{x}|\Omega_1)$ und $P(\mathbf{x}|\Omega_2)$ seien a-priori gegeben. Sie sind im jeweils linken Bild dargestellt. Die Wahrscheinlichkeiten, mit der die einzelnen Klassen (und damit die beiden Laute) ausgewählt werden, sind im Beispiel mit $P(\Omega_1) = 0,6$ und $P(\Omega_2) = 0,4$ gewählt (ebenfalls A-Priori-Wahrscheinlichkeiten). Dies würde bedeuten, dass der Laut 1 etwas häufiger vorkommt als Laut 2. Aus ihnen können die Verbundwahrscheinlichkeiten $P(\mathbf{x}, \Omega_1)$ und $P(\mathbf{x}, \Omega_2)$ (Mitte) nach Gl. (5.1) berechnet werden. Die A-Posteriori-

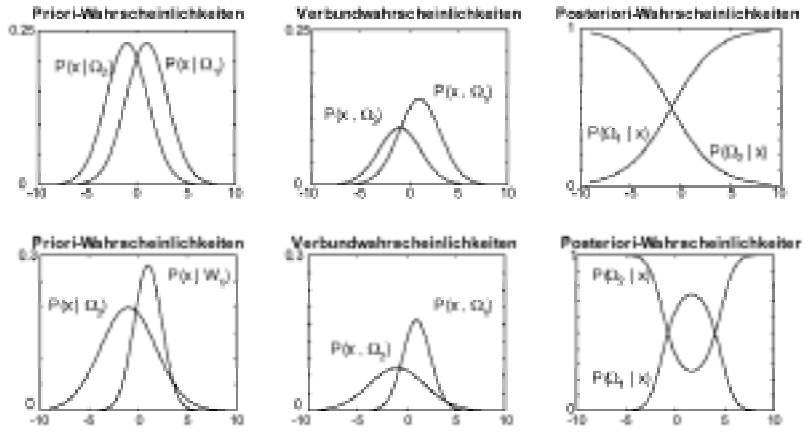


***Abbildung 5.1:*** *Priori-/ Verbund-/ Posteriori Wahrscheinlichkeiten an 2 Beispielen*

Wahrscheinlichkeiten $P(\mathbf{x}|\Omega_1)$ und $P(\mathbf{x}|\Omega_2)$ sind im rechten Bild dargestellt (nach der Bayes-Formel (5.2) berechnet).

In Worten ist die klassenabhängige Wahrscheinlichkeit $P(\mathbf{x}|\Omega)$ in diesem Beispiel: „Die Wahrscheinlichkeit, dass ein bestimmter Vektor $\mathbf{x}$ erzeugt wurde, *wenn* die Klasse 1 (oder 2) ausgewählt wurde“. Die Verbundwahrscheinlichkeiten bedeuten in Worten: „Die Wahrscheinlichkeit, dass Klasse 1 (oder 2) ausgewählt wurde *und* der Vektor $\mathbf{x}$ von dieser Klasse erzeugt wurde.“ Was wir jedoch bei der Klassifikation wissen wollen, sind die Posteriori-Wahrscheinlichkeiten: „Gegeben sei ein beliebiger Vektor $\mathbf{x}$, mit welcher Wahrscheinlichkeit wurde er von Klasse 1 (oder 2) erzeugt?“

## 5.2 Zielsetzung des Kapitels „Klassifikation“

In den letzten Kapiteln wurde erläutert, wie wir Schritt für Schritt mit verschiedenen Verfahren aus einem Audiosignal mit Hilfe des Vokaltrakt-Modells nach entsprechender Filterung im Cepstralbereich eine lineare Vorhersage bzw. eine Schätzung des Vokaltraktfrequenzganges vorgenommen hatten. Das Ergebnis dieser Schätzung ist ein Merkmalsvektor für jeden Zeitabschnitt einer zu erkennenden Äußerung. Durch die Hinzunahme der Kontexte (Nachbarschaften) und der Ableitungen für die einzelnen Komponenten der Merkmalsvektoren erhält man einen hochdimensionalen Vektor, dessen Dimensionalität mit entsprechenden Raumtransformationen reduziert werden kann.

Hat jeder Merkmalsvektor z.B. 76 Komponenten, so repräsentiert jeder Merkmalsvektor einen Punkt im 76-dimensionalen Merkmalsraum. Um Sprache zu erkennen, nehmen wir nun an, dass ähnlich ausgesprochene Laute auch ähnliche (also dicht beieinander liegende) Merkmalsvektoren produzieren. Ein Problem, dem wir jedoch begegnen müssen, ist, dass ein und dasselbe Wort zweimal ausgesprochen in der Regel verschiedene Merkmalsvektoren erzeugen wird. Dies ist leicht einzusehen, denn die Aussprache eines