Model Deployment (Nasazení modelu)

Model Deployment je proces integrace modelu strojového učení do stávajícího produkčního prostředí. Zatímco trénování modelu je zaměřeno na přesnost, nasazení se soustředí na **dostupnost, rychlost (latenci), stabilitu a škálovatelnost**.

1. Typy strategií nasazení

Podle toho, jakým způsobem aplikace potřebuje výsledky, volíme různé architektury:

A. Online Inference (Real-time)

Model běží jako služba (často v kontejneru) a odpovídá na požadavky přes API (REST nebo gRPC).

- **Vhodné pro:** Doporučovací systémy v e-shopech, detekci podvodů při platbě kartou.
- **Výhoda:** Okamžitá odpověď.

B. Batch Inference (Dávkové zpracování)

Model zpracovává velké balíky dat najednou v pravidelných intervalech (např. jednou za noc). Výsledky se uloží do databáze.

- **Vhodné pro:** Generování měsíčních reportů, hromadné bodování zákazníků (scoring).
- **Výhoda:** Vysoká propustnost, nižší náklady na infrastrukturu.

C. Edge Deployment

Model běží přímo na zařízení uživatele (mobil, IoT senzor, auto).

- **Vhodné for:** Rozpoznávání obličeje v telefonu, autonomní řízení.
- **Výhoda:** Soukromí, funguje bez internetu, nulová síťová latence.

2. Techniky bezpečné aktualizace modelu

Při nasazování nové verze modelu musíme minimalizovat riziko chyby:

Strategie	Popis
Blue-Green	Máte dvě identická prostředí. Nový model (Green) se otestuje a pak se na něj naráz přepne veškerý provoz z Blue.
Canary Deployment	Nový model dostane nejdříve jen malé procento provozu (např. 5 %). Pokud jsou výsledky dobré, podíl se postupně zvyšuje.

Strategie	Popis
Shadow Mode	Nový model běží na pozadí, dostává reálná data, ale jeho predikce se uživateli neukazují. Pouze se porovnávají s produkčním modelem.
A/B Testing	Část uživatelů vidí výsledky modelu A, část modelu B. Sleduje se, který model má lepší obchodní výsledky (např. vyšší prodeje).

3. Nástroje pro nasazení

Dnes se k nasazení využívají technologie, které zajišťují stabilitu:

- **Kontejnery (Docker):** Zabalí model se všemi knihovnami.
- **Orchestrace (Kubernetes):** Spravuje běh mnoha instancí modelu.
- **Model Servery:** Specializované nástroje jako **TFServing**, **TorchServe** nebo **NVIDIA Triton**, které optimalizují využití GPU/TPU.

4. Monitorování po nasazení

Nasazením práce nekončí. Je nutné sledovat:

- **Latenci:** Jak dlouho trvá jedna predikce.
- **Data Drift:** Zda se data v reálném světě nezačala lišit od těch, na kterých se model učil.
- **Využití zdrojů:** CPU, RAM a GPU paměť.

5. Formáty modelů pro produkci

Při přechodu z vývoje do produkce se modely často převádějí do formátů optimalizovaných pro rychlost:

- **ONNX (Open Neural Network Exchange):** Univerzální formát pro přenos mezi různými frameworky.
- **TensorRT:** Optimalizace pro NVIDIA grafické karty.
- **TensorFlow Lite:** Pro mobilní zařízení.

> **Zajímavost:** Existuje pojem **„Model Decay“** (rozklad modelu). Je to jev, kdy model v produkci postupně ztrácí svou přesnost jednoduše proto, že se mění okolní svět. Průměrný model pro predikci chování uživatelů na webu může začít zastarávat již po několika týdnech bez aktualizace.

Zpět na AI rozcestník