Stabilitet i topologisk dataanalyse

Andreas M. Kristensen

February 21, 2025

Contents

1 Introduksjon

Målet med denne oppgaven er å gi en innledning til topologisk data analyse og stabilitetsteoremet.

Først går oppgaven innom litt førkunnskaper om topologiske rom, grunnlegende homologi, vektorrom og kategoriteori. Disse emnene vil være essensielle for å kunne studere data på en topologisk måte.

Senere går vi igjennom persistensmoduler og barkoder. Det er disse som kommer til å være hovedfokuset av oppgaven. Vi definerer pseudometrikker på barkoder og persistensmoduler som forteller oss hvor like barkodene og persistensmodulene er.

Disse metrikkene vil senere lede til målet med selve oppgaven som er stabilitetsteoremet.

2 Forkunnskaper

Her går vi igjennom noen nødvendige definisjoner

2.1 Topologiske rom

Definisjon 2.1.1. *Et par $(X, \mathcal{T})$ hvor X er en mengde og $\mathcal{T} \subset \mathcal{P}(X)$ slik at*

- $X, \emptyset \in \mathcal{T}$
- *Gitt en vilkårelig samling av mengder $\{U_\alpha\}_\alpha$ så er $\bigcup_\alpha U_\alpha \in \mathcal{T}$*
- *For en endelig samling av mengder $\{U_1, \ldots, U_n\} \in \mathcal{T}$ så er snittet $U_1 \cap \cdots \cap U_n \in \mathcal{T}$*

Vi kaller mengden $\mathcal{T}$ for topologien på X og mengdene i $\mathcal{T}$ for åpne mengder.

Når topologien $\mathcal{T}$ på en mengde X er kjent eller ikke viktig lar vi være å skrive det topologiske rommet som et par $(X, \mathcal{T})$ og skriver bare X. Alle funksjoner mellom topologiske rom vil være kontinuerlige. Til slutt så vil et "rom" bety et topologisk rom.

Eksempel 2.1.1. *Euklidisk rom $(\mathbb{R}^n, \mathcal{T})$ er et topologisk rom med åpne mengder unioner av vilkårlig mange mengder av typen*

$$\mathcal{B}(x, \delta) = \{y \in \mathbb{R}^n \mid \|x - y\| < \delta\}$$

kalt åpne baller. Euklidisk rom er som regel alltid bare skrevet $\mathbb{R}^n$ siden det er den topologien på $\mathbb{R}^n$ som er antatt.

Definisjon 2.1.2. *La $(X, \mathcal{T}_X)$ og $(Y, \mathcal{T}_Y)$ være topologiske rom. En funksjon $f : X \to Y$ er kalt kontinuerlig hvis for en hver $V \in \mathcal{T}_Y$ så er $f^{-1}(V) \in \mathcal{T}_X$.*

Eksempel 2.1.2. *La $f : \mathbb{R} \to \mathbb{R}$ være funksjonen $f(x) = 2x + 1$, da er f kontinuerlig. La $V = (a, b)$ et opent intervall, da blir $f^{-1}(V) = \{x \in \mathbb{R} \mid 2x+1 \in V\} = \left(\frac{a}{2} - 1, \frac{b}{2} - 1\right)$ som også er et åpent intervall.*

Definisjon 2.1.3. *La $(X, \mathcal{T})$ være et topologisk rom og la $A \subset X$ da er det en naturlig topologi $\mathcal{T}_A$ vi kan sette på A definert ved*

$$U \in \mathcal{T}_A \iff \exists\, V \in \mathcal{T} \quad s.a \quad V \cap A = U$$

Vi kaller $(A, \mathcal{T}_A)$ et underrom av $(X, \mathcal{T})$ og vi kaller $\mathcal{T}_A$ underromstopologien på A.

Når vi senere ser på punktskyer i $\mathbb{R}^n$ og spesifkt simplisial kompleksene

2.2 Homotopi

Definisjon 2.2.1. *La X og Y være topologiske rom og la $f, g : X \to Y$ være funksjoner. En homotopi mellom f og g er en funksjon*

$$F : X \times [0,1] \to Y.$$

Slik at $F(x,0) = f(x)$ og $F(x,1) = g(x)$. Hvis det eksisterer en homotopi mellom en funksjon f og g sier vi at de er homotope og vi skriver at $f \simeq g$.

Definisjon 2.2.2. *To topologiske rom X og Y er homotopiekvivalente hvis det eksisterer funksjoner $f : X \to Y$ og $g : Y \to X$ slik at*

$$g \circ f \simeq \mathrm{id}_X \quad f \circ g \simeq \mathrm{id}_Y$$

2.3 Δ-Komplekser

En måte å lage topologiske rom er å starte med enkle byggeklosser og lime dem sammen. Dette kan vi gjøre ved å bruke n-dimensjonale trekanter kalt n-simplekser.

definisjonene er fra Hatcher [2002]

Definisjon 2.3.1. *Standardsimplekset Δ^n er definert ved*

$$\Delta^n = \{\boldsymbol{x} \in \mathbb{R}^{n+1} \mid \sum_{i=0}^{n+1} x_i = 1, x_i \geq 0 \; \forall i = 0, \ldots, n\}$$

En side på et n-simpleks er en $(n-1)$-simpleks.

Definisjon 2.3.2. *En Δ-kompleks struktur på et rom X er en samling av kontinuerlige funksjoner $\sigma_\alpha : \Delta^n \to X$ med n avhengig av α slik at:*

1. *Restriksjonen $\sigma_\alpha \mid \mathring{\Delta}^n$ er injektiv og hvert punkt i X er i bildet av nøyaktig en slik Restriksjon.*
2. *Hver restriksjon av σ_α til en side av Δ^n er en av funksjonene $\sigma_\beta : \Delta^{n-1} \to X$*
3. *En mengde $A \subset X$ er åpen hvis og bare hvis $\sigma_\alpha^{-1}(A)$ er åpen for hver σ_α*

2.4 Vektorrom

For å definere hva et vektorrom er må vi først gå igjennom hva en kropp er.

Definisjon 2.4.1. *En mengde K sammen med binære operatorer $+, \cdot : K \times K \to K$ er en kropp hvis gitt $a, b, c \in K$ så holder det følgende*

- $(a + b) + c = a + (b + c)$
- $a + b = b + a$

- $a \cdot (b + c) = a \cdot b + a \cdot c$
- *Det eksisterer et element* $0 \in K$ *slik at* $a + 0 = a = 0 + a$
- *Det eksisterer et element* $1 \in K$ *slik at* $a \cdot 1 = a = 1 \cdot a$
- *Det eksisterer et element* $-a \in K$ *slik at* $a + (-a) = 0$
- *Det eksisterer et element* $a^{-1} \in K$ *slik at* $a \cdot a^{-1} = 1$.

Ofte lar vi være å skrive $a + (-b)$ og skriver heller $a - b$ vi lar også være å skrive $a \cdot b$ og skriver heller ab

Definisjon 2.4.2. *Et vektorrom* V *over en kropp* K *er en mengde med binære operatorer* $+ : V \times V \to V$ *og* $\cdot : K \times V \to V$*, kalt skalarmultiplikasjon, slik at for elementer* $u, v, w \in V$ *og* $a, b, c \in K$ *så holder det følgende*

- $(u + v) + w = u + (v + w)$
- $u + v = v + u$
- *Det eksisterer et element* $0 \in V$ *slik at* $u + 0 = u = 0 + u$
- *Det eksisterer et element* $-u \in V$ *slik at* $u + (-u) = 0$
- $(a + b) \cdot u = a \cdot u + b \cdot u$
- $a \cdot (u + v) = a \cdot u + a \cdot v$
- $a \cdot (b \cdot u) = (ab) \cdot u$.

Vi kaller elementer $v \in V$ *for vektorer og elementer* $a \in K$ *for skalarer.*

Igjen skriver vi ofte $v + (-u)$ som $v - u$ og $a \cdot v$ som av.

Definisjon 2.4.3. *La* V *og* W *være vektorrom over en kropp* K *og la* $f : V \to W$ *være en funksjon. Vi kaller* f *lineær hvis for vektorer* $u, v \in V$ *og en skalar* $a \in K$ *så holder det følgende*

- $f(u + v) = f(u) + f(v)$
- $f(av) af(v)$

Vi kaller også slike funksjoner lineære avbildinger/transformasjoner/funksjoner

Eksempel 2.4.1. *Rommet* $V = \mathbb{R}^n$ *over kroppen* $\mathbb{R}$ *er et vektorrom med punktvis addisjon, og skalarmultiplikasjon*

$$(a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n)$$

og

$$c(a_1, \ldots, a_n) = (ca_1, \ldots, ca_n).$$

Eksempel 2.4.2. *funksjonen* $f : \mathbb{R}^2 \to \mathbb{R}^2$ *definert ved* $f(v) = 2v$ *er lineær siden gitt* $u, v \in V$ *og* $a \in \mathbb{R}$ *så er*

- $f(u + v) = 2(u + v) = 2u + 2v = f(u) + f(v)$
- $f(av) = 2(av) = (2a)v = (a \cdot 2)v = af(v)$

2.5 Homologi

Det er mange spørsmål om topologiske rom som er vanskelige å svare på om man ikke har de rette verktøyene.

Et eksempel på et teorem som er vanskelig å bevise rent topologisk er Borsuk-Ulam teoremet

Teorem 2.5.1. *La $f : S^n \to \mathbb{R}^n$ være en kontinuerlig funksjon, da eksisterer det et punkt $x \in S^n$ slik at $f(x) = f(-x)$.*

Vi kan studere topologiske rom ved bruk av algebra. Dette kan vi gjøre med homologigruppene

Definisjon 2.5.1. *La X være et simplisialkompleks og la*

$$C_n(X) = \langle [v_0, \ldots, v_n] \in X \rangle_k$$

3 Persistensmoduler og Barkoder

3.1 Persistensmoduler

Et sentralt tema for å kunne forstå stabilitet og topologisk dataanalyse er ideen om persistensmoduler. I dette kapitellet går vi gjennom en litt abstrakt introduksjon og så ser vi på hvorfor de er viktige innenfor topologisk dataanalyse. Definisjonen på en Persistensmodul er kort og enkel.

Definisjon 3.1.1. *En persistensmodul M er en funktor $M : \boldsymbol{R} \to \mathbf{vect}_k$.*

Vi skriver M_t for vektorromet $M(t)$ (det t-ende vektorromet) for å unngå fremtidig forvirring. Siden en persistensmodul M er en funktor fra pomengden $\mathbf{R}$ til $\mathbf{vect}_k$ så har vi for hver $s \leq t$ en lineær avbilding $\varphi_M(s,t) : M_s \to M_t$ som vi kaller overgangsavbildinger.

Gitt to persistensmoduler M og N kan vi definere en morfi $f : M \to N$ som en samling av lineære avbildinger $\{f(s) : M_s \to N_s \mid s \in \mathbb{R}\}$ slik at diagrammet

$$\begin{array}{ccc} M_s & \xrightarrow{f(s)} & N_s \\ \downarrow{\scriptstyle \varphi_M(s,t)} & & \downarrow{\scriptstyle \varphi_N(s,t)} \\ M_t & \xrightarrow{f(t)} & N_t \end{array}$$

kommuterer. Vi kan komponere morfiene på den åpenbare måten; gitt morfier $f : M \to N$ og $g : M \to P$ er $g \circ f$ definert som samlingen $\{g_s \circ f_s : M_s \to P_s \mid s \in \mathbb{R}\}$.

Siden vi har objekter, persistensmoduler, og vi har morfier mellom dem kan vi definere kategorien av persistens moduler

Definisjon 3.1.2. *Kategorien $\mathbf{vect}_k^{\boldsymbol{R}}$ er kategorien av persistensmodulene med persistensmodul-morfier mellom dem.*

Eksempel 3.1.1. *Gitt en filtrering $F_\bullet X = \{F_t X\}_{t\in\mathbb{R}}$ av et topologisk rom X kan vi definere persistensmodulen $H_n(F_\bullet X) = \{H_n(F_t X; k)\}_{t\in\mathbb{R}}$ med overgangsavbildinger $\varphi_{H_n(F_\bullet X)}(s,t) = i_*(s,t)$*

Et eksempel på en særlig enkel, men viktig persistensmodul er intervall-persistensmodulen definert som følgende.

La $I \subset \mathbb{R}$ være et intervall da definerer vi intervall-persistensmodulen $C(I)$ som følger:

$$C(I)_t = \begin{cases} k, & t \in I \\ 0, & \text{ellers} \end{cases}$$

med overgangsavbildinger definert ved

$$\varphi_{C(I)}(s,t) = \begin{cases} id_k, & s,t \in I \\ 0, & \text{ellers} \end{cases}$$

Denne persistensmodulen er nyttig når vi skal definere barkoder snart.

3.1.1 Interleaving-distanse

Stabilitet av persistensmoduler innebærer relasjonen mellom to typer distanser, Bottleneck distansen mellom barkoder og Interleaving distansen mellom persistensmoduler. Her definerer vi interleaving distansen mellom to Persistensmoduler.

For å definere distansen må vi gjennom noen få steg.

Definisjon 3.1.3. *En δ-forskyvning av en persistensmodul er en funktor*

$$(\cdot)(\delta) : \mathbf{vect}_k^{\boldsymbol{R}} \to \mathbf{vect}_k^{\boldsymbol{R}}$$

Som tar en persistensmodul M til $M(\delta)$ hvor $M(\delta)_t = M_{t+\delta}$ og tar persistensmodulmorfier $f : M \to N$ til $f(\delta) : M(\delta) \to N(\delta)$.

Denne funktorer gir oss konseptet av δ-interleavinger.

Definisjon 3.1.4. *La M og N være persistensmoduler. Vi sier at M og N er δ-interleavet hvis det eksisterer persistensmodulmorfier $f : M \to N(\delta)$ og $g : N \to M(\delta)$ slik at*

$$g(\delta) \circ f = \varphi_M(t, t+2\delta), \quad f(\delta) \circ g = \varphi_N(t, t+2\delta)$$

Vi kaller $\varphi_M^{\varepsilon}(t) = \varphi_M(t, t+\varepsilon)$. Bemerk at $\varphi_M^0 = \mathrm{id}_M$ fordi $\varphi_M^0(t) = \varphi_M(t, t+0) = \varphi_M(t,t) = \mathrm{id}_M$.

Definisjon 3.1.5. *For M og N persistensmoduler definerer vi interleaving-distansen d_I ved*

$$d_I(M,N) = \inf\{\delta \in [0,\infty) \mid M \text{ og } N \text{ er } \delta\text{-interleavet}\}$$

Denne avstanden gir et tall på hvor "isomorfe" to persistensmoduler er.

Proposisjon 3.1.1. *For M og N persistensmoduler så holder*

$$d_I(M,N) = 0 \iff M \cong N$$

Proof. " $\implies$ "

Hvis $d_I(M,N) = 0$ så finnes det en 0-interleaving mellom M og N altså det eksisterer persistensmodulmorfier $f : M \to N(0) = N$ og $g : N \to M(0) = M$ slik at $g(0) \circ f = g \circ f = \varphi_M^0 = \mathrm{id}_M$ og $f(0) \circ g = \varphi_N^0 = \mathrm{id}_N$. Dermed er f og g inverser av hverandre og er dermed isomorfier.

" $\Longleftarrow$ "

Hvis $M \cong N$ så eksisterer det persistensmodulmorfier $f : M \to N$ og $g : N \to M$ slik at $g \circ f = \mathrm{id}_M = \varphi_M^0$ og $f \circ g = \mathrm{id}_N = \varphi_N^0$. Så det eksisterer en 0-interleaving og dermed er $d_I(M, N) = 0$. □

3.2 Multimengder

Mengder er begrenset i og med at de ikke inneholder repitisjoner, mengden $\{a, a, b\}$ er regnet som mengden $\{a, b\}$. For oss vil vi ha muligheten for at en mengde kan inneholde mange like elementer.

Dermed definerer vi en multimengde.

Definisjon 3.2.1. *Vi definerer en multimengde som et par $\mathcal{S} = (S, m)$, hvor S er en mengde og en funksjon $m : S \to \mathbb{N}$.*

Multimengder er derimot vanskelige å jobbe med, derfor jobber vi med deres representasjoner

$$\mathrm{Rep}(\mathcal{S}) = \{(s, k) \in S \times \mathbb{N} \mid k \leq m(s)\}.$$

3.3 Barkoder

En barkode $\mathcal{B}$ er en representasjon av en multimengde av intervaller. Elementer i en barkode er dermed par (I, k) der I er et intervall og $k \in \mathbb{N}$. Ofte når indeksen k er nødvendig skriver vi bare I for et intervall i barkoden.

I Bauer and Lesnick [2015] sier forfatte at gitt en persistensmodul M som kan skrives

$$M \cong \bigoplus_{I \in \mathcal{B}_M} C(I)$$

Da er $\mathcal{B}_M$ unikt bestemt.

Dette er en konsekvens av følgende teorem

Teorem 3.3.1. *Hvis $\bigoplus_{I \in \mathcal{B}} C(I) \cong \bigoplus_{J \in \mathcal{C}} C(J)$ så er $\mathcal{B} \cong \mathcal{C}$*

Proof. Anta at det ikke eksisterer en bijeksjon mellom $\mathcal{B}$ og $\mathcal{C}$ f.eks. □

4 Persistent Homologi

En grunn til å bry seg om persistensmoduler er fordi de er en generalisering av homologien av en filtrering av et topologisk rom.

Definisjon 4.0.1. *La X være et topologisk rom da er en filtrering på X en følge $F_\bullet X = \{F_t X\}_{t\in\mathbb{R}}$ slik at hvis $s \leq t$ så er $F_s X \subset F_t X$ og $F_\infty X = X$.*

Siden det er en naturlig inklusjon $i_{F_\bullet X}(s,t) : F_s X \hookrightarrow F_t X$ når $s \leq t$ kan vi se på en filtrering som en funktor $\mathbf{R} \to \mathbf{Top^{ink}}$ hvor $\mathbf{Top^{ink}}$ er kategorien av topologiske rom hvor morfiene er inklusjoner. Vi kan også definere morfier mellom filtreringer:

La $F_\bullet X$ og $G_\bullet X$ være filtreringer av X da er en morfi $f : F_\bullet X \to G_\bullet X$ en følge $\{f(t) : F_t X \to G_t X\}$ slik at diagrammet

$$\begin{array}{ccc} F_s X & \xrightarrow{f(s)} & G_s X \\ \downarrow{\scriptstyle i_{F_\bullet X}(s,t)} & & \downarrow{\scriptstyle i_{G_\bullet X}(s,t)} \\ F_t X & \xrightarrow{f(t)} & G_t X \end{array}$$

kommuterer. Da er filtreringer av et rom en kategori som vi kaller $\mathbf{Filt}(X)$.

Akkurat som topologiske rom kan vi ta homologien på filtreringer ved komposisjonen $\mathbf{R} \xrightarrow{F_\bullet X} \mathbf{Top^{ink}} \xrightarrow{H_i} \mathbf{vect}$. Eksplisitt blir dette følgende:

Gitt et rom X la $F_\bullet X$ være en filtrering. Da er $H_i(F_\bullet X)$ en persistensmodul definer ved

$$H_i(F_\bullet X)_t = H_i(F_t X)$$

med overgangsavbildinger

$$\varphi_{H_i(F_\bullet X)}(s,t) = (i_{F_\bullet X}(s,t))_*$$

Dette er metoden man bruker i topologisk dataanalyse for å studere "formen" på en punktsky av data, noe vi kommer tilbake til i anvendelsene. Før vi kan diskutere slike anvendelser må vi først vite hvordan vi lager topologiske rom ved en punktskyer.

4.1 Topolgiske rom fra punktskyer

En punktsky $P \subset \mathbb{R}^d$ er en diskret mengde av punkter i $\mathbb{R}^d$. Dette kan være data om farger eller gråtoner på bilder, nerver i en hjerne osv.

Det er ikke mye topologisk informasjon vi kan få ut a skyen i seg selv gitt at den er en diskret mengde, men vi kan lage simplisialkomplekser av skyen. Dette kan gjøres på mange måter, men det er to hovedmetoder å gjøre dette på.

4.1.1 Cech-komplekser

En måte å lage simplisialkomplekser av en punktsky er ved å lage en k-simpleks mellom $k+1$ punkter hvis snittet av ε-ballene i punktene snitter hverandre.

Definisjon 4.1.1. *La $P \subset \mathbb{R}^d$ være en punktsky vi definerer Cech-komplekset ved*

$$\mathcal{C}_\varepsilon(P) = \left\{ (x_i)_i \mid \bigcap_i \bar{\mathcal{B}}(x_i, \varepsilon) \neq \emptyset \right\}.$$

Problemet med dette komplekset er at en må telle hvor mange sirkler som snitter hverandre.

Eksempel 4.1.1. *La $P = \{(-1,1),(1,1),(1,-1),(-1,-1)\}$ for forskjellige verdier av ε får vi forskjellige simplisialkomplekser gitt her, for å forkorte mengden av simplisialkompleksene skriver vi $v_0 = (-1,-1), v_1 = (1,-1), v_2 = (1,1), v_3 = (-1,1)$ i stedet for punktene*

- *Når $0 \leq \varepsilon < \frac{1}{2}$ så er*

$$\mathcal{C}_\varepsilon(P) = \{[v_0],[v_1],[v_2],[v_3]\}.$$

- *Når $\frac{1}{2} \leq \varepsilon < \sqrt{2}$ så er*

$$\mathcal{C}_\varepsilon(P) = \{[v_0],[v_1],[v_2],[v_3],[v_0,v_1],[v_0,v_3],[v_2,v_3],[v_1,v_2]\}.$$

- *Når $\varepsilon \geq \sqrt{2}$ så er*

$$\mathcal{C}_\varepsilon(P) = \{[v_0],[v_1],[v_2],[v_3],[v_0,v_1],[v_0,v_2],[v_0,v_3],[v_1,v_2],[v_1,v_3],[v_2,v_3],[v_0,v_1,v_2],[v_0,v_1,v_3],[v_0,v_2,v_3]$$

4.2 Rips-komplekser

En annen måte å få et simplisialkompleks av en punktsky er å lage et k-simpleks mellom $k+1$ punkt hvis de er ε nærme hverandre.

Definisjon 4.2.1. *La $P \subset \mathbb{R}^d$ være en punktsky, da er Rips-komplekset definert ved*

$$\mathcal{R}_\varepsilon(P) = \{(x_i)_i \mid |x_i - x_j| \leq \varepsilon\}$$

Cech- og Rips-kompleksene er begge filtreringer av simplisialkomplekset der hver kombinasjon av punkter i punktskyen har en simpleks. Dermed er de også filtreringer av topologiske rom. Dette gir oss persistensmodulene

$$H_i(\mathcal{C}_\bullet(P)) = \{H_i(\mathcal{C}_\varepsilon(P))\}_{\varepsilon \in \mathbb{R}}$$

og

$$H_i(\mathcal{R}_\bullet(P)) = \{H_i(\mathcal{R}_\varepsilon(P))\}_{\varepsilon \in \mathbb{R}}.$$

Det er disse man bruker når man studerer de topologiske egenskapene til data.

5 Algebraisk Stabilitet

References

Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-79160-X; 0-521-79540-0.

Ulrich Bauer and Michael Lesnick. Induced Matchings and the Algebraic Stability of Persistence Barcodes. *Journal of Computational Geometry*, pages 162–191 Pages, March 2015. doi:10.20382/jocg.v6i2a9.

[Referanser]