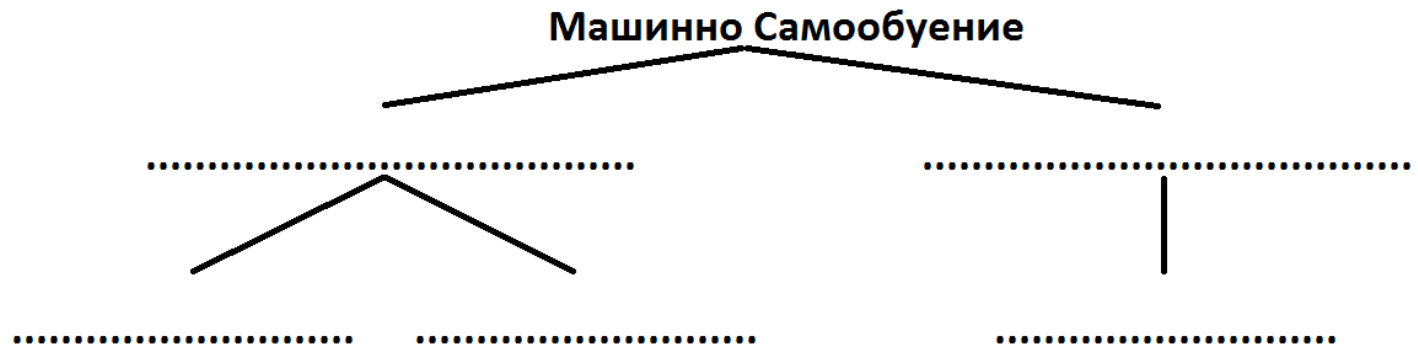
Какво представлява Машинното самообучение?

На колко вида се дели Машинното самообучение? Моля, попълнете следната йерархия и дайте определение за всеки един от елементите в нея:



Учене с учител
- Класификация
- Регресия

Учене без учител
- клъстеризация

Какво представлява процесът на предварителна обработка на данните? Защо е необходим?
Data Cleaning: Липсващи данни; Неконсистентни данни; Data Transformation: Нормализация (стойностите трябва да са в специфичен диапазон); Агрегация; Генерализация; Конструкция на атрибутите; Data Reduction: Редуцуране броя на атрибутите, техните стойности и броя на инстанциите.
Изброите от 3 до 5 алгоритъма за класификация/регресия:

Изброите от 3 до 5 алгоритъма за клъстеризация:

Можете ли да посочите алгоритъм, който да може да бъде приложен и за класификация и за клъстеризация?

Каква е разликата между Глобални и Локални подходи при обучаване на модел?

Дайте определение за мързеливо учене (lazy learning) и нетърпеливо учене (eager learning). Сравнете ги. Дайте пример за минимум по един алгоритъм за всеки един от типовете учене.

Обяснете как работи алгоритъма за k най-близки съседи kNN.

Обяснете как работят Дървета на решенията. Посочете какви имплементации знаете за Дървета на решенията и какви са разликите между тях.

Какъв проблем имат дърветата на решенията? Как избираме хипотези?

Прекомерното нагаждане (overfitting) е значително практическо затруднение за моделите на дърветата на решение, както и за много други предсказващи модели. Прекомерно нагаждане се случва, когато обучаващият се алгоритъм продължава да развива хипотези, които намаляват грешката по време на обучение, но повишат грешката по време на тестване. Има няколко подхода за избягване на прекомерно нагаждане при изграждането на дървета на решенията.

Предварително отрязване/обрязване (Pre-pruning) - спира да строй дървото преди то перфектно да класифицира тренировъчните данни.

По-следващо отрязване/обрязване (Post-pruning) - построява дърво, което перфектно да класифицира тренировъчните данни и след това го отрязва/обрязва.

На практика, вторият подход е по-успешен, защото не е лесно да се прецени кога да се спре да се строи дървото.

Хипотези избираме по метода на Окам (Бръснача на Окам) – „Между конкуренти хипотези, трябва да се избере тази с най-малко предположения“.

Напишете формулите за ентропия и гейн. Има ли нещо специфично при тях?

Ентропия – мярка на несигурността на случайна променлива. Колкото по-малка е ентропията, толкова по-сигурна е информацията, която имаме.

Гейн (функция на печалбата) – най-малката оставаща ентропия след теста.

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

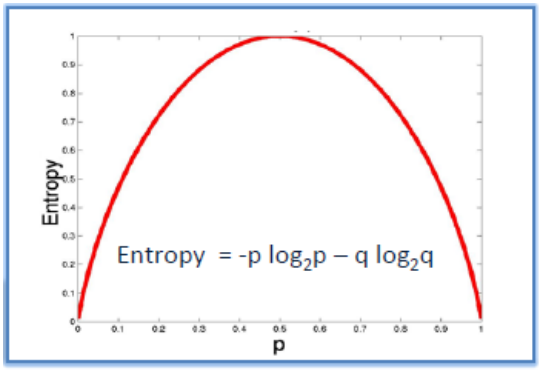
Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)***E**(3,2) + **P**(Overcast)***E**(4,0) + **P**(Rainy)***E**(2,3)
= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971
= 0.693

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			



Entropy = -0.5 $\log_2$0.5 – 0.5 $\log_2$0.5 = 1

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)
= 0.940 – 0.693 = 0.247

E(a,b) = - a log a - b log b
-1/2 log 1/2 - 1/2 log 1/2 = 1
-0/2 log 0/2 - 2/2 log 2/2 = 0
-1/3 log 1/3 - 2/3 log 2/3 = 0.39 + 0.52 ~ 0.9

Обяснете как работи Наивният Бейсов Класификатор? Защо е наивен? Напишете теоремата на Бейс и я разпишете! Как бихте се справили с нулеви вероятности?

Обяснете как работи kMeans. Какви са характерните особености за този алгоритъм? На коя група алгоритми прилича?

Какви са Йерархичните подходи за клъстеризация? Обяснете ги и ги сравнете.

На какви видове се делят невронните мрежи според слоевете, които имат? Каква е разликата между тях? Сравнете ги.

Какви други видове невронни мрежи знаете? Как работят невронните мрежи?

Какво е Backpropagation?

Какво представляват асоциативните правила?
Учене обосновано на асоциативни правила е метод за откриване на интересни връзки между променливи в големи бази данни. Предназначено е за откриването/идентифицирането на строги правила, открити в базата данни използвайки мерки оценяване на интереси.
За да изберете интересни правила от множеството с всички правила се използват ограничения върху различните мерки за оценяване на значението и интереса. Най-добре познати ограничения са минималния праг на поддръжка (означение колко често дадена група (itemset) присъства в базата данни) и доверие (означение, колко често дадено правило е установено като истина).

Обяснете как работи алгоритмът Apriori.

На какви подмножества може да се раздели един набор от данни (dataset)? Обяснете предназначението на всяко едно подмножество.

Как може да оцените даден алгоритъм.

Какво е крос валидация и за какво се ползва?

Посочете ако знаете някакви други практики подходи или каквото и да било характерно за машинното самообучение.