

Modelarea Statistică Secvențială: Modele Markov Ascunse (HMM)

3.1. Introducere

Recunoașterea automată a vorbirii (ASR) implică modelarea unor secvențe temporale complexe, în care unitățile fundamentale ale limbajului – fonemele – sunt produse într-o succesiune guvernată atât de reguli fonetice, cât și de variabilitatea acustică a fiecărui vorbitor. Înainte de apariția modelelor neuronale profunde, cea mai importantă paradigmă statistică utilizată pentru rezolvarea acestei probleme a fost reprezentată de modelele Markov ascunse (HMM).

Acestea au dominat cercetarea și aplicațiile practice în ASR timp de mai multe decenii, datorită capacității lor de a modela formal procesele secvențiale, de a face inferențe eficiente și de a oferi o interpretare probabilistă clară. Chiar dacă în prezent HMM-urile sunt adesea combinate sau înlocuite de rețele neuronale, ele rămân esențiale pentru înțelegerea istorică și conceptuală a modului în care vorbirea poate fi reprezentată matematic și recunoscută automat.

3.2. Fonetica Computațională și Conceptul de Fonem

Unitatea de bază a vorbirii este **fonemul**, definit în lingvistică drept cea mai mică unitate sonoră care distinge sensuri în cadrul unei limbi. Din punct de vedere acustic, fonemele nu sunt constante, ci prezintă variații semnificative în funcție de contextul fonetic, accent și particularitățile fiecărui vorbitor. În recunoașterea automată, aceste variații trebuie surprinse printr-un model statistic flexibil, care să fie capabil să descrie atât natura secvențială a fonemelor, cât și variabilitatea observațiilor acustice.

HMM-urile oferă un cadru matematic potrivit pentru această sarcină. Ele fac ipoteza că secvența fonemelor poate fi reprezentată ca o succesiune de **stări ascunse**, care emit observații măsurabile, de regulă vectori de trăsături acustice precum coeficienții MFCC.

3.3. Structura unui Model Markov Ascuns

Un HMM este un model statistic care descrie un sistem ce evoluează printr-o serie de stări discrete, neobservabile direct (ascunse), dar care generează la fiecare pas o observație vizibilă (Rabiner, 1989). Formal, un HMM este definit de un set de parametri $\lambda = (A, B, \pi)$:

1. **Setul de stări ascunse**, $Q = \{q_1, q_2, \dots, q_N\}$: Un set finit de N stări. În ASR, acestea pot corespunde fonemelor, sub-fonemelor (început, mijloc, sfârșit de fonem) sau unităților dependente de context (trifoane).

2. **Matricea probabilităților de tranziție**, $A = \{a_{ij}\}$: $a_{ij} = P(s_{t-1} = q_i)$, unde a_{ij} este probabilitatea de a tranziționa de la starea q_i la starea q_j . Matricea A are dimensiunea $N \times N$ și respectă constrângerea $\sum_{j=1}^N a_{ij} = 1$ pentru fiecare i .

3. **Distribuția probabilităților de emisie**, $B = \{b_j(o_t)\}$: $b_j(o_t) = P(s_t = q_j)$, unde $b_j(o_t)$ este probabilitatea de a genera observația o_t (un vector de trăsături acustice) fiind în starea q_j . În ASR, unde observațiile sunt vectori continui (ex: MFCC), aceste distribuții sunt modelate, de obicei, prin Amestecuri de Distribuții Gaussiene (Gaussian Mixture Models - GMM).

4. **Distribuția probabilităților inițiale**, $\pi = \{\pi_i\}$: $\pi_i = P(s_1 = q_i)$, unde π_i este

probabilitatea ca secvența să înceapă în starea q_i .

Un HMM se bazează pe două ipoteze simplificatoare fundamentale:

- **Ipoteza Markov:** Probabilitatea unei tranziții la o anumită stare depinde doar de starea anterioară: $P(s_1, \dots, s_{t-1}) = P(s_{t-1})$.
- **Ipoteza de independență a ieșirilor:** Probabilitatea unei observații la un moment dat depinde doar de starea curentă la acel moment, nu de stările sau observațiile anterioare: $P(s_1, \dots, s_t, o_1, \dots, o_{t-1}) = P(s_t)$.

3.4. Cele Trei Probleme Fundamentale ale HMM-urilor

Utilizarea practică a HMM-urilor, așa cum a fost formalizată în tutorialul de referință al lui Rabiner (1989), presupune rezolvarea a trei probleme esențiale.

3.4.1. Problema 1: Evaluarea (Calculul Verosimilității)

Problema: Dată o secvență de observații $O = (o_1, o_2, \dots, o_T)$ și un model HMM $\lambda = (A, B, \pi)$, cum calculăm eficient probabilitatea $P(\lambda)$?

Această problemă este crucială în recunoașterea cuvintelor izolate, unde fiecare cuvânt este modelat de un HMM distinct. Pentru a recunoaște un enunț, calculăm probabilitatea ca secvența de observații să fi fost generată de fiecare model de cuvânt și alegem cuvântul corespunzător modelului cu cea mai mare probabilitate.

O abordare naivă ar fi să însumăm probabilitățile tuturor secvențelor de stări posibile $Q=(s_1,s_2,\dots,s_T)$:

$$P(\lambda) = \sum_{\text{toate } Q} P(\lambda) = \sum_{\text{toate } Q} P(O|Q, \lambda) P(Q|\lambda)$$

Complexitatea acestei abordări este de ordinul $O(N^T \cdot T)$, ceea ce este computațional prohibitiv.

Soluția: Algoritmul Forward

Algoritmul Forward este o procedură de programare dinamică care calculează $P(O | \lambda)$ cu o complexitate de $O(N^2 \cdot T)$. Acesta se bazează pe definirea variabilei forward, $\alpha_{t(i)}$, ca fiind probabilitatea de a observa secvența parțială $o_1, \dots, o_t$ și de a fi în starea q_i la momentul t .

Algoritmul constă în trei pași:

Inițializare ($t=1$):

$$\alpha_1(i) = \pi_i b_i(o_1), \text{ pentru } 1 \leq i \leq N.$$

Recursie (pentru $t = 1, \dots, T - 1$):

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1}), \text{ pentru } 1 \leq j \leq N.$$

Terminare:

$$P(\lambda) = \sum_{i=1}^N \alpha_T(i).$$

3.4.2. Problema 2: Decodarea

Problema: Dată o secvență de observații O și un model HMM λ , care este cea mai probabilă secvență de stări ascunse $Q = (s_1, \dots, s_T)$ care a generat observațiile?

Aceasta este problema centrală în ASR, deoarece secvența de stări ascunse corespunde secvenței de foneme recunoscute.

Soluția: Algoritmul Viterbi

Algoritmul Viterbi este un alt algoritm de programare dinamică, foarte similar cu algoritmul Forward, dar care utilizează operația de maximizare în loc de însumare pentru a găsi cea mai probabilă cale unică prin trellis-ul de stări (Wikipedia – Viterbi, 2025).

Algoritmul definește variabila $\delta_t(i)$ ca fiind cea mai mare probabilitate de-a lungul unei singure căi care se termină în starea q_i la momentul t . Pentru a putea reconstrui calea, se stochează și un pointer $\psi_t(j)$ care reține starea de la momentul $t - 1$ ce a condus la maximul pentru starea q_j la momentul t .

1. Inițializare (t=1):

$$\begin{aligned}\delta_1(i) &= \pi_i b_i(o_1) \\ \psi_1(i) &= 0\end{aligned}$$

2. Recursie (pentru $t = 2, \dots, T$):

$$\delta_t(j) = [\delta_{t-1}(i) a_{ij}] b_j(o_t) \quad \psi_t(j) = [\delta_{t-1}(i) a_{ij}]$$

3. Terminare și Backtracking:

$$P^* = s_T^* = \operatorname{argmax}_{1 \leq i \leq N}$$

Apoi, se reconstruiește calea optimă mergând înapoi de la $t=T$ la 1:

$$s_t^* = \psi_{t+1}(s_{t+1}^*)$$

3.4.3. Problema 3: Învățarea

Problema: Cum ajustăm parametrii modelului $\lambda = (A, B, \pi)$ pentru a maximiza $P(\lambda)$, având la dispoziție un set de secvențe de observații de antrenament?

Aceasta este problema de antrenare a modelului. Nu există o soluție analitică pentru a găsi global maximul lui $P(\lambda)$ în funcție de λ .

Soluția: Algoritmul Baum-Welch (Forward-Backward)

Algoritmul Baum-Welch este o instanță a algoritmului Expectation-Maximization (EM) și este utilizat pentru a găsi o soluție local optimă pentru parametrii HMM (Kwok, 2019). Acesta funcționează iterativ:

1. **Pasul E (Expectation):** Se calculează probabilitățile forward, $\alpha_t(i)$, și backward, $\beta_t(i)$. Variabila backward $\beta_t(i) = P(s_t = q_i, \lambda)$ este probabilitatea de a observa restul secvenței, dat fiind că ne aflăm în starea q_i la momentul t . Se calculează recursiv, pornind de la sfârșitul secvenței.

2. **Pasul M (Maximization):** Se re-estimează parametrii modelului pe baza așteptărilor calculate la pasul E. De exemplu, noua probabilitate de tranziție a_{ij} este raportul dintre numărul așteptat de tranziții de la starea q_i la q_j și numărul total așteptat de tranziții din starea q_i .

Acești pași se repetă până la convergență.

3.5. Limitări ale HMM-urilor și Tranziția către Modele Hibride

În ciuda succesului lor istoric, HMM-urile au limitări importante:

- **Ipoteze simplificatoare:** Ipoteza Markov de ordinul întâi și ipoteza de independență a ieșirilor sunt prea restrictive pentru a modela complexitatea reală a vorbirii, unde există dependențe pe termen lung.

- **Modelarea emisiei:** GMM-urile, deși flexibile, nu sunt modele optime pentru distribuțiile acustice, care sunt adesea neliniare și corelate.

Aceste limitări au condus la dezvoltarea modelelor hibride DNN-HMM, unde GMM-urile sunt înlocuite de rețele neuronale profunde, capabile să învețe reprezentări mai bune ale datelor acustice. Această tranziție va fi subiectul detaliat al capitolului următor.

În ciuda succesului lor, HMM-urile prezintă și limite importante. Ele se bazează pe ipoteza de independență condițională între observații, ceea ce simplifică prea mult realitatea vorbirii, unde dependențele de lungă durată sunt semnificative. În plus, utilizarea distribuțiilor Gaussiene poate fi restrictivă în fața variabilității acustice complexe. Aceste limitări au pregătit terenul pentru adoptarea rețelelor neuronale, capabile să surprindă relații mai complexe și să modeleze dependențele de lungă durată.

Cu toate acestea, chiar și în arhitecturile moderne, idei fundamentale din HMM-uri, precum structura secvențială și conceptul de decodare optimă, continuă să joace un rol esențial.

3.6. Concluzii

HMM-urile au reprezentat timp de decenii paradigma dominantă în recunoașterea automată a vorbirii. Prin structura lor simplă, dar puternică, ele au permis modelarea secvențelor de foneme, au oferit soluții eficiente pentru evaluare, decodare și învățare și au pus bazele sistemelor ASR moderne. Înțelegerea acestor modele rămâne indispensabilă pentru oricine dorește să aprofundeze domeniul procesării vorbirii, chiar dacă astăzi acestea sunt integrate sau înlocuite de rețele neuronale profunde.

