0 Evaluace modelů

Jedna z klíčových aspektů strojového učení je, aby natrénovaný model byl schopný generalizace – fungovat dobře na nových vstupech.

Při trénování můžeme natrénovat celou řadu modelů (např. ze stejné rodiny ale s různými sadami hyperparametrů). Pro to, abychom mohli modely porovnat, potřebujeme mít nějakou kvantitativní míru výkonnosti – metriku. To jakou zvolíme často záleží na charakteru problému. Někdy chceme minimalizovat drobné chyby, někdy celkovou chybovost, apod.

0.1 Ztrátová funkce

Uvažujme model natrénovaný na vstupu $\boldsymbol{X}$ s vysvětlovanou proměnnou Y. Takový model zpravidla není dokonalý a pro $\boldsymbol{X}$ predikuje nějaké $\hat{Y} \equiv \hat{Y}(\boldsymbol{X})$. Funkci L, měřící chybu této predikce, nazýváme ztrátovou funkcí (loss function).

Regrese V případě regrese je typická kvadratická ztrátová funkce (squared error):

$$L(Y, \hat{Y}) = (Y - \hat{Y})^2,$$

případně L_1 ztrátová funkce měřící absolutní chybu (absolute error):

$$L(Y, \hat{Y}) = |Y - \hat{Y}|.$$

Klasifikace U binární klasifikace se často odhaduje pravděpodobnost

$$\hat{p} = \hat{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}),$$

pro kterou se nabízí následující ztrátová funkce (binary cross-entropy loss function)

$$L(Y, \hat{Y}) = -Y \log \hat{p} - (1 - Y) \log(1 - \hat{p})$$

0.1.1 Trénovací chyba

Při trénování se snažíme minimalizovat průměrnou hodnotu ztrátové chyby přes všechny prvky v trénovací množině (trénovací chybu, test error):

$$\overline{\mathrm{err}}_{\mathrm{train}} = \mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} L(Y_i, \hat{Y}(\boldsymbol{x}_i))$$

V případě regrese se tedy bude jednat například o

$$\mathrm{MSE}_{\mathrm{train}} = \frac{1}{N} \sum_{i=1}^{N} (Y - \hat{Y})^2 \quad \text{nebo} \quad \mathrm{MAE}_{\mathrm{train}} = \frac{1}{N} \sum_{i=1}^{N} |Y - \hat{Y}|$$

A u binární klasifikaci o

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}[-Y_i \log \hat{p}(\boldsymbol{x}_i) - (1 - Y_i)\log(1 - \hat{p}(\boldsymbol{x}_i))]]$$

Řešení minimalizující trénovací chybu (test error) lze u některých modelů spočíst explicitně (closed-form solution), např. u lin. regrese. Pro většinu modelů to nelze a musíme jej počítat numericky iterativními metodami (např. gradientním sestupem).

0.1.2 Testovací chyba

Testovací chyba (test error) je střední chyba na novém vstupu $\boldsymbol{X}$ při dané trénovací množině $\mathcal{D}$:

$$\mathrm{Err}_{\mathcal{D}} = \mathrm{E}(L(Y, \hat{Y}(\boldsymbol{X})) \mid \mathcal{D})$$

kterou můžeme odhadnout

$$\overline{\mathrm{err}}_{\mathrm{test}} = \frac{1}{N_{\mathrm{test}}}\sum_{i=1}^{N_{\mathrm{test}}}(L(Y_i, \hat{Y}(\boldsymbol{x}_i))$$

Nejobecnější mírou schopnosti modelu generalizovat je očekávaná testovací chyba (expected test error), která je střední hodnotou testovací chyby pro náhodný výběr trénovací množiny $\mathcal{D}$:

$$\mathrm{Err} = \mathrm{E}(\mathrm{Err}_{\mathcal{D}}) = \mathrm{E}(L(Y, \hat{Y}(\boldsymbol{X})))$$

Jedná se tedy odhad $\mathrm{Err}_{\mathcal{D}}$ s neznámým $\mathcal{D}$.

0.2 Evaluační scénáře

Z pohledu evaluace máme dva úkoly:

- Výběr modelu - odhadnout chybu různých modelů za účelem výběru nejlepšího.
- Ohodnocení modelu - odhadnout testovací chybu finálního modelu.

0.2.1 Hold-out

Pokud máme dostatek dat, dělíme data na část

- Trénovací - k natrénování konkrétních modelů.
- Validační - k výběru nejlepší sady hyperparametrů.
- Testovací - k odhadu testovací chyby, kterou očekáváme na nových datech.

Pro získání neoptimistického odhadu výkonnosti modelu, musí být testovací část skutečně nový vstup, který nijak neovlivnil parametry ani volbu modelu.

0.2.2 k-fold cross-validation

Pokud nemáme dostatek dat, je často nerozumné je dělit na trénovací, validační a testovací. Vedlo by to na nedostatek dat pro správně natrénovaní, dobrou volbu správného modelu a spolehlivého odhadu testovací chyby.

S křížovou validací se obejdeme bez validační množiny.

Algorithm Cross-validation

Require: $2 \leq k \leq N$

1: Trénovací data $\mathcal{D}$ rozděl na k podobně velkých podmnožin $\mathcal{D}_1, \ldots, \mathcal{D}_k$.
2: Pro $j = 1, \ldots, k$ natrénuj model s danými hyperparametry na $\mathcal{D} \setminus \mathcal{D}_j$.
3: Na $\mathcal{D}_j$ odhadni chybu modelu e_j.
4: **return** cross-validation error

$$\hat{e} = \frac{1}{k} \sum_{i=1}^{k} e_i$$

Takto odhadneme cross-validační chybu pro všechny uvažované sady hyperparametrů a zvolíme model s nejnižší takovou chybou. Zvolený model s nejlepšími hyperparametry natrénujeme na celé množině $\mathcal{D}$.

Volba k je kvůli výpočetním nárokům obvykle malá (5, 10). Pro velmi malé datasety může však být únosná i volba $k = N$ (leave-one-out cross-validation).

Cross-validation error Cross-validation error je odhad očekávané testovací chyby Err a ne testovací chyby $\text{Err}_{\mathcal{D}}$.

Při skutečně malém datasetu je možné zvolit dvoustupňovou křížovou validaci, která si neodkládá testovací sadu, ale zanořuje se s křížovou validací o jednu stupeň níž, přičemž vnitřní cross-validační chyba se používá pro výběr modelu a vnější pro odhad očekávané chyby.

Vnější cross-validation chyba je pouze jedna ("společná") a odpovídá očekávané testovací chybě celé procedury pro výběr nejlepšího modelu, ale nejlepší model teprve musíme získat (natrénování zvoleného modelu na celé množině $\mathcal{D}$).

Finální model Ve všech metodách lze po výsledném odhadu testovací chyby přetrénovat nejlepší model na celém datasetu a pro odhad výkonnosti použít testovací chybu, resp. očekávanou testovací chybu, z předchozího kroku. U hold-out lze před evaluací na trénovacích datech ještě přetrénovat model na (trénovací + validační) množině.

0.3 Evaluace regrese

Nejčastější volba ztrátové funkce u regresních úloh je MSE (mean squared error) nebo MAE (mean absolute error):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \qquad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i|$$

Zatímco MSE je citlivější na velká residua, MAE se spíše soustředí na celkovou chybu a odlehlým hodnotám se chová "spravedlivě". Jednotky MSE jsou v kvadrátech, proto se také často udává přeškálované RMSE, které má interpretovatelné jednotky vysvětlované proměnné:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}$$

Pro nezáporné hodnoty lze použít ztrátovou funkcí RMSLE, které se soustředí na relativní míru odchylek:

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log Y_i - \log \hat{Y}_i)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\log^2 \frac{Y_i}{\hat{Y}_i}}$$

Koeficient determinace R^2 (coefficient of determination) vyjadřuje podíl variability cílové proměnné, kterou model vysvětluje:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\Sigma_{i=1}^{N}(\log Y_i - \log \hat{Y}_i)^2}{\Sigma_{i=1}^{N}(\log Y_i - \log \bar{Y})^2},$$

kde TSS je total sum of squares (RSS modelu, který predikuje $\bar{Y}$ pro každý datový bod).

0.4 Evaluace klasifikace

Matice záměn U klasifikace je konstrukce a interpretace ztrátových funkcí obecně problematické. Využívá se proto matice záměn:

		Skutečnost		
		$Y = 1$	$Y = 0$	Σ
Predikce	$\hat{Y} = 1$	TP	FP	$\hat{N}_+$
	$\hat{Y} = 0$	FN	TN	$\hat{N}_-$
	Σ	N_+	N_-	N

(T/F - True/False; P/N - Positive/Negative)

Z matice záměn můžeme vyvodit následující míry:

		Skutečnost	
$\mathrm{P}(\hat{Y} \mid Y)$		$Y = 1$	$Y = 0$
Predikce	$\hat{Y} = 1$	$\mathrm{TPR} = \frac{\mathrm{TP}}{N_+}$	$\mathrm{FPR} = \frac{\mathrm{FP}}{N_+}$
	$\hat{Y} = 0$	$\mathrm{FNR} = \frac{\mathrm{FN}}{N_-}$	$\mathrm{TNR} = \frac{\mathrm{TN}}{N_-}$

- True positive rate (TPR): senzitivita (recall).
- False positive rate (FPR): type I error rate.
- False negative rate (FNR): type II error rate.
- True negative rate (TNR): specificita (specificity).

Často používanou mírou (např. ve výpočtu F_1) je také odhad $P(Y = 1 \mid \hat{Y} = 1)$ (precision):

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\hat{N}_+}$$

0.4.1 Evaluační míry binární klasifikace

Nejpoužívanější mírou je přesnost (accuracy) – odhad $P(Y = \hat{Y})$:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{N}$$

Tato míra však není spolehlivá pro nevybalancovaná data, kdy modelu stačí predikovat pouze majoritní třídu. V takovém případě se používá F_1 score:

$$F_1 = \frac{2}{\mathrm{PPV}^{-1} + \mathrm{TPR}^{-1}},$$

kde hodnota $P(Y = 1)$ je velmi malá.

0.4.2 ROC a AUC

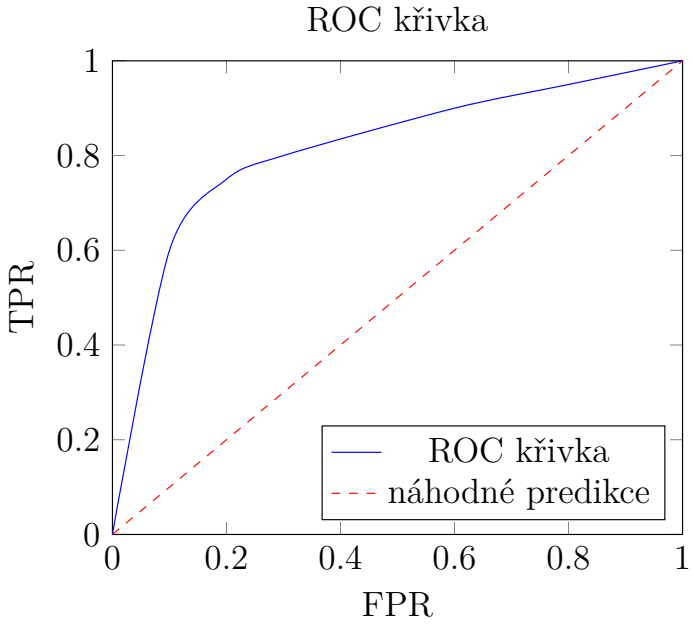
Modely binární klasifikace často odhadují $p(\boldsymbol{X}) = P(Y = 1 \mid \boldsymbol{X})$ a predikují

$$\hat{Y} = \mathbb{1}_{\hat{p} > 0.5}$$

Hodnotu 0.5 však můžeme parametrizovat pomoc $\tau \in [0, 1]$ a získat tak pro různé hranice různé predikce:

$$\hat{Y}_\tau = \mathbb{1}_{\hat{p} > \tau}$$

ROC S hodnotou τ od 0 do 1 se různě mění TPR a FPR, jejichž vztah lze vykreslit a následně dále zkoumat pomocí ROC křivky (receiver operating characteristic curve).



Pro dobrý model se křivka přimyká k levé a horní ose (strmě roste).

AUC Kvalita modelu, pro kterou máme ROC křivku, vyhodnocujeme plochou pod křivkou AUC (area under the curve). Model s náhodnými predikcemi má AUC = 0.5, dokonalý model by měl AUC = 1.