1. Datová analýza a strojové učení

1.1. STROJOVÉ UČENÍ

Stále se vyvíjející platformy a nástroje nabízejí možnost vyvíjet modely, které se učí ze stávajících vzorů mezi proměnnými a které tyto modely následně používají ke spolehlivé předpovědi, co se stane v budoucnu.

Například počítačový program je navržen streamovací službou tak, aby jednotlivým uživatelům doporučoval filmy, které by se jim mohly líbit. Algoritmus analyzuje filmy, které diváci již viděly a filmy, které lidé s podobnými preferencemi sledování vysoce hodnotily.

Metody strojového učení jsou používány v celé řadě aplikací, včetně rozpoznávání řeči, lékařské diagnostice, autonomních vozidlech, doporučování zboží na internetu a mnoho dalších.

Algoritmy zlepšují svůj výkon u konkrétních úloh na základě opakovaného plnění těchto úloh.

1.2. TYPY ALOGORITMŮ STROJOVÉHO UČENÍ

1.2.1. **UČENÍ POD DOHLEDEM** (supervised)

- Jsou nejčastěji používanými algoritmy strojového učení pro prediktivní analýzu.
- Opírají se o soubory dat, které byly zpracovány lidskými odborníky (odtud slovo „dohled“).
- Poté se naučí, jak provádět stejné úlohy zpracování samostatně na nových sadách dat.
- Používají se zejména k řešení **regresních** a **klasifikačních** problémů

1.2.1.1 Regresní algoritmy (predikují hodnotu nového datového bodu na základě historických dat.)

Jedná se o odhad matematických vztahů mezi spojitou proměnnou a jednou nebo více dalšími proměnnými. Tento matematický vztah pak lze použít k výpočtu hodnot jedné neznámé proměnné vzhledem ke známým hodnotám ostatních.

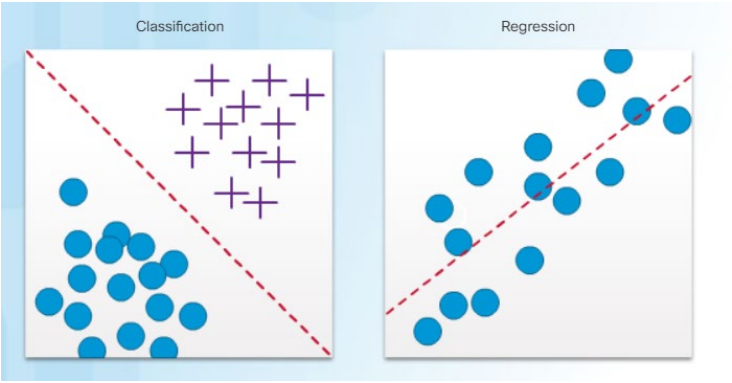
- Odhad polohy a rychlosti automobilu pomocí GPS
- Předpověď trajektorie tornáda pomocí údajů o počasí
- Předpověď budoucí hodnoty populace pomocí historických dat a dalších zdrojů informací
- Předpověď ceny nemovitosti se dvěma ložnicemi příští rok ve městě
- Předpověď kolik na kliniku přijde pacientů v úterý

1.2.1.2 Klasifikační algoritmy (využívají prediktivních výpočtů k zařazení dat do přednastavených kategorií)

Používají se u diskrétních proměnných. Problém spočívá v odhadu, ke které z předem definovaných tříd patří konkrétní vzorek.

- Rozpoznávání obrazu (tváří, tvarů, apod)
- Diagnostika patologií z lékařských testů
- Je tato emailová zpráva nevyžádanou poštou?
- Jaké je zabarvení (kladné, záporné nebo neutrální) daného textu?

Algoritmus se „učí“ na příkladech a mapuje si tak hraniční křivku mezi jednotlivými třídami. Ta může být následně použita pro klasifikaci nových příkladů.



1.2.2. **UČENÍ BEZ DOHLEDU** (unsupervised)

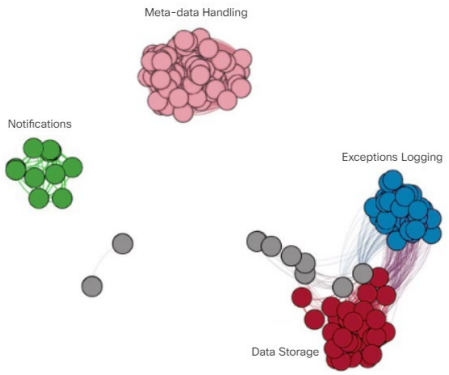
- Nevyžadují lidského odborníka k označení datových bodů
- Algoritmus sám uspořádá data a popíše jejich strukturu (užitečné, pokud nevíme, jak by měl výsledek vypadat)
- Algoritmus autonomně objevuje vzory v datech
- Příklad problémů řešených těmito metodami jsou algoritmy clusteringu a asociace (association)

1.2.2.1 **Zjištění podobností** (clustering)

Ty lze považovat za automatické objevování skupin vzorků, které mají podobné vlastnosti. Rozdělují tedy data do více skupin určením úrovně podobností mezi datovými body.

Algoritmy shlukování se například používají k identifikaci skupin uživatelů na základě jejich online historie nákupů a poté se každému členu zasílají cílené reklamy.

- Kteří diváci mají rádi stejné typy filmů?
- Které modely tiskárny selžou stejným způsobem?



1.2.2.2 Asociační metody (association)

Objevování skupin položek, které jsou často pozorovány společně. Používají se k navrhování dalších nákupů uživateli na základě obsahu jeho nákupního košíku.

1.3. PROCES STROJOVÉHO UČENÍ

Vývoj řešení s využitím strojového učení je nelineární proces a vyžaduje několik pokusů a omylů k dosažení požadovaného výsledku.

Krok 1. Proces přípravy dat, převodu dat na strukturovaná, řešení chybějících dat, odstranění odlehlých pozorování apod.

Krok 2. Vytvoření **tréninkové sady dat** (learning set), která bude použita k trénování modelu.

Krok 3. Vytvoření **testovací sady dat** (test set), která bude použita ke zhodnocení výkonu modelu (pouze v případě učení pod dohledem).

Krok 4. Smyčka. Je vybrán vhodný algoritmus na základě daného problému.

V závislosti na vybraném algoritmu můžou být potřebné další kroky, **předzpracování** (**pre-processing**), jako je extrahování prvků z datové množiny, které jsou relevantní vzhledem k řešenému problému.

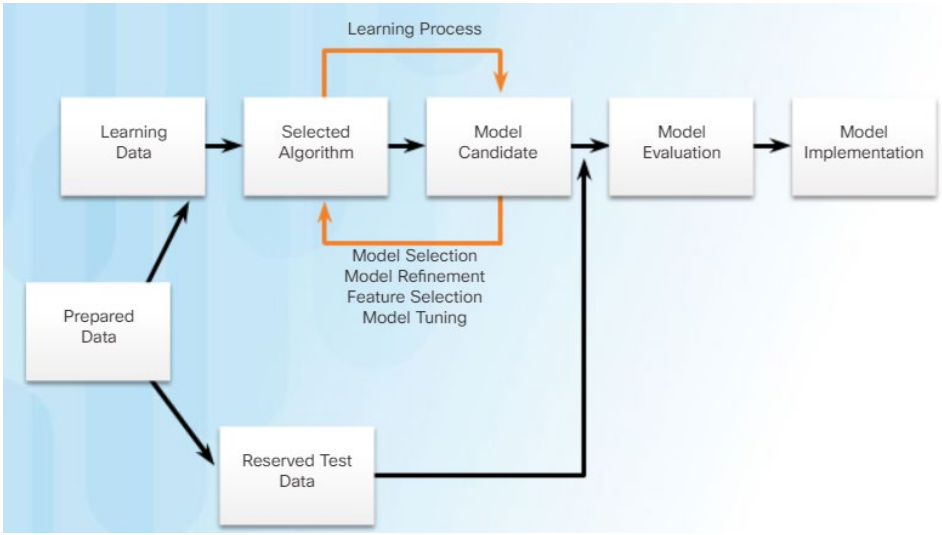
Pokud se například pokoušíte analyzovat úroveň aktivity osoby na základě fitness náramku, lze z nezpracovaných měření senzoru extrahovat funkce, jako je počet kroků, nadmořská výška, maximální zrychlení a tak dále.

V tomto kroku lze také provést **kroky následného zpracování** (**post-processing**), jako je jemné doladění parametrů modelu/algoritmu.

Pokud algoritmus a model dosáhnou dostatečného výkonu na tréninkové sadě dat, je řešení validováno na testovací sadě dat. Jinak je navržen nový model a/nebo algoritmus a proces učení se opakuje.

Krok 5. Otestování modelu na testovacích datech se nazývá **vyhodnocení modelu** (**model evaluation**). Dobrý výkon modelu na tréninkových datech není nutně zárukou dobrého výkonu na testovacích datech.

Krok 6. Když model dosáhne uspokojivého výkonu na testovacích datech, model může být následně **implementován**.



1.4. REGRESNÍ ANALÝZA

Regresní analýza je jednou z nejstarších a nejčastěji používaných statistických metod pro analýzu dat.

Hlavní myšlenkou regrese je kvantifikovat matematický vztah mezi jednou, nebo více **nezávislými proměnnými** (*predictor*) a **závislou proměnnou** (target, cílovou proměnnou).

Když je tento matematický vztah (**regresní funkce**) získán, lze jej použít k odhadu hodnot závislé proměnné mimo rozsah pozorovaných hodnot. Jinými slovy, regresní model umožňuje analytikovi extrapolovat mimo dostupný soubor dat.

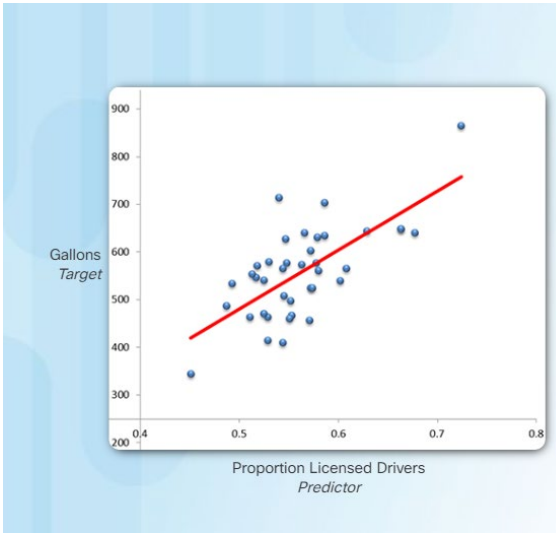
Například při práci s daty časové řady umožňuje regrese analytikovi předpovídat budoucí hodnoty z historických dat. V zásadě se regrese snaží najít vztah mezi libovolnými typy spojitých proměnných.

Zejména se snaží odpovědět na obecnou otázku: *„Jak moc se změní proměnná V1, pokud se proměnná/proměnné V2 (V3, V4, V5) změní/změnily o veličinu X*? “

Jednoduchý způsob, jak vizualizovat regresní funkci, je přednastavit si množinu bodů ve dvou rozměrech (viz obrázek).

Nezávislá proměnná (predictor), vynesena dle konvencí **na osu x**, je podíl řidičů s licencí v různých geografických oblastech. **Na ose y**, běžně používanou pro **závislou proměnnou** (cílovou), je odpovídající spotřeba benzínu. V tomto případě je možná **regresní funkce znázorněná červenou čarou**.

Skutečnost, že se v tomto příkladu jedná o přímku, naznačuje velmi intuitivní výsledek: *nárůst licencovaných řidičů v oblasti způsobí úměrný nárůst benzínu.* Zatímco jednoduchá vizuální kontrola distribuce datových bodů naznačuje, že přímka je v tomto případě nejvhodnější, regrese neomezuje na tvar regresní funkce.



Gasoline Consumption as Related to Percentage of Licensed Drivers

Proportion of Population with Driver's Licenses	Consumption of Gasoline (Millions of Gallons)
0.525	541
0.572	524
0.58	561
0.529	414
0.544	410
0.571	457
0.451	344
0.553	467
0.529	464
0.552	498
0.53	580
0.525	471
0.574	525

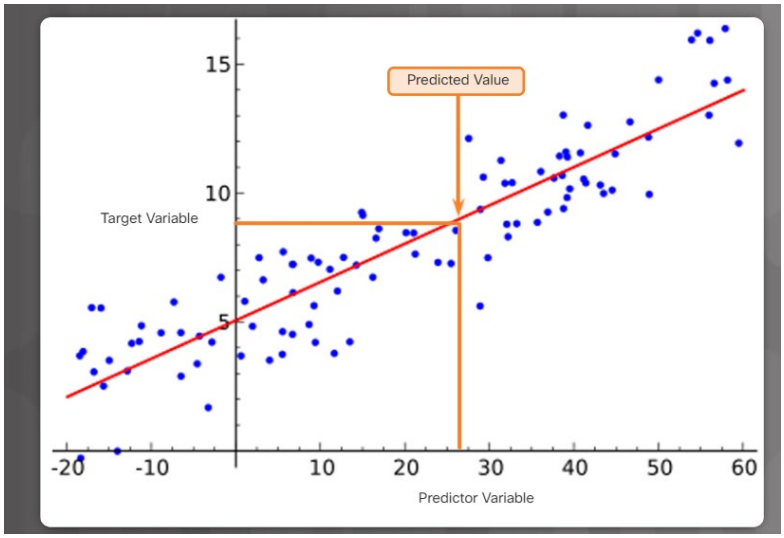
Obr. předpověď spotřeby benzínu v závislosti na počtu vydaných řidičských oprávnění

1.4.1 LINEÁRNÍ REGRESE

Nejběžnější regresní metody se nazývají lineární regrese. Ty *jsou z výpočetního i matematického hlediska nejjednodušší*. Proto představují první volbu pro datového analytika, který má problém s regresí.

Navzdory slovu lineární v názvu, lineární regrese neznamená proložení přímky datovými body. Termín lineární znamená, že *regresní funkce se bude vždy snažit přizpůsobit data pomocí váženého průměru* jiných funkcí, ať již jsou tyto funkce lineární nebo ne.

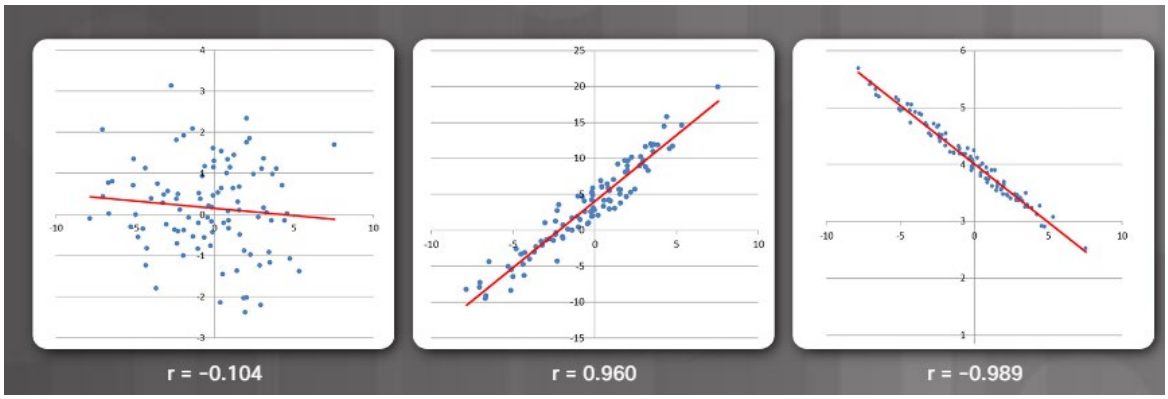
Vlastnost linearity zjednodušuje výpočet parametrů regresního modelu a zároveň umožňuje použít prakticky jakýkoli tvar pro přizpůsobení pozorování. Nejjednodušší případ lineární regrese spočívá v proložení přímky. Tomu se také říká **jednoduchý lineární model**.



Obr. proložení datových bodů přímkou, která by měla vyjadřovat lineární závislost

Vysoká **Pearsonova korelace** naznačuje, že mezi daty existuje silná závislost. Následující obrázek ukazuje příklady silně pozitivních a negativních korelovaných pozorování.

- Hodnota korelačního koeficientu **-1** značí zcela nepřímou závislost (antikorelaci), tedy čím více se zvětší hodnoty v první skupině znaků, tím více se zmenší hodnoty v druhé skupině znaků.
- Pokud je korelační koeficient roven **0** (nekorelovanost), pak mezi znaky není žádná statisticky zjistitelná lineární závislost.
- **+1** značí zcela přímou závislost



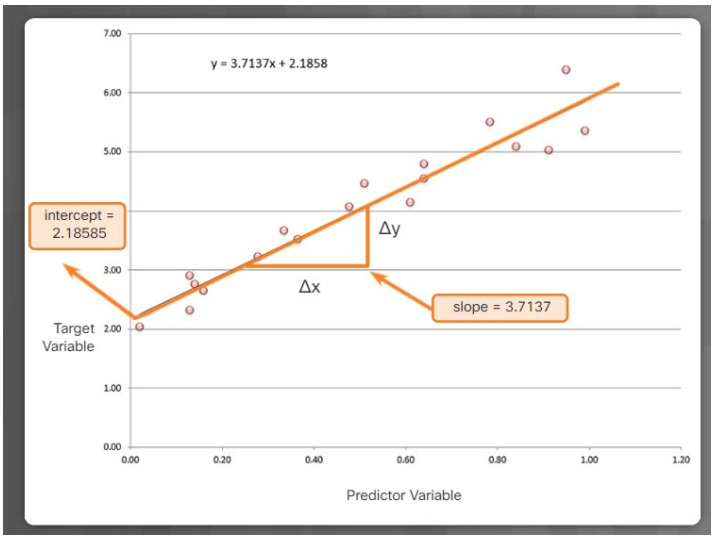
Obr. regresní přímka ve vztahu ke korelaci

Regresní proces v tomto případě spočívá v *nalezení sklonu* (slope) a průsečíku přímky, který minimalizuje součet vzdáleností mezi přímkou a všemi datovými body (viz obrázek). Při použití lineárních modelů je nejběžněji používaným algoritmem odhad těchto optimálních parametrů modelu, nazývá se **metoda nejmenších čtverců**. Podrobné informace o metodě nejmenších čtverců.

$Rovnice\ přímky\ v\ tzv.\ směrnicovém\ tvaru: y = kx + q$

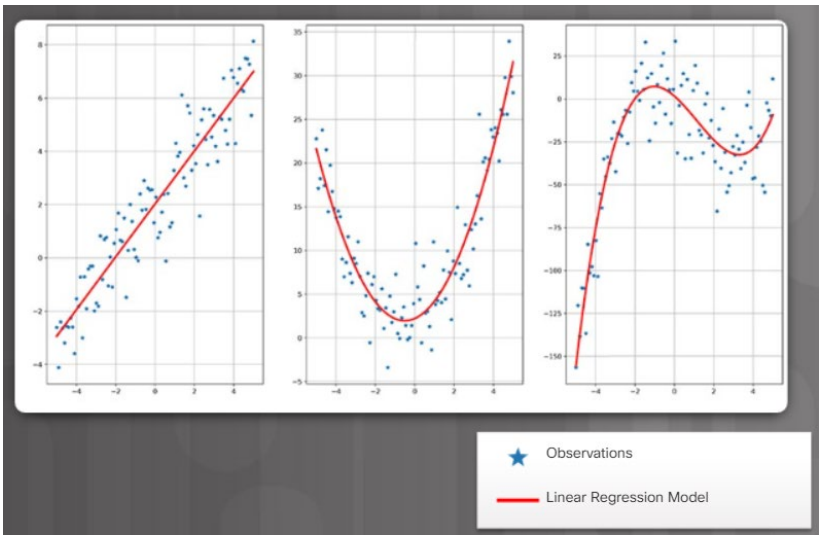
q – bod v němž přímka protíná osu y

k – je směrnice, která nám říká, jaký sklon má přímka vůči ose x (jaké stoupání je na jednotku délky nezávislé proměnné)



Obr. vzorec přímky ve směrnicovém tvaru

Na dalším obrázku můžeme vidět tři datové soubory, z nichž každý má jednu závislou proměnnou (target) a jednu nezávislou proměnnou (predictor). Ve všech třech případech lze pozorovat to, jak navzdory šumu ovlivňujícímu pozorování existuje jasná křivka zachycující vztah mezi proměnnými. Červená čára představuje lineární regresní model, který minimalizuje vzdálenost od všech pozorování.



1.4.2 APLIKACE LINEÁRNÍ REGRESE

Regresní analýza má mnoho využití. Je často aplikována v obchodní a finanční analýze s historickými daty, aby poskytla informace pro budoucí strategie. Může být používána k předpovědi trendů v ekonomice a může být základem pro politická rozhodnutí, která řídí hospodářský růst. Chování zákazníků lze také předvídat, aby bylo možné rozpoznat např. podvodné chování v oblasti pojištění, nebo spotřebitelských úvěrů.

Ve zdravotnictví lze pomocí vícenásobné regrese vyhodnotit, která z mnoha proměnných může ovlivnit cílovou proměnnou. Například by mohl být analyzován vztah mezi výběrem životního stylu, jako je kouření, množství cvičení, stravovací návyky, aby bylo možné určit, jak ovlivňují zdravotní stav (krevní tlak, cukrovka, nebo dokonce očekávaná délka života).

1.5. KLASIFIKACE

Klasifikace je dalším běžným problémem strojového učení, který zapadá do kategorie učení pod dohledem. V posledním desetiletí bylo dosaženo masivních zlepšení, zejména v oblasti *rozpoznávání obrazu*. Klasifikaci lze chápat, jako regresní problém, kde je *cílová proměnná diskrétní* a představuje třídu, do které lidský expert klasifikuje vzorek dat.

V problémech klasifikace je běžné poskytovat nejen sadu příkladů datových bodů pro každou třídu, ale také určit, které vlastnosti každého datového bodu jsou důležité pro odhad odpovídající třídy.

Tyto vlastnosti mohou být dostupné přímo z hodnot senzorů, ale častěji je potřeba je vypočítat, nebo extrahovat z nezpracovaných dat, než je využijeme v algoritmu. *Definování klíčových vlastností je zásadním krokem*, který s výjimkou velmi pokročilých algoritmů, jako je Deep Learning, spoléhá na lidské odborné znalosti.

Například cestovní kancelář má zájem poskytovat zákazníkům na webu hodnocení spolehlivosti letů. Metodou pokusů a omylů různých použitých modelů bylo určeno, které proměnné ze všech v souboru dat jsou pro klasifikace nejrelevantnější. Pouze tyto relevantní vlastnosti jsou extrahovány z dat a použity k tréninku klasifikátoru. Hodnocení spolehlivosti je pak navrženo tak, aby sdělovalo míru pravděpodobnosti, že let poletí včas, bude zpožděn, nebo zrušen.

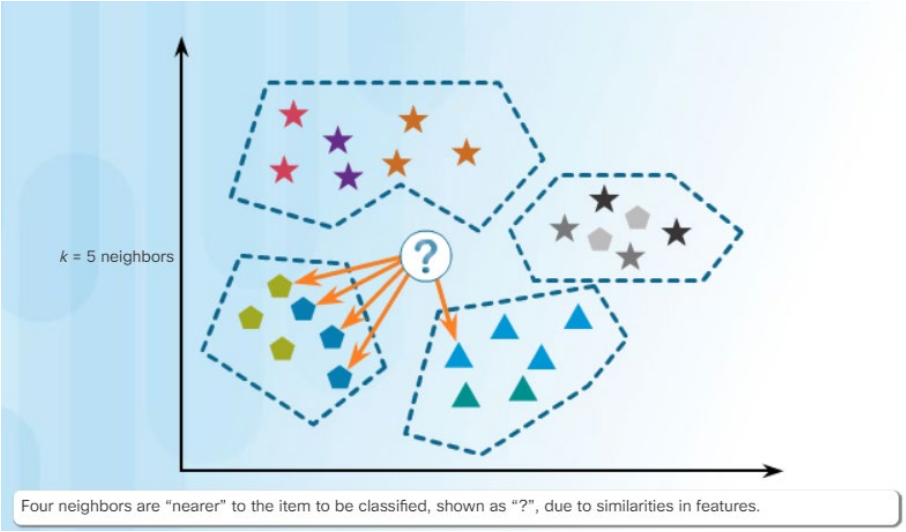
1.5.1 KLASIFIKAČNÍ ALGORITMY

Existuje mnoho klasifikačních algoritmů, které jsou oblíbené pro různé účely. Stručně popíšeme tři z nich.

k-Nearest Neighbor (k-NN)

Jedná se možná o nejjednodušší klasifikátor, který používá vzdálenost mezi tréninkovými daty, jako míru podobnosti. Chcete-li si představit, jak funguje k-NN, představte si, že každý vzorek má dvě vlastnosti, jejichž hodnoty lze vyjádřit 2D grafem.

Na obrázku jsou datové body každé třídy označeny jiným symbolem. Vzdálenost mezi body představuje rozdíl mezi hodnotami jeho vlastností. Při vložení nového datového bodu se klasifikátor podívá na nejbližší tréninkové body. *Nový bod bude přiřazen do třídy, která je nejběžnější třídou mezi sousedy*.

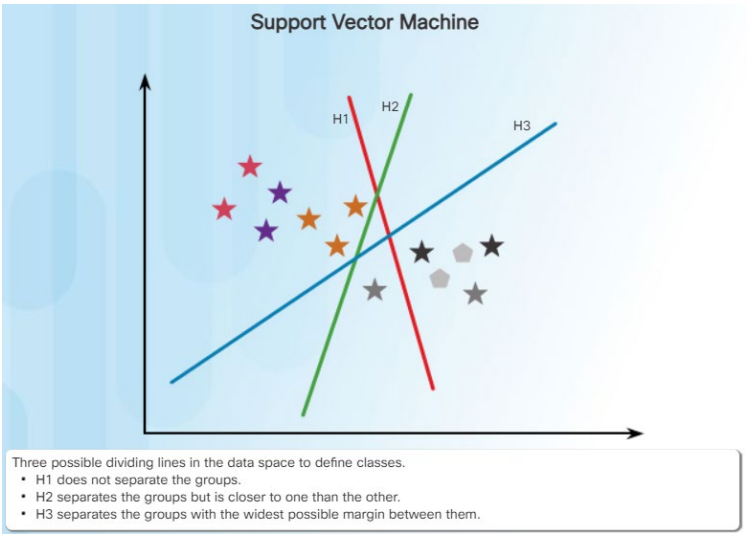


Obr. algoritmus k-Nearest Neighbor

Support vector machines (SVM)

SVM je příklad klasifikátoru strojového učení pod dohledem. Spíše než zakládat přiřazení ke kategoriím na základě vzdálenosti od jiných bodů, *SVM vypočítá hranici neboli nadrovinu, která lépe odděluje jednotlivé skupiny*.

Na obrázku je H3 nadrovina, která maximalizuje vzdálenost mezi tréninkovými body dvou tříd, označenými barevně, nebo černobíle. Pokud bude přidán nový datový bod, bude klasifikován podle toho, zda leží na jedné, nebo druhé straně H3.

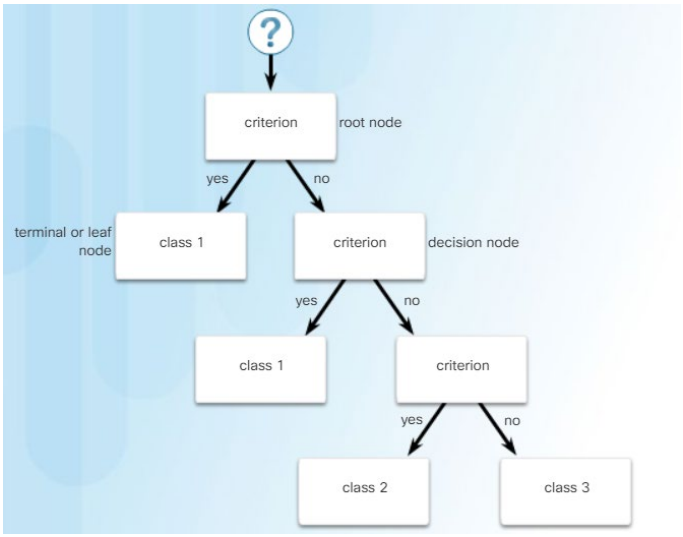


Obr. Support Vector Machine

Decision Tree

Rozhodovací stromy představují klasifikační problém jako soubor rozhodnutí založených na hodnotách vlastností.

Každý uzel stromu představuje prahovou hodnotu prvku a rozděluje data do dvou menších sad. Rozhodovací proces se opakuje u všech prvků, přičemž strom roste, dokud není vypočítán optimální způsob rozdělení vzorků. *Klasifikaci nového vzorku pak lze získat sledováním větví stromu na základě jeho vlastností.*



Obr. rozhodovací strom

1.5.2 APLIKACE KLASIFIKACE

Hodnocení rizik – klasifikační systémy mohou být použity k určení, které z mnoha faktorů přispívají k pravděpodobnosti různých rizik. Například pro klasifikaci řidičů do kategorie s nízký, středním a vysokým rizikem pojištění automobilů. Pro úpravu pojistného, které řidič platí, lze použít řadu faktorů podle úrovně rizika.

Lékařská diagnostika – klasifikační systémy mohou používat řízené otázky k sestavení rozhodovacího stromu, který může pomoci diagnostikovat různé nemoci a rizika onemocnění. Klasifikační systémy strojového učení by také mohly provádět předběžnou analýzu velkého počtu diagnostických snímků a označovat podezřelé stavy ke kontrole lékaři.

Rozpoznávání obrazu – například při rozpoznávání rukopisu může systém pracovat s úkolem identifikovat ručně psané číslice. Číslice 0 - 9 lze považovat za třídy. Klasifikátor je opatřen velkým vzorkem ručně psaných číslic, z nichž každá byla označena skutečnou zastoupenou číslicí. Klasifikátor by hledal vlastnosti, které jsou s největší pravděpodobností přítomné a jedinečné pro každé z čísel.

1.6. POUŽITÁ LITERATURA

[1] CISCO Big Data & Analytics: Advanced Data Analytics and Machine Learning. *Www.netacad.com* [online]. [cit. 2021-11-24]. Dostupné z: https://contenthub.netacad.com/legacy/IoTFBDA/2.01/en/index.html#4.0.1.1

[2] Algoritmy strojového učení: Úvod do matematiky a logiky stojící za strojovým učením. *Azure* [online]. [cit. 2021-11-24]. Dostupné z: https://azure.microsoft.com/cs-cz/overview/machine-learning-algorithms/#overview

[3] Korelace: Korelace ve statistice. *Wikipedie* [online]. [cit. 2021-11-24]. Dostupné z: https://cs.wikipedia.org/wiki/Korelace