Kapitola 1

Logistická regrese

Předpokládám, že už jsme zavedli základní pojmy jako rysy a že už máme nějaké značení

Velkost trenovacich dat a pocet parametru

Motivační povídání ... jeden z nejpoužívanějších modelů ... v jistém smyslu lze logistickou regresi považovat za základ neuronových sítí, ke kterým se dostaneme později.

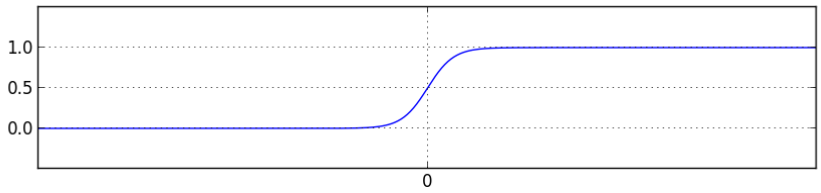
Logistická regrese, přestože se jmenuje z historických důvodů regrese, je klasifikační model, který ve své základní podobě klasifikuje objekt popsaný vektorem rysů do tříd 1 (ano) a 0 (ne).

1.1 Model

Stejně jako v případě perceptronu (viz. kapitola **??**, využijeme vážený součet rysů. Váhy pro jednotlivé rysy nám říkají, jak moc je který rys přispívá tomu, abychom objekt klasifikovali kladně. Váhy tedy mohou být i záporné, v takovém případě, čím silnější rys je, tím spíše bude objekt klasifikován záporně.

Jmenuje se podle logistické funkce, která se používá pro rozhodování, do jaké třídy objekt patří. Logistickou funkci můžeme popsat vzorcem:

$$f(z) = \frac{1}{1 + \mathrm{e}^{-z}}. \tag{1.1}$$



Obrázek 1.1: Průběh logistické funkce na intervalu $[-30; 30]$

RR: Přidat čáru $x = 0$, $y = 0.5$

Jak vidíme z průběhu funkce (viz Obrázek **??**), chová se jako jako jakási prahová funkce. Pokud dostane dostatečně velké číslo, vrací číslo, které se blíží jedničce, naopak při dostatečně nízkém vstupu vrací číslo, které se velice blíží nule. Všimněte si také, že pro $z = 0$ je její hodnota $\frac{1}{2}$.

Za z do logistické funkce dosadíme vážený součet rysů. Rovnicí zapíšeme takto:

$$z = w_0 + \sum_{i=1}^{n} w_i x_i \tag{1.2}$$

To znamená, že čím větší vážený součet rysů je, tím je logistická funkce bližší jedničce a naopak. V hodnotě jedna polovina se tedy láme, zda více věříme 1 nebo 0.

Z toho už je krůček k formálnímu popisu modelu:

$$t = \begin{cases} 1 & \text{pokud } \frac{1}{1+\mathrm{e}^{-w_0-\sum_{i=1}^{n} w_i x_i}} \geq \frac{1}{2} \\ 0 & \text{jinak} \end{cases} \tag{1.3}$$

Bystrý čtenář si možná všiml, že takto zadaný model se příliš neliší od perceptronu. Model jako takový je skutečně ekvivalentní – ptáme se na to, jestli je vážený součet rysů větší nebo menší, než nějaká pevně stanovená mez. Co se v tomto případě liší, je způsob učení modelu, který se v případě logistické regrese odvozuje pomocí teorie pravděpodobnosti.

JL: Teď popsat, jak se to učí

JL: Napsat, že existuje SGD

JL: Regularizace ... nejprve natrenovat model bez regularizace a ukazat learning curves a overfitting

JL: Pak pridat regularizaci jako penaltu za vysoke vahy a ukazat rozdil

JL: Plot: velikost regularizacniho parametru a trenovaci a testovaci chyba

Následující dvě části obsahují více matematiky.

1.2 ∗ Pravděpodobnostní odvození

JL: Z tohodle textu smazat většinu

Logistickou regresi lze odvodit také jako odhad podmíněné pravděpodobnosti klasifikace hodnotou 1, je-li dán vektor rysů. Dřív, než se dostaneme k náznaku tohoto odvození, je potřeba varovat, že pravděpodobnostní interpretace logistické regrese, může být v mnohém zavádějící. Výstup logistické funkce je sice z čistě formálního matematického hlediska pravděpodobností, v praxi ale zdaleka nemusí odpovídat tomu, co si pod pravděpodobností intuitivně představujeme. Lepší odhad toho, s jakou jistotou model funguje, nám dává jeho úspěšnost na testovacích datech, byť i zde musíme být opatrní, pokud chceme toto číslo používat jako pravděpodobnost.

Předpokládáme totiž, že když už máme nějaký vektor rysů (to je ta podmínka, proto podmíněná pravděpodobnost), je rozhodnutí, jaká bude klasifikace, náhodné – tzv. Bernoulliho

proces. Ten si můžeme představit jako hod neférovou mincí, kde padá orel (kladná klasifikace) s pravděpodobností p. V případě logistické regrese se pokusíme to, jak by takový proces mohl vypadat, popsat následující (autoři vědí, že bizarní) alegorií.

Představujeme si, že někde existuje nějaký stroj, ve kterém je naprogramovaný náš model. Když dostane na vstupu vektor rysů pro nějaký objekt, spočítá pravděpodobnost, že bude klasifikován kladně. Potom vyrobí minci, na které s touto pravděpodobností padá orel a mincí si hodí. Pokud padne orel, řekne 1, v opačném případě 0.

Je jistě zřejmé, že jen málokterý jev ve světě se dá popsat, takovýmto pravděpodobnostním mechanizmem. Tento předpoklad nám umožňuje, aby byl model stále relativně jednoduchý a v praxi se ukazuje, že takové učení dává dobré výsledky. Výstupu logistické funkce se obvykle říká míra jistoty nebo konfidence modelu (anglicky confidence).

Nyní už slibovaný náznak pravděpodobnostního odvození. Pravděpodobnost nějakého jevu (třeba že na minci padne orel) běžně odhadujeme jako poměr počtu případů, kdy jev nastal, a počtu všech pokusů. Pokud hodím tisíckrát mincí a jenom 420 krát padne orel, mohu pravděpodobnost toho, že padl orel, odhadnout na $420/1000 = 0,42$.

Jak ale odhadneme pravděpodobnost v případě že máme k dispozici pouze vážené součty rysů? Pomůžeme si tím, že si zavedeme skóre, kterému se někdy také říká energie podle fyzikálních souvislostí. Budeme mít zvlášť váhy pro situaci, kdy bychom objekt klasifikovali kladně $(w_0^1, \ldots, w_n^1)$ a záporně $(w_0^0, \ldots w_n^1)$. Energii pro kladnou klasifikaci zavedu jako

$$E(1|\mathbf{x}) = \mathrm{e}^{w_0^1 + \sum_{i=1}^n w_i^1 x_i} \tag{1.4}$$

a analogicky pro zápornou.

Na rozdíl od prostého váženého součtu, máme záruku, že toto skóre bude mít vždy kladnou hodnotu. Pravděpodobnost potom můžu odhadnout nikoli jako poměr počtů, ale jako poměr těchto skóre:

$$\mathsf{P}(y=1|\mathbf{x}) = \frac{\mathrm{e}^{w_0^1 + \sum_{i=1}^n w_i^1 x_i}}{\mathrm{e}^{w_0^0 + \sum_{i=1}^n w_i^0 x_i} + \mathrm{e}^{w_0^1 + \sum_{i=1}^n w_i^1 x_i}}; \quad \mathsf{P}(y=0|\mathbf{x}) = 1 - P(y=1|x). \tag{1.5}$$

Jedničku v pravděpodobnosti záporné klasifikace můžeme vyjádříme jako $1 - y$ – v takovém případě vychází pro kladnou klasifikaci nula – a mínus před pravděpodobností vyjádříme jako násobení hodnotou $(-1 + 2y)$ – pro kladnou klasifikace dává $+1$. Nyní už můžeme vyjádřit model obecně pro obě možné hodnoty y jedním vzorcem:

$$\mathsf{P}(y|\mathbf{x}) = (1-y) + (2y-1)\frac{\mathrm{e}^{w_0^1 + \sum_{i=1}^n w_i^1 x_i}}{\mathrm{e}^{w_0^0 + \sum_{i=1}^n w_i^0 x_i} + \mathrm{e}^{w_0^1 + \sum_{i=1}^n w_i^1 x_i}} \tag{1.6}$$

Podívejme se nyní detailněji na zlomek v rovnicích 1.5 a 1.6. Zlomek nejprve rozšíříme hodnotou v čitateli a dostáváme:

$$\frac{\mathrm{e}^{w_0^1 + \sum_{i=1}^n w_i^1 x_i}}{\mathrm{e}^{w_0^0 + \sum_{i=1}^n w_i^0 x_i} + \mathrm{e}^{w_0^1 + \sum_{i=1}^n w_i^1 x_i}} \cdot \frac{\frac{1}{\mathrm{e}^{w_0^1 + \sum_{i=1}^n w_i^1 x_i}}}{\frac{1}{\mathrm{e}^{w_0^1 + \sum_{i=1}^n w_i^1 x_i}}} = \frac{1}{1 + \frac{\mathrm{e}^{w_0^0 + \sum_{i=1}^n w_i^0 x_i}}{\mathrm{e}^{w_0^1 + \sum_{i=1}^n w_i^1 x_i}}} \tag{1.7}$$

Zlomek ve jmenovateli upravíme pomocí pravidel pro počítaní s exponenciálními funkcemi, zjednodušíme a dostaneme výraz

$$\frac{1}{1 + \mathrm{e}^{(w_0^0 - w_0^1) + \sum_{i=1}^n (w_i^0 - w_i^1) x_i}}. \tag{1.8}$$

To už je vlastně logistická funkce. Vidíme také, že nepotřebuje zvlášť znát váhy pro kladnou a zápornou klasifikace, jediné co potřebujeme je jejich rozdíl. Právě tento rozdíl jsou váhy, které využívá logistická regrese – pokládáme $w_i = w_i^0 - w_i^1$.

Funkce, kterou používáme v logistické regresi, jistým způsobem odhaduje pravděpodobnost, že bude objekt bude klasifikován kladně. Toho se využívá při učení model, jak ukážeme v další část, ovšem za nerealistického předpokladu, že klasifikace se ve skutečném světě Bernoulliho proces.

1.3 ∗ Odvození učení

Učení modelu využívá jeho pravděpodobností interpretaci. Z té odvodíme takzvanou věrohodnost parametrů – vah rysů (anglicky likelihood). Ukážeme si algoritmus, kterým se dá postupně zvyšovat věrohodnost modelu, až dosáhne svého maxima.

Mějme tedy trénovací data $\mathcal{D}$, která tvoří N dvojic vektorů a správných klasifikací $(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)$ a model s vahami součtu, které budeme pro jednoduchost zápisu označovat jako $\mathbf{w}$.

Když definujeme věrohodnost modelu, díváme se pravděpodobnost jakoby z druhé strany, než je logické. Pro váhy rysů $\mathbf{w}$ se ptáme, jak pravděpodobné by bylo, že vznikla trénovací data, která používáme – za předpokladu, že by model, co používáme, byl fixně daný.

Pokusíme tuto krkolomnou úvahu vysvětlit trochu pomaleji. Vrátíme se k představě podivného stroje z minulé části. Když se ptáme, jak věrohodné jsou určité váhy, vyrobíme nejprve stroj pro tyto váhy a potom se zeptáme, jak pravděpodobné by bylo, že kdybychom tomu stroji dodali stejné vektory rysů, jako máme v trénovacích data, dostali bychom totožné klasifikace. Vlastně se ptáme, jak moc dobře je možné, že by model sám připravil naše trénovací data – a předpokládáme, že čím lepší model, tím spíše se takto chová.

Abychom toto mohli odhadnout, předpokládáme navíc, že jednotlivé trénovací příklady jsou na sobě navzájem nezávislé. Protože pravděpodobnost toho, že současně nastane několik nezávislých jevů, je rovna součinu jejich pravděpodobností, můžeme vyjádřit věrohodnost jako součin pravděpodobností pro jednotlivé trénovací příklady. Z předchozí části víme, že tuto pravděpodobnost můžeme vyjádřit právě logistickou funkcí.

$$\mathcal{L}'(\mathbf{w}) = \prod_{i=1}^{N} \mathsf{P}(t_i|x_i) = \prod_{i=1}^{N} (1 - t_i) + (2t_i - 1)\frac{1}{1 + \mathrm{e}^{-w_0 - \sum_{i=1}^{n} w_i x_i}} \tag{1.9}$$

Některým čtenářům se může celý tento obrat zdát podezřelý. Na místě je jistě otázka, jak to věrohodnost vah počítáme jako podmíněnou pravděpodobnost dat, jsou-li dány váhy modelu. Není to nějaký nesmysl? Nemělo by to být naopak? Odpověď je, že by to klidně mohl být naopak, ale výsledek by byl stejný. Tyto dva pohledy totiž zachycují rozdíl mezi klasickým (tzv. frekvetistickým) a Bayesovským pohledem na statistiku. Pro uklidnění zvídavých čtenář ukážeme, proč jsou v tomto případě oba pohledy ekvivalentní, pro stručnost pouze ve zjednodušené notaci.

Podmíněnou pravděpodobnost vah $\mathbf{w}$, jsou-li dána trénovací data $\mathcal{D}$ rozepíšeme pomocí Bayesova pravidla:

$$\mathsf{P}(\mathbf{w}|\mathcal{D}) = \frac{\mathsf{P}(\mathcal{D}|\mathbf{w})\mathsf{P}(\mathbf{w})}{\mathsf{P}(\mathcal{D})} \tag{1.10}$$

Protože dopředu nevíme o vahách vůbec nic, má každá sada vah stejnou pravděpodobnost. Co je pravděpodobnost dat $\mathsf{P}\mathcal{D}$, je na delší povídání, ale můžeme se spokojit s tím, že trénovací data jsou jenom jedna a proto je to vždy stejné číslo. Tyto dva členy jsou tedy konstantní bez ohledu na to, jaké váhy zvolíme. Z toho plyne, že když věrohodnost jako podmíněnou pravděpodobnost vah modelu, jsou-li dána data, stačí maximalizovat obrácenou podmíněnou pravděpodobnost, jak jsme věrohodnost zavedli v rovnici 1.9.

Pokud bychom věrohodnost počítali v této formě na počítači, docházelo by často k numerickým chybám. Jedná se o součin mnoha čísel mezi nulou a jednou, takže výsledek byl velice malé číslo, které by bylo vzhledem k vlastnostem procesorů zaokrouhleno na nulu. Využijeme toho, že logaritmus je prostá funkce a tedy $\mathcal{L}'$ zlogaritmujeme, nezmění to, v jakých bodech je její minimum a maximum. Můžeme tedy psát:

$$\mathcal{L}(\mathbf{w}) = \ln \prod_{i=1}^{N} (1 - t_i) + (2t_i - 1) \frac{1}{1 + \mathrm{e}^{-w_0 - \sum_{i=1}^{n} w_i x_i}}$$

Když provedeme úpravy podle pravidel pro počítání s logaritmy, dostaneme:

Ti, co se již setkali s diferenciálním počtem, vědí, že maximum funkce lze najít tak, že nejprve spočítáme derivaci funkce a potom hledáme body, kdy je derivace rovna nule.

JL: Dokončit odvození derivace