BI-VZD přednáška 9

Alexander Kovalenko

FIT ČVUT

04. 04. 2022

Autoři: Karel Klouda, Juan Pablo Maldonado Lopez, Daniel Vašata.
Problémy, návrhy apod. hlaste v GitLabu.
Verze souboru: 4. dubna 2022 10:22.

Co bude v dnešní přednášce

- Připomenutí problému klasifikace
- Základní myšlenka logistické regrese
- Maximálně věrohodný odhad: myšlenka
- Výpočet odhadu

Úvodní poznámky

- Přestože se metoda **Logistická regrese** jmenuje tak, jak se jmenuje, je to metoda určená pro *klasifikaci*.
- Pro připomenutí: V případě klasifikace vysvětlovaná proměnná Y může nabývat jen několik málo hodnot.
- **My se omezíme na binární klasifikaci**, kdy má Y hodnotu buď 0 nebo 1.
- Rozdíl mezi regresí (spojitá vysvětlovaná proměnná) a klasifikací se projevuje zejména v tom, jak vypadá tvar hledané závislosti mezi příznaky $X_1, X_2, \ldots, X_p$ a vysvětlovanou proměnnou.
- Zatímco pro regresi může (ale často nemá) tato závislost vztah připomínající klasické funkce známé z analýzy, u funkcí, jejichž obor hodnot je dvouprvková množina, se musí použít nějaký „trik“.

Připomenutí známých metod

Připomeňme si metody, které už známe:

- Rozhodovací stromy (zvládají klasifikaci i regresi): Funkční závislost je komplikovaná a daná průchodem daným stromem. V případě regrese má výsledná funkce tvar po (velmi malých) částech konstantní funkce.

- KNN (zvládá klasifikaci i regresi): Funkční závislost je opět velice komplikovaná a nemá žádný explicitní tvar. Rozhodnutí o hodnotě Y je velice „lokální“.

- Lineární regrese (jen pro regresi): Jedná se o parametrickou metodu, kde se hledá závislost v zadaném tvaru „lineární kombinace příznaků“

$$Y \approx w_0 + w_1 x_1 + \cdots + w_p x_p,$$

kde x_i jsou nějaké konkrétní hodnoty příznaků X_i a w_i jsou neznámé koeficienty.

Příklad (rýmička)

- Uvažujme trochu modifikovaný příklad „rýmičkového“ příkladu z druhé přednášky o stromech:
  - Vysvětlovaná proměnná Y má dvě hodnoty: $Y = 1$ značí, že má daná osoba rýmičku, $Y = 0$ značí, že jí nemá.
  - X_1 je binární příznak pohlaví (1 = žena, 0 = muž).
  - X_2 je numerický příznak věku (v celých letech).
  - X_3 je teplota ve stupních Celsia.

- Souvislost logistické regrese a lineární regrese spočívá v tom, že i u logistické regrese se rozhodnutí konstruuje pomocí lineární kombinace příznaků, tedy v našem případě výrazu

$$w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3.$$

- Jak ale donutit tento výraz, aby z něho padalo rozhodnutí o tom, jestli je $Y = 1$ nebo $Y = 0$?

Myšlenka logistické regrese

- Logistická regrese stojí na triku, který z diskrétního problému dělá problém spojitý: **Namísto hodnoty vysvětlované proměnné $Y \in \{0, 1\}$ se snažíme predikovat pravděpodobnost, že Y má hodnotu 1, tj. číslo $P(Y = 1)$ z intervalu $[0, 1]$.**
- Přesněji řečeno hledáme *funkční předpis*, který pro dané hodnoty x_i příznaků X_i a dané hodnoty koeficientů w_i, vrátí číslo z intervalu $[0, 1]$, které bude odhadem toho, že daná osoba má rýmičku $(Y = 1)$.
- Tuto získanou pravděpodobnost budeme značit následovně:

$$P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{w}),$$

  čímž je vyjádřeno to, že je závislá na hodnotách příznaků $\boldsymbol{x} = (x_1, x_2, x_3)$ a koeficientů $\boldsymbol{w} = (w_0, w_1, w_2, w_3)$.
- Jelikož součet pravděpodobností, že někdo má a nemá rýmičku musí být jedna, platí

$$P(Y = 0 \mid \boldsymbol{x}, \boldsymbol{w}) = 1 - P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{w})$$

  a nám tedy opravdu stačí najít pouze model pro $P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{w})$.

Logistická funkce (Sigmoida)

- Přešli jsme tedy od funkce s oborem hodnot $\{0, 1\}$ k funkci se spojitým oborem hodnot $[0, 1]$.
- Jak ale donutit lineární výraz

$$w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3,$$

  aby neutekl z tohoto intervalu?
- Zde přichází na řadu druhý trik: **Číslo $w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$ dosadíme do vhodně zvolené funkce, jejíž obor hodnot je podmnožinou intervalu $[0, 1]$.** Tím zajistíme, že pro jakékoli hodnoty koeficientů i příznaků dostaneme číslo smysluplně vyjadřující pravděpodobnost.
- Obvyklou volbou této funkce je **sigmoida** (angl. **sigmoid function**), což je speciální případ **logistické funkce**:

$$f(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}.$$

Logistická funkce (Sigmoida): vlastnosti

$$f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}.$$

- Definičním oborem je celá množina $\mathbb{R}$, za x tedy můžeme dosadit cokoli; my budeme dosazovat číslo $w_0 + w_1x_1 + w_2x_2 + w_3x_3$.
- Oborem hodnot je interval $(0, 1)$, nikdy se nám tedy nestane, že by pravděpodobnost $Y = 1$ byla jedna (jistý jev) nebo nula (nemožný jev).
- Funkce ostře rostoucí na $\mathbb{R}$ a tedy prostá. Inverzní funkcí je

$$f^{-1}(x) = \ln \frac{x}{1-x}.$$

- Limity pro $x \to -\infty$ a $x \to +\infty$ jsou 0 resp. 1, platí $f(0) = \frac{1}{2}$ a funkce $f(x) - \frac{1}{2}$ je lichá.
- Číslo $f(x)$ bude pro nás pravděpodobnost jevu $Y = 1$, opačný jev $Y = 0$ tak bude mít pravděpodobnost

$$1 - f(x) = \frac{1}{1+e^x}.$$

Jak to tedy bude celé fungovat?

- Uvažujme náš rýmičkový příklad se třemi příznaky. Z technických důvodů přidáme nultý příznak X_0, který bude mít vždy hodnotu $x_0 = 1$ (viz pojem *intercept* z minulé přednášky).
- Předpokládejme dále, že koeficienty mají tyto hodnoty:

$$\boldsymbol{w} = (w_0, w_1, w_2, w_3) = (0.1, -0.3, -0.2, 0.2),$$

  zatím ponechme stranou, jak se koeficienty hledají.
- Předpokládejme, že máme 35letého muže, který má teplotu 37.2 stupně Celsia. Tj. příslušný vektor s hodnotami příznaků je

$$\boldsymbol{x} = (x_0, x_1, x_2, x_3) = (1, 0, 35, 37.2).$$

- Pravděpodobnost toho, že má rýmičku dostaneme výpočtem výrazu

$$\boldsymbol{w}^T \boldsymbol{x} = w_0 x_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 = 0.54\,.$$

  a jeho dosazením do sigmoidy

$$P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{w}) = \frac{e^{0.54}}{1 + e^{0.54}} = 0.631812,$$

  tedy odhad pravděpodobnosti rýmičky je přes 63 %, tedy je rozumnější rozhodnout, že rýmičku daná osoba má.

Logistická regrese: popis modelu

Jak tedy vypadá model logistické regrese pro binární klasifikaci:

- Máme binární vysvětlovanou proměnnou Y s hodnotami 0 a 1 a p příznaků $X_1, X_2, \ldots, X_p$ s konstantním $X_0 = 1$.

- Hledáme model pro odhad pravděpodobnosti, který pro dané hodnoty $\boldsymbol{x} = (x_0, x_1, \ldots, x_p)$ a pro koeficienty $\boldsymbol{w} = (w_0, w_1, \ldots, w_p)$ má tvar

$$P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{w}) = \frac{e^{\boldsymbol{w}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{w}^T \boldsymbol{x}}}.$$

- Predikce se pak dělá následovně: Pro daná data $\boldsymbol{x}$ se spočítá odhad pravděpodobnosti $P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{w})$. Je-li tato větší než $\frac{1}{2}$, rozhodneme se pro $Y = 1$, je-li menší než 0.5, pak pro $Y = 0$.

Logistická regrese: hranice rozhodnutí

- Jelikož máme p číselných příznaků, je každý datový bod vlastně bodem v prostoru $\mathbb{R}^p$.
- Máme-li pevně zvolené parametry modelu $\boldsymbol{w}$, můžeme (teoreticky) pro každý bod spočítat hodnotu

$$P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{w}) = \frac{e^{\boldsymbol{w}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{w}^T \boldsymbol{x}}}.$$

  a rozhodnout, jestli je větší než 0.5 ($Y = 1$) nebo menší ($Y = 0$).
- Jak vypadají tyto dvě podmnožiny $\mathbb{R}^p$?
- Nebo jinými slovy: Jak vypadá hranice mezi těmito množinami daná rovnicí

$$P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{w}) = \frac{e^{\boldsymbol{w}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{w}^T \boldsymbol{x}}} = \frac{1}{2}\,?$$

- Z toho, co jsme si řekli o sigmoidě víme, že řešením této rovnice je

$$\boldsymbol{w}^T \boldsymbol{x} = w_0 x_0 + w_1 x_1 + \cdots + w_p x_p = 0,$$

  což není nic jiného, než **nadrovina** (angl. **hyperplane**) v prostoru $\mathbb{R}^p$.
- Jazykem lineární algebry je to lineární varieta dimenze $p - 1$.

Logistická regrese: hranice rozhodnutí

- Ukázali jsme si, jak model logistické regrese vypadá a funguje, ale zatím nevíme, jak se učí, tj. jak se vybírá hodnota parametrů $\boldsymbol{w}$ na základě trénovacích dat, u kterých známe vedle příznaků i hodnoty Y.
- U modelů, kde odhadujeme přímo hodnotu proměnné Y, se obvykle postupuje tak, že se rozhodneme pro nějakou míru chyby a parametry vybíráme tak, abychom tuto chybu minimalizovali. Vzpomeňme na minulou přednášku o lineární regresi.
- U logistické regrese ale odhadujeme pravděpodobnosti hodnot proměnné Y; měřit chybu je tedy těžké.
- Musíme postupovat jinak!
- Pro znalé statistiky by stačilo říci, že parametry $\boldsymbol{w}$ odhadneme metodou MLE (**maximálně věrohodný odhad**, angl. **maximum likelihood estimate**).
- Jelikož ale znalost této metody zatím předpokládat nemůžeme, ukážeme si myšlenku MLE na jednoduchém příkladě.

MLE odhad pro házení mincí (1/4)

- Zapomeňme chvilku na logistickou regresi a uvažujme následující příklad.
- Máme minci kterou házíme a zaznamenáváme si, co nám padlo: 1 = hlava, 0 = orel. Označme si tuto náhodnou veličinu jako Y.
- Máme důvod si myslet, že mince není férová, tedy že je možné, že $P(Y = 1) \neq \frac{1}{2}$.
- Označme pravděpodobnost $P(Y = 1) = p$. Chceme na základě trénovacích dat nějak odhadnout hodnotu p.
- Hodíme desetkrát mincí a padne nám sedmkrát $Y = 1$ a třikrát $Y = 0$.
- Naučit model v tomto případě znamená určit hodnotu p, protože to je jediný parametr. Jak na to?

MLE odhad pro házení mincí (2/4)

- Pro každou hodnotu p umíme spočítat, s jakou pravděpodobností nám při deseti hodech padnou naše trénovací data. Např. pro férovou minci s $p = \frac{1}{2}$ je to[1]

$$\left(\frac{1}{2}\right)^{10} = 0.0009765625.$$

- Například pro $p = 0.6$ je ale tato pravděpodobnost vyšší:

$$0.6^7(1 - 0.6)^3 = 0.0017915904.$$

- Z toho důvodu bereme $p = 0.6$ jako lepší model našich trénovacích dat než $p = 0.5$: Pro $p = 0.6$ jsou totiž **trénovací data pravděpodobnější**!
- **Odhad metodou MLE pak odpovídá hodnotě p, pro která jsou trénovací data nejpravděpodobnější možná!**
- Jedná se tedy o optimalizaci (maximalizujeme pravděpodobnost), v tomto případě funkce jedné proměnné $p \in [0, 1]$ která udává pravděpodobnost trénovacích dat:

$$p^7(1 - p)^3.$$

[1] Viz Bernoulliho a geometrické rozdělení v BI-PST.

MLE odhad pro házení mincí (3/4)

- Hledáme tedy maximum funkce[2]

$$L(p) = p^7(1-p)^3$$

na intervalu $[0, 1]$.

- Jedná se o reálnou funkci jedné proměnné, takže můžeme funkci klasicky zderivovat a najít nulové body.
- Jedná se o polynom desátého stupně v součinovém tvaru, takže by derivace byla celkem pracná. Proto použijeme obvyklý trik a funkci zlogaritmujeme.
- Jelikož je logaritmus ostře rostoucí funkce, mají funkce

$$L(p) = p^7(1-p)^3 \quad \text{a} \quad \ell(p) = \ln p^7(1-p)^3 = 7\ln p + 3\ln(1-p)$$

extrémy ve stejných bodech.

[2]Sice je to pravděpodobnost, ale obvykle se značí písmenem L od slova likelihood (česky věrohodnost). Více viz BI-PST.

MLE odhad pro házení mincí (4/4)

- Derivace funkce $\ell(p)$ je ale mnohem „hezčí“ funkce:

$$\ell'(p) = \frac{7}{p} - \frac{3}{1-p}.$$

- Snadno spočítáme, že jediný nulový bod derivace je maximum

$$\hat{p}^{(MLE)} = \frac{7}{10}.$$

- Pro tuto hodnotu p je pravděpodobnost trénovacích dat

$$0.7^7(1-0.7)^3 = 0.0022235661,$$

  což je maximum možného, pro žádné jiné p není tato pravděpodobnost vyšší.

- Tato vlastnost dává volbě $p = 0.7$ důvěryhodnost! (Toto si rozmyslete!)

MLE odhad pro logistickou regresi (1/6)

- S logistickou regresí je to velmi podobné jako pro minci: Máme také jen dvě hodnoty $Y = 1$ a $Y = 0$.
- Máme také daný vzorec pro výpočet pravděpodobnosti; ten ovšem nezávisí na jediném parametru ale na $p + 1$ parametrech $w_0, w_1, \ldots, w_p$. Označme pro úsporu místa

$$p_1(\boldsymbol{x}, \boldsymbol{w}) = P(Y = 1 \mid \boldsymbol{x}, \boldsymbol{w}) = \frac{e^{\boldsymbol{w}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{w}^T \boldsymbol{x}}}$$

a

$$p_0(\boldsymbol{x}, \boldsymbol{w}) = P(Y = 0 \mid \boldsymbol{x}, \boldsymbol{w}) = 1 - \frac{e^{\boldsymbol{w}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{w}^T \boldsymbol{x}}} = \frac{1}{1 + e^{\boldsymbol{w}^T \boldsymbol{x}}}.$$

- Máme-li i-tý datový bod s hodnotou vysvětlované proměnné Y_i a s hodnotami příznaků $\boldsymbol{x}_i = (x_0, x_1, x_2, \ldots, x_p)$, lze pro zadané hodnoty parametrů $\boldsymbol{w}$ označit pravděpodobnost tohoto datového bodu jako

$$p_{Y_i}(\boldsymbol{x}_i, \boldsymbol{w}).$$

MLE odhad pro logistickou regresi (2/6)

- Předpokládejme, že máme N bodů v trénovacích datech, každý trénovací bod sestává z vysvětlované proměnné Y_i a hodnot příznaků

$$\boldsymbol{x}_i = (x_{i;0}, x_{i;1}, \ldots, x_{i;p}),$$

  kde $i = 1, \ldots, N$ a $x_{i;0} = 1$ (intercept).
- Tyto hodnoty zapíšeme do vektoru $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)^T \in \mathbb{R}^N$ a do matice

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{1;1} & x_{1;2} & \cdots & x_{1;p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N;1} & x_{N;2} & \cdots & x_{N;p} \end{pmatrix}.$$

- S tímto značením můžeme konečně napsat, jaká je pravděpodobnost konkrétních trénovacích dat při parametrech $\boldsymbol{w}$. Předpokládáme (většinou oprávněně), že jednotlivé datové body jsou navzájem nezávislé a pravděpodobnost tak lze napsat jako součin pravděpodobností jednotlivých bodů:

$$L(\boldsymbol{w}) = \prod_{i=1}^{N} p_{Y_i}(\boldsymbol{x}_i, \boldsymbol{w}).$$

MLE odhad pro logistickou regresi (3/6)

- Snažíme se tedy maximalizovat funkci

$$L(\boldsymbol{w}) = \prod_{i=1}^{N} p_{Y_i}(\boldsymbol{x}_i, \boldsymbol{w}),$$

  což je reálná funkce $p+1$ proměnných $\mathbb{R}^{p+1} \to \mathbb{R}$ vyjadřující **pravděpodobnost trénovacích dat**.

MLE odhad pro logistickou regresi (4/6)

- Podobně jako u hodu mincí budeme derivovat místo součinu pravděpodobností logaritmus:

$$\ell(\boldsymbol{w}) = \ln L(\boldsymbol{w}) = \sum_{i=1}^{N} \ln p_{Y_i}(\boldsymbol{x}_i, \boldsymbol{w}) =$$

$$= \sum_{i=1}^{N} \left(Y_i \ln p_1(\boldsymbol{x}_i, \boldsymbol{w}) + (1 - Y_i) \ln p_0(\boldsymbol{x}_i, \boldsymbol{w})\right) =$$

$$= \sum_{i=1}^{N} \left(Y_i \ln \left(\frac{e^{\boldsymbol{w}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{w}^T \boldsymbol{x}}} \right) + (1 - Y_i) \ln \left(\frac{1}{1 + e^{\boldsymbol{w}^T \boldsymbol{x}}} \right) \right) =$$

$$= \{\text{trocha čarování s logaritmem}\} =$$

$$= \sum_{i=1}^{N} \left(Y_i \boldsymbol{w}^T \boldsymbol{x} - \ln \left(1 + e^{\boldsymbol{w}^T \boldsymbol{x}} \right) \right).$$

MLE odhad pro logistickou regresi (5/6)

- Hledáme tedy maximum funkce

$$\ell(\boldsymbol{w}) = \sum_{i=1}^{N} \left(Y_i \boldsymbol{w}^T \boldsymbol{x} - \ln\left(1 + e^{\boldsymbol{w}^T \boldsymbol{x}}\right)\right).$$

- Stejně jako v případě lineární regrese, kde se minimalizoval součet čtverců residuí $RSS(\boldsymbol{w})$, se postupuje tak, že se najde gradient, tj. vektor složený z parciálních derivací podle všech proměnných $w_0, w_1, \ldots, w_p$:

$$\frac{\partial \ell}{\partial w_j}(\boldsymbol{w}) = \sum_{i=1}^{N} \left(x_{i;j}(Y_i - p_1(\boldsymbol{x}_i, \boldsymbol{w})\right), \quad j = 0, 1, \ldots, p.$$

- Pomocí maticového násobení lze pak gradient napsat ve tvaru

$$\nabla \ell(\boldsymbol{w}) = \mathbf{X}^T(\boldsymbol{Y} - \boldsymbol{P}), \quad \text{kde } \boldsymbol{P} = (p_1(\boldsymbol{x}_1, \boldsymbol{w}), p_1(\boldsymbol{x}_2, \boldsymbol{w}), \ldots, p_1(\boldsymbol{x}_N, \boldsymbol{w}))^T.$$

MLE odhad pro logistickou regresi (6/6)

- Teorie říká[3], že maximum bychom měli nalézt mezi řešeními rovnice „gradient se rovná nule“, tedy
$$\nabla \ell(\boldsymbol{w}) = \mathbf{X}^T(\boldsymbol{Y} - \boldsymbol{P}) = 0.$$

- Tato rovnice nevypadá nijak strašně, dokud si neuvědomíme, co se skrývá pod nevinným označeným $\boldsymbol{P}$: vektor plný sigmoid se všemi těmi exponenciálami.

- **Na rozdíl od lineární regrese neumíme najít explicitní řešení**, tedy neexistuje vzorec, do kterého bychom dosadili trénovací data a vypadly by nám z něho hodnoty koeficientů $\boldsymbol{w}$.

- **Je tedy nutné použít numerické aproximativní metody**.

- Používají se buď vícerozměrná verze Newtonovy metody, nebo gradientní vzestup: v obou případech se konstruuje posloupnost $\boldsymbol{w}_0, \boldsymbol{w}_1, \boldsymbol{w}_2, \ldots$ o které lze předpokládat, že konverguje k nějakému lokálnímu maximu[4].

[3] Více se dozvíte v kurzu MI-MPI nebo BI-VMM

[4] Pro logistickou regresi lze ukázat, že jediné lok. maximum je to hledané globální.

Různé poznámky

- Logistická regrese je přímým použitím metod parametrické statistiky, o kterých budete mluvit podrobně v BI-PST[5].
- Celá metoda stojí na předpokladu, že se chování dat dá zachytit ve tvaru daném sigmoidou jakožto funkcí $w_0 + w_1 x_1 + \cdots + w_p x_p$.
- Díky parametrům $w_0, w_1, w_2, \ldots, w_p$ má tento model poměrně dost volnosti, ale jestli tato volnost stačí k tomu, aby se model mohl dostatečně přiblížit ke skutečnosti, je důležitá a těžko zodpověditelná otázka, kterou je třeba mít na paměti.
- Logistická regrese je výpočetně náročná a příslušný odhad není dán explicitně: To co nám vrátí počítač je tedy aproximace a je i možné, že nám nevrátí nic.
- Stejně jako u regrese a jiných modelů můžeme možnosti modelu značně rozšířit, použijeme-li i odvozené příznaky od příznaků dostupných v datech (druhé a vyšší mocniny, součiny více různých příznaků, atp.).

[5] Tak dávejte pozor!