Stavový model a Kalmanův filtr

20. prosince 2013

Stav je veličina, kterou neznáme, ale chtěli bychom znát. Dozvídáme se o ní zprostředkovaně prostřednictvím výstupů.

Příkladem může být např. nějaký zašuměný signál, který měříme. Stavem pak je onen signál nezašuměný, který se snažíme odhadnout.

Jiným příkladem může být zdravotní stav pacienta, který odhadujeme na základě výstupů různých vyšetření.

Ještě dalším příkladem může být stav křižovatky, který odhadujeme na základě výstupů několika čidel.

Pokud proces popisujeme lineárním modelem s normálním šumem, můžeme tento model převést na vektorový stavový model prvního řádu. To je velmi důležité, protože na stavový model prvního řádu s normálním šumem mohu uplatnit velmi účinný algoritnus, který dělá bodové odhady stavu a který se jmenuje Kalmanův filtr.

Převod lineárního regresního modelu do stavového tvaru

Popis stavového modelu

Stavový model se skládá ze dvou částí. Za prvé z modelu dynamiky stavu, který říká, jak závisí stav v příštím okamžiku na stavu v okamžiku minulém. Tento lineární model prvního řádu vypadá takto:

$$x_n = M \cdot x_{n-1} + N \cdot u_n + w_n.$$

Veličina x značí stav, u je řízení a w je normální bílý šum. Koeficienty M a N jsou čísla respektive matice, pokud jsou ostatní veličiny vektory.

Druhou součástí stavového modelu je model měření výstupu, který říká, jak závisí aktuální výstup na aktuálním stavu:

$$y_n = A \cdot x_n + B \cdot u_n + v_n.$$

Koeficienty A a B jsou čísla respektive matice, pokud jsou ostatní veličiny vektory. Veličina v je normální bílý šum.

Převod na stavový model

Nyní si na příkladu ukážeme, jak převést lineární regresní model libovolného řádu na stavový model prvního řádu:

Máme lineární regresní model druhého řádu s dvěmi řízeními s a t:

$$y_n = \alpha s_n + \beta t_n + C y_{n-1} + D s_{n-1} + E t_{n-1} + F y_{n-2} + G s_{n-2} + H t_{n-2} + K + e_n.$$

Bílý šum má normální rozdělení s rozptylem R: $e_n \sim N(0, R)$.

Vezmeme vektor vstupů a oddělíme část, která patří aktuálním veličinám. Tím dostaneme stavový vektor pro minulý čas. Ve sloupečkové podobě:

$$x_{n-1} = \begin{pmatrix} y_{n-1} \\ s_{n-1} \\ t_{n-1} \\ y_{n-2} \\ s_{n-2} \\ t_{n-2} \\ 1 \end{pmatrix}.$$

Zvýšením indexů o jedna získáme stavový vektor pro aktuální čas:

$$x_n = \begin{pmatrix} y_n \\ s_n \\ t_n \\ y_{n-1} \\ s_{n-1} \\ t_{n-1} \\ 1 \end{pmatrix}.$$

Vztah mezi x_n a x_{n-1} určuje čtvercová matice M, kterou vytvoříme takto:

V prvním řádku budou koeficienty příslušné veličinám z x_{n-1}. Další řádky jen zajišťují, aby se příslušné veličiny převedly. Tedy y_{n-1} na y_{n-1}, t_{n-1} na t_{n-1}, 1 na 1, apod.:

$$M = \begin{pmatrix} C & D & E & F & G & H & K \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Vektor aktuálního řízení vyrobíme také sloupečkový:

$$u_n = \begin{pmatrix} s_n \\ t_n \end{pmatrix}.$$

Matice N tedy bude mít dva sloupce. Dosadíme příslušné konstanty tak, abychom na prvním řádku dostali původní rovnici pro výstup (zatím bez šumu)

a aby se příslušná řízení objevila na příslušných místech stavového vektoru, tedy na místě druhém a třetím:

$$N = \begin{pmatrix} \alpha & \beta \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Šum umístíme na první místo vektoru šumu w_n:

$$w_n = \begin{pmatrix} e_n \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Náš model převedený do stavového tvaru vypadá tedy takto:

$$\begin{pmatrix} y_n \\ s_n \\ t_n \\ y_{n-1} \\ s_{n-1} \\ t_{n-1} \\ 1 \end{pmatrix} = \begin{pmatrix} C & D & E & F & G & H & K \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} y_{n-1} \\ s_{n-1} \\ t_{n-1} \\ y_{n-2} \\ s_{n-2} \\ t_{n-2} \\ 1 \end{pmatrix} + \begin{pmatrix} \alpha & \beta \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} s_n \\ t_n \end{pmatrix} + \begin{pmatrix} e_n \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Druhá rovnice pro výstup nám pak pouze vypreparuje příslušný výstup:

$$y_n = P \cdot x_n = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} y_n \\ s_n \\ t_n \\ y_{n-1} \\ s_{n-1} \\ t_{n-1} \\ 1 \end{pmatrix}.$$

Poznámka o šumu

My jsme šum dosadili do prvé rovnice. Mohli jsme jej však dosadit do rovnice druhé. Jak to tedy je? Rozhodovat musíme případ od případu. Vezměme si např. zašuměný signál. Šum v první rovnici odpovídá šumu, který bychom naměřili i v nezašuměném signálu. Prostě proto, že náš model odpovídá datům jen přibližně. Šum v druhé rovnici odpovídá šumu, který vzniká např. při přenosu či měření signálu.

Náš šum můžeme rozdělit na šumy dva, do první i druhé rovnice. Jaké mají mít tyto šumy vlastnosti? Normální rozdělení, nulovou střední hodnotu a rozptyly, jež dají v součtu celkový rozptyl R. Tyto dva šumy označíme f a g.

Pokud dáme rozptyl i do druhé rovnice, výstup z této rovnice už nebude pouhou první položkou stavového vektoru. Bude zašuměnou první položkou stavového vektoru. Proto tyto dva výstupy odlišíme. Ten zašuměný, který přímo měříme, budeme značit y. Onen nezašuměný, který bychom chtěli znát, budeme značit Y.

Stavový model s oběma šumy bude vypadat takto:

$$\begin{pmatrix} Y_n \\ s_n \\ t_n \\ Y_{n-1} \\ s_{n-1} \\ t_{n-1} \\ 1 \end{pmatrix} = \begin{pmatrix} C & D & E & F & G & H & K \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} Y_{n-1} \\ s_{n-1} \\ t_{n-1} \\ Y_{n-2} \\ s_{n-2} \\ t_{n-2} \\ 1 \end{pmatrix} + \begin{pmatrix} \alpha & \beta \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} s_n \\ t_n \end{pmatrix} + \begin{pmatrix} f_n \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

$$y_n = P \cdot x_n + v_n = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} Y_n \\ s_n \\ t_n \\ Y_{n-1} \\ s_{n-1} \\ t_{n-1} \\ 1 \end{pmatrix} + \begin{pmatrix} g_n \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

V datech máme k dispozici zašuměný výstup y_n, rádi bychom ale znali nezašuměný výstup Y_n. Jak na to?

Filtrace a predikce

Chceme tedy na základě zašuměných výstupů odhadovat stav. Jak budou postupně přicházet data, budou se v našem odhadování neustále opakovat dva kroky - filtrace a predikce.

Filtrace

Jsme v n-tém kole. Z minulého kola máme odhad aktuálního stavu pro n-té kolo, ovšem z dat z $(n-1)$. kola. Nyní dorazila data pro n-té kolo. Na základě těchto nových dat opravíme náš odhad aktuálního stavu pro n-té kolo. Tomuto se říká filtrace.

Z Bayesovy věty můžeme pro filtraci odvodit následující vzorec:

$$f\left(x_n | d\left(n\right)\right) = \frac{f\left(y_n | x_n, u_n\right) \cdot f\left(x_n | d\left(n-1\right)\right)}{f\left(y_n | u_n, d\left(n-1\right)\right)}.$$

Tento vzorec je obdobou vzorce pro rozdělení parametrů. Tentokrát jsme jej napsali i s příslušnou konstantou ve jmenovateli. Význam jednotlivých výrazů je následující:

- $f\left(x_n|d\left(n-1\right)\right)$.....odhad stavu na základě $\left(n-1\right)$. dat. Pokud je stav vektorem, jde o sdružené rozdělení všech položek tohoto vektoru.
- $f\left(x_n|d\left(n\right)\right)$.....opravený odhad stavu na základě n-tých dat.
- $f\left(y_n|x_n,u_n\right)$.....model ve tvaru hustoty pravděpodobnosti. Získáme jej z modelu měření výstupu, tedy z druhé rovnice pro stavový model. Za u_n a y_n dosadíme aktuální data.
- $f\left(y_n|u_n,d\left(n-1\right)\right)$...normalizační konstanta. V tomto výrazu se nevyskytuje neznámý stav. Za u_n, y_n a $d\left(n-1\right)$ dosadíme. Výsledkem je tedy konstanta, která zapříčiní, aby integrál přes celou hustotu pravděpodobnosti byl opravdu 1.

Predikce

Jsme stále v n-tém kole. Máme odhad pro aktuální stav x_n. Chceme odhadnout stav pro příští, $\left(n+1\right)$. kolo. Ze sdružené hustoty pro stavy x_n a x_{n+1} na základě n-tých dat odvodíme vzorec pro hustotu pravděpodobnosti pro příští stav x_{n+1}:

$$f\left(x_{n+1}|d\left(n\right)\right)=\int_{\Omega}f\left(x_{n+1}|x_n,u_n\right)\cdot f\left(x_n|d\left(n\right)\right)dx_n.$$

Tento vzorec je obdobou vzorce pro rozdělení výstupu. Význam jednotlivých výrazů je následující:

- $f\left(x_n|d\left(n\right)\right)$.....odhad aktuálního stavu na základě n-tých dat. Pokud je stav vektorem, jde o sdružené rozdělení všech položek tohoto vektoru. Je to výsledek z minulého vzorce.
- $f\left(x_{n+1}|d\left(n\right)\right)$.....odhad příštího stavu na základě n-tých dat.
- $f\left(x_{n+1}|x_n,u_n\right)$.....model příštího stavu ve tvaru hustoty pravděpodobnosti. Získáme jej z modelu dynamiky stavu, tedy z první rovnice pro stavový model.
- $\int_{\Omega}\ldots dx_n$.....integrujeme přes všechny položky aktuálního stavového vektoru. Je to integrál určitý, přes celý definiční obor.

Poznámka

Výpočet dle těchto vzorců, zejména integrace, může být poměrně náročná.

Pokud jsou všechny šumy gaussovské, můžeme místo vývoje funkcí počítat jen vývoj středních hodnot a rozptylů, což je nesrovnatelně jednodušší. Normalita rozdělení se zachová. Algoritmus, který toto provádí, se jmenuje Kalmanův filtr.

Kalmanův filtr

Kalmanův filtr má dvě části. Filtraci a predikci.

Výstupem filtrace je bodový odhad pro střední hodnoty položek stavového vektoru. Tento vektor středních hodnot budeme značit $x_{n|n}$. První n znamená, že je to odhad stavu pro n-té kolo, druhé n znamená, že vycházíme z n-tých dat. Právě tento vektor je pro nás důležitý, protože jeho první položku budeme interpretovat jako odhad nezašuměného výstupu.

Druhým výstupem filtrace je kovarianční matice pro položky stavového vektoru. Tuto kovarianční matici budeme značit $R_{n|n}$.

Připomeneme, co je kovarianční matice. Na diagonálách najdeme rozptyly pro jednotlivé položky. Tyto rozptyly, resp. názorněji směrodatné odchylky z nich vytvořené, nám říkají, jak moc bude sdružená gaussovka v tom či onom směru široká. Čím budou tyto rozptyly menší, tím přesněji jsou příslušné položky stavového vektoru určeny. Členy mimo diagonálu nám říkají, jak jsou jednotlivé položky mezi sebou závislé. Nula na i-tém řádku a v j-tém sloupci znamená nezávislost i-té a j-té položky. Kladné číslo znamená kladnou korelaci, záporné číslo zápornou. V grafu sdružené hustoty pravděpodobnosti se tyto nediagonální členy projeví tím, že sdružená gaussovka bude protáhlá v nějakém šikmém směru.

Druhou částí Kalmanova filtru je predikce. Jejím výstupem budou odhady pro střední hodnoty a kovarianční matice pro příští kolo. Vektor bodových odhadů středních hodnot budeme značit $x_{n+1|n}$, kovarianční matici pro položky stavového vektoru $R_{n+1|n}$. Druhé n připomíná, že stále vycházíme pouze z n-tých dat.

Vstupem do Kalmanova filtru jsou bodové odhady středních hodnot $x_{n|n-1}$ a kovarianční matice $R_{n|n-1}$ z minulého kola.

Pro první kolo volím vektor $x_{1|0}$ např. nulový a kovarianční matici $R_{1|0}$ s velkými diagonálními členy (např. 1000). To znamená, že svému odhadu prvního stavu $x_{1|0}$ přikládám jen velmi malou váhu a že velkou váhu naopak budou mít přicházející data. Správný stav se pak nastaví rychleji.

Dalšími vstupy jsou matice M, N, A a B z modelu.

Dalšími vstupy jsou matice kovariancí šumů r_w a r_v.

Matice kovariancí šumů

U těchto matic se na chvíli zastavíme. V našem modelu měla jen první položka šumových vektorů význam, ostatní byly nulové. V principu ale můžeme zavést šumy i do ostatních položek šumových vektorů. Např. šum v druhé a třetí položce vektoru w znamená, že řízení s a t k nám doráží také zašuměné.

Pokud jsou šumy na první, druhé a třetí pozici šumového vektoru nezávislé, bude mít kovarianční matice šumů r_w na prvních třech diagonálních místech rozptyly příslušných vektorů, jinde budou samé nuly.

Pokud bychom ale chtěli vyjádřit, že např. řízení s a t jsou zašuměny sice náhodným, ale stejným šumem, byla by situace trochu náročnější. Podrobněji to zde rozebírat nebudeme.

Kalmanův filtr

Algoritmus Kalmanova filtru vypadá takto:

Filtrace:

$y_p = Ax_{n\|n-1} + Bu_n$	...	předpověď výstupu
$R_p = r_v + A \cdot R_{n\|n-1} \cdot A$'	...	kovariance výstupu
$R_{n\|n} = R_{n\|n-1} - R_{n\|n-1} \cdot A' \cdot R_p^{-1} \cdot A \cdot R_{n\|n-1}$	...	přepočet kovariance stavu
$K = R_{n\|n} \cdot A' \cdot r_v^{-1}$	...	Kalmanův gain
$x_{n\|n} = x_{n\|n-1} + K \cdot (y_n - y_p)$	...	datová oprava stavu

Predikce

$x_{n+1\|n} = M \cdot x_{n\|n} + N \cdot u_n$	...	předpověď stavu
$R_{n+1\|n} = r_w + M \cdot R_{n\|n} \cdot M'$	...	předpověď kovariance stavu

Slabina Kalmanova filtru

K výpočtu Kalmanova filtru potřebujeme znát matice M, N, A a B a matice kovariancí šumů r_w a r_v. Ty v reálu neznáme.

Bodový odhad parametrů ještě zvládneme pomocí datové matice. To umíme. Takže matice M, N, A a B nejsou takovým problémem.

I celkový šum pomocí datové matice spočteme. Ale matice kovariancí šumů?

Pokud máme v rovnici měření výstupu jen jeden šum, tak můžeme uvažovat, že právě toto je cca ten šum, který měříme z dat a že šum v nezašuměném signálu bude podstatně menší. Ten ale musíme odhadnout.

Nepříjemné je, že běh Kalmanova filtru na volbě matic kovariancí šumů poměrně závisí.

Nezbývá tedy, než dobře analyzovat situaci, odhadovat a zkoušet.