1. **Co to są drzewa decyzyjne (decission tree)? MJ**

<https://en.wikipedia.org/wiki/Decision_tree_learning>

1. **Co to jest i po co robi się regularyzację? MJ**

asd

<https://en.wikipedia.org/wiki/Regularization_(mathematics)>

1. **Różnice pomiędzy L1 i L2? MJ**
2. **Co to jest Random Forrest? NK**
3. **Co to jest ensembling? jakie algorytmy znasz? MJ**
4. **Co to jest boosting? MJ**

<https://datascience.stackexchange.com/questions/16904/gbm-vs-xgboost-key-differences>

1. **Jak działa gbm?**

<https://en.wikipedia.org/wiki/Gradient_boosting>

1. **Jak działa catboost?**
2. **Jak działa xgboost? MJ**
3. **Główna wada Random Forest. NK**
4. **Co to jest SVD? NK**
5. **Co to jest t-sne?**
6. **Jak wykorzystuje się autoencodery do redukcji wymiarowości?**
7. **Co to jest MaxPooling?**
8. **Co to jest PCA? Opisz działanie NK**
9. **Dlaczego redukuje się wymiary zbioru danych. NK**
10. **Metody redukcji wymiarów. Jakie znasz? Czym są? Po co się je stosuje? NK**
11. **Co to jest embedding MJ**
12. **Na czym polega Bias-Variance tradeoff? MJ**
13. **Zdefiniuj bias MJ**
14. **Zdefiniuj variance MJ**
15. **Co to jest dropout? NK**
16. **Co to jest i dlaczego stosuje się one-hot encoding? MJ**
17. **Co to jest i jak sobie radzić skewed data? MJ**
18. **Jak sobie poradzić z niezbalansowanym binarną klasyfikacja (Jest o wiele więcej jednej klasy niz drugiej w zbiorze danych)? MJ**
19. **Jak sobie radzić z overfittingiem? MJ**
20. **Różnice SVM v Random Forrest? NK**
21. **Co to jest problem interpretowalności (interpretability) w algorytmach ml? MJ**

<https://arxiv.org/pdf/1606.03490.pdf>

1. **Wyprowadzenie Metoda Najmniejszych Kwadratów NK**

<http://fa.bianp.net/teaching/2018/eecs227at/>

1. **Co to jest generator w pythonie?**
2. **Jakie problemy mogą być z hash mapami? Jak sobie z nimi radzic?**
3. **Pandas- omów : groupby, agregacja, merge**
4. **Omów regresję liniową NK**
5. **Co to jest LSE? NK**

<https://towardsdatascience.com/ml-notes-why-the-least-square-error-bf27fdd9a721>

<https://stats.stackexchange.com/questions/118/why-square-the-difference-instead-of-taking-the-absolute-value-in-standard-devia>

<https://stats.stackexchange.com/questions/127598/square-things-in-statistics-generalized-rationale/128619>

1. **Co to jest gradient descent?**

<http://mccormickml.com/2014/03/04/gradient-descent-derivation/>

<http://fa.bianp.net/teaching/2018/eecs227at/>

<https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3>

Jest to algorytm służący do minimalizacji funkcji (kosztu).

Rozwiązywania problemów typu:

znajdź minimum funkcji f(x), gdzie x jest z przestrzeni , a f jest funkcją róźniczkowalną

algorytm:

dla początkowej wartości x = x0 powtarzaj:

x += - alfa \* gradient(f(x))

gdzie alfa to wartość kroku dostosowana empirycznie

1. **Co to jest stochastic gradient descent?**

Przy obliczaniu kolejnego kroku pochodne cząstkowe składające się na gradient liczone są na podstawie pojedynczego przykładu ze zbioru treningowego w odróżnieniu od wyliczenia na podstawie całego zbioru treningowego. Przykłady ze zbioru treningowego są w losowej kolejności. Stosuje się zazwyczaj od 1 do 10 przejść przez cały zbiór treningowy. W ten sposób złożoność obliczeniowa pojedynczego kroku jest istotnie zredukowana kosztem bardziej chaotycznej zbieżności (mogą wystąpić kroki algorytmu które oddalają bieżące rozwiązanie od optimum). Zazwyczaj też algorytm nie jest zbieżny do pojedynczego punktu optimum lecz od pewnego momentu podawane rozwiązania zaczynają krążyć w ograniczonym obszarze w pobliżu optimum.

Stosuje się w przypadku, kiedy np. nie możemy zmieścić całego zbioru treningowego w pamięci, albo chcemy trenować model w sposób online-owy.

1. **regresja logistyczna i czym sie różni od liniowej NK**

rl służy do rozwiązywania problemów klasyfikacyjnych

TODO: logistyczna tez do regresji mozna wykorzystywac - zobacz naa stronie sklearna

1. **Co to jest optymum (minimum) lokalne? Jak sobie z nim radzić (jak unikać)? MJ**
2. **Co to jest True Positive/Negative, False Positive/Negative?**

Są to miary przy ewaluacji klasyfikatorów binarnych.

true positive (TP) - sklasyfikowano: True, na prawdę: True

false positive (FP) - sklasyfikowano: True, na prawdę: False

true negative (TN) - sklasyfikowano: False, na prawdę: False

false negative (FN) - sklasyfikowano: False, na prawdę: True

<https://en.wikipedia.org/wiki/Confusion_matrix>

<https://pl.wikipedia.org/wiki/Tablica_pomy%C5%82ek>

1. **Cot to jest AUCROC?**

<https://www.youtube.com/watch?v=OAl6eAyP-yo&feature=youtu.be>

<https://stats.stackexchange.com/questions/132777/what-does-auc-stand-for-and-what-is-it>

1. **Jakie znasz metody ewaluacji klasyfikatorów binarnych?**

<https://en.wikipedia.org/wiki/Precision_and_recall>

<http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf>

<http://people.inf.elte.hu/kiss/11dwhdm/roc.pdf>

<https://en.wikipedia.org/wiki/Receiver_operating_characteristic>

1. **Co to jest gold standard? MJ**
2. **Co to jest long tailed distribution?**
3. **Co to jest latent semantic indexing?**
4. **Co to jest resampling methods?**
5. **Jak jest różnica pomiędzy NN z funkcja aktywacyjną typu SoftMax, regresja logistyczną i maximum entropy classifier?**
6. **Co to jest sparse/dense matrix? MJ**
7. **Co to jest feature vector?**

Feature jest jedną zmierzoną cechą charakterystyczną jakiejś obserwacji.

Feature vector to zbiór tych zmierzonych cech, dotyczących jednej obserwacji, wyrażonych w formie wektora.

<https://en.wikipedia.org/wiki/Feature_(machine_learning)>

<https://brilliant.org/wiki/feature-vector/>

1. **Co to jest funkcja kosztu (loss/cost function)?**

Jest to funkcja reprezentująca różnice pomiędzy wartością estymowaną, a rzeczywistą.

1. **Co to jest regresja liniowa?**

Regresja liniowa jest modelem służącym do estymowania wartości zmiennej y na podstawie innych zmiennych x. Regresja liniowa do estymacji wykorzystuje jakąś funkcję liniową (albo wielomian), którego parametry są obliczane na podstawie obserwacji ze zbioru treningowego.

<https://pl.wikipedia.org/wiki/Regresja_liniowa>

1. **Co to jest Ordinary Least Squares (Metoda Najmniejszych Kwadratów)?**

Jest to jedna z metod dopasowywania (nauczenia) parametrów regresji liniowej poprzez minimalizację funkcji kosztu. Najbardziej klasyczna. Polega na minimalizacji funkcji kosztu będącej kwadratem różnicy wartości przewidywanej przez model i wartości rzeczywistej.

Ma wady, najbardziej widoczna to mała odpornośc na elementy odstające (outliers), najbardziej znanymi przykładami jest Kwartet Anscomb’a.

Złożoność obliczeniowa: , gdzie *n* to liczba obserwacji, a *p* to liczba ficzerów

<http://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares>

<https://en.wikipedia.org/wiki/Ordinary_least_squares>

1. **Co to jest Ridge Regression?**

Jest to jedna z metod dopasowywania parametrów regresji liniowej, oparta na metodzie najmniejszych kwadratów. Różni się funkcją kosztu od MNK tym, że zastosowano w niej wyrażenie regularyzacyjne

<http://scikit-learn.org/stable/modules/linear_model.html#ridge-regression>

1. **Co to jest Lasso? MJ**

Jest to jedna z metod dopasowywania parametrów regresji liniowej, oparta na metodzie najmniejszych kwadratów. Różni się funkcją kosztu od MNK tym, że zastosowano w niej wyrażenie regularyzacyjne

1. **rozwiązywanie problemów optymalizacyjne (analiza mat) - przykład NK**
2. **bagging MJ**
3. **logloss MJ**
4. **Znane algorytmy uczenia maszynowego do procesu klasyfikacji.**

**sklearn NK**

1. **Czym jest współliniowość i dlaczego powoduje problemy. NK**
2. **Coś o SVM NK**
3. **Co to jest kernel w SVM? NK**
4. **Dlaczego SVM musi zmaksymalizować odstęp pomiędzy support vectorami? NK**
5. **Opisz znane Ci kernele w SVM? NK**
6. **Co to jest LDA? NK**
7. **Co to jest QDA? NK**
8. **Sieć neuronowa – jaką wykorzystywałem. NK**
9. **Co to jest kernel trick, dlaczego wykorzystuje się go w SVM, a nie mówi się o nim w sieciach neuronowych? NK**
10. **Na podstawie obrazów predykcji w zagadnieniu klasyfikacji ocenić, która metoda została wybrana? NK**
11. **jak zbudować algorytm rekomendujący użytkownikowi treści na stronie: a. dane po stronie użytkowników, b. dane na stronie , c. segmentacja samych produktów.** MJ

asdsad

<https://www.kaggle.com/fabiendaniel/film-recommendation-engine/notebook>

1. **Co to jest bootstrap. NK**
2. **Co to jest kroswalidacja NK**
3. **Interpretacja zmiennych w regresji logistycznej. NK**
4. **Co to jest stacking MJ**
5. **Co to jest blending MJ**
6. **Co to jest momentum MJ**

<https://distill.pub/2017/momentum/>

1. **Co to jest i jak działa adam optimizer? NK**
2. **Co to jest Clustering, jakie metody znasz, opisz je. NK**
3. **Jakie znasz funkcje kosztu MJ**

<https://medium.com/towards-data-science/ml-notes-why-the-least-square-error-bf27fdd9a721>

1. **Scoring (ewaluacja) modelu, jakie znasz funkcje, wady/zalety MJ**

<https://towardsdatascience.com/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4>

1. **Co to jest word2vec? NK**
2. **Co to jest backprop? Jak działa? NK**

<https://www.kdnuggets.com/2016/06/visual-explanation-backpropagation-algorithm-neural-networks.html>

1. **Co to jest CNN? Co to jest convolution (splot)? NK**
2. **Co to jest RNN? NK**
3. **Co to jest LSTM? NK**
4. **Jak działa YOLO? NK**
5. **Jak działa Faster R-CNN? NK**
6. **Jakie znasz funkcje aktywacji? Jakie są ich wady / zalety? MJ**

<https://www.reddit.com/r/MachineLearning/comments/7leg0v/d_has_there_been_any_studies_on_of_relu_still/?st=JBSVSOIN&sh=48370c8f>

1. **Dlaczego używa się GPU przy trenowaniu sieci neuronowych? NK**
2. **Co to jest reinforced learning? MJ**
3. **Co to jest q-learning? MJ**
4. **Co to jest bag-of-words? MJ**
5. **Co to jest perceptron? NK**
6. **Co to jest TF-IDF? MJ**
7. **Co to jest HMM? MJ**
8. **Co to jest GAN? NK**
9. **Co to jest Bayesian Deep Network? NK**
10. **Co to jest algorytm L-BFGS? MJ**

<http://aria42.com/blog/2014/12/understanding-lbfgs>

1. **Co to jest batch normalization? MJ**
2. **Co to jest learning rate decay? MJ**

<https://arxiv.org/abs/1711.00489>

<https://www.coursera.org/learn/deep-neural-network/lecture/hjgIA/learning-rate-decay>

1. **Co to jest normalizacja danych? Po co się ją stosuje? MJ**
2. **Co to jest preprocessing? MJ**
3. **Co to jest Gaussian Mixture Models? NK**
4. **Co to jest, jak wykryc i jak sobie poradzic z anomaliami w danych (outliers)? MJ**
5. **Skąd wiesz czy znaleziony trend nie jest spowodowany anomalia? MJ**
6. **Co to jest feature selection? Jakie znasz metody? MJ**
7. **Czy feature selection stosuje sie w sieciach neuronowych? NK**
8. **Co to jest K-means? MJ**
9. **Co to jest i jak sobie radzić ze sparse data? MJ**
10. **Co to jest supervised/unsupervised learning? MJ**
11. **Jak sobie radzić z missing lub bad data? MJ**
12. **Jak utworzył byś nowe feature w oparciu o już istniejące? MJ**
13. **Co to jest map-reduce? NK**
14. **Co to jest problem badania sentymentu? Jakie algorytmy staraja sie go rozwiazac? MJ**
15. **Co to jest non-normal probability distribution i jak je stosować? NK**

<http://www.statisticshowto.com/probability-and-statistics/non-normal-distributions/>

1. **Co to jest Heteroskedastyczność i jak sobie z nią radzić? NK**

<https://en.wikipedia.org/wiki/Heteroscedasticity>

1. **Co to jest Multicollinearity (współliniowość) i jak sobie z nią radzić? NK**
2. **Co to jest Naive Bayes? NK**
3. **Co to są p-wartości (p-values)**

<https://en.wikipedia.org/wiki/P-value>

1. **Co to są przedziały ufności (Confidence interval) NK**

<https://en.wikipedia.org/wiki/Confidence_interval>

1. **Co to jest Central Limit Theorem? Jak go udowodnić i jakie są jego zastosowania? NK**
2. **Napisz funkcje która mnoży macierze (macierz-macierz)? MJ**
3. **Co to jest ETL (Extract-Transform-Load)? MJ**
4. **Co to jest lift w data science? MJ**

<https://en.wikipedia.org/wiki/Lift_(data_mining)>

1. **Co to jest KPI? MJ**
2. **Co to jest model robustness? MJ**
3. **Co to jest model fitting? MJ**
4. **Jakie znasz rozwiązania ml do znajdowanie plagiatów? MJ**
5. **Jak sobie radzić z brakującymi danymi (missing values)? MJ**
6. **Co to jest algorytm zillow? MJ**
7. **Jakie znasz rodzaje wizualizacji danych? MJ**
8. **Co to są n-gramy? Gdzie się je stosuje? MJ**
9. **Co to jest bk-tree i do czego służy? MJ**

<https://en.wikipedia.org/wiki/BK-tree>

1. **Co to są hyperparametry? Jakie znasz metody dobierania ich? MJ**
2. **Co to są sieci syjamskie? MJ**

**Cheat Sheats:**

* <https://becominghuman.ai/cheat-sheets-for-ai-neural-networks-machine-learning-deep-learning-big-data-678c51b4b463>
* <https://www.oreilly.com/ideas/the-current-state-of-machine-intelligence-3-0>
* <http://ml-cheatsheet.readthedocs.io/en/latest/>
* <https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/?utm_source=linkedin&utm_medium=cpc&utm_campaign=analytics-global&utm_content=US_skills_audience-network>
* <https://www.kdnuggets.com/2017/09/essential-data-science-machine-learning-deep-learning-cheat-sheets.html?utm_content=buffer3602f&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer>