PPM –
Prediction by Partial Matching

Inhalt

1. Kurzer Überblick
2. Arithmetische Codierung
3. PPM
4. Zusammenfassung & Ausblick

PPM: Steckbrief

- statistisches Verfahren

  (wie auch z.B. Arithmetische Codierung und

  Huffman-Codierung)

- entwickelt von Cleary und Witten (1984)
- verfeinert und implementiert von Moffat 1990

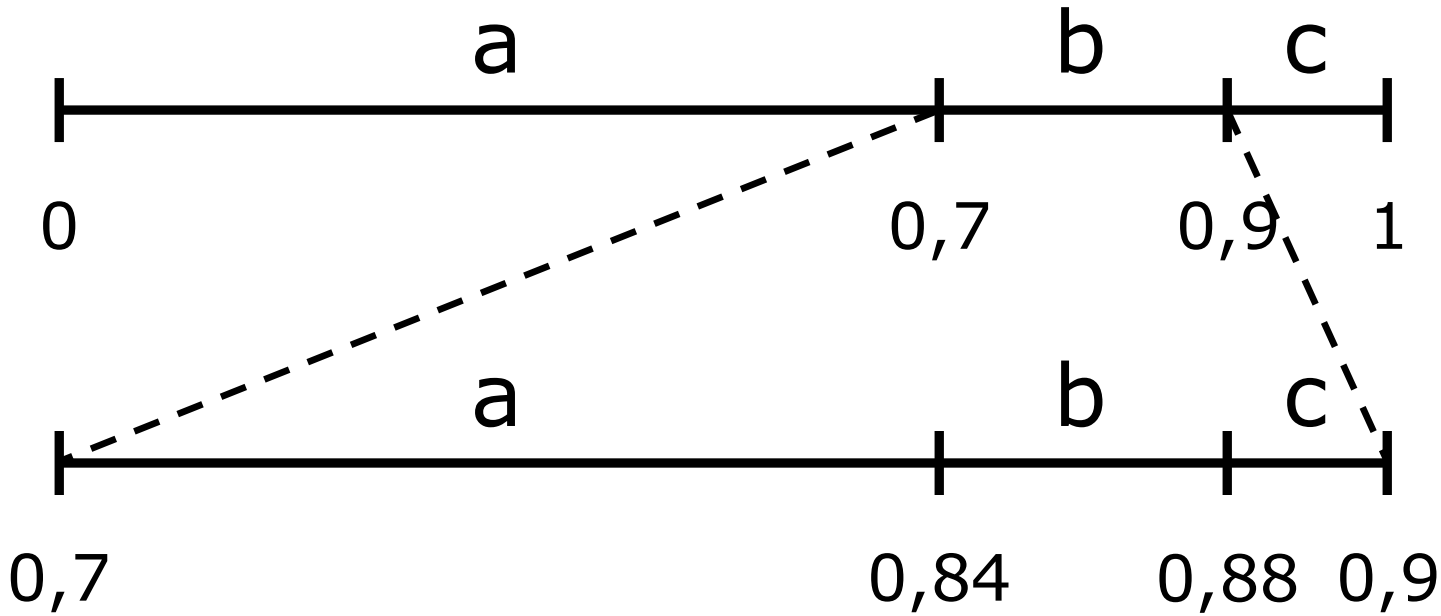
PPM + Arithmetische Codierung

- Arithmetische Codierung komprimiert Daten gemäß Wahrscheinlichkeiten
- Wahrscheinlichkeitsverteilung ist ein Modell
- PPM: eine Möglichkeit, Modell zu erzeugen

  → PPM allein ist *kein* Kompressionsverfahren

Wdh.: Arithmetische Codierung

P(a) = 0,7
P(b) = 0,2
P(c) = 0,1



Wdh.: Arithmetische Codierung

- Problem: Nullwahrscheinlichkeit

  → jedes Zeichen muss Wahrscheinlichkeit ≠ 0

  haben, sonst nicht codierbar

- AC ist asymptotisch optimal

  → Juniors Vortrag

Adaptive Arithmetische Codierung

- im Prinzip wie AC
- Modell (d.h. Wahrscheinlichkeiten) wird nach

  Codierung eines Zeichens angepasst

Adaptive Arithmetische Codierung (Bsp.)

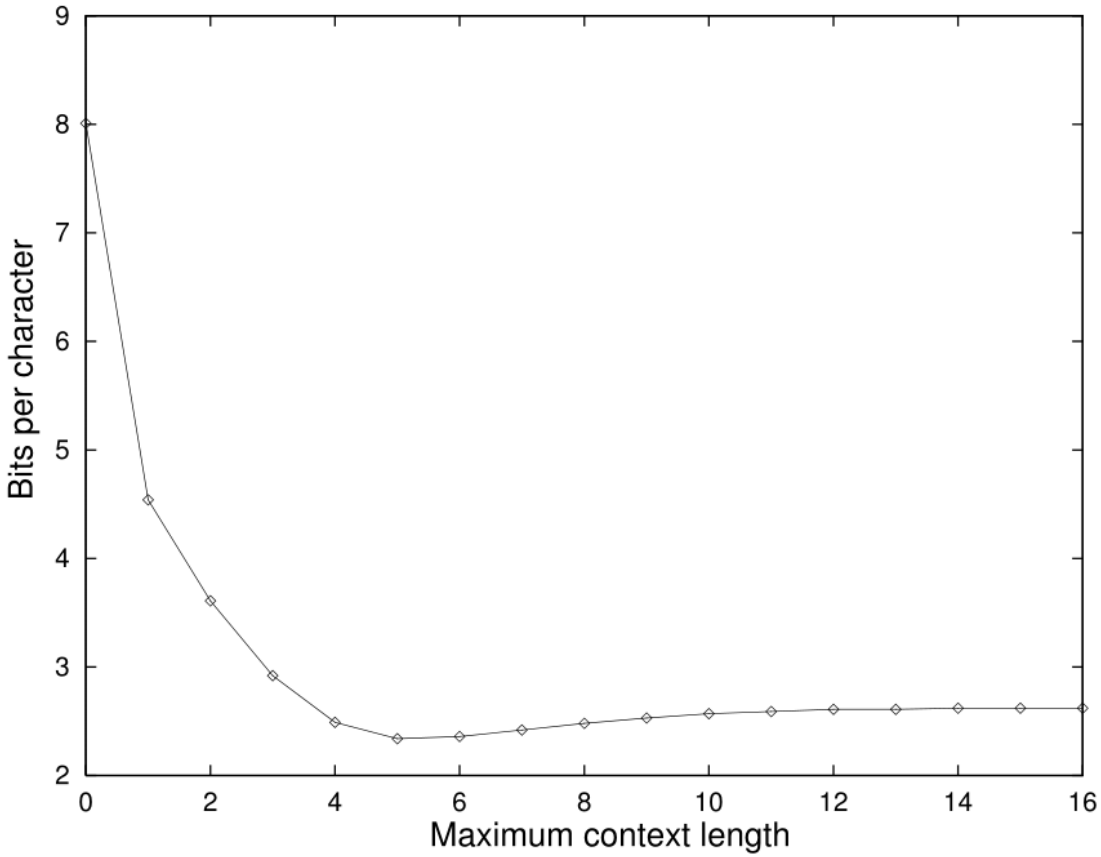
aabacba

Kontext

- Idee: betrachte n Zeichen vor aktuellem Zeichen
- Hoffnung: Zeichenwahrscheinlichkeit abhängig vom Kontext (z.B. ie, th)
- Vermutung:

  längerer Kontext → bessere Vorhersage
- Problem: längerer Kontext → Speicher, Laufzeit

Kontext



Beispiel für Zusammenhang zwischen Kontextlänge und Kompression (book1)
© Cleary, Teahan, Witten – Unbounded Length Contexts for PPM

Kontext-Wechsel

- Zeichen im gegenwärtigen Kontext unbekannt?

  → Wechsel auf kürzeren Kontext

- wiederholen bis Match, oder Kontext Länge 0

  → umgeht Problem der Nullwahrscheinlichkeit

Escape-Symbol

- wie erkennt Decoder Kontext-Wechsel?

  → Escape-Symbol

- auch nötig wegen Nullwahrscheinlichkeiten

Unbekannte Symbole

- Kontext der Länge 0, aber Symbol unbekannt?

  → Wechsel auf Kontext der Länge „-1“

- Kontext der Länge -1:

  → alle Symbole gleich wahrscheinlich

Unbekannte Symbole (Bsp.)

aababa

Verbesserung

- Wahrscheinlichkeit des nächsten Zeichens möglichst hoch → bessere Kompression
- Idee: bestimmte Möglichkeiten bei Wechsel auf kürzeren Kontext ausschließen

Ausschluss (Bsp.)

… abc …

2		1		0	-1
…		…		…	…
ab	a → 5/7	b	a → 9/17		
	b → 2/7		b → 3/17		
…			c → 5/17		
		…			

Zusammenfassung

- verwendete Kompression ist AC

  → asymptotisch optimal

- PPM-Modelle in der Praxis ziemlich gut

- benötigt viel Speicher + Laufzeit

Vergleich

	en-wik9*	Text	Bild	Audio
unkomprimiert	1'000'000'000	249'339	4'114'090	69'795'007
DURILCA	127'377'411	60'856 (-t2) 69'933 (-t0)	4'002'938	69'527'163
7zip	178'965'454	73'420 (PPMd) 88'747 (LZMA2) 88'741 (LZMA)	4'038'342 4'085'819 4'085'871	70'721'916 69'699'312 70'222'141
WinRAR	198'454'545	73'364	4'069'840	69'795'089
bzip2	253'977'839	80'302	4'094'989	69'965'765
gzip	322'591'995	99'763	4'111'149	69'656'728

* http://mattmahoney.net/dc/text.html

Ausblick

- unterschiedliche Modelle für Escape-Wahrscheinlichkeit (→ Poisson-Verteilung)
- Kontexte unendlicher Länge

  → deterministische Kontexte
- PAQ (nächste Woche)