

# Obsah

<b>1</b>	<b>Teória strojového učenia I</b>	<b>2</b>
1.1	Matematický model . . . . .	2
1.2	Analýza veľkostí chýb . . . . .	3
1.2.1	Teoretické limity . . . . .	3
1.2.2	Bias-variance tradeoff . . . . .	4
1.2.3	Bias-variance tradeoff, verzia 2. . . . .	7
1.3	Ako sa vysporiadať s preučeníím/podučeníím? . . . . .	8

# Kapitola 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 tréningovej množiny. Proces, ktorým  $h$  zostrojíme, si môžeme predstaviť ako algoritmus, ktorý berie ako vstupy tréningové 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 tréningové dáta  $(x_1, y_1), \dots, (x_t, y_t)$  a na výstup vráti funkciu  $h : X \rightarrow Y$ . Túto funkciu budeme volať *hypotéza*. Množinu všetkých možných funkcií, ktoré môže náš algoritmus vrátiť, budeme volať *množina hypotéz* a značiť ho  $H$ .

**Chyba hypotézy.** Ako vyjadriť mieru toho, že sa funkcií “dobré darí”? Správime tak pomocou *chybovej funkcie*  $\text{err} : Y \times Y \rightarrow \mathbb{R}^+$ , 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  $\text{err}$ , nasledovne:

$$\text{err}(h) = \mathbb{E}_{x,y} [\text{err}(h(x), y)]$$

Pod  $\mathbb{E}_{x,y}$  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

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

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 tréningových dát  $T = \{(x_1, y_1), \dots, (x_t, y_t)\}$ , 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 tréningových dát, je rovná

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

Pod  $\mathbb{E}_T$  sa rozumie stredná hodnota cez všetky možné  $t$ -tice tréningových dát  $T$ , brané nezávisle z pravdepodobnostnej distribúcie  $P$ .

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

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

Priemerná tréningová chyba z pohľadu algoritmu bude

$$\mathbb{E}_T [\text{err}_T(\hat{h})].$$

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 tréningová chyba priemernej hypotézy) od veľkosti tréningovej 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

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

### 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

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

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:

$$\text{err}(h) = \mathbb{E}_{x,y} [(h(x) - y)^2] \tag{1.1}$$

$$= \mathbb{E}_x \left[ \mathbb{E}_{y|x} [(h(x) - y)^2] \right] \tag{1.2}$$

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  $c = E[y]$ . Takže

$$h^\square(x) = E_{y|x}[y],$$

a jeho priemerná chyba je

$$\text{err}(h^\square) = E_x \left[ E_{y|x} [(y - E[y])^2] \right] = E_x \left[ \text{Var}(y) \right].$$

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

### 1.2.2 Bias-variance tradeoff

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  $h^*$ , teda

$$h^* = \arg \min_h (\text{err}(h)).$$

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

$$\text{chyba algoritmu} = E_T [\text{err}(\hat{h})] \quad (1.3)$$

$$= E_T \left[ E_{x,y} [(\hat{h}(x) - y)^2] \right] \quad (1.4)$$

$$= E_T \left[ E_{x,y} \left[ \left( (\hat{h}(x) - h^*(x)) + (h^*(x) - y) \right)^2 \right] \right] \quad (1.5)$$

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

$$\text{chyba algoritmu} = E_T \left[ E_{x,y} [(\hat{h}(x) - h^*(x))^2] \right] + E_T \left[ E_{x,y} [(h^*(x) - y)^2] \right]$$

Druhý zo sčítancov sa dá ešte zjednodušiť. Keďže  $h^*$  ani  $y$  nezávisia od tréningových dát, môžeme sa zbaviť vonkajšej strednej hodnoty. Dostávame tak výslednú rovnosť

$$\text{chyba algoritmu} = \underbrace{E_T \left[ E_{x,y} [(\hat{h}(x) - h^*(x))^2] \right]}_{\text{rozptyl}} + \underbrace{E_{x,y} [(h^*(x) - y)^2]}_{\text{výchylka}}.$$

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  $h^*$  ani  $y$  nezávisia od tréningovej množiny  $T$ , merať chybu na testovacích dátach  $x, y$  je to isté, ako merať ju na tréningových dátach  $x_i, y_i$ , berúc ich náhodný výber. Teda

$$\text{výchylka} = E_T \left[ E_{x_i, y_i} [(h^*(x_i) - y_i)^2] \right] \quad (1.6)$$

$$= E_T \left[ E_{x_i, y_i} \left[ \left( (h^*(x_i) - \hat{h}(x_i)) + (\hat{h}(x_i) - y_i) \right)^2 \right] \right] \quad (1.7)$$

Opäť, použitím toho istého technického kroku dostaneme:

$$\text{výchylka} = \underbrace{\mathbb{E}_T \left[ \mathbb{E}_{x_i, y_i} \left[ (h^*(x_i) - \hat{h}(x_i))^2 \right] \right]}_{\text{trénovací rozptyl}} + \underbrace{\mathbb{E}_T \left[ \mathbb{E}_{x_i, y_i} \left[ (\hat{h}(x_i) - y_i)^2 \right] \right]}_{\text{priemerná trénovacia chyba}} \quad (1.8)$$

Trénovací rozptyl vyjadruje, ako ďaleko je naša hypotéza  $\hat{h}$  od najlepšej možnej  $h^*$  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.

**Záver.** 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

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

**Veta 1.** *Predpokladajme, že vstupom do hypotéz sú vektory reálnych čísel (tj.  $X = \mathbb{R}^n$ ), cieľom je predpovedať jedno reálne číslo (tj.  $Y = \mathbb{R}$ ), a že pravdepodobnostné rozdelenie  $P$  je spojité.*

*Nech množina hypotéz  $H$  je uzavretá na lineárne kombinácie a na limity (teda ak postupnosť funkcií v  $H$  konverguje, jej limita je tiež v  $H$ ).*

*Ďalej predpokladajme, že trénovací algoritmus vždy vráti takú funkciu  $\hat{h} \in H$ , ktorá minimalizuje trénovacia chybu. Inak zapísané,*

$$\hat{h} = \arg \min_{h \in H} \left( \mathbb{E}_T [\text{err}_T(h)] \right).$$

*Potom platí*

$$\mathbb{E}_T \left[ \mathbb{E}_{x, y} \left[ \left( (\hat{h}(x) - h^*(x)) + (h^*(x) - y) \right)^2 \right] \right] = \mathbb{E}_T \left[ \mathbb{E}_{x, y} \left[ (\hat{h}(x) - h^*(x))^2 \right] \right] + \mathbb{E}_T \left[ \mathbb{E}_{x, y} \left[ (h^*(x) - y)^2 \right] \right]$$

*Poznámka 1.* Dokazovaná rovnosť je ekvivalentná s nasledovnou, stručnejšou:

$$\mathbb{E}_T \left[ \mathbb{E}_{x, y} \left[ (\hat{h}(x) - h^*(x)) \cdot (h^*(x) - y) \right] \right] = 0.$$

Túto kratšiu verziu získame roznásobením a použitím linearity strednej hodnoty. V dôkaze budeme dokazovať túto rovnosť.

*Poznámka 2.* Všimnite si, že potrebujeme uzavretosť množiny  $H$  na limity na to, aby bolo  $\arg \min_{h \in H} (\dots)$  dobre definované. Vo všeobecnosti nemusí existovať taká funkcia, ale môže existovať nekonečná postupnosť funkcií, každá ďalšia lepšia, ako tá predchádzajúca. (Inak povedané, neexistuje minimum, iba infimum.)

*Poznámka 3.* Veta by sa dala rozšíriť aj na iné množiny  $X, Y$ , napríklad keď predpovedaná premenná je vektor ( $Y = \mathbb{R}^m$ ), ... Možno ani  $P$  nemusí byť spojitá. Pre jednoduchosť argumentu ale budeme uvažovať vetu tak, ako je popísaná vyššie.

*Poznámka 4.* Predpoklady vety sú značne obmedzujúce. Napríklad si uvedomte, že ju nie je možné použiť na klasifikáciu, či dokonca ani na ľubovoľnú ohraničenú regresiu (kde rozumné hodnoty  $y$  sú ohraničené). Ale taká je teória.

Pri našom dôkaze využijeme niekoľko vlastností funkcií, ktoré uvádzame v nasledujúcom odseku. Skúsený čitateľ-matematik ho môže preskočiť.

**Definícia 1.** (Skalárny súčin.) Nech  $f, g$  sú funkcie z  $X$  do  $\mathbb{R}$ , z nejakej príjemne sa správajúcej množiny funkcií (tj. rovnomerne spojitých, ..., čokoľvek, aby nasledujúce argumenty prešli). Definujeme ich skalárny súčin  $\langle \cdot, \cdot \rangle$  ako

$$\langle f, g \rangle = \int f(x) \cdot g(x) d\rho x \quad (1.9)$$

$$= \mathbb{E}_x [f(x) \cdot g(x)], \quad (1.10)$$

kde  $\rho$  je hustota pravdepodobnosti distribúcie  $P$ . Rozmyslite si, že takto definovaný skalárny súčin má všetky vlastnosti, ktoré sa bežne požadujú od skalárnych súčinov:

- Je symetrický od svojich argumentov, teda  $\langle f, g \rangle = \langle g, f \rangle$ .
- Je lineárny:  $\langle f, g + h \rangle = \langle f, g \rangle + \langle f, h \rangle$  a tiež  $\langle k \cdot f, g \rangle = k \cdot \langle f, g \rangle$ .
- $\langle f, f \rangle \geq 0$  pre ľubovoľné  $f$ , pričom rovnosť nastáva práve vtedy, keď je  $f$  konštantne nulové.

**Definícia 2.** (Kolmost'.) Dve funkcie  $f, g$  sú na seba kolmé, ak ich skalárny súčin je 0. Značíme  $f \perp g$ .

**Definícia 3.** (Norma.) Podľa skalárneho súčinu definujeme normu funkcie (jej "dĺžku"):

$$\|f\| = \sqrt{\langle f, f \rangle} = \sqrt{\mathbb{E}_x [f^2(x)]}$$

Splňa *trojuholníkovú nerovnosť*: pre ľubovoľné funkcie  $f, g$  platí

$$\|f\| + \|g\| \geq \|f + g\|.$$

Definuje nám teda (euklidovskú) metriku nad funkciami, podľa ktorej definujeme limity a konvergenciu.

**Lemma 1.** (Pytagorova veta.) Nech  $f \perp g$ . Potom platí:

$$\|f\|^2 + \|g\|^2 = \|f + g\|^2$$

*Dôkaz.* Pozrime sa na pravú stranu. Iba v nej zapíšeme normu ako skalárny súčin a využijeme jeho linearitu a symetriu:

$$\|f + g\|^2 = \langle f + g, f + g \rangle \quad (1.11)$$

$$= \langle f, f \rangle + \langle g, g \rangle + 2 \cdot \langle f, g \rangle \quad (1.12)$$

Pretože  $f \perp g$ , posledný sčítanec je nulový, čím dostávame dokazované tvrdenie.  $\square$

**Definícia 4.** (Projekcia na množinu.) Nech  $H$  je množina funkcií, ktorá je uzavretá na lineárne kombinácie a na limity, a nech  $f$  je funkcia. *Projekciu* funkcie  $f$  na množinu  $H$  budeme označovať  $f_H$  a budeme pod ňou rozumieť nasledovný výraz:

$$f_H = \arg \min_{h \in H} d(f, h)$$

**Lemma 2.** (Kolmost' projekcie.) Pre ľubovoľnú funkciu  $h \in H$  platí  $h \perp f - f_H$ .

*Dôkaz.* Sporom, predpokladajme, že  $h \not\perp f - f_H$ . Takže  $\langle h, f - f_H \rangle \neq 0$ . Ukážeme, že potom existuje v  $H$  funkcia, ktorá je k funkcii  $f$  bližšie, ako funkcia  $f_H$ . To bude hľadaný spor s definíciou  $f_H$ .

Pozrime sa na všetky funkcie, ktoré ležia na priamke  $f_H + \Delta \cdot h$ . Tieto funkcie sú v množine  $H$ , pretože  $f_H, h \in H$  a množina  $H$  je uzavretá na lineárne kombinácie. Každú z týchto funkcií vieme asociovať s jedným reálnym číslom  $\Delta$ . Pozrime sa na ich vzdialenosti od funkcie  $f$ , vyjadrené ako funkcia od  $\Delta$ :

$$\text{dist}(\Delta) = d(f, f_H + \Delta \cdot h) \quad (1.13)$$

$$= \langle (f - f_H) + \Delta \cdot h, (f - f_H) + \Delta \cdot h \rangle \quad (1.14)$$

$$= \langle f - f_H, f - f_H \rangle + 2\Delta \cdot \langle h, f - f_H \rangle + \Delta^2 \cdot \langle h, h \rangle \quad (1.15)$$

Pozrime sa na deriváciu tejto funkcie. Podľa definície  $f_H$  by malo byť  $f - f_H$  najkratšie možné, teda pre  $\Delta = 0$  by mala funkcia  $\text{dist}$  nadobúdať minimum, a teda mať tam nulovú deriváciu. Uvidíme, že tomu tak nie je:

$$\frac{\partial \text{dist}}{\partial \Delta}(0) = \lim_{\Delta \rightarrow 0} \left( \frac{\text{dist}(\Delta) - \text{dist}(0)}{\Delta} \right) \quad (1.16)$$

$$= \lim_{\Delta \rightarrow 0} \left( \frac{2\Delta \cdot \langle h, f - f_H \rangle + \Delta^2 \cdot \langle h, h \rangle}{\Delta} \right) \quad (1.17)$$

$$= 2 \cdot \langle h, f - f_H \rangle \quad (1.18)$$

To je nenulové, nakoľko  $h \not\perp f - f_H$ . Čo je hľadaný spor.  $\square$

**TODO dokončiť dôkaz**

### 1.2.3 Bias-variance tradeoff, verzia 2.

V literatúre pod názvom *bias-variance tradeoff* vystupuje aj podobný, ale predsa odlišný výsledok, ako bolo uvedené vyššie. Ukážeme a odvodíme si ho.

**Veta 2.** *Nech  $y : X \rightarrow \mathbb{R}$  je funkcia, ktorú sa snažíme modelovať. Predpokladajme, že sa dá rozložiť na časti:  $y = f(x) + \varepsilon$ , kde  $\varepsilon$  hrá rolu šumu: je nezávislý od všetkého a  $\mathbb{E}[\varepsilon] = 0$ . Označíme jeho pravdepodobnostnú distribúciu  $E$ .*

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

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

*Tvrdíme, že sa dá rozložiť na tri nasledovné časti:*

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

*Poznámka 5.* 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 distribúcií.

*Poznámka 6.* Funkcia  $f$  hrá v podstate tú istú rolu, čo najlepšia možná hypotéza spomedzi všetkých funkcií (nielen tých v množine hypotéz),  $h^\square$ .

*Poznámka 7.* V tomto znení bias-variance tradeoff-u názvy *rozptyl* a *výchylka* zodpovedajú príslušným štatistickým/pravdepodobnostným pojmom.

*Poznámka 8.* Na rozdiel od predchádzajúcej verzie bias-variance tradeoff-u, tu nebudeme potrebovať žiadne dodatočné predpoklady od algoritmu ani od jeho množiny hypotéz. (Nemusí teda vracieť hypotézu, ktorá je spomedzi hypotéz v  $H$  najlepšia na daných tréningových dátach. Takisto od množiny hypotéz nepožadujeme žiadne vlastnosti.)

*Dôkaz.* Upravujeme pôvodný výraz.

$$\text{chyba algoritmu} = \mathbb{E} \left[ (\hat{f}(x) - y)^2 \right] \quad (1.19)$$

$$= \mathbb{E} \left[ (\hat{f}(x) - f(x) - \varepsilon)^2 \right] \quad (1.20)$$

$$= \mathbb{E} \left[ (\hat{f}(x) - f(x))^2 \right] + \mathbb{E} [\varepsilon^2] - 2 \cdot \mathbb{E} [\varepsilon \cdot (\hat{f}(x) - f(x))] \quad (1.21)$$

$$= \mathbb{E} \left[ (\hat{f}(x) - f(x))^2 \right] + \mathbb{E} [\varepsilon^2] \quad (1.22)$$

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

$$\text{prvý sčítanec} = \mathbb{E} \left[ (\hat{f}(x) - f(x))^2 \right] \quad (1.23)$$

$$= \mathbb{E}[\hat{f}(x)^2] + \mathbb{E}[f(x)^2] - 2 \cdot \mathbb{E}[\hat{f}(x) \cdot f(x)] \quad (1.24)$$

$$= (\text{Var}(\hat{f}(x)) + \mathbb{E}[\hat{f}(x)]^2) + (\text{Var}(f(x)) + \mathbb{E}[f(x)]^2) - 2 \cdot \mathbb{E}[\hat{f}(x) \cdot f(x)] \quad (1.25)$$

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

$$\text{prvý sčítanec} = \text{Var}(\hat{f}(x)) + \text{Var}(f(x)) + (\mathbb{E}[\hat{f}(x)] - \mathbb{E}[f(x)])^2 \quad (1.26)$$

$$+ 2 \cdot \mathbb{E}[\hat{f}(x)] \cdot \mathbb{E}[f(x)] - 2 \cdot \mathbb{E}[\hat{f}(x) \cdot f(x)] \quad (1.27)$$

$$= \text{Var}(\hat{f}(x)) + \text{Var}(f(x)) + (\mathbb{E}[\hat{f}(x)] - \mathbb{E}[f(x)])^2 - 2 \cdot \text{Cov}(\hat{f}(x), f(x)) \quad (1.27)$$

$$= \text{Var}(\hat{f}(x) - f(x)) + (\mathbb{E}[\hat{f}(x)] - \mathbb{E}[f(x)])^2 \quad (1.28)$$

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

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

□

### 1.3 Ako sa vysporiadať s preučeníím/podučením?

TODO regularizácia

TODO holdout testing

TODO  $k$ -fold cross validation

TODO best practices