Státnice 2000, otázka 1/7

Tomáš Hlavatý, xhlavaty@fel.cvut.cz

Leden 2000

Základy teorie informace, informační entropie, kódování, kapacita diskrétního i spojitého přenosového kanálu, redundance, vztah termodynamické a informační entropie.

Bayesovské rozhodování, diskriminační funkce, ztrátová funkce, minimalizace ztrát.

1 Teorie informace

Hlavní principy Teorie informace [1] definoval po 2. světové válce C. E. Shannon. Analyzuje problematiku přenosu zpráv včetně poruch, parametrů kanálu i přenášeného signálu, definuje kvantitativní míru informace. Zabývá se rovněž optimální transformací dat do přenosové formy — kódováním a současně také zvýšením odolnosti přenášeného signálu vůči poruchám. Základní obecnou myšlenkou teorie informace je způsob sběru, přenosu, zpracování a archivace informace. S informací se pracuje jako s libovolnou fyzikální veličinou.

Graficky můžeme množství přenášené informace zobrazit jako kvádr s hranami v osách počtu bitů, šířky pásma a času.

Přenos informace lze podle Shannonovy teorie rozdělit do pěti vrstev:

statistika — Shannonova teorie používá statistické zpracování. Tato teorie dovoluje popisování vlastností informačních znaků a jejich kvantifikaci.

syntaxe — Řeší problematiku znaků, slov a vět.

sémantika — Přenosová informace se může přenášet různými médii, podle nichž se volí typy kódů.

pragmatika — Při přenosu informace mezi vysílačem a přijímačem dochází k navázání volného spojení, kdy např. řeč není synchronizována; spojení v omezeném výběru; spojení s využitím inteligentního řízení.

apobetika — Nejvyšší vrstva zajišťující cíleně orientované programové vybavení.

1.1 Zpráva, data a údaje

zpráva — Sdělování ucelené myšlenky vyjadřující stav nějakého objektu a jeho chování v minulosti, přítomnosti nebo budoucnosti. Je to výběr souboru prvků z určité množiny. Jestliže je tato množina konečná/neohraničená, hovoříme o diskrétním/spojitém zdroji zpráv.

data — Zprávy určené pro strojní zpracování nebo výsledek tohoto zpracování.

údaje — Zprávy získané jako produkt nějakého postupu (výstup měřícího přístroje nebo senzoru). Při číslicovém zpracování se stávají daty.

1.2 Informační význam pravděpodobnosti jevu

Přijetí zprávy o velmi pravděpodobném jevu přinese málo informace a naopak. Množství informace je tedy nepřímo úměrné pravděpodobnosti výskytu zprávy u příjemce před jejím přijetím. Přijetím zprávy D získáme informaci

$$I(D) = \gamma[\frac{1}{P(D)}].$$

Jestliže je D akceptována přijetím zpráv A a B, pak $I(D) = I(A)+I(B)$, a při nezávislosti jevů A a B platí $P(D) = P(A)P(B)$. Z těchto rovnic už plyne, že závislost γ musí být log. Množství přenášené informace je tedy

$$I(D) = -\log P(D).$$

1.3 Entropie

Je definována jako střední hodnota informačního obsahu zprávy.

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log_a f(x) dx,$$

kde $f(x)$ je hustota pravděpodobnosti.

Entropie vyjadřuje průměrný obsah jednoho prvku. Jestliže je základ logaritmu $a = 2$, jedná se o binární zprávy a jednotka je bit/prvek. Jestliže je výraz přirozeným logaritmem, je jednotkou nit a pro dekadický logaritmus je jednotkou Hartley.

Pro spojitý náhodný proces s normálním rozdělením hustoty pravděpodobnosti je entropie

$$H(X) = \frac{1}{2} \log(2\pi e \sigma_x^2).$$

1.4 Kapacita spojitého přenosového kanálu

Přenášíme-li signál náhodného procesu $x(t)$ kanálem se šumem $\xi(t)$ s normálním rozdělením hustoty pravděpodobnosti, je rychlost přenosu informace

$$C = f_{max} \log_2 [\frac{P_X}{P_\xi} + 1] \qquad [bit/s].$$

Kapacitu kanálu je možné zvětšit zvýšením maximální frekvence, resp. větší šířkou frekvenčního pásma, ale také zvýšením poměrů výkonů signálu k šumu.

Pokud bude rušivý signál bílý šum s konstantní výkonovou spektrální hustotou v daném frekvenčním pásmu ($P_\xi = P_0 f_{max}$), bude kapacita kanálu při $f_{max} \to \infty$

$$C_{max} = 1.44 \frac{P_X}{P_0} \qquad [bit/s].$$

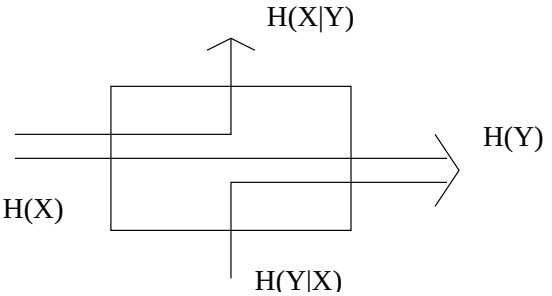
1.5 Kapacita diskrétního přenosového kanálu

Diskrétní kanál je popsán množinou vstupních znaků X, množinou výstupních znaků Y a pravděpodobnostmi $p(y_j|x_i)$, že při vstupu znaku x_i do kanálu vystoupí znak y_j. Tyto pravděpodobnosti se nazývají přímé, zatímco $p(x_i|y_j)$ se nazývají zpětné (pro dekódování) [2].

Množství informace, průměrně obsažené v jednom znaku zprávy, je entropie vstupního rozdělení $H(X) = -\sum p(x_i) \log p(x_i)$. Při přenosu diskrétním kanálem se ztratí množství informace

$$H(X|Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i|y_j)$$

průměrně na jeden znak. Zbývá $H(X) - H(X|Y)$ přenesené informace (obr. 1).



Obrázek 1: Přenosový kanál s poruchami

Jestliže entropii počítáme v bitech a známe průměrnou dobu T, kterou kanál spotřebuje na přenos jednoho znaku, je rychlost přenosu

$$R(X,Y) = \frac{H(X) - H(X|Y)}{T} \qquad [bit/s].$$

Kapacita kanálu je maximální rychlost přenosu při jakémkoli rozdělení pravděpodobnosti znaků

$$C = \max_{p(x_i)} R(X,Y) \qquad [bit/s].$$

Určení kapacity je snadné, jakmile známe optimální rozdělení $p(x_i)$ daného kanálu (což nebývá snadné najít).

1.6 Kódování

Kódování lze chápat jako jednoznačné přiřazení symbolů jedné množiny symbolům druhé množiny. Kódy můžeme dělit podle:

- počtu prvků v kódové kombinaci: jednoprvkové, víceprvkové;
- možnosti záměny kódových kombinací: záměnné, nezáměnné — nevyžadují oddělovací znaky mezi jednotlivými kódovými bajty;
- délky kódových skupin: rovnoměrné, nerovnoměrné — mají různou délku kódových skupin a jsou navrženy tak, aby byl přenos maximálně efektivní (nejužívanější znaky mají nejkratší délky — optimální Shannon-Fanonův kód, Huffmanův kód);
- zabezpečení přenosu před náhodnými poruchami: s minimální redundancí — např. parita, redundantní (bezpečnostní) — detekční a samoopravné (korekční);

Bezpečnostní kódy ještě můžeme dělit podle délky a struktury kódového slova na systematické a nesystematické — kódové slovo nelze rozdělit na informační a zabezpečovací část (např. konstantní počet jedniček v každém přenášeném slově); spojité, blokové (zabezpečující prvky jsou na konci zabezpečeného bloku); lineární (zabezpečení se provede lineární kombinací informačních prvků), cyklické (vhodné proti shlukům chyb, jednoduché).

1.7 Vztah termodynamické a informační entropie

Oba pojmy udávají neuspořádanost systému. Zatím co informační entropie popisuje přenos zpráv a informace, termodynamická entropie popisuje termodynamický systém. Entropie termodynamického systému přirozeně roste. Naší snahou bývá její zmenšení, k čemuž musíme dodat energii. Informační entropii se snažíme naopak zvětšovat, abychom získali co nejvíce informace.

2 Bayesovské rozhodování

Úloha rozpoznávání (klasifikace) spočívá v zařazování objektů, jevů a situací reálného světa do tříd [3]. Stroj, který klasifikaci provádí, se nazývá **klasifikátor**. Klasifikátor zobrazuje množinu vektorů příznaků na množinu jmen (indikátorů) tříd — definuje **rozhodovací pravidlo**.

Tato kapitola se zabývá bayesovským rozpoznáváním, které patří mezi příznakové metody rozpoznávání, v nichž jsou obrazy reprezentovány vektorem číselných hodnot (příznaků). Bayesovské rozpoznávání patří mezi metody statistické a je založeno na Bayesovské statistice.

2.1 Diskriminační funkce

Rozhodovací pravidlo je funkce

$$\omega = d(x),$$

která přiřazuje každému vektoru příznaků x indikátor třídy ω. Rozhodovací pravidlo vymezuje v příznakovém prostoru R vzájemně disjunktních množin, kde R je počet tříd. Nadplochy, které jsou společné dvěma množinám nazýváme **rozdělující nadplochy**.

Rozdělující nadplochy lze definovat pomocí R skalárních funkcí vektorového argumentu $g_i(x), i = 1 \dots R$, které nazýváme **diskriminační funkce**. Každá z diskriminačních funkcí je přiřazena jedné z tříd. Diskriminační funkce se vybírají tak, aby platilo

$$\forall x \in \omega_i \; g_i(x) > g_j(x), \; j = 1 \dots R, \; j \neq i.$$

Klasifikátor klasifikuje vektor příznaků x do třídy ω_i, pokud platí

$$g_i(x) = \max_j g_j(x).$$

Rozdělující nadplochy mohou být lineární, po částech lineární či nelineární. Nelineární diskriminační funkce se po částech linearizují nebo se provede nelineární transformace příznakového prostoru, která umožní použití lineárního klasifikátoru.

Klasifikace do dvou tříd se nazývá **dichotomie**. V tomto případě stačí posuzovat znaménko

rozdílu $g(x) = g_1(x) - g_2(x)$ (např. Rosenblattův perceptron).

2.2 Ztrátová funkce

V praktických úlohách se množiny obrazů jednotlivých tříd mohou překrývat (nejsou disjunktní) nebo nemusíme být schopni jednoznačně posoudit, do které třídy vektor příznaků patří. Každé chybné rozhodnutí představuje jistou ztrátu [4]. Definujeme ztrátovou funkci

$$\lambda(\omega_i|\omega_j),$$

což je ztráta vznikající při klasifikaci obrazu ze třídy ω_j do třídy ω_i.

Při hledání nejmenší možné ztráty při daných ztrátových funkcích a pravděpodobnostech výskytu obrazů je nezbytné vyhodnotit **střední ztrátu** v závislosti na parametrech nastavení klasifikátoru q rozhodovacího pravidla $\omega = d(x, q)$. Hodnota q^*, při níž nastává minimum všech ztrát, se nazývá optimální parametr a funkce

$$\omega = d(x, q^*)$$

se nazývá **optimální rozhodovací pravidlo**.

2.3 Kritérium minimálních ztrát

Úkolem je nastavit klasifikátor tak, aby ztráty způsobené chybným rozhodnutím byly minimální (minimalizace ztrát, risku nebo rizika).

Předpokládejme, že známe apriorní pravděpodobnosti $P(\omega_i),\ i = 1 \dots R$, podmíněné hustoty pravděpodobnosti $p(x|\omega_i),\ i = 1 \dots R$ a matici ztrátových funkcí $\lambda = [\lambda(\omega_i, \omega_j)],\ i, j = 1 \dots R$. Optimální rozhodovací pravidlo minimalizuje střední ztrátu:

$$\begin{aligned} J(q^*) &= \min_q J(q) \\ &= \min_q \int \sum_{i=1}^{R} \lambda(d(x,q)|\omega_i) p(x|\omega_i) P(\omega_i) dx \end{aligned}$$

2.4 Kritérium minimální chyby

Často se používá speciální tvar ztrátových funkcí

$$\lambda(\omega_i|\omega_j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} .$$

Kritérium minimálních ztrát pak přejde na kritérium minimální chyby. Kritérium minimální chyby se také nazývá Bayesovské kritérium.

Při volbě těchto **jednotkových ztrátových funkcí** představuje střední ztráta pravděpodobnost chybného rozhodnutí. Optimální klasifikátor ve smyslu kritéria minimální chyby s jednotkovými ztrátovými funkcemi tedy klasifikuje s **nejmenší pravděpodobností chybného rozhodnutí**. Diskriminační funkce se díky jednotkovým ztrátovým funkcím zjednoduší na tvar $g_i(x) = P(\omega_i|x)$, který lze užitím Bayesova vztahu upravit na

$$g_i(x) = P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)},$$

kde $p(x) = \sum_{i=1}^{R} p(x|\omega_i)P(\omega_i)$ je hustota pravděpodobnosti rozložení vektoru příznaků x v příznakovém prostoru bez ohledu na třídu.

Pomocí aposteriorních pravděpodobností stanovíme rozhodovací pravidlo: vektor x zařadíme do takové třídy i, pro kterou platí

$$P(\omega_i|x) = \max_j P(\omega_j|x).$$

Toto pravidlo můžeme také vyjádřit pomocí diskriminačních funkcí. Jejich výběr není jednoznačný. Kromě dříve uvedené můžeme např. volit jednu z funkcí

$$g_i(x) = p(x|\omega_i)P(\omega_i)$$

nebo

$$g_i(x) = \log p(x|\omega_i) + \log P(\omega_i).$$

Speciálním případem kritéria minimální chyby je **klasifikátor podle minima vzdálenosti**. U tohoto klasifikátoru je $p(x|\omega_i)$ ve tvaru normálního rozložení s jednotkovým rozptylem

a $P(\omega_i)$ stejným pro všechny třídy. Rozhodovací pravidlo je ve tvaru

$$|x - \mu_i| = \min_j |x - \mu_j|.$$

2.5 Bayesovská statistika

Klasická statistika pracuje s pojmem **pravděpodobnost** ve smyslu existující limity relativních četností výskytu uvažovaného jevu. Bayesovská statistika naproti tomu používá pojem pravděpodobnosti ke kvantitativnímu popisu **neurčitosti**. Pojem náhodný může být interpretován jako neurčitý a hustota pravděpodobnosti je pak interpretována jako subjektivní míra důvěry racionálně a konzistentně uvažující osoby o hodnotě odhadovaného parametru.

Význam subjektivní interpretace hustoty pravděpodobnosti je v tom, že umíme formálně popsat a kvantifikovat proces akumulování informace získané pozorováním či experimentem. To zachycuje **Bayesův vztah**

$$P(\omega|x) = \frac{p(x|\omega)P(\omega)}{\int p(x|\omega)P(\omega)d\omega},$$

který popisuje korekci subjektivní apriorní hustoty pravděpodobnosti objektivními daty.

Velkým problémem Bayesovského přístupu je závislost výsledků na apriorní informaci. Pokud nemáme žádnou apriorní informaci k dispozici, je nutné volit tzv. **neinformativní apriorní hustotu pravděpodobnosti**, která není obecně rovnoměrně rozdělená.

V předcházejícím textu jsme předpokládali, že známe apriorní pravděpodobnosti $P(\omega_i),\ i = 1 \ldots R$. Pokud bychom je neznali, museli bychom je volit rovnoměrně rozdělené ($\forall i\ P(\omega_i) = \frac{1}{R}$). Tím bychom získali speciální případ Bayesovského klasifikátoru — **ML klasifikátor** (maximálně věrohodný) [5].

Literatura

[1] Petr Kocourek: Přenos informace. ČVUT, 1996.

[2] Jiří Adámek: Stochastické procesy a teorie informace — úlohy. ČVUT, 1989.

[3] Mařík, Štěpánková, Lažanský a kolektiv: Umělá inteligence 1. Academia, 1993.

[4] Kotek, Brůha, Chalupa, Jelínek: Adaptivní a učící se systémy. SNTL, 1980.

[5] Přednášky z předmětu Rozpoznávání, 1999.