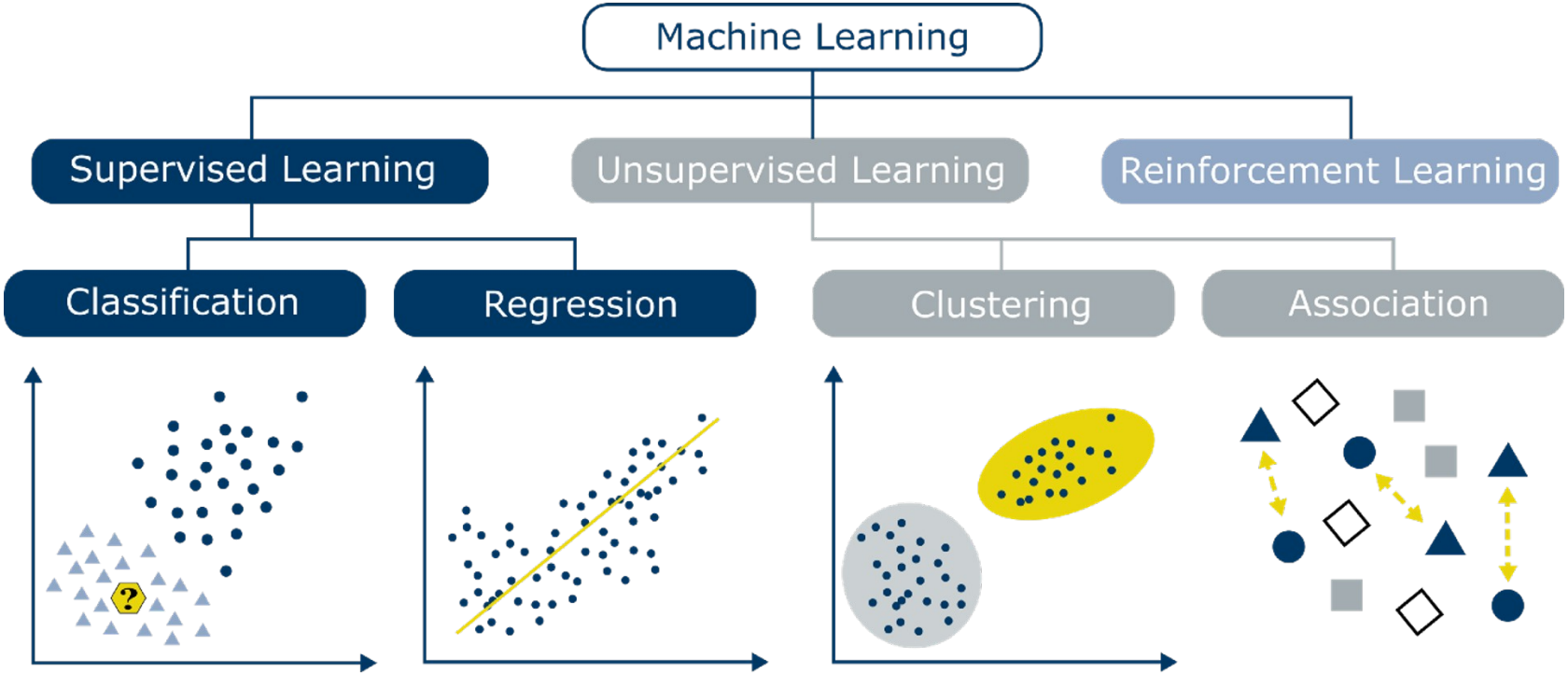
Katedra Geomatiky

Strojové učení v dálkovém průzkumu Země

155YUSU

Lekce 4: Generalizace ML modelu
{přeučení a nedoučení modelu, diagnostika modelu, generalizace}

Přednášející: Lukáš Brodský lukas.brodsky@natur.cuni.cz, Ondřej Pešek ondrej.pesek@fsv.cvut.cz a Martin Landa martin.landa@fsv.cvut.cz


Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Classification
Regression
Clustering
Association
?

Cíl učení

- Model strojového učení transformuje svá vstupní data na smysluplné výstupy;
- Jinými slovy, **učí se reprezentace vstupních dat** - reprezentace, které nás přiblíží očekávanému výstupu.

Jak se učí?

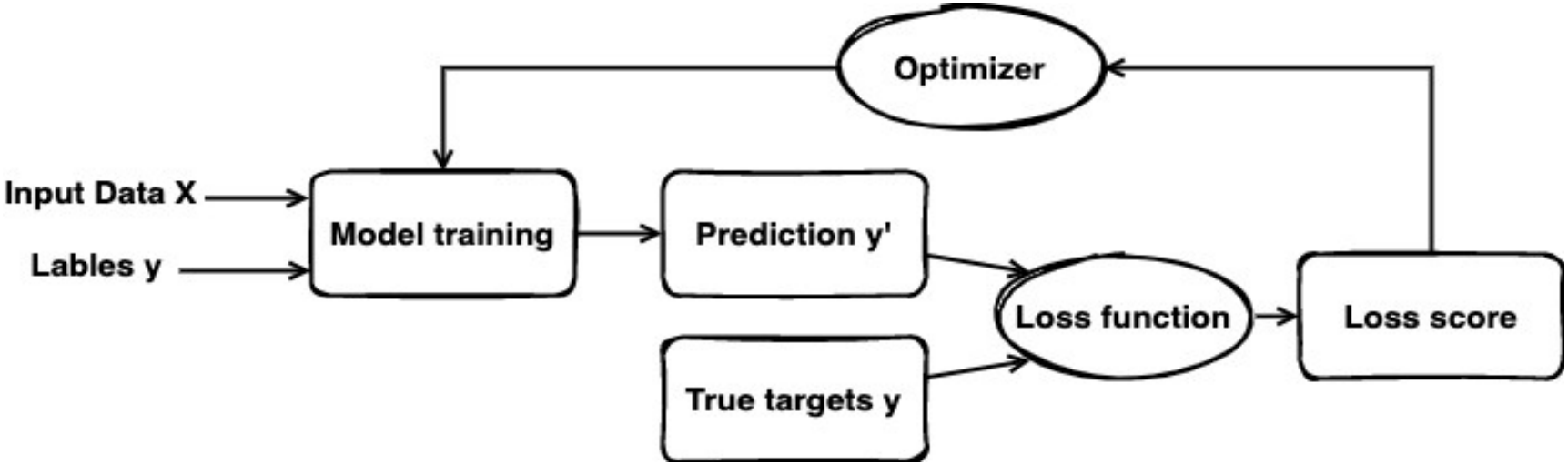
- **Zapamatování (memorování)**:
  - **hromadění jednotlivých faktů**
  - omezené časem na pozorování faktů a pamětí na jejich ukládání

-> deklarativní znalosti (co, ale ne jak)

- **Zobecnění (generalizace)**:
  - **odvozování nových skutečností ze starých skutečností**
  - omezeno přesností procesu dedukce, v podstatě prediktivní činnost

=> imperativní znalost (jak najít odpověď)

Učící se algoritmus



Optimalizace

$$w' = argmin_w \, L(w),$$

daná ztrátová funkce .. *L a* váhy.. *w*

- **Cílem** optimalizačního algoritmu **je minimalizovat ztráty (chyby);**
- **Nikoliv zobecnění modelu**!

  => PROBLÉM

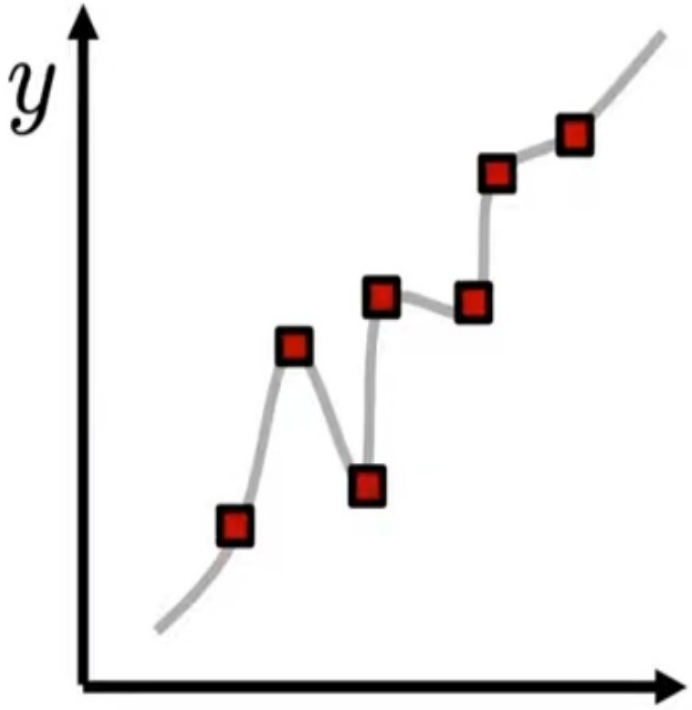
Přeučení / Nedoučení

(*Overfitting / Underfitting*)

Regrese

Příklad:

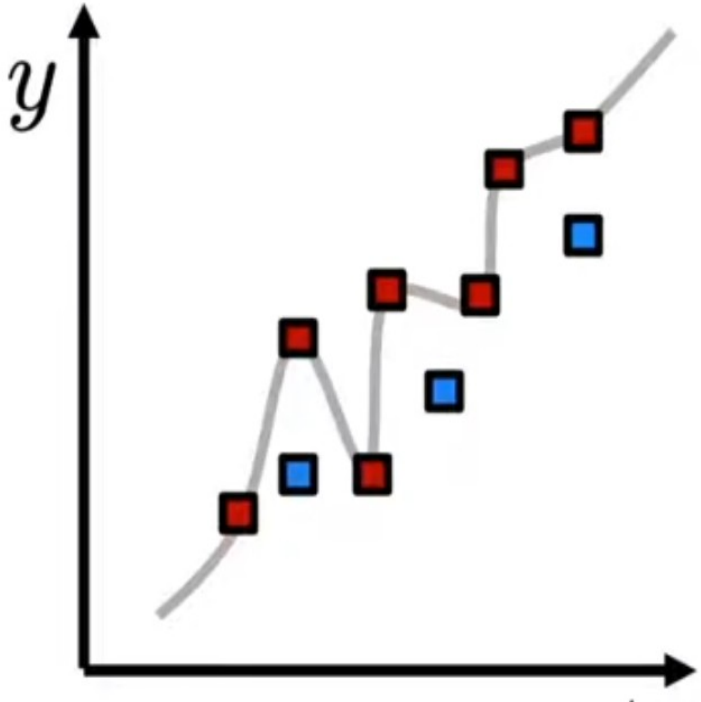
- nulová chyba při učení



Regrese

Příklad:

- nízká chyba tréninku (červené vzorky)
- vysoká testovací chyba (modré vzorky)

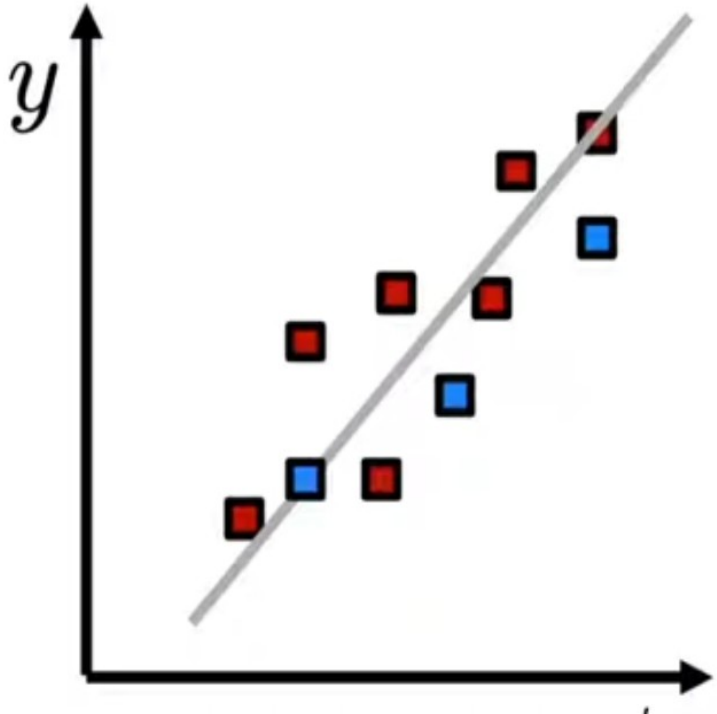


Regrese

Příklad:

- nízká chyba trénování
- nízká chyba při testování!

Dobrá generalizace



Regrese

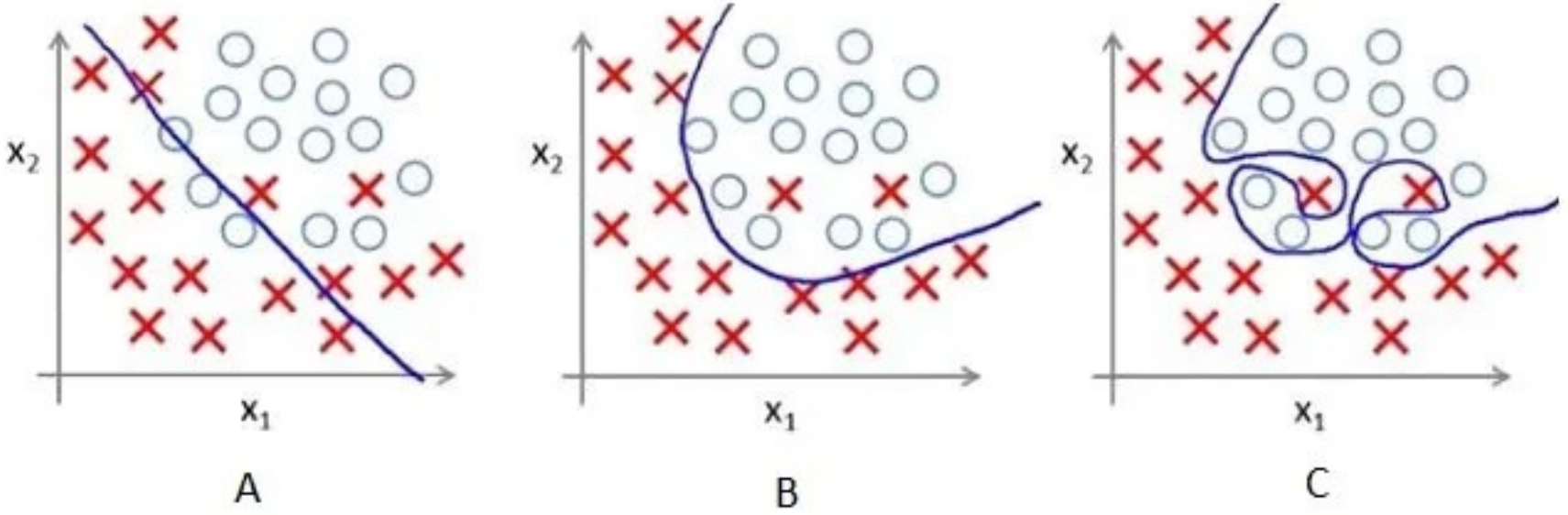
Příklad:

- vysoká chybovost tréninku
- vysoká chybovost testů

-> ... ?

Klasifikace

Příklad: A ?, B ?, C?



Zkreslení (bias):

- Tato část chyby zobecnění je způsobena **nesprávnými předpoklady**,
- jako je předpoklad, že data jsou lineární, zatímco ve skutečnosti jsou kvadratická.
- Model s vysokým zkreslením s největší pravděpodobností nedostatečně odpovídá trénovacím datům.

Model není schopen zachytit relevantní charakteristiky dat (základní jevy).

Rozptyl (variance):

- Předpokládáme, že máme množinu možných modelů: rozptyl je rozdíl mezi množinou modelů;
- Tato část je způsobena **přílišnou citlivostí** modelu **na malé odchylky v trénovacích datech**.
- Model s **mnoha stupni volnosti** (např. polynomický model vysokého stupně) bude mít pravděpodobně velkou disperzi, a bude tedy **příliš vyhovovat** trénovacím datům.

Neodstranitelná chyba:

- Tato část je způsobena **šumem** samotných dat.
- Jediným způsobem, jak tuto část chyby snížit, je data vyčistit (např. opravit zdroje dat, jako jsou nefunkční senzory, nebo odhalit a odstranit odlehlé hodnoty).

Co je nedoučení a přeučení?

Přeučení

Přeučení je situace, kdy **se** náš **model naučí “veškerý rozptyl dat”** (šum v datech). Odborníci říkají, že si model začne šum pamatovat, místo aby se model učil.

Nedostatečné učení (nedoučení)

Nedostatečné učení je stav, kdy model **nemůže dostatečně modelovat existující vztah v trénovacích datech**.

Nedostatečné učení: přílišné zjednodušení problému. Model je příliš jednoduchý (příklad: aplikace lineárního modelu na nelineární problém).

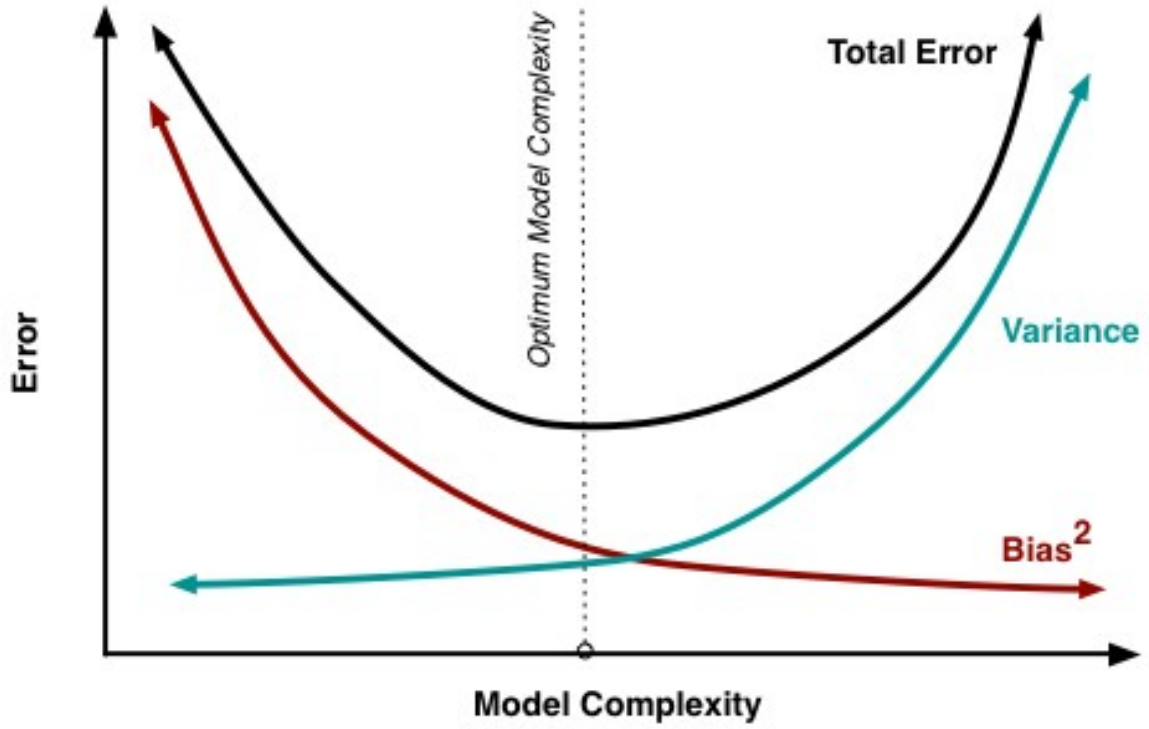
Generalizace modelu

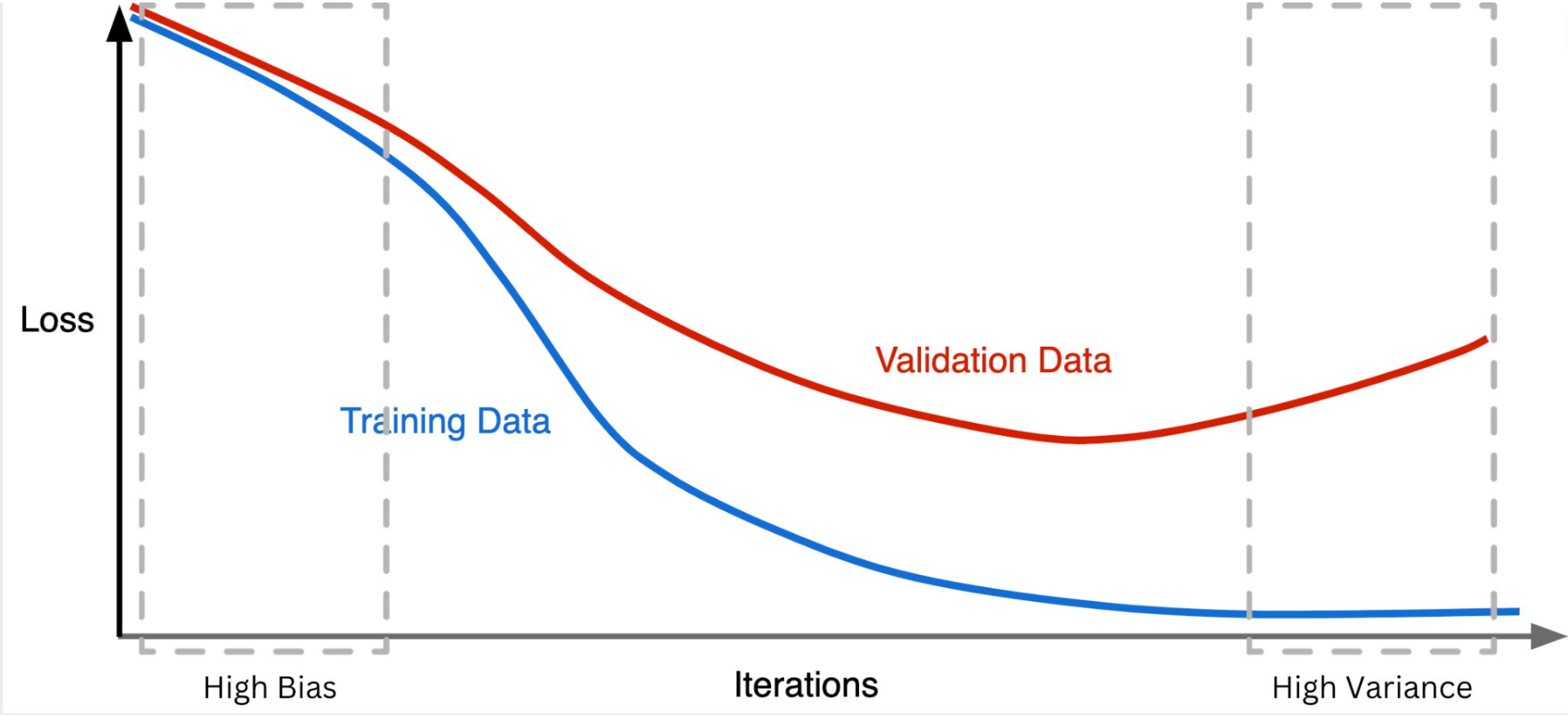
- Zvyšování složitosti modelu obvykle zvyšuje jeho rozptyl a snižuje jeho zkreslení.
- Naopak snížení složitosti modelu zvyšuje jeho zkreslení a snižuje jeho rozptyl.

-> **Který z nich?**

Model by měl udržovat rovnováhu mezi těmito dvěma typy chyb. To je známé jako kompromisní řízení chyb (trade-off) zkreslení a odchylkou.

Kompromis mezi zkreslením a odchylkou (Bias-variance trade-off)




Loss
Validation Data
Training Data
High Bias
Iterations
High Variance

Diagnostika

Evaluace modelu

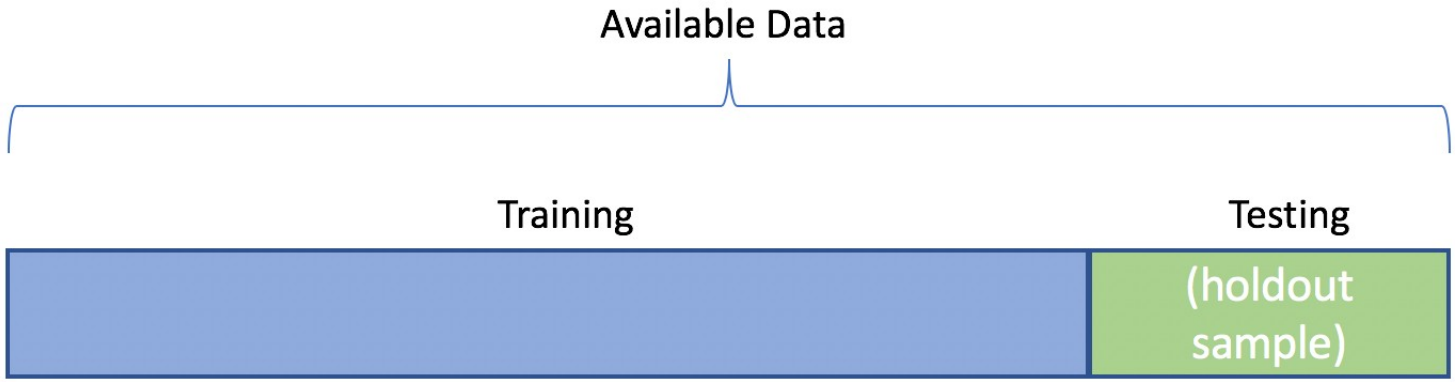
Jediný způsob, jak zjistit, jak dobře se model zobecní na nové případy, je skutečně **ho vyzkoušet na nových případech**.

Evaluace modelu

Rozdělení dat:

- Data, která používáme ke **konstrukci** našich modelů (**trénovací** soubor)
- Data, která používáme k **testování** našich modelů (**testovací** soubor)

Evaluace modelu

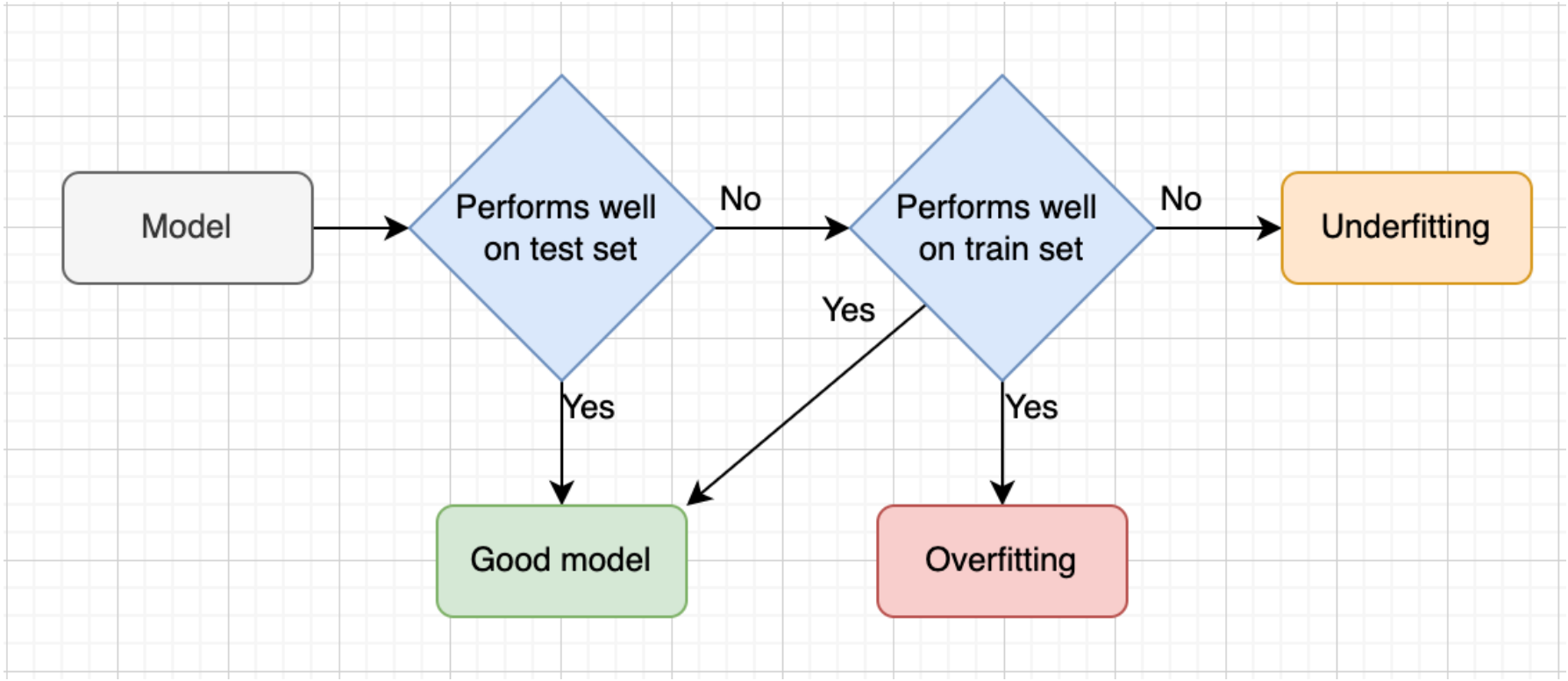


Evaluace modelu

- **Chyba trénování** (chyba ve vzorku): chyba vypočtená **na stejných** vzorcích **dat**, které jste použili k učení modelu;
- **Chyba testu** (chyba mimo vzorek): chyba **na nezkoumaných** vzorcích **dat**;

Chybovost na nových případech se nazývá **generalizační chyba**, a vyhodnocením vašeho modelu na testovacím souboru získáte odhad této chyby.

Evaluace modelu



Evaluace modelu

Případ A/

1/ Vyladěný model (hyperparametry), nízká chyba trénování a testování (chyba generalizace <5 %);

2/ -> Model do produkce!

Bohužel však nefunguje tak dobře, jak se očekávalo, a produkuje 15% chybu.

Co se stalo?

Evaluace modelu

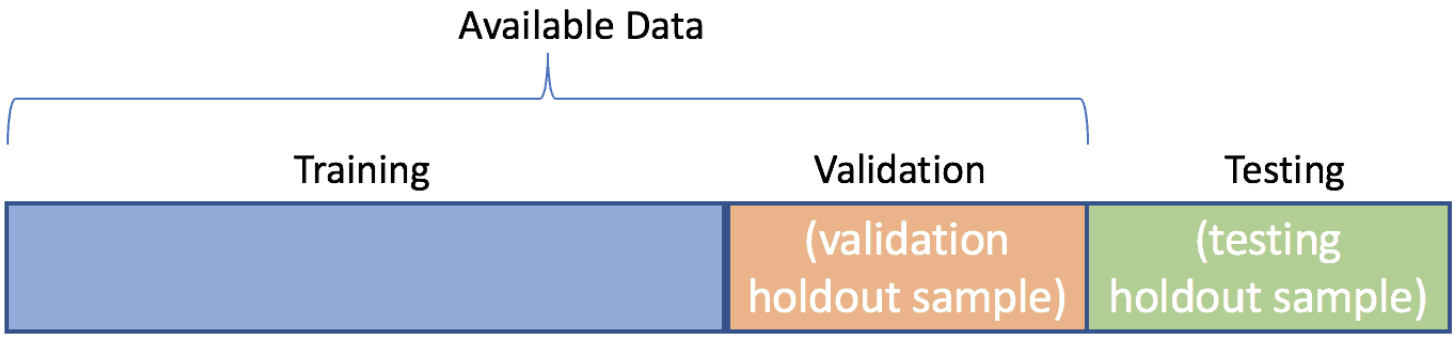
Problém spočívá v tom, že jsme **opakovaně měřili chybu generalizace na testovací množině**, a **přizpůsobili jsme model a hyperparametry tak, abychom pro tuto množinu vytvořili nejlepší model**. To znamená, že model pravděpodobně nebude fungovat stejně dobře na nových datech.

Evaluace modelu

Řešení:

- Mějte **další množinu, která se nazývá validační množina**.
- Pomocí trénovací množiny natrénujete několik modelů s různými hyperparametry,**vyberete model a hyperparametry, které nejlépe fungují na validační množině**, a když jste se svým modelem spokojeni, **provedete jeden závěrečný test proti testovací množině**, abyste získali odhad generalizační chyby.

Evaluace modelu



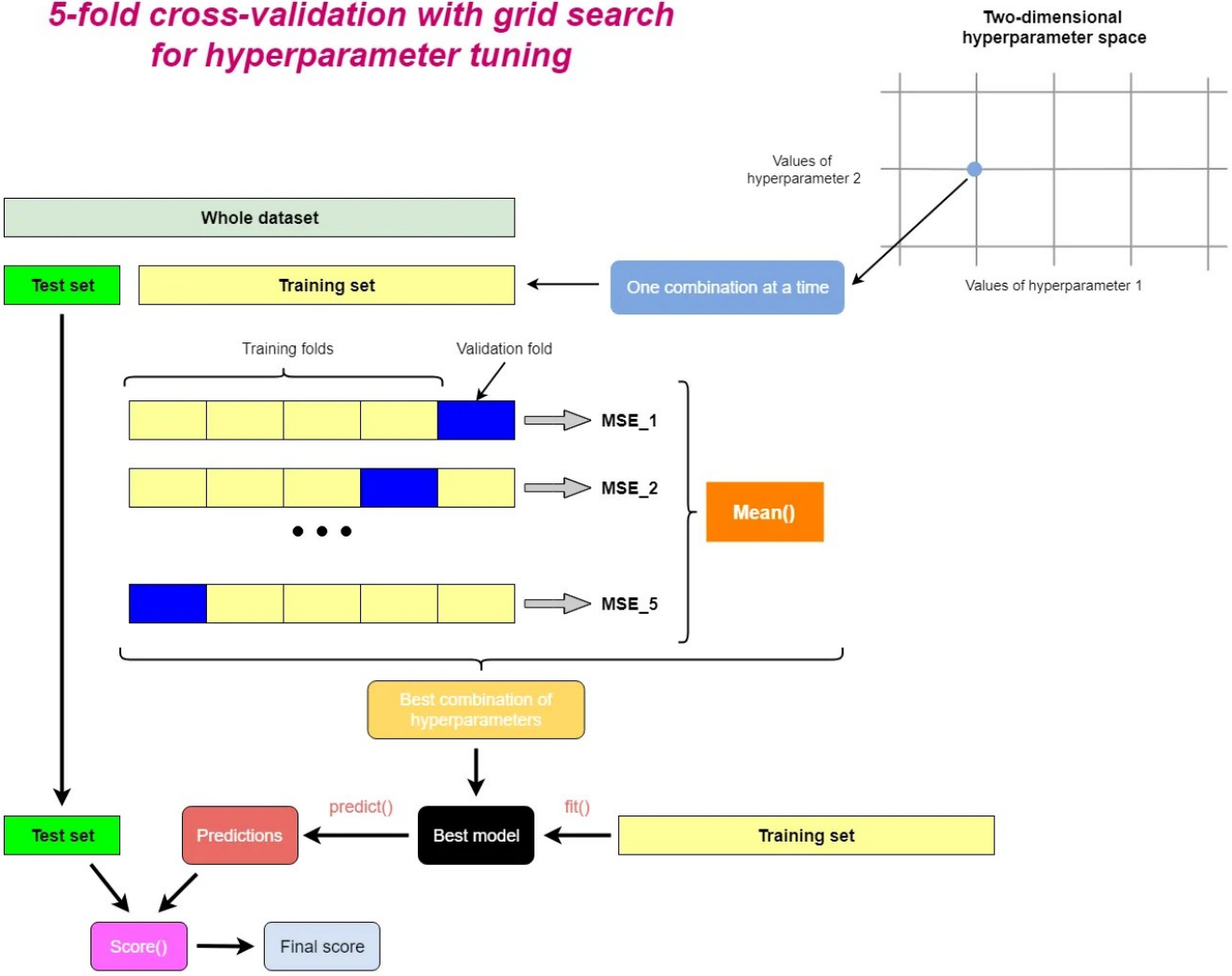
Evaluace modelu

Případ: **nedostatek dat!**

Řešení: Aby se zabránilo „plýtvání“ příliš velkým množstvím trénovacích dat ve validačních sadách, je běžnou technikou použití **křížové validace**.

Trénovací množina se rozdělí **na doplňující se podmnožiny**, každý model se natrénuje proti jiné kombinaci těchto podmnožin a validuje se proti zbývajícím částem.

5-fold cross-validation with grid search for hyperparameter tuning



Křížová validace

- Pro **trénink** používáme **100 % trénovacích + validačních dat**, což vyhlazuje problémy, kdy byla počáteční trénovací sada možná velmi zkreslená.
- Můžeme (po částech) **validovat všechna data**: nestaneme se obětí případů s vysokou odchylkou v malé validační množině.
- Zprůměrujeme náš výkon na všech datech, což nám dává mnohem větší jistotu v odhadu dovedností modelu a také skutečný **obraz toho, jak je model proměnlivý vůči výkyvům ve** vstupních datech.
- Jsme automaticky nuceni sestavit mnohem **méně nadhodnocený** (a tím i zobecnitelný) model, protože se snažíme **maximalizovat průměrnou výkonnost** v mnoha validačních sadách, nikoli v jedné konkrétní validační sadě.

Proč dochází k nedoučení a přeučení?

Případy

- Jednoduchý model
- Nedostatek dat
- Nedostatek rozptylu v trénovacích datech

=> **Nedoučení**

Případy

- Nevyčištěná data
- Složitost modelu je vysoká
- Málo trénovacích dat
- Dlouhé trénování na jediném vzorku dat

=> **přeučení**

Strategie pro generalizaci modelu

Strategie generalizace

1. **Zvyšování počtu tréninkových dat**: Více dat může pomoci modelu lépe zobecnit, protože poskytuje širší pohled na problém. Nebo **rozšiřování dat**: technika, která mírně mění vzorová data pokaždé, když je model zpracovává.
2. **Čištění dat**: odstranění odlehlých hodnot
3. **Křížová validace**: Místo toho, abyste k trénování použili celou sadu dat, rozdělte ji na trénovací a validační sadu.
4. **Zjednodušení modelu**: Pokud je problémem overfitting, zkuste použít méně složitý model nebo omezte počet funkcí.
5. **Přidání smysluplných příznaků**: při nedoučení;
6. **Výběr vhodných příznaků**: odstraňte irelevantní nebo nadbytečné příznaky, které mohou způsobit nadměrné přizpůsobení modelu.

Strategie generalizace

7. **Regularizace**: Techniky jako L1 nebo L2 **penalizují složitost modelu** a odrazují od nadměrného přeučení.

8. **Snížení hloubky** nebo složitosti rozhodovacích stromů, aby se zabránilo šumu při fitování.

9. **Včasné zastavení** (*early stopping*): zastavte trénování, když se výkon modelu na validační množině začne zhoršovat.

10. **Ansámbly**: učení pomocí ansámblu je technika strojového učení, která zvyšuje přesnost a odolnost předpovědí sloučením předpovědí z více modelů.

Dotazy?