Přehled algoritmů Machine Learningu

Tato stránka slouží jako katalog základních i pokročilých algoritmů strojového učení. Výběr správného algoritmu závisí na typu dat, velikosti datasetu a požadovaném výstupu (predikce čísla, zařazení do kategorie, nalezení struktury).

1. Učení s učitelem (Supervised Learning)

Algoritmy se učí na datech, která mají známý výsledek (label). Cílem je naučit model předpovídat tento výsledek pro nová, neznámá data.

A. Regrese (Predikce čísel)

Používá se, pokud je výstupem spojitá hodnota (cena bytu, teplota, prodeje).

- **Lineární regrese (Linear Regression):** Hledá přímku (nebo nadrovinu), která nejlépe prokládá data. Je to základní, rychlý a snadno interpretovatelný model.
- **Polynomiální regrese:** Umožňuje proložit data křivkou (ne jen přímkou) pomocí mocninných funkcí.

B. Klasifikace (Zařazování do tříd)

Používá se, pokud je výstupem kategorie (spam/nespam, kočka/pes/auto).

- **Logistická regrese (Logistic Regression):** Navzdory názvu jde o klasifikátor. Predikuje pravděpodobnost (0 až 1), že daný vzorek patří do určité třídy.
- **Support Vector Machines (SVM):** Hledá ideální hranici (nadrovinu) mezi třídami tak, aby byla mezera (margin) mezi nimi co největší. Velmi efektivní pro komplexní data.
- **Naivní Bayes (Naive Bayes):** Pravděpodobnostní klasifikátor založený na Bayesově větě. Extrémně rychlý, často používaný pro filtrování spamu a analýzu textu.
- **k-Nearest Neighbors (k-NN):** „Líný“ algoritmus. Nový bod zařadí tam, kam patří většina jeho *k* nejbližších sousedů.

C. Rozhodovací stromy a Ensembling

Populární metody díky své schopnosti zachytit nelineární vztahy.

- **Decision Trees (Rozhodovací stromy):** Vytváří sadu podmínek (if-else), podle kterých data dělí. Jsou snadno vizualizovatelné, ale náchylné k přeučení (overfitting).
- **Random Forest (Náhodný les):** Vytvoří stovky stromů a nechá je „hlasovat“. Eliminuje chyby jednotlivých stromů a je velmi robustní.
- **Gradient Boosting (XGBoost, LightGBM, CatBoost):** Staví stromy postupně, kde každý nový strom opravuje chyby toho předchozího. Dnes **nejvýkonnější metoda** pro tabulková data (vítězí v soutěžích Kaggle).

2. Učení bez učitele (Unsupervised Learning)

Data nemají žádné labely. Algoritmus v nich hledá skrytou strukturu.

A. Shlukování (Clustering)

- **K-Means:** Rozdělí data do *K* skupin (clusterů) podle podobnosti. Vyžaduje předem určit počet skupin.
- **DBSCAN:** Shlukuje body, které jsou blízko u sebe, a osamocené body označuje jako šum (outliers). Nemusíte znát počet skupin předem.

B. Redukce dimenze

Slouží ke zjednodušení dat (snížení počtu sloupců/rysů) při zachování důležitých informací.

- **PCA (Principal Component Analysis):** Matematická transformace, která najde nové osy (komponenty), v nichž mají data největší rozptyl. Umožňuje vizualizovat vícerozměrná data ve 2D nebo 3D.

3. Tahák: Který algoritmus vybrat?

Jednoduchý průvodce pro výběr správného nástroje:

Úloha	Typ dat	Doporučený algoritmus	Poznámka
Predikce hodnoty	Lineární závislost	Lineární regrese	Začněte zde, pokud chcete jednoduchost.
Predikce hodnoty	Komplexní vztahy	Random Forest / XGBoost	Zlatý standard pro tabulková data.
Ano / Ne	Textová data	Naivní Bayes	Rychlý pro NLP.
Ano / Ne	Málo dat, vysoká přesnost	SVM	Dobře funguje ve vyšších dimenzích.
Segmentace	Zákazníci, produkty	K-Means	Pro rozdělení do skupin.
Obrázky/Zvuk	Pixely, vlnové formy	**Neurální sítě (CNN/RNN)**	Viz článek neural_networks.

Ukázka kódu (Python - Scikit-Learn)

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split

# 1. Příprava dat
X, y = load_data()  # X = rysy, y = labely
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```python
# 2. Inicializace a trénink modelu
clf = RandomForestClassifier(n_estimators=100)
clf.fit(X_train, y_train)

# 3. Predikce
prediction = clf.predict(X_test)
```

Tagy: ml algoritmy python data_science statistika