

kapitola II

Kolika parametry nakreslíš slona?

Slon se čtyřmi parametry

Jaro 1953. Dyson veze Fermimu dva roky výpočtů, co krásně sedí na data. Fermi se zeptá: „Kolik volných parametrů?„ — „Čtyři.“

*„Se čtyřmi parametry ti nafituju slona, a s pěti mu rozhýbu chobot.
„ — von Neumann*

Krásná shoda s daty byla iluze koupená za čtyři čísla.

Papoušek a student

Papoušek se naučil loňský test nazpaměť — i s tečkami. Student pochopil, proč jsou odpovědi takové.

Na loňském testu vyhraje papoušek. Na letošním — s otázkami, které nikdo neviděl — se papoušek zhroutí a student projde.

Rozdíl nepoznáš na tom, co model viděl. Jen na tom, co neviděl.

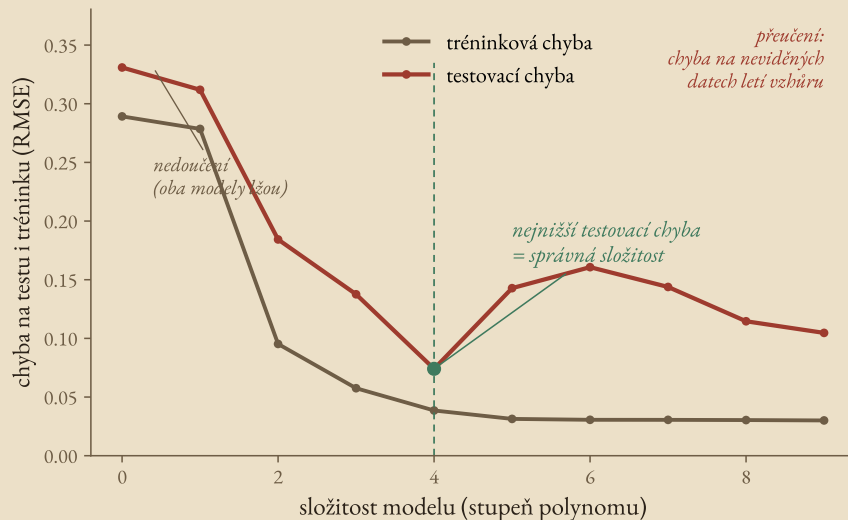
Letošní test

- **Přeučení** (overfitting) — model se naučil data nazpaměť i se šumem. Papoušek.
- Než začneš učit, schovej **pětinu dat** do trezoru.
- Zajímá tě **chyba venku**, ne doma. Křížovou validací ji změříš spolehlivě.

Dvě nemoci, jedno údolí

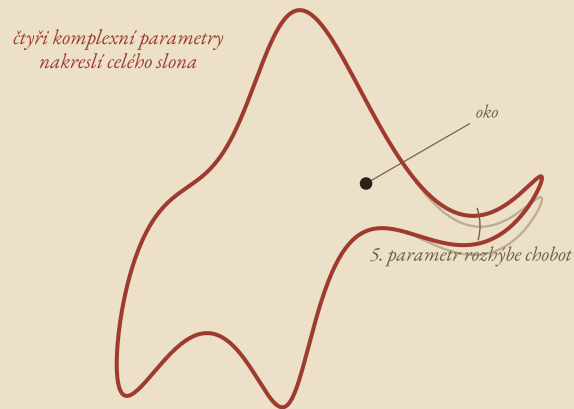
- **Nedoučení** — přímka přes vlnu. Model je tak chudý, že netrefí ani tvar. Vysoké vychýlení.
- **Přeučení** — had přes šum. Prolézá každý bod včetně náhody. Vysoký rozptyl.
- **Správná složitost** leží v údolí mezi nimi.
- Regularizace: nech model složitý, ale **zdaň mu složitost**.

Kde se papoušek prozradí



Test (sanguine) klesne ke dnu a pak stoupá. To dno je správná složitost.

A teď doopravdy



věrná rekonstrukce dle Mayer, Khairy & Howard, Am. J. Phys. 78, 648 (2010)

Mayer, Khairy & Howard (2010): čtyři komplexní čísla nakreslí slona.

„Jak bych poznal, že se pletu?„

Než uvěříš, že model něco umí: viděl ta data při tréninku?
A hlavně — jaká je jeho chyba na tom, co neviděl?