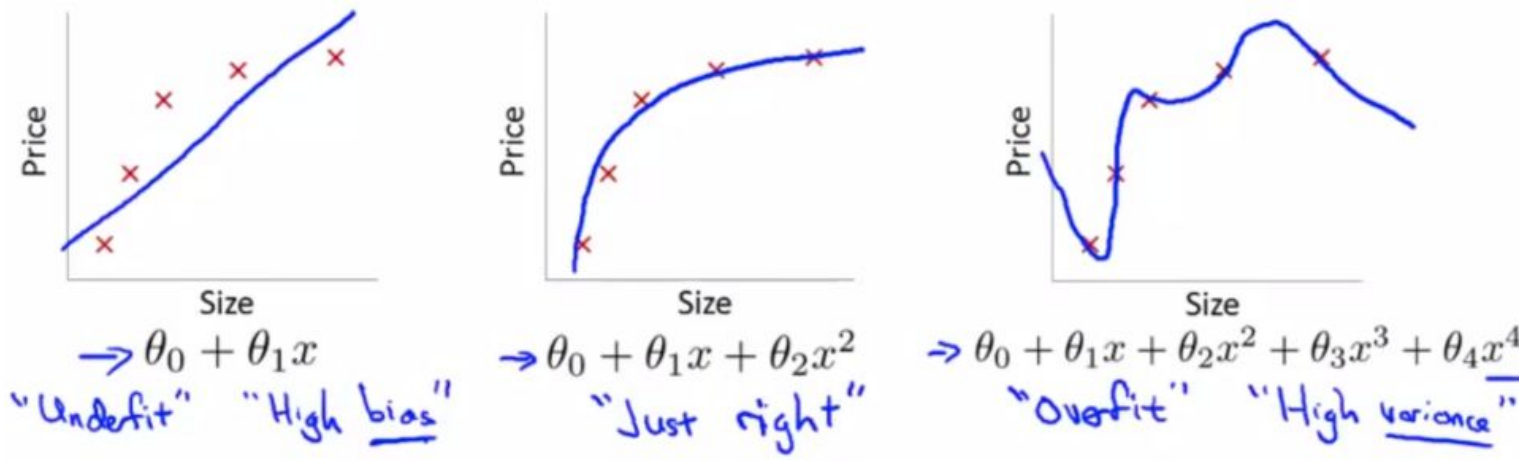
Обучение с учителем: Регуляризация в линейных моделях. Метод Ближайших Соседей (KNN)

Екатерина Кондратьева


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

Переобучение (model overfitting)



Здесь theta (θ) - β

Регуляризация

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

Регрессия:

МНК функция потерь:

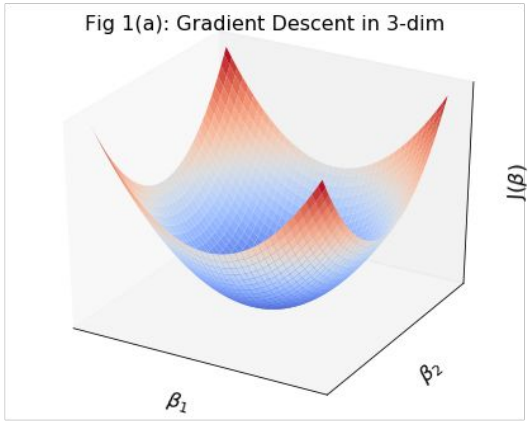
$$\begin{aligned} \mathrm{RSS}(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \Big)^2 . \end{aligned}$$

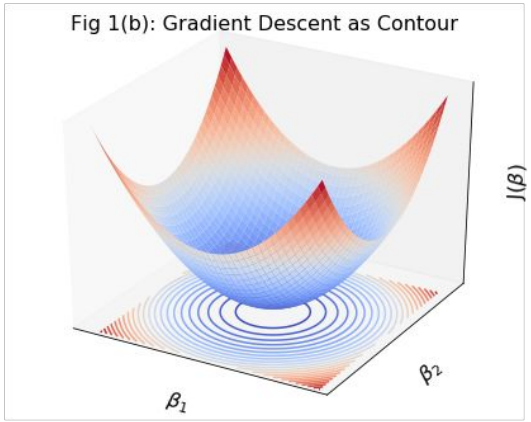
N—number of samples
p—number of independent variables or features
x—feature
y—actual target or dependent variable
f(x)—estimated target

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

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

Регрессия





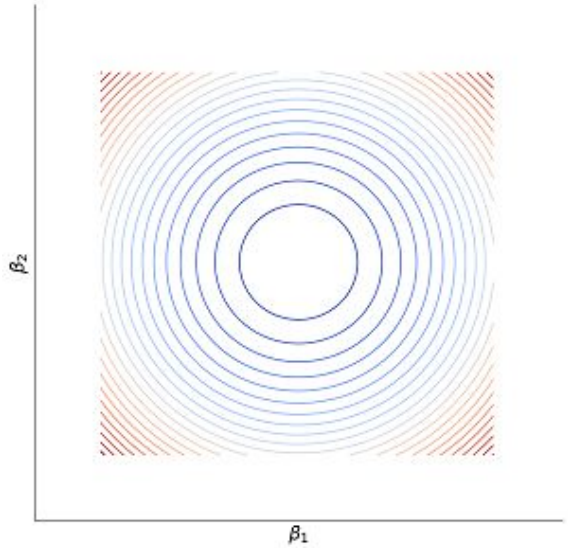


Fig 2: Gradient Descent on axes of β1 and β2

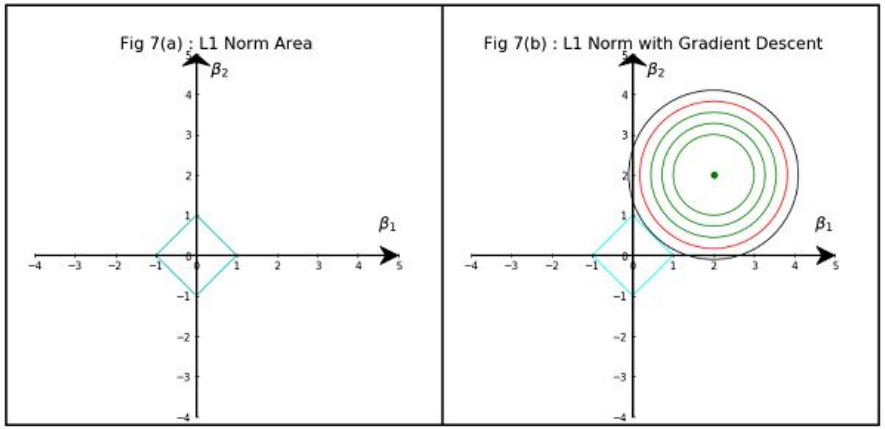
https://towardsdatascience.com/regularization-in-machine-learning-connecting-the-dots-c6e030bfaddd

L1 Norm or Lasso Regression

L1 Norm is of the form |β1| + |β2|.

Modified Cost function for L1 Regularization is as follows:

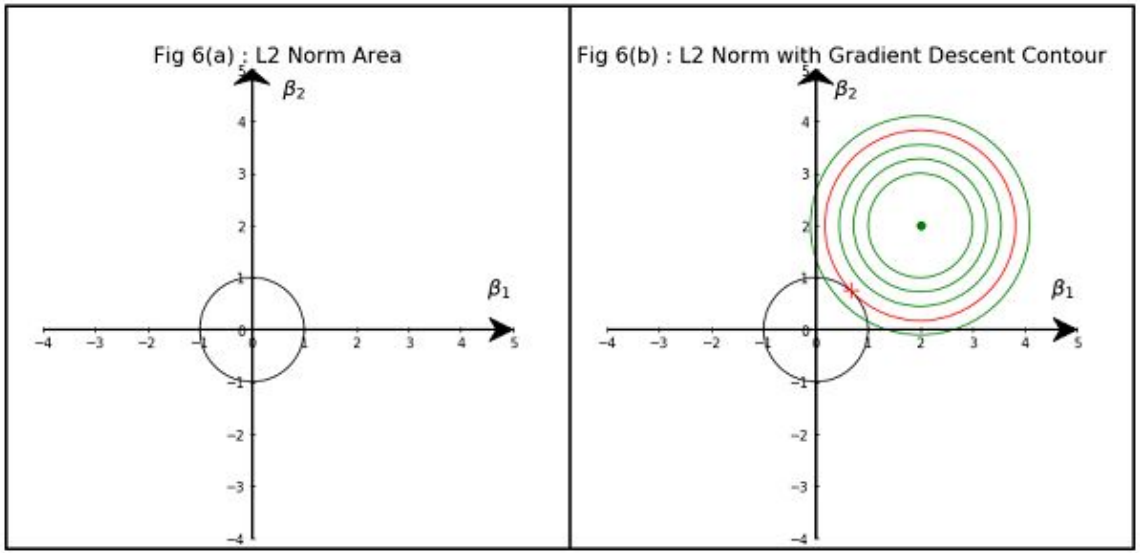
$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$



L2 Norm or Ridge Regression

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

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



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

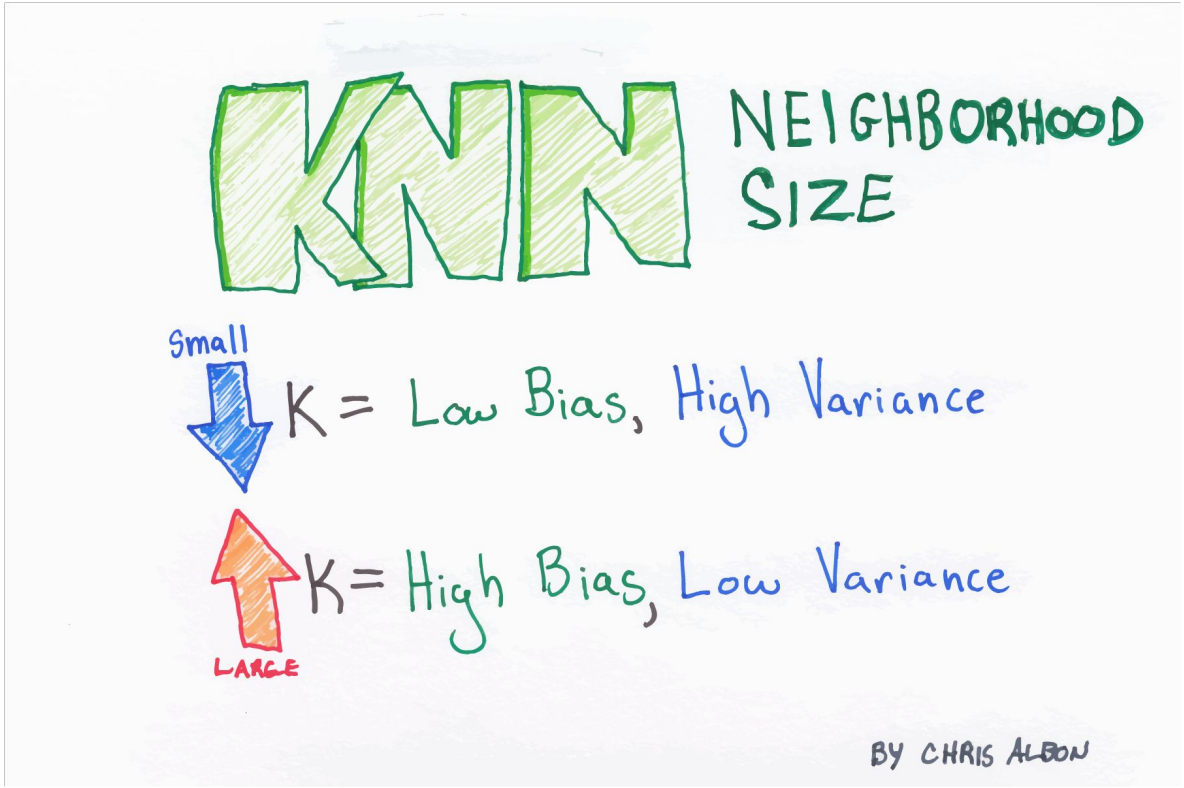
При таком способе во внимание принимается не только количество попавших в область определенных классов, но и их удаленность от нового значения. Для каждого класса *j* определяется оценка близости:

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

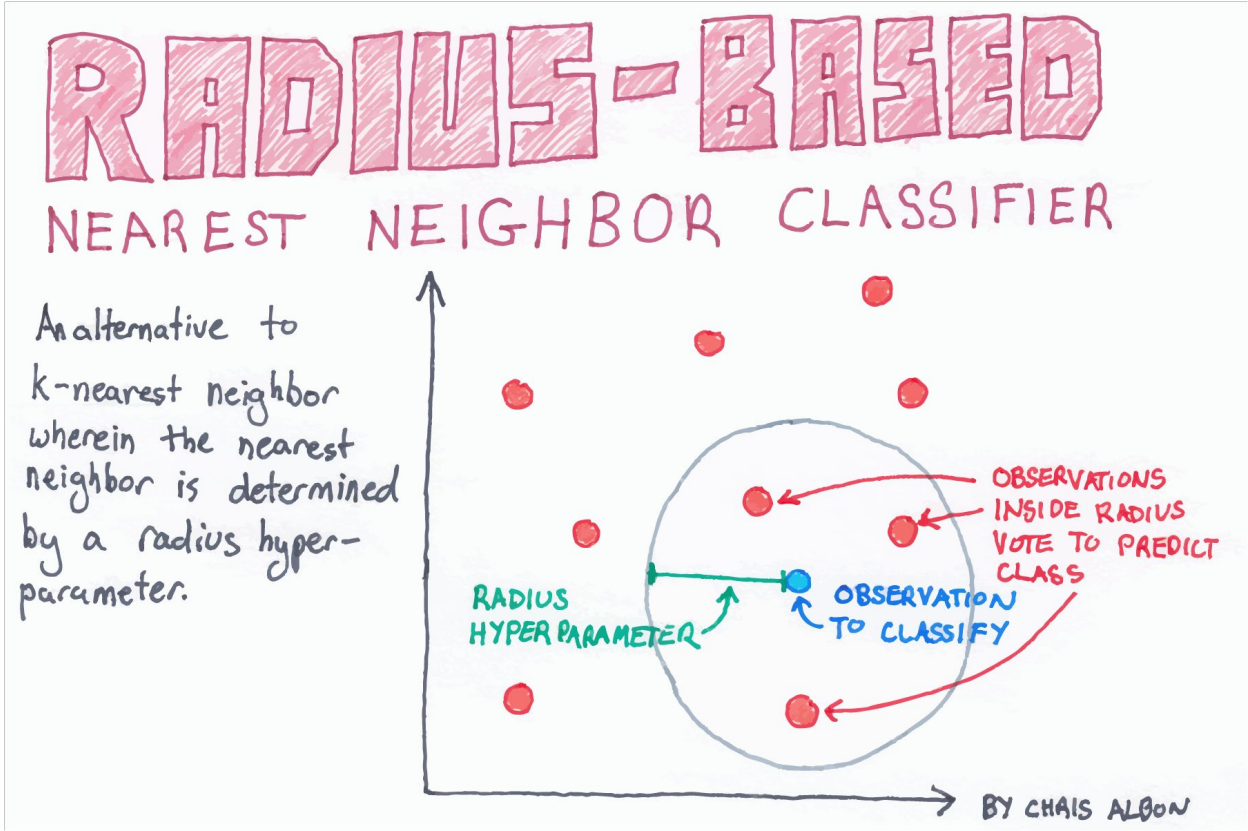
У какого класса выше значение близости, тот класс и присваивается новому объекту.

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

Метод k ближайших соседей



https://chrisalbon.com/



Часто метрики дистанции используются для снижения размерности:
https://www.stat.berkeley.edu/~bickel/mldim.pdf


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 коэффициент регуляризации нельзя подбирать по обучающей выборке?
- Что такое кросс-валидация, чем она лучше использования отложенной выборки?

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=KIoz_aa1ed4
7. https://github.com/esokolov/ml-course-hse/blob/master/2018-fall/lecture-notes/lecture03-linregr.pdf