Алгоритъм на k-най-близките съседи (kNN)

В тази лекция е направено въведение в алгоритъма k-Nearest Neighbors (kNN). Алгоритъмът kNN е един от най-известните алгоритми за машинно обучение.

По-долу е разгледан алгоритъма kNN както на теория, така и на практика. Въпреки че често се пропуска теоретичната част и се фокусират само върху използването на библиотеки, не искате да сте зависими от автоматизирани пакети за вашето машинно обучение. Важно е да научите за механиката на алгоритмите за машинно обучение, за да разберете техния потенциал и ограничения.

В същото време е важно да разберете как да използвате даден алгоритъм на практика. Имайки това предвид, във втората част на тази лекция ще се съсредоточите върху използването на kNN в библиотеката на Python scikit-learn, с разширени съвети за повишаване на производителността до максимум.

В тази лекция ще научите как да:

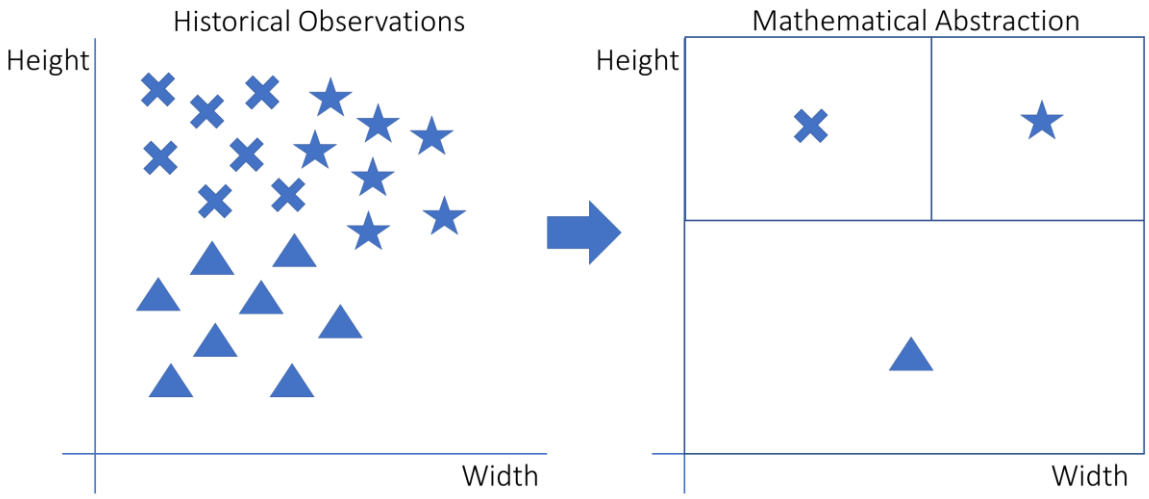
- Обяснете алгоритъма kNN както интуитивно, така и математически
- Настройте хиперпараметрите на kNN с помощта на GridSearchCV
- Добавете пакетиране към kNN за по-добро представяне.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

1. Основи на машинното обучение

За да се включите, струва си да направите крачка назад и да направите бързо проучване на машинното обучение като цяло. Тук ще получите въведение в основната идея зад машинното обучение и ще видите как алгоритъмът kNN се свързва с други инструменти за машинно обучение.

Общата идея на машинното обучение е да се получи модел за научаване на тенденции от данни по всяка тема и възможност за възпроизвеждане на тези тенденции върху сравними данни в бъдеще. Ето диаграма, очертаваща основния процес на машинно обучение



Тази графика е визуално представяне на модел на машинно обучение, който е монтиран върху данни. Отляво са оригиналните наблюдения с три

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

променливи: височина, ширина и форма. Формите са звезди, кръстове и триъгълници.

Формите са разположени в различни области на графиката. Вдясно виждате как тези първоначални наблюдения са преведени в правило за вземане на решение. За ново наблюдение трябва да знаете ширината и височината, за да определите в кой квадрат попада. Квадратът, в който попада, от своя страна определя каква форма е най-вероятно да има.

За тази задача могат да се използват много различни модели. Моделът е математическа формула, която може да се използва за описание на точки от данни. Един пример е линейният модел, който използва линейна функция, дефинирана от формулата $y = ax + b$.

Ако оценявате или пасвате модел, намирате оптималните стойности за фиксираните параметри, като използвате някакъв алгоритъм. В линейния модел параметрите са a и b.

След като моделът бъде оценен, той се превръща в математическа формула, в която можете да попълните стойности за вашите независими променливи, за да направите прогнози за вашата целева променлива.

2. Отличителни черти на kNN

Сега, след като разбирате основната идея зад машинното обучение, следващата стъпка е да разберете защо има толкова много налични модели. Линейният модел, който току-що видяхте, се нарича линейна регресия.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Линейната регресия работи в някои случаи, но не винаги прави много точни прогнози. Ето защо математиците са измислили много алтернативни модели за машинно обучение, които можете да използвате. Алгоритъмът k-най-близки съседи е един от тях.

Всички тези модели имат своите особености. Ако работите върху машинно обучение, трябва да имате дълбоко разбиране за всички тях, за да можете да използвате правилния модел в правилната ситуация. Ще разберете защо и кога да използвате kNN, след това ще разгледате как kNN се сравнява с други модели за машинно обучение.

kNN е контролиран алгоритъм за машинно обучение

Първото определящо свойство на алгоритмите за машинно обучение е разделението между контролирани и неконтролирани модели. Разликата между контролираните и неконтролираните модели е постановката на проблема.

В контролираните модели имате два типа променливи едновременно:

1. Целева променлива, която също се нарича зависима променлива или променлива *y*.

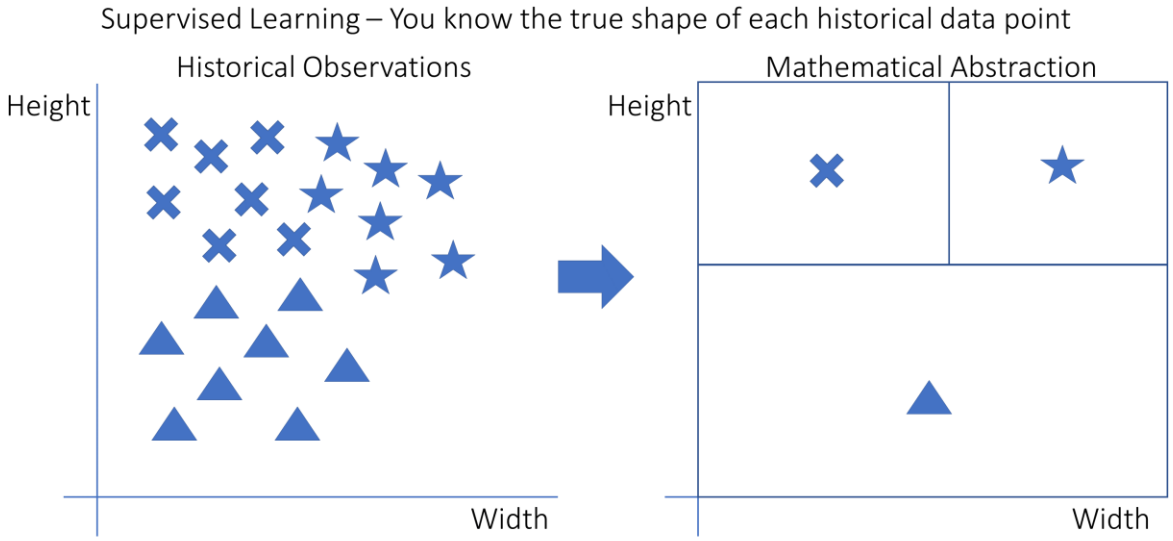
2. Независими променливи, които също са известни като *x* променливи или обяснителни променливи.

Целевата променлива е променливата, която искате да предвидите. Зависи от независимите променливи и не е нещо, което знаете предварително.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Независимите променливи са променливи, които знаете предварително. Можете да ги включите в уравнение, за да предвидите целевата променлива. По тази начин той е относително подобен на случая $y = ax + b$.

В графиката, която сте виждали преди, и следващите графики, целевата променлива е формата на точката от данни, а независимите променливи са височината и ширината. Можете да видите идеята зад контролираното обучение в следната графика:



В тази графика всяка точка от данни има височина, ширина и форма. Има кръстове, звезди и триъгълници. Вдясно е правило за вземане на решение, което моделът на машинно обучение може да е научил.

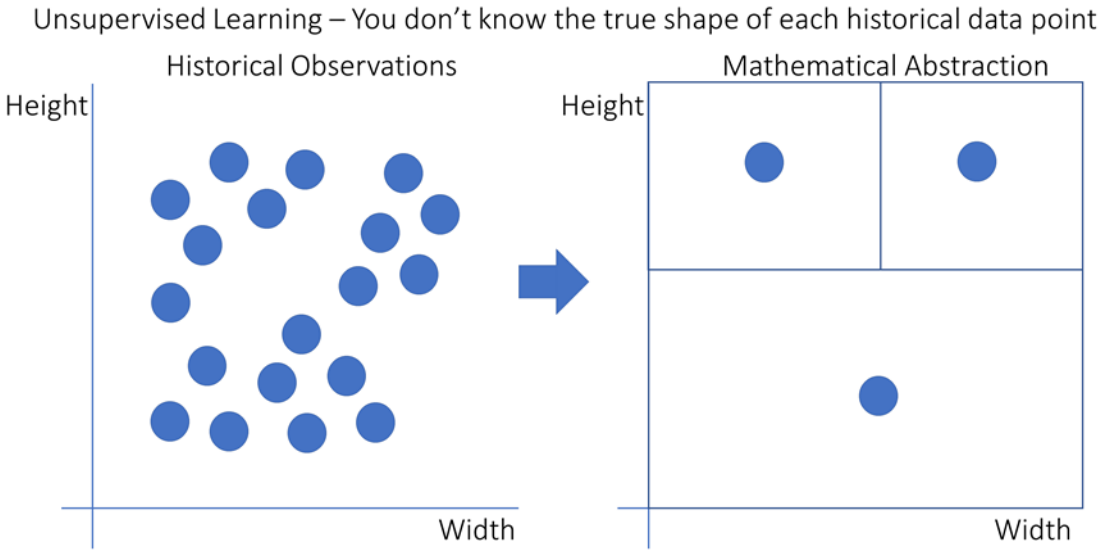
В тази случай наблюденията, отбелязани с кръст, са високи, но не широки. Звездите са едновременно високи и широки. Триъгълниците са къси,

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

но могат да бъдат широки или тесни. По същество моделът е научил правило за вземане на решение, за да реши дали едно наблюдение е по-вероятно да бъде кръст, звезда или триъгълник въз основа само на неговата височина и ширина.

При моделите без надзор няма разделение между целеви променливи и независими променливи. Неконтролираното обучение се опитва да групира точки от данни, като оценява тяхното сходство.

Както можете да видите в примера, никога не можете да сте сигурни, че групираните точки от данни по същество си принадлежат една към друга, но стига групирането да има смисъл, то може да бъде много ценно на практика. Можете да видите идеята зад обучението без надзор в следната графика



В тази графика наблюденията вече нямат различни форми. Всички те са кръгове. И все пак те все още могат да бъдат групирани в три групи въз

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

основа на разстоянието между точките. В тази конкретен пример има три групи от точки, които могат да бъдат разделени въз основа на празното пространство между тях.

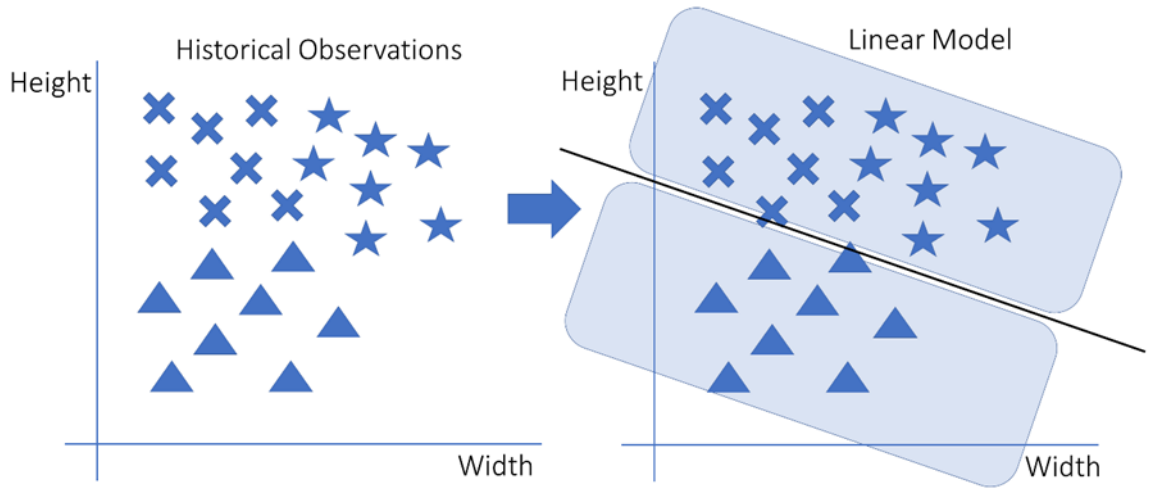
Алгоритъмът kNN е модел за контролирано машинно обучение. Това означава, че прогнозира целева променлива, използвайки една или множество независими променливи.

kNN е нелинеен алгоритъм за обучение

Второ свойство, което прави голямата разлика в алгоритмите за машинно обучение, е дали моделите могат или не да оценяват нелинейни връзки.

Линейните модели са модели, които предвиждат с помощта на линии или хиперравнини. На изображението моделът е изобразен като линия, начертана между точките. Моделът $y = ax + b$ е класическият пример за линеен модел. Можете да видите как линеен модел може да пасне на примерните данни в следния схематичен чертеж:

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.



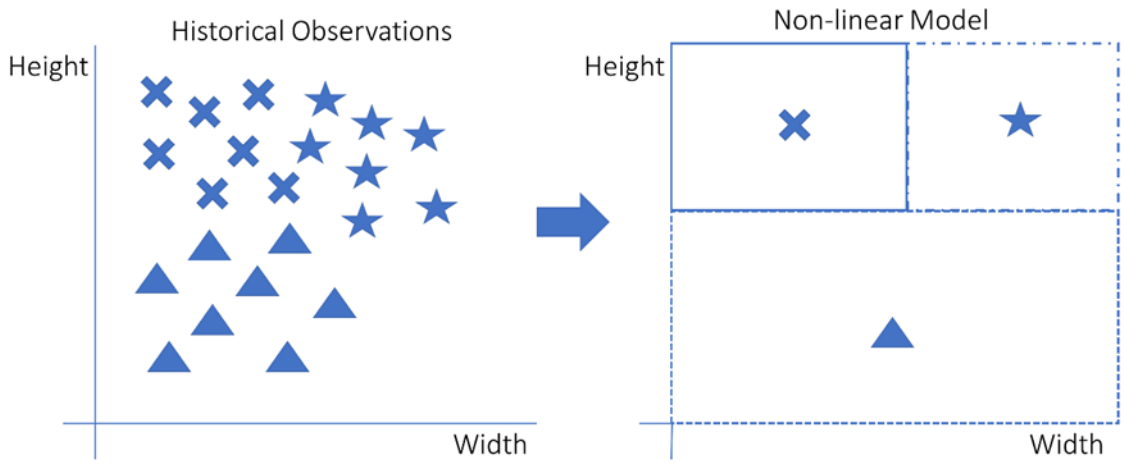
На тази графика точките с данни са изобразени отляво със звезди, триъгълници и кръстове. Вдясно е линеен модел, който може да разделя триъгълниците от нетриъгълниците. Решението е линия. Всяка точка над линията е нетриъгълник, а всичко под линията е триъгълник.

Ако искате да добавите друга независима променлива към предишната графика, ще трябва да я начертаете като допълнително измерение, като по тази начин създадете куб с фигурите вътре в него. И все пак една линия не би могла да разреже куб на две части. Многомерният двойник на линията е хиперравнината. Следователно линейният модел е представен от хиперравнина, която в случай на двумерно пространство е линия.

Нелинейните модели са модели, които използват всеки подход, различен от линия, за да разделят своите случаи. Добре известен пример е

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

дървото на решенията, което в основата си е дълъг списък от оператори if... else. В нелинейната графика операторите if... else биха ви позволили да нарисувате квадрати или всяка друга форма, която искате да нарисувате. Следващата графика изобразява нелинеен модел, приложен към примерните данни:



Тази графика показва как едно решение може да бъде нелинейно. Правилото за решение се състои от три квадрата. Кутията, в която попада нова точка от данни, ще определи нейната прогнозирана форма. Обърнете внимание, че не е възможно да се побере това наведнъж с помощта на линия: Необходими са две линии. Тази модел може да бъде пресъздаден с оператори if ... else, както следва:

• Ако височината на точката с данни е ниска, това е триъгълник.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

• В противен случай, ако ширината на точката с данни е ниска, това е кръст.

• В противен случай, ако нито едно от горните не е вярно, тогава е звезда.

kNN е пример за нелинеен модел. По-късно в тази лекция ще се върнете към точния начин, по който моделът е изчислен.

kNN е контролиран обучаем както за класификация, така и за регресия.

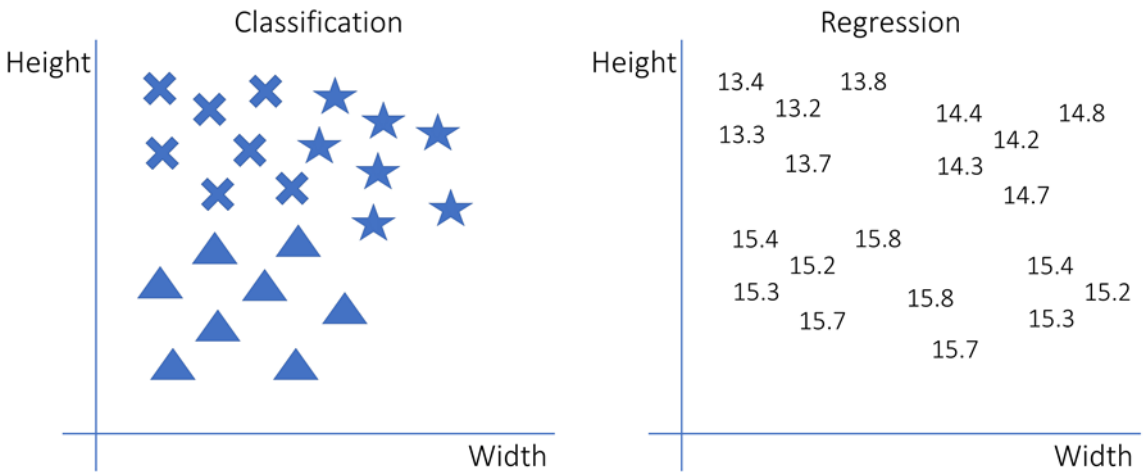
Алгоритмите за контролирано машинно обучение могат да бъдат разделени на две групи въз основа на типа целева променлива, която могат да предвидят:

1. Класификацията е задача за прогнозиране с категорична целева променлива. Класификационните модели научават как да класифицират всяко ново наблюдение. Тази присвоен клас може да бъде правилен или грешен, а не по средата. Класически пример за класификация е наборът от данни за ириса, в който използвате физически измервания на растения, за да предвидите техния вид. Известен алгоритъм, който може да се използва за класификация, е логистичната регресия.

2. Регресията е задача за прогнозиране, в която целевата променлива е числова. Известен пример за регресия е предизвикателството за цените на жилищата. В това състезание за машинно обучение участниците се опитват да предскажат продажните цени на къщи въз основа на множество независими променливи.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

В следващата графика можете да видите как биха изглеждали регресия и класификация, използвайки предишния пример:



Лявата част на това изображение е класификация. Целевата променлива е формата на наблюдението, която е категорична променлива. Дясната част е регресия. Целевата променлива е числова. Правилата за вземане на решения могат да бъдат абсолютно еднакви за двата примера, но техните интерпретации са различни.

За една единствена прогноза класификациите са правилни или грешни, докато регресиите имат грешка в непрекъсната скала. Наличието на числена мярка за грешка е по-практично, така че много класификационни модели предвиждат не само класа, но и вероятността да бъдете в който и да е от класовете.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Някои модели могат да правят само регресия, други могат да правят само класификация, а някои могат да правят и двете. Алгоритъмът kNN безпроблемно се адаптира към класификация и регресия.

kNN е бърз и интерпретируем.

Като последен критерий за характеризиране на моделите за машинно обучение, трябва да вземете предвид сложността на модела. Машинното обучение и особено изкуственият интелект в момента процъфтяват и се използват в много сложни задачи, като разбиране на текст, изображения и реч, или за самоуправляващи се автомобили.

Ще са необходими много повече данни, за да паснете на по-сложен модел, а данните не винаги са налични. Не на последно място, по-сложните модели са по-трудни за тълкуване от нас, хората, и понякога тази интерпретация може да бъде много ценна.

Тук се крие силата на модела kNN. Той позволява на своите потребители да разберат и интерпретират какво се случва вътре в модела и се развива много бързо. Това прави kNN чудесен модел за много случаи на използване на машинно обучение, които не изискват много сложни техники.

Недостатъци на kNN

Справедливо е да бъдем честни и за недостатъците на алгоритъма kNN. Както беше споменато по-рано, истинският недостатък на kNN е способността му да се адаптира към много сложни връзки между независими и зависими

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

променливи. kNN е по-малко вероятно да се представи добре при сложни задачи като компютърно зрение и обработка на естествен език.

Можете да опитате да увеличите производителността на kNN доколкото е възможно, потенциално чрез добавяне на други техники от машинно обучение. В определен момент на сложност обаче kNN вероятно ще бъде по-малко ефективен от други модели, независимо от начина, по който е бил настроен.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.