

Gernot A. Fink

# Mustererkennung mit Markov-Modellen

Theorie - Praxis -  
Anwendungsgebiete



Teubner

B. G. Teubner Stuttgart • Leipzig • Wiesbaden

# Inhalt

<b>1</b>	<b>Einleitung</b>	<b>13</b>
1.1	Thematischer Kontext . . . . .	14
1.2	Funktionsprinzipien von Markov-Modellen . . . . .	15
1.3	Zielsetzung und Aufbau . . . . .	17
<b>2</b>	<b>Anwendungen</b>	<b>19</b>
2.1	Sprache . . . . .	19
2.2	Schrift . . . . .	25
2.3	Biologische Sequenzen . . . . .	33
2.4	Ausblick . . . . .	37
<b>I</b>	<b>Theorie</b>	<b>39</b>
<b>3</b>	<b>Grundlagen der Statistik</b>	<b>41</b>
3.1	Zufallsexperiment, Ereignis und Wahrscheinlichkeit . . . . .	41
3.2	Zufalls variable und Wahrscheinlichkeitsverteilungen . . . . .	43
3.3	Parameter von Wahrscheinlichkeitsverteilungen . . . . .	45
3.4	Normal Verteilungen und Mischverteilungsmodelle . . . . .	46
3.5	Stochastische Prozesse und Markov-Ketten . . . . .	47
3.6	Prinzipien der Parameterschätzung . . . . .	49
3.6.1	Maximum-Likelihood-Schätzung . . . . .	49
3.6.2	Maximum-a-posteriori-Schätzung . . . . .	51
3.7	Literaturhinweise . . . . .	51
<b>4</b>	<b>Vektorquantisierung</b>	<b>53</b>
4.1	Definition . . . . .	53
4.2	Optimalität . . . . .	55
4.3	Algorithmen zum Design von Vektorquantisierern . . . . .	57

	<i>Lloyd-Algorithmus</i> . . . . .	58
	<i>LBG-Algorithmus</i> . . . . .	59
	<i>k-means-Algorithmus</i> . . . . .	61
4.4	Schätzung von Mischverteilungsmodellen . . . . .	62
	<i>EM-Algorithmus</i> . . . . .	63
4.5	Literaturhinweise . . . . .	66
<b>5</b>	<b>Hidden-Markov-Modelle</b>	<b>67</b>
5.1	Definition . . . . .	67
5.2	Emissionsmodellierung . . . . .	69
5.3	Verwendungskonzepte . . . . .	70
5.4	Notation . . . . .	72
5.5	Bewertung . . . . .	73
5.5.1	Die Produktions Wahrscheinlichkeit . . . . .	73
	<i>Forward-Algorithmus</i> . . . . .	74
5.5.2	Die "optimale" Produktionswahrscheinlichkeit . . . . .	76
5.6	Dekodierung . . . . .	79
	<i>Viterbi-Algorithmus</i> . . . . .	80
5.7	Parameterschätzung . . . . .	81
5.7.1	Grundlagen . . . . .	82
	<i>Forward-Backward-Algorithmus</i> . . . . .	83
5.7.2	Trainingsverfahren . . . . .	85
	<i>Baum-Welch-Algorithmus</i> . . . . .	85
	<i>ViteTbi-Training</i> . . . . .	91
	<i>Segmental</i> $k$ <i>-Means</i> .....	93
5.7.3	Mehrere Observationsfolgen . . . . .	95
5.8	Modellvarianten . . . . .	96
5.8.1	Alternative Algorithmen . . . . .	96
5.8.2	Alternative Modellarchitekturen . . . . .	97
5.9	Literaturhinweise . . . . .	97
<b>6</b>	<b>n-Gramm-Modelle</b>	<b>99</b>
6.1	Definition . . . . .	99
6.2	Verwendungskonzepte . . . . .	100
6.3	Notation . . . . .	101
6.4	Bewertung . . . . .	102
6.5	Parameterschätzung . . . . .	104
6.5.1	Umverteilung von Wahrscheinlichkeitsmasse . . . . .	105
	<i>Discounting</i> . . . . .	105

Inhalt		9
6.5.2	Einbeziehung allgemeinerer Verteilungen . . . . .	107
	<i>Interpolation</i> . . . . .	108
	<i>Backing-Off</i> . . . . .	110
6.5.3	Optimierung verallgemeinerter Verteilungen. . . . .	111
6.6	Modellvarianten. . . . .	113
6.6.1	Kategoriemodelle. . . . .	113
6.6.2	Längere zeitliche Abhängigkeiten. . . . .	115
6.7	Literaturhinweise. . . . .	116
<b>II</b>	<b>Praxis</b>	<b>117</b>
<b>7</b>	<b>Rechnen mit Wahrscheinlichkeiten</b>	<b>119</b>
7.1	Logarithmische Wahrscheinlichkeitsrepräsentation . . . . .	119
7.2	Untere Schranken für Wahrscheinlichkeiten . . . . .	122
7.3	Codebuchauswertung für semi-kontinuierliche HMMs. . . . .	123
7.4	Wahrscheinlichkeitsverhältnisse. . . . .	124
<b>8</b>	<b>Konfiguration von Hidden-Markov-Modellen</b>	<b>127</b>
8.1	Modelltopologien . . . . .	127
8.2	Modelluntereinheiten. . . . .	128
8.2.1	Kontextunabhängige Wortuntereinheiten. . . . .	129
8.2.2	Kontextabhängige Wortuntereinheiten. . . . .	130
8.3	Verbundmodelle. . . . .	131
8.4	<i>Proßle-HMMs</i> . . . . .	133
8.5	Emissionsmodellierung. . . . .	135
<b>9</b>	<b>Robuste Parameterschätzung</b>	<b>137</b>
9.1	Merkmalsoptimierung . . . . .	139
9.1.1	Dekorrelation . . . . .	140
	<i>Hauptachsentransformation I</i> . . . . .	141
	<i>Whitening</i> . . . . .	145
9.1.2	Dimensionsreduktion. . . . .	146
	<i>Hauptachsentransformation II</i> . . . . .	146
	<i>Lineare Diskriminanzanalyse</i> . . . . .	147
9.2	<i>Tying</i> . . . . .	151
9.2.1	Modelluntereinheiten. . . . .	152
9.2.2	Zustandstying. . . . .	155
9.2.3	<i>Tying</i> in Mischverteilungsmodellen. . . . .	159
9.3	Parameterinitialisierung . . . . .	161

10		Inhalt
<b>10</b>	<b>Effiziente Modellauswertung</b>	<b>163</b>
10.1	Effiziente Auswertung von Mischverteilungen . . . . .	.163
10.2	<i>Beam Search</i> . . . . .	.165
10.3	Effiziente Parameterschätzung . . . . .	.168
10.3.1	<i>Forward-Backward-Pruning</i> . . . . .	.168
10.3.2	Segmentweiser Baum-Welch-Algorithmus . . . . .	.169
10.3.3	Training von Modellhierarchien . . . . .	.170
10.4	Baumförmige Modellorganisation . . . . .	.171
10.4.1	Präfixbaum für HMMs . . . . .	.171
10.4.2	Baumrepräsentation für n-Gramm-Modelle . . . . .	.172
<b>11</b>	<b>Modellanpassung</b>	<b>177</b>
<b>11.1</b>	Grundprinzipien . . . . .	.177
11.2	Adaption von Hidden-Markov-Modellen . . . . .	.178
	<i>Maximum-Likelihood Linear-Regression</i> . . . . .	.180
11.3	Adaption von n-Gramm-Modellen . . . . .	.182
11.3.1	Cache-Modelle . . . . .	.183
11.3.2	Dialogschrittabhängige Modelle . . . . .	.183
11.3.3	Topic-basierte Sprachmodelle . . . . .	.184
<b>12</b>	<b>Integrierte Suchverfahren</b>	<b>185</b>
12.1	HMM-Netzwerke . . . . .	.188
12.2	Mehrphasensuche . . . . .	.189
12.3	Suchraumkopien . . . . .	.190
12.3.1	Kontextbasierte Suchraumkopien . . . . .	.190
12.3.2	Zeitbasierte Suchbaumkopien . . . . .	.191
12.3.3	<i>Language-Model Look-Ahead</i> . . . . .	.192
12.4	Zeitsynchrone parallele Modelldekodierung . . . . .	.193
12.4.1	Generierung von Segmenthypothesen . . . . .	.194
12.4.2	Sprachmodellbasierte Suche . . . . .	.195
<b>III</b>	<b>Systeme</b>	<b>197</b>
<b>13</b>	<b>Spracherkennung</b>	<b>200</b>
13.1	Erkennungssystem der RWTH Aachen . . . . .	.200
13.2	BBN-SpracherkennungssystemBYBLOS . . . . .	.202
13.3	ESMERALDA . . . . .	.203

<b>Inhalt</b>	<b>11</b>	
<b>14</b>	<b>Schrifterkennung</b>	<b>207</b>
14.1	OCR-System von BBN . . . . .	207
14.2	Duisburger on-line Handschrifterkennungssystem . . . . .	209
14.3	ESMERALDA off-line Erkennungssystem . . . . .	210
<b>15</b>	<b>Analyse biologischer Sequenzen</b>	<b>213</b>
<b>15.1</b>	<b>HMMER</b> . . . . .	<b>213</b>
<b>15.2</b>	<b>SAM</b> . . . . .	<b>214</b>
<b>Literaturverzeichnis</b>		<b>216</b>
<b>Sachverzeichnis</b>		<b>230</b>