1 Teória strojového učenia I

Chceme sa naučiť na základe nejakých vstupných dát x predikovať y. Môžeme si to predstaviť tak, že príroda vie poskytovať pozorovania, každé v tvare dvojice (x, y). Dostali sme od nej sadu t pozorovaní, na základe ktorých chceme navrhnúť nejakú funkciu h, ktorá predpovedá y na základe x. Dobrá funkcia je taká, ktorá je schopná *zovšeobecňovať*, teda sa jej "dobré darí" aj na dátach mimo trénovacej množiny. Proces, ktorým h zostrojíme, si môžeme predtaviť ako algoritmus, ktorý berie ako vstupy trénovacie dáta a vráti nám funkciu.

1.1 Matematický model

Z matematického hľadiska, prírodu vieme formalizovať ako pravdepodobnostnú distribúciu P. Množinu všetkých možných x označíme X, množinu možných y označíme Y.

V tejto časti sa nebudeme zaoberať výpočtovou stránkou strojového učenia, od detailov ako časová zložitosť, ..., abstrahujeme. Algoritmus si teda predstavíme iba ako niečo, čo vezme ako vstup trénovacie dáta $(x_1, y_1), \ldots, (x_t, y_t)$ a na výstup vráti funkciu $h : X \to Y$. Túto funkciu budeme volať *hypotéza*. Množinu všetkých možných funkcii, ktoré môže náš algoritmus vrátiť, budeme volať *množina hypotéz* a značiť ho H.

Chyba hypotézy. Ako vyjadriť mieru toho, že sa funkcii "dobre darí"? Spravíme tak pomocou *chybovej funkcie* $\text{err} : Y \times Y \to \mathbb{R}^+$, ktorej význam je nasledovný: $\text{err}(y, y')$ vyjadruje, ako veľmi sa od seba líšia y a y'. Pomocou tejto funkcie vieme odmerať priemernú chybu hypotézy h, ktorú budeme tiež označovať err, nasledovne:

$$\text{err}(h) = \mathop{\mathrm{E}}_{x,y}\left[\text{err}(h(x), y)\right]$$

Pod $\mathrm{E}_{x,y}$ sa rozumie stredná hodnota cez (x, y) z pravdepodobnostnej distribúcie P, teda $(x, y) \sim P$. Pri klasifikácii sa zvykne používať chybová funkcia

$$\text{err}(y, y') = \begin{cases} 0, & \text{ak } y = y' \\ 1, & \text{inak} \end{cases}$$

a potom zrejme

$$\mathop{\mathrm{E}}_{x,y}\left[\text{err}(h(x), y)\right] = \mathop{\mathrm{P}}_{x,y}\left(h(x) \neq y\right).$$

Pri regresii máme viacero možností, bežné voľby sú kvadratická chyba $(y - y')^2$ a absolútna chyba $|y - y'|$.

Chyba algoritmu. Ako vyjadriť chybu celého učiaceho algoritmu? Uvedomme si, že výstup algoritmu je závislý od trénovacích dát $T = \{(x_1, y_1), \ldots, (x_t, y_t)\}$, ktoré dostane. Takže výstupná funkcia je od nich závislá, budeme ju označovať $\hat{h}$. Potom priemerná chyba algoritmu (alebo inak *priemerná chyba priemernej hypotézy*), braná cez všetky možné vzorky trénovacích dát, je rovná

$$\mathop{\mathrm{E}}_{T}\left[\text{err}(\hat{h})\right] = \mathop{\mathrm{E}}_{T}\left[\mathop{\mathrm{E}}_{x,y}\left[\text{err}(\hat{h}(x), y)\right]\right].$$

Pod E_T sa rozumie stredná hodnota cez všetky možné t-tice trénovacích dát T, brané nezávisle z pravdepodobnostnej distribúcie P.

Trénovacia chyba. Pri vyššie uvedených chybách sme vždy merali vzhľadom na skutočnú distribúciu P. Môže nás ale zaujímať, aká je priemerná chyba hypotézy na trénovacích dátach T. Túto chybu budeme označovať $\text{err}_T(h)$, a vypočítame ju ako

$$\text{err}_T(h) = \mathop{\mathrm{E}}_{x_i,y_i}\left[\text{err}(h(x_i), y_i)\right] = \frac{1}{t} \cdot \sum_{i=1}^{t} \text{err}(h(x_i), y_i).$$

Priemerná trénovacia chyba z pohľadu algoritmu bude

$$\mathop{\mathrm{E}}_{T}\left[\mathrm{err}_T(\hat{h})\right].$$

V nasledujúcom texte budeme vynechávať premenné, cez ktoré prebiehajú stredné hodnoty, všade tam, kde budú zrejmé z kontextu.

1.2 Analýza veľkostí chýb

V tejto časti sa podrobnejšie pozrieme na to, ako závisia vyššie uvedené štatistiky (tj. priemerná testovacia a trénovacia chyba priemernej hypotézy) od veľkosti trénovacej množiny T a od veľkosti množiny hypotéz H.

V celej časti budeme predpokladať, že úloha je regresného charakteru a chyba sa meria ako kvadratická odchýlka, teda

$$\mathrm{err}(y, y') = (y - y')^2.$$

1.2.1 Teoretické limity.

Najprv sa ale pozrieme na teoretické limity toho, ako dobrá vôbec môže nejaká funkcia byť. Označme $h^\square$ najlepšiu možnú funkciu, nemusí byť nutne z H. Teda

$$h^\square = \arg\min_h (\mathrm{err}(h)) = \arg\min_h \left(\mathop{\mathrm{E}}_{x,y}\left[(h(x) - y)^2\right]\right).$$

Jediné obmedzenia kladené na h sú, že je to funkcia: pre každé x musí vrátiť vždy jednu a tú istú hodnotu. Distribúcia P ale nemusí pre dané x vždy vrátiť to isté y: môže byť zašumená, alebo jednoducho x neobsahuje dostatočnú informáciu. Napríklad, ak podľa plochy bytu určujeme jeho cenu, niektoré dva byty môžu mať rovnakú plochu a predsa rôznu cenu. Ako uvidíme, tento nedeterminizmus je jediný dôvod, prečo hypotéza $h^\square$ nemusí mať nulovú chybu.

Chybu ľubovoľnej hypotézy h vieme upraviť nasledovne:

$$\begin{aligned} \mathrm{err}(h) &= \mathop{\mathrm{E}}_{x,y}\left[(h(x) - y)^2\right] &(1)\\ &= \mathop{\mathrm{E}}_{x}\left[\mathop{\mathrm{E}}_{y|x}\left[(h(x) - y)^2\right]\right] &(2)\end{aligned}$$

Pozrime sa na vnútornú strednú hodnotu. V nej je x konštanta, a teda aj $h(x) = c$ je konštanta. Aká konštanta minimalizuje danú strednú hodnotu? Nie je ťažké vidieť (napríklad zderivovaním), že minimum sa nadobúda pre $c = \mathrm{E}[y]$. Takže

$$h^\square(x) = \mathop{\mathrm{E}}_{y|x}[y],$$

a jeho priemerná chyba je

$$\mathrm{err}(h^\square) = \mathop{\mathrm{E}}_{x}\left[\mathop{\mathrm{E}}_{y|x}[(y - \mathrm{E}[y])]\right] = \mathop{\mathrm{E}}_{x}\left[\mathop{\mathrm{Var}}_{y|x}(y)\right].$$

Vidíme teda, že pokiaľ je y jednoznačne určené x-om, tak $h^\square$ bude mať nulovú chybu.

1.2.2 Bias-variance tradeoff, verzia 1.

V tomto odseku si ukážeme zaujímavý výsledok, ktorý nám za určitých predpokladov umožňuje vyjadriť chyby pomocou iných, jasnejších veličín: tzv. *výchylky* a *rozptylu*.

Odvodenie. Označme najlepšiu hypotézu z množiny H ako $h^\star$, teda

$$h^\star = \arg\min_h (\mathrm{err}(h)) .$$

Budeme upravovať výraz reprezentujúci priemernú chybu priemernej hypotézy $\hat{h}$.

$$\begin{aligned}
\text{chyba algoritmu} &= \mathop{\mathrm{E}}_T \left[\mathrm{err}(\hat{h})\right] && (3)\\
&= \mathop{\mathrm{E}}_T \left[\mathop{\mathrm{E}}_{x,y} \left[(\hat{h}(x) - y)^2\right]\right] && (4)\\
&= \mathop{\mathrm{E}}_T \left[\mathop{\mathrm{E}}_{x,y} \left[\left((\hat{h}(x) - h^\star(x)) + (h^\star(x) - y)\right)^2\right]\right] && (5)
\end{aligned}$$

V tomto momente prichádza netriviálny technický krok, ktorý si vyžaduje dodatočné predpoklady. Tieto technické detaily prenecháme na koniec časti, sústreďme sa na to hlavné.

$$\text{chyba algoritmu} = \mathop{\mathrm{E}}_T \left[\mathop{\mathrm{E}}_{x,y} \left[(\hat{h}(x) - h^\star(x))^2\right]\right] + \mathop{\mathrm{E}}_T \left[\mathop{\mathrm{E}}_{x,y} \left[(h^\star(x) - y)^2\right]\right]$$

Druhý zo sčítancov sa dá ešte zjednodušiť. Kedže $h^\star$ ani y nezávisia od trénovacích dát, môžeme sa zbaviť vonkajšej strednej hodnoty. Dostávame tak výslednú rovnosť

$$\text{chyba algoritmu} = \underbrace{\mathop{\mathrm{E}}_T \left[\mathop{\mathrm{E}}_{x,y} \left[(\hat{h}(x) - h^\star(x))^2\right]\right]}_{\text{rozptyl}} + \underbrace{\mathop{\mathrm{E}}_{x,y} \left[(h^\star(x) - y)^2\right]}_{\text{výchylka}} .$$

Prvý zo sčítancov budeme volať *rozptyl*. Vyjadruje, ako ďaleko je naša funkcia od najlepšej možnej, vrámci množiny hypotéz H. Druhý zo sčítancov budeme volať *výchylka*. Vyjadruje chybu, ktorá je spôsobená výberom množiny hypotéz.

Výchylku vieme upraviť ďalej. Pretože hypotéza $h^\star$ ani y nezávisia od trénovacej množiny T, merať chybu na testovacích dátach x, y je to isté, ako merať ju na trénovacích dátach x_i, y_i, berúc ich náhodný výber. Teda

$$\begin{aligned}
\text{výchylka} &= \mathop{\mathrm{E}}_T \left[\mathop{\mathrm{E}}_{x_i,y_i} \left[(h^\star(x_i) - y_i)^2\right]\right] && (6)\\
&= \mathop{\mathrm{E}}_T \left[\mathop{\mathrm{E}}_{x_i,y_i} \left[\left((h^\star(x_i) - \hat{h}(x_i)) + (\hat{h}(x_i) - y_i)\right)^2\right]\right] && (7)
\end{aligned}$$

Opäť, použitím toho istého technického kroku dostaneme:

$$\text{výchylka} = \underbrace{\mathop{\mathrm{E}}_T \left[\mathop{\mathrm{E}}_{x_i,y_i} \left[(h^\star(x_i) - \hat{h}(x_i))^2\right]\right]}_{\text{trénovací rozptyl}} + \underbrace{\mathop{\mathrm{E}}_T \left[\mathop{\mathrm{E}}_{x_i,y_i} \left[(\hat{h}(x_i) - y_i)^2\right]\right]}_{\text{priemerná trénovacia chyba}} \qquad (8)$$

Trénovací rozptyl vyjadruje, ako ďaleko je naša hypotéza $\hat{h}$ od najlepšej možnej $h^\star$ z H. Na rozdiel od rozptylu ale túto vzdialenosť meriame na trénovacích dátach, nie na testovacích. To spraví rozdiel, nakoľko $\hat{h}$ je závislé od trénovacích dát. Priemerná trénovacia chyba je priemerná chyba, ktorej sa dopustí výstup z algoritmu $\hat{h}$ na tých istých dátach, pomocou ktorých sme $\hat{h}$ zostrojili.

Závery. Podarilo sa nám teda rozložiť chybu algoritmu na dve, prípadne tri časti. Načo je to ale dobré? Ukážeme si, ako pomocou nich vieme získať intuíciu o tom, ako sa správa chyba algoritmu v závislosti od veľkosti trénovacej množiny a veľkosti (tj. zložitosti) množiny hypotéz.
TODO obrázok kriviek učenia, vysvetlenie
TODO podučenie, preučenie

	vektory	funkcie
súradnice	$y_1, y_2, \dots, y_n$	$f(x)$
dĺžka	$\lVert y \rVert = \sqrt{\sum_{i=1}^n y_i^2}$	$\lVert f \rVert = \sqrt{\int f^2(x)\ d\rho x} = \sqrt{\mathrm{E}_x\left[f^2(x)\right]}$
skalárny súčin	$\langle y, w \rangle = \sum_{i=1}^n y_i \cdot w_i$	$\langle f, g \rangle = \int f(x) \cdot g(x)\ d\rho x = \mathrm{E}_x\left[f(x) \cdot g(x)\right]$
vzdialenosť	$d(y, w) = \lVert y - w \rVert$	$d(f, g) = \lVert(\lVert f - g)$

Tabuľka 1: Vektory/funkcie.

Technické detaily. Nakoniec sa vyjadríme k spomínanému technickému kroku. Začneme jeho znením a potom uvedieme jeho predpoklady.

Po prvé, pre jednoduchosť budeme predpokladať, že vstup je vektor reálnych čísel (tj. $X = \mathbb{R}^n$), snažíme sa predpovedať jedno reálne číslo (tj. $Y = \mathbb{R}$), a že pravdepodobnostné rozdelenie P je spojité. Jeho hustotu pravdepodobnosti označíme ρ.

Po druhé, predpokladáme, že trénovací algoritmus vráti takú funkciu $\hat{h} \in H$, ktorá minimalizuje trénovaciu chybu. Inak zapísané,

$$\hat{h} = \arg\min_{h \in H} \left(\mathrm{err}_T(h)\right).$$

Po tretie, kladieme obmedzenia na množinu hypotéz H: musí byť uzavretá na lineárne kombinácie a na limity. Dá sa na ňu teda pozerať ako na (nie nutne konečnorozmerný) vektorový priestor, ktorý je navyše uzavretý: ak postupnosť vektorov (v našom prípade funkcii) konverguje, tak jej limita je tiež v tom priestore.

V tabuľke 1 uvádzame na jednej strane vlastnosti konečnorozmerných vektorov, na druhej strane vlastnosti funkcii ako vektorov.

TODO dokončiť dôkaz

1.2.3 Bias-variance tradeoff, verzia 2.

Ak ste si dali vyhľadávať tento pojem, určite ste narazili na kopu materiálov, v ktorých sa výsledok vôbec nepodobal na to, čo sme ukazovali vyššie. Ide totiž o veľmi príbuzný, avšak predsa odlišný výsledok, ukážeme a odvodíme si ho.

Veta 1. *Nech $y : X \to \mathbb{R}$ je funkcia, ktorú sa snažíme modelovať. Predpokladajme, že sa dá rozložiť na časti: $y = f(x) + \varepsilon$, kde ε hrá rolu šumu: je nezávislý od všetkého a $\mathrm{E}[\varepsilon] = 0$. Označíme jeho pravdepodobnostnú distribúciu E.*

Nech výstupom trénovacieho algoritmu je $\hat{f}$. Za chybovú funkciu zvoľme kvadratickú chybu. Chybu algoritmu vieme teda vypočítať nasledovne:

$$\text{chyba algoritmu} = \mathop{\mathrm{E}}_{(x,y)\sim P, T\sim P^t, \varepsilon\sim E}\left[(\hat{f}(x) - y)^2\right].$$

Tvrdíme, že sa dá rozložiť na tri nasledovné časti:

$$\text{chyba algoritmu} = \underbrace{\mathrm{Var}(\hat{f}(x) - f(x))}_{\text{rozptyl}} + \underbrace{(\mathrm{E}[\hat{f}(x)] - \mathrm{E}[f(x)])^2}_{\text{výchylka}^2} + \underbrace{\mathrm{Var}(\varepsilon)}_{\text{šum}}$$

Poznámka 1. V poslednej rovnici sme kvôli stručnosti vynechali pri stredných hodnotách a rozptyloch premenné a distribúcie, z ktorých ich berieme. V dôkaze budeme vždy brať všetky premenné z ich príslušných distribúcii.

Poznámka 2. Funkcia f hrá v podstate tú istú rolu, čo najlepšia možná hypotéza spomedzi všetkých funkcii (nielen tých v množine hypotéz), $h^{\square}$.

Poznámka 3. V tomto znení bias-variance tradeoff-u názvy *rozptyl* a *výchylka* zodpovedajú príslušným štatistickým/pravdepodobnostným pojmom.

Poznámka 4. Na rozdiel od predchádzajúcej verzie bias-variance tradeoff-u, tu nebudeme potrebovať žiadne dodatočné predpoklady od algoritmu ani od jeho množiny hypotéz. (Nemusí teda vracať hypotézu, ktorá je spomedzi hypotéz v H najlepšia na daných trénovacích dátach. Takisto od množiny hypotéz nepožadujeme žiadne vlastnosti.)

Dôkaz. Upravujme pôvodný výraz.

$$\begin{aligned}
\text{chyba algoritmu} &= \mathrm{E}\left[(\hat{f}(x) - y)^2\right] && (9)\\
&= \mathrm{E}\left[(\hat{f}(x) - f(x) - \varepsilon)^2\right] && (10)\\
&= \mathrm{E}\left[(\hat{f}(x) - f(x))^2\right] + \mathrm{E}\left[\varepsilon^2\right] - 2 \cdot \mathrm{E}\left[\varepsilon \cdot (\hat{f}(x) - f(x))\right] && (11)\\
&= \mathrm{E}\left[(\hat{f}(x) - f(x))^2\right] + \mathrm{E}\left[\varepsilon^2\right] && (12)
\end{aligned}$$

Výraz sme upravili, roznásobili a využili linearitu strednej hodnoty. V poslednom kroku sme použili $\mathrm{E}[ab] = \mathrm{E}[a]\cdot\mathrm{E}[b]$, ktorý platí pre ľubovoľné nezávislé premenné, s $a := \varepsilon$, $b := \hat{f}(x) - f(x)$. Zamerajme sa ďalej na prvý sčítanec.

$$\begin{aligned}
\text{prvý sčítanec} &= \mathrm{E}\left[(\hat{f}(x) - f(x))^2\right] && (13)\\
&= \mathrm{E}[\hat{f}(x)^2] + \mathrm{E}[f(x)^2] - 2 \cdot \mathrm{E}[\hat{f}(x) \cdot f(x)] && (14)\\
&= (\mathrm{Var}(\hat{f}(x)) + \mathrm{E}[\hat{f}(x)]^2) + (\mathrm{Var}(f(x)) + \mathrm{E}[f(x)]^2) - 2 \cdot \mathrm{E}[\hat{f}(x) \cdot f(x)] && (15)
\end{aligned}$$

V poslednom kroku sme využili vzťah $\mathrm{Var}(a) = \mathrm{E}[a^2] - \mathrm{E}[a]^2$. Pokračujme ďalej v úpravách.

$$\begin{aligned}
\text{prvý sčítanec} &= \mathrm{Var}(\hat{f}(x)) + \mathrm{Var}(f(x)) + (\mathrm{E}[\hat{f}(x)] - \mathrm{E}[f(x)])^2 \\
&\quad + 2 \cdot \mathrm{E}[\hat{f}(x)] \cdot \mathrm{E}[f(x)] - 2 \cdot \mathrm{E}[\hat{f}(x) \cdot f(x)] && (16)\\
&= \mathrm{Var}(\hat{f}(x)) + \mathrm{Var}(f(x)) + (\mathrm{E}[\hat{f}(x)] - \mathrm{E}[f(x)])^2 - 2 \cdot \mathrm{Cov}(\hat{f}(x), f(x)) && (17)\\
&= \mathrm{Var}(\hat{f}(x) - f(x)) + (\mathrm{E}[\hat{f}(x)] - \mathrm{E}[f(x)])^2 && (18)
\end{aligned}$$

Využili sme najprv vzťah $\mathrm{Cov}(a, b) = \mathrm{E}[ab] - \mathrm{E}[a] \cdot \mathrm{E}[b]$, a potom $\mathrm{Var}(a - b) = \mathrm{Var}(a) + \mathrm{Var}(b) - 2 \cdot \mathrm{Cov}(a, b)$. Keď to teda celé dáme do jednej rovnice, dostaneme

$$\text{chyba algoritmu} = \underbrace{\mathrm{Var}(\hat{f}(x) - f(x))}_{\text{rozptyl}} + \underbrace{(\mathrm{E}[\hat{f}(x)] - \mathrm{E}[f(x)])^2}_{\text{výchylka}^2} + \underbrace{\mathrm{Var}(\varepsilon)}_{\text{šum}}$$

□

1.3 Ako sa vysporiadať s preučením/podučením?

TODO regularizácia
TODO holdout testing
TODO k-fold cross validation
TODO best practices