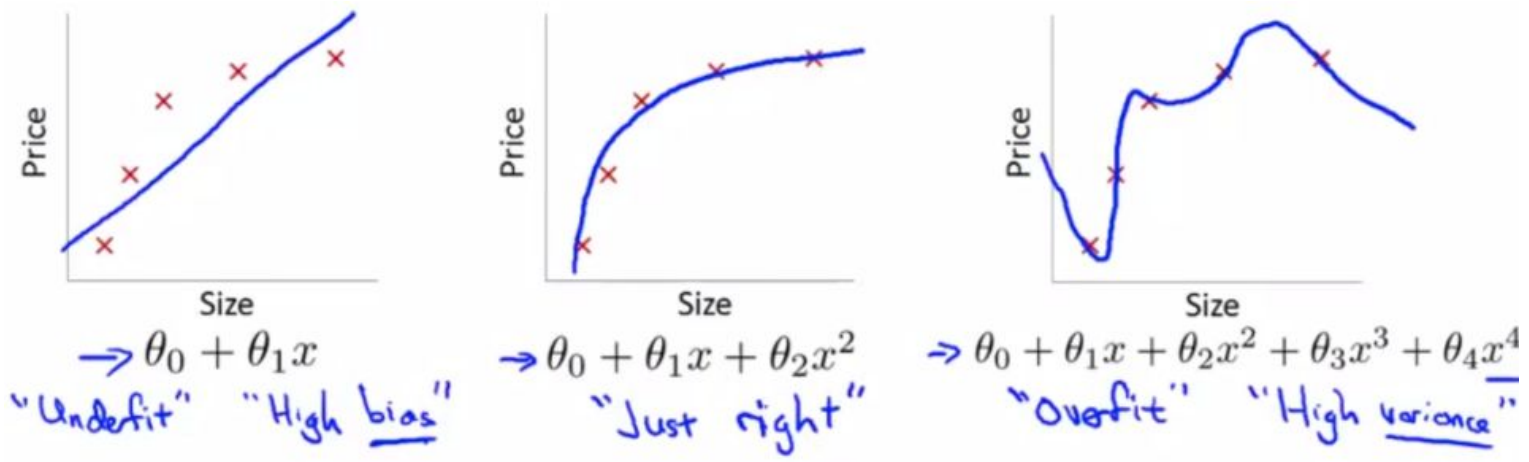
Обучение с учителем: Регуляризация в линейных моделях. Метод Ближайших Соседей (KNN)

Екатерина Кондратьева


NOT SURE IF GOOD MODEL...
...OR JUST OVERFITTING
memegenerator.net

Переобучение (model overfitting)



*Здесь theta (θ) - β

*В предыдущей лекции это были a и b

Регуляризация

Используется для улучшения обобщающей способности модели, то есть уменьшения эффекта переобучения, на практике часто рассматривается логистическая регрессия с регуляризацией.

Регрессия:

МНК функция потерь:

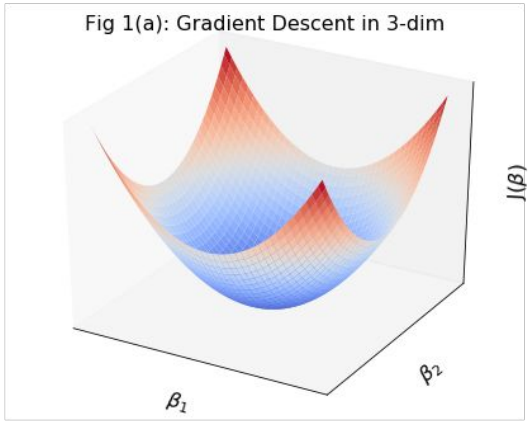
$$\begin{aligned} \mathrm{RSS}(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \Big)^2 . \end{aligned}$$

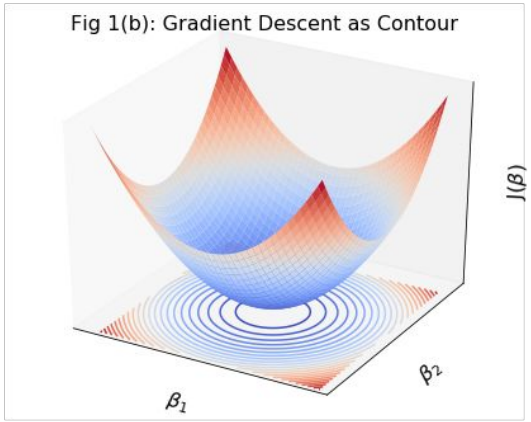
N—number of samples
p—number of independent variables or features
x—feature
y—actual target or dependent variable
f(x)—estimated target

β—coefficient or weight corresponding to each feature or independent var.

https://web.stanford.edu/~hastie/ElemStatLearn/

Регрессия





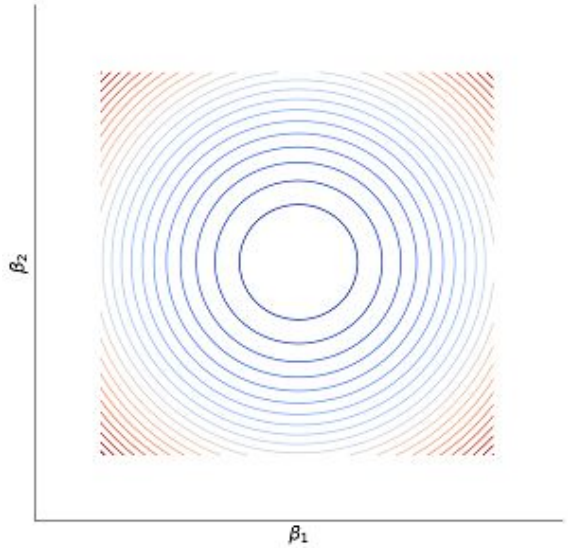


Fig 2: Gradient Descent on axes of β1 and β2

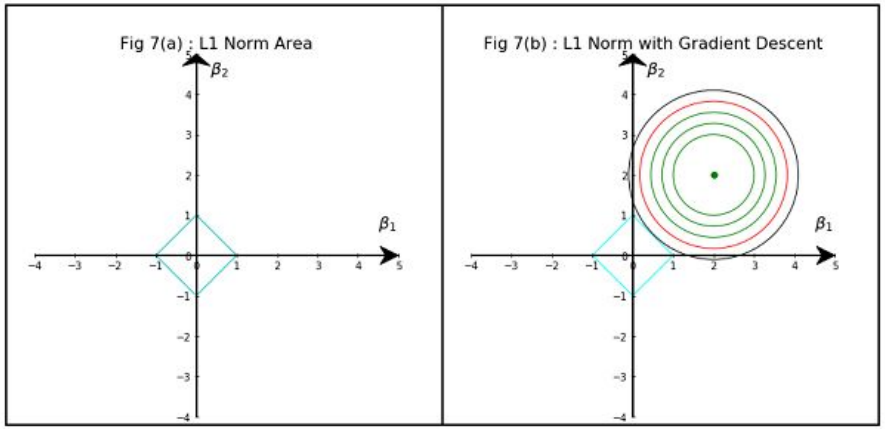
https://towardsdatascience.com/regularization-in-machine-learning-connecting-the-dots-c6e030bfaddd

L1 Norm or Lasso Regression

L1 Norm is of the form |β1| + |β2|.

Modified Cost function for L1 Regularization is as follows:

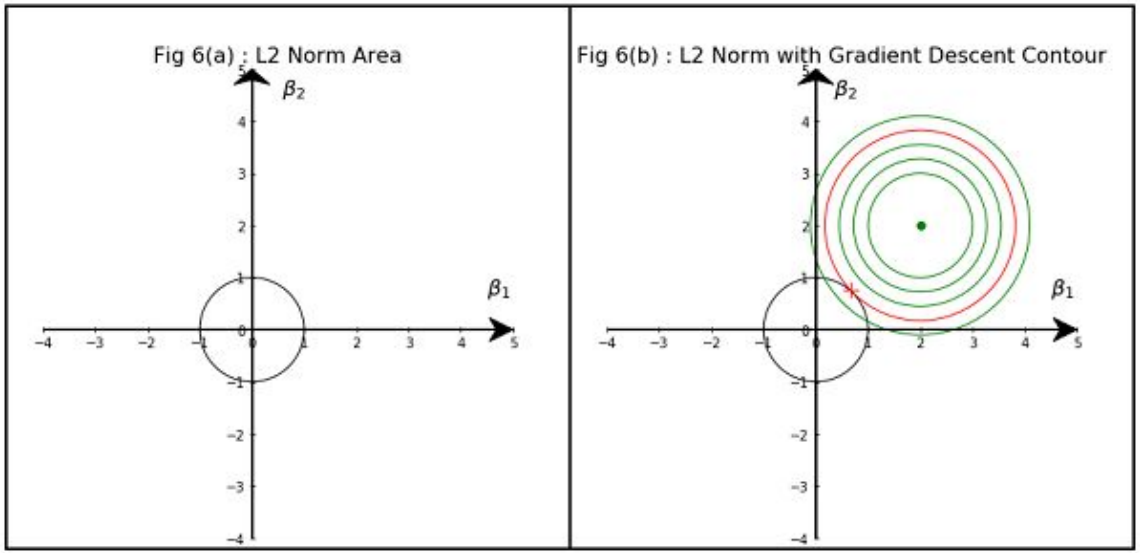
$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$



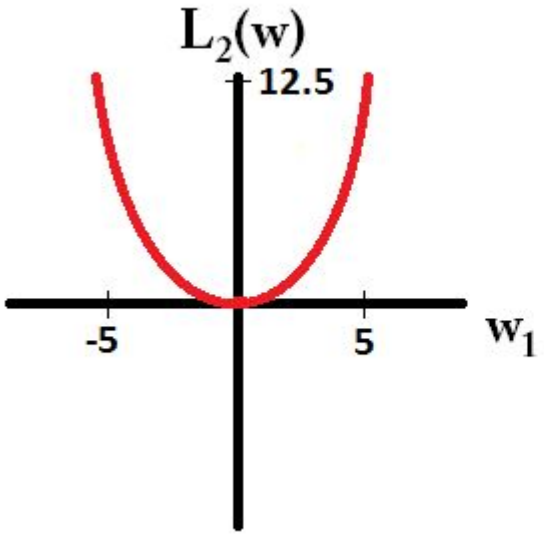
L2 Norm or Ridge Regression

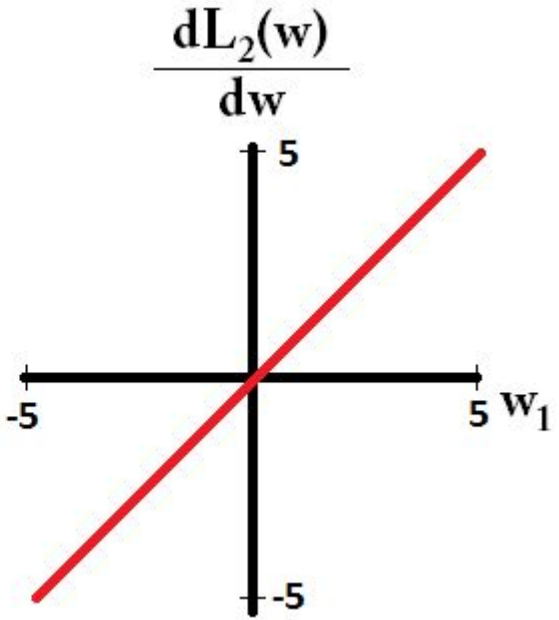
L2 Norm is Euclidean distance norm of the form $|\beta 1|^2 + |\beta 2|^2$.

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$



L2 подробнее



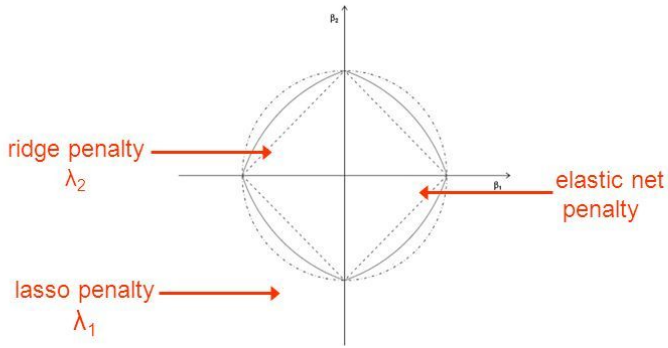


https://stats.stackexchange.com/questions/45643/why-l1-norm-for-sparse-models

- субградиенты, - изменение learning rate

Elastic Net (L1 + L2 Norm)

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1).$$



https://en.wikipedia.org/wiki/Elastic_net_regularization

Метод k Ближайших Соседей

Метод k ближайших соседей

Метод *k*-ближайших соседей (*k-nearest neighbors algorithm,* k-NN) — метрический алгоритм для автоматической классификации объектов или регрессии.

- В случае использования метода **для классификации** объект присваивается тому классу, который является наиболее распространённым среди соседей данного элемента, классы которых уже известны.
- В случае использования метода **для регрессии**, объекту присваивается среднее значение по ближайшим к нему объектам, значения которых уже известны

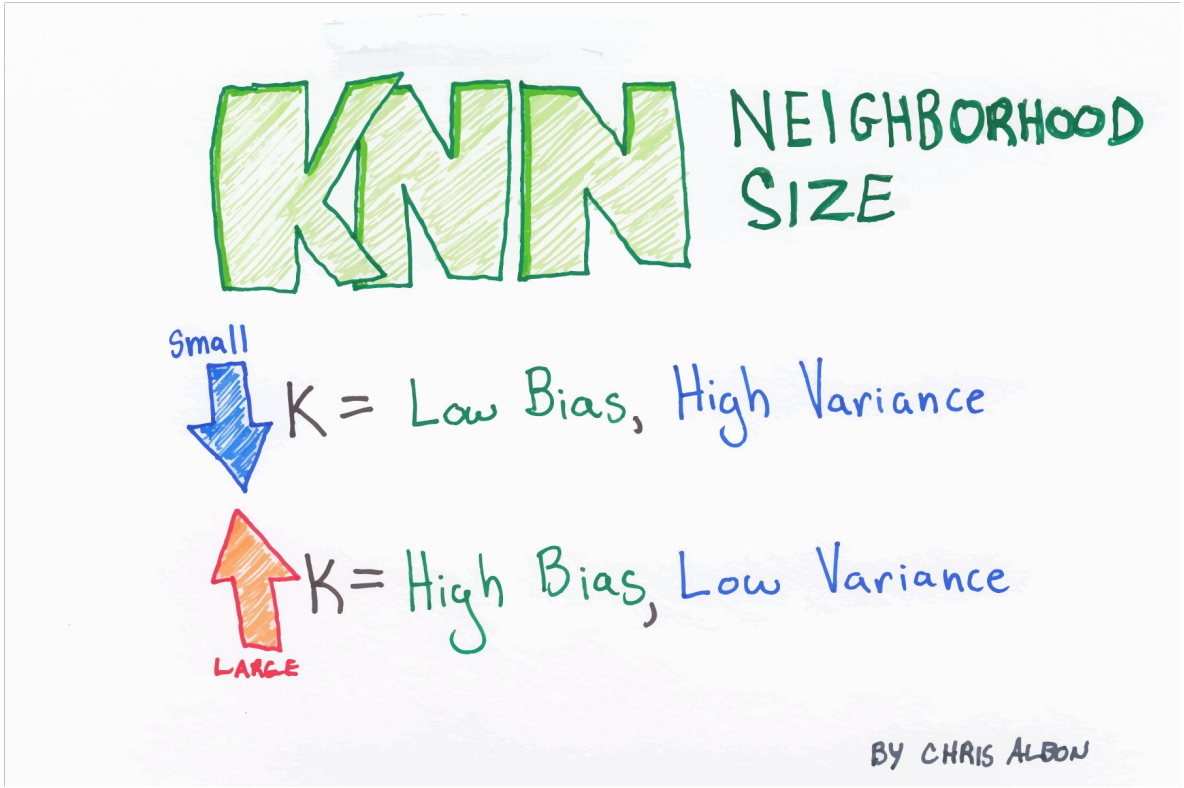
При таком способе во внимание принимается не только количество попавших в область определенных классов, но и их удаленность от нового значения. Для каждого класса *j* определяется оценка близости:

$$Q_j = \sum_{i=1}^{n} \frac{1}{d(x, a_i)^2}$$, где *d(x, a)* — дистанция от нового значения *x* до объекта *a*.

У какого класса выше значение близости, тот класс и присваивается новому объекту.

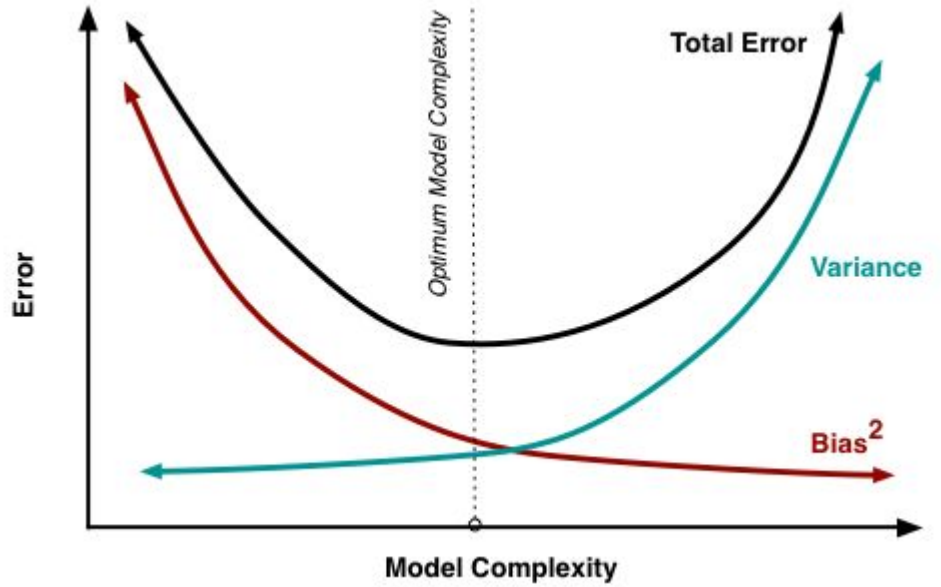
Лекция: https://ru.coursera.org/lecture/vvedenie-mashinnoe-obuchenie/mietod-blizhaishikh-sosiediei-jCkvu

Метод k ближайших соседей



https://chrisalbon.com/

Bias/variance tradeoff. Дилемма смещения/дисперсии

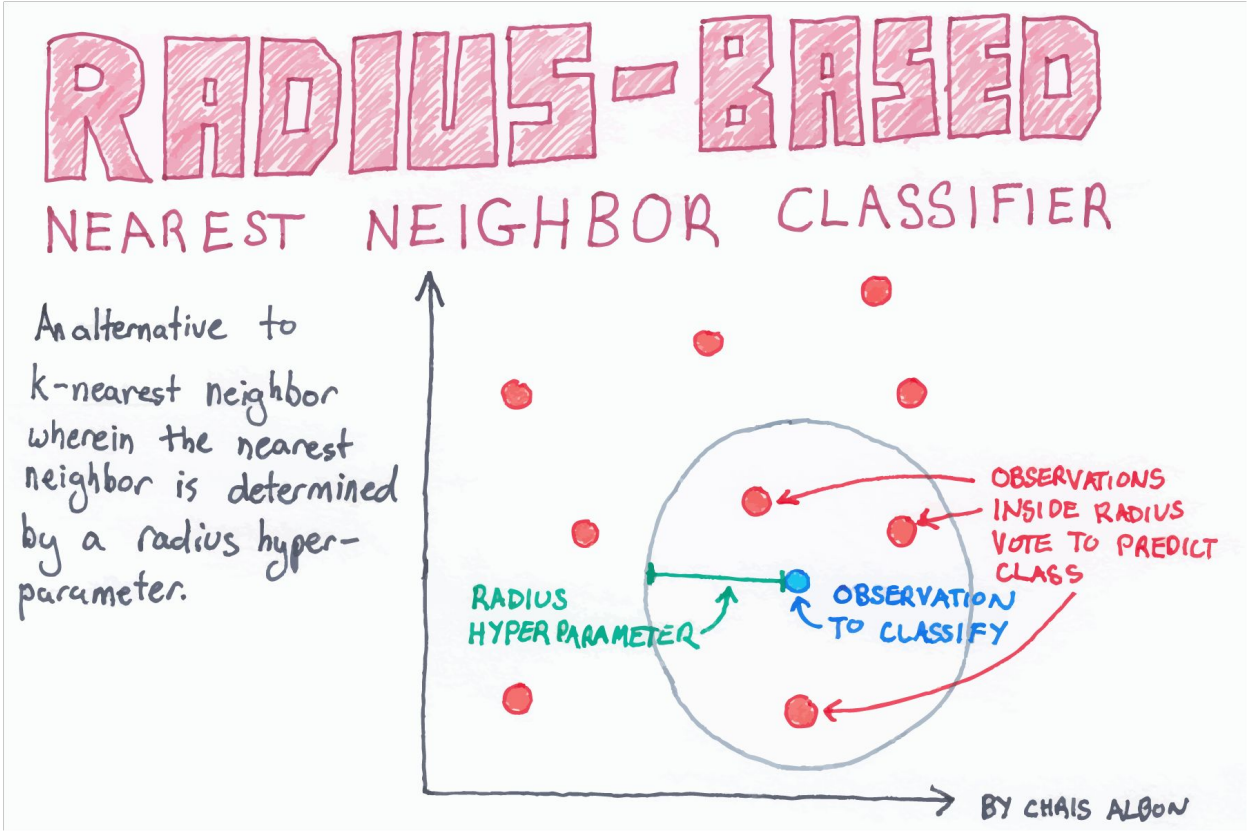


Простыми словами:

- если модель идеально описывает все данные (подфитилась), не факт, что она хорошо генерализуется на других данных
- если модель плохо описывает данные, то она не переобучилась, но, возможно, и не обучилась совсем :)

Смещение (bias)— это ошибка, возникающая в результате ошибочного предположения в алгоритме обучения. В результате большого смещения алгоритм может пропустить связь между признаками и выводом **(недообучение)**.

Дисперсия (variance)— это ошибка чувствительности к малым отклонениям в тренировочном наборе. При высокой дисперсии алгоритм может как-то трактовать случайный шум[en] в тренировочном наборе, а не желаемый результат **(переобучение)**.

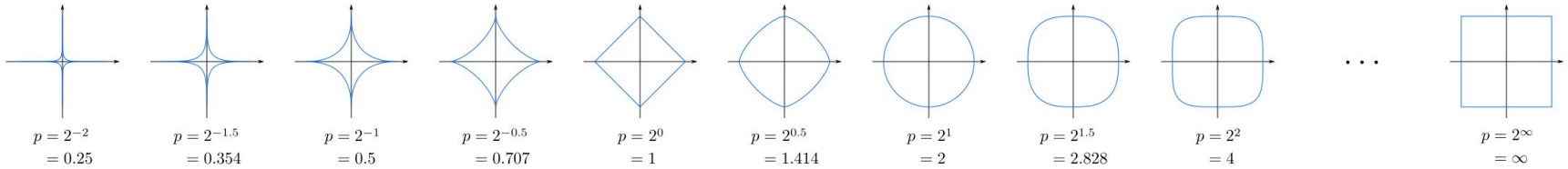


https://chrisalbon.com/

Метрики “близости”

Как расстояние между соседями измеряет sklearn:

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html



https://en.wikipedia.org/wiki/Minkowski_distance

Метрики близости

А если не числовые объекты:

- Редакторское расстояние Левенштейна
- BLEU score (для переводчиков текста)

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html


KNN BE LIKE
"Show me your friends, and I'll tell you who you are."

Вопросы для самопроверки:

- Почему L1-регуляризация производит отбор признаков?
- Почему может быть сделан выбор в сторону L2- регуляризации?
- Почему коэффициент регуляризации нельзя подбирать по обучающей выборке?

Источники:

1. https://towardsdatascience.com/regularization-in-machine-learning-connecting-the-dots-c6e030bfaddd
2. https://github.com/esokolov/ml-course-hse/
3. https://chrisalbon.com/
4. https://github.com/SlinkoIgor/express_ml
5. https://docplayer.ru/41305484-Lekciya-2-obobshchennye-lineynye-modeli-regulyarizaciya-obucheniya.html
6. https://www.youtube.com/watch?v=Kloz_aa1ed4
7. https://github.com/esokolov/ml-course-hse/blob/master/2018-fall/lecture-notes/lecture03-linregr.pdf