

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

Екатерина Кондратьева

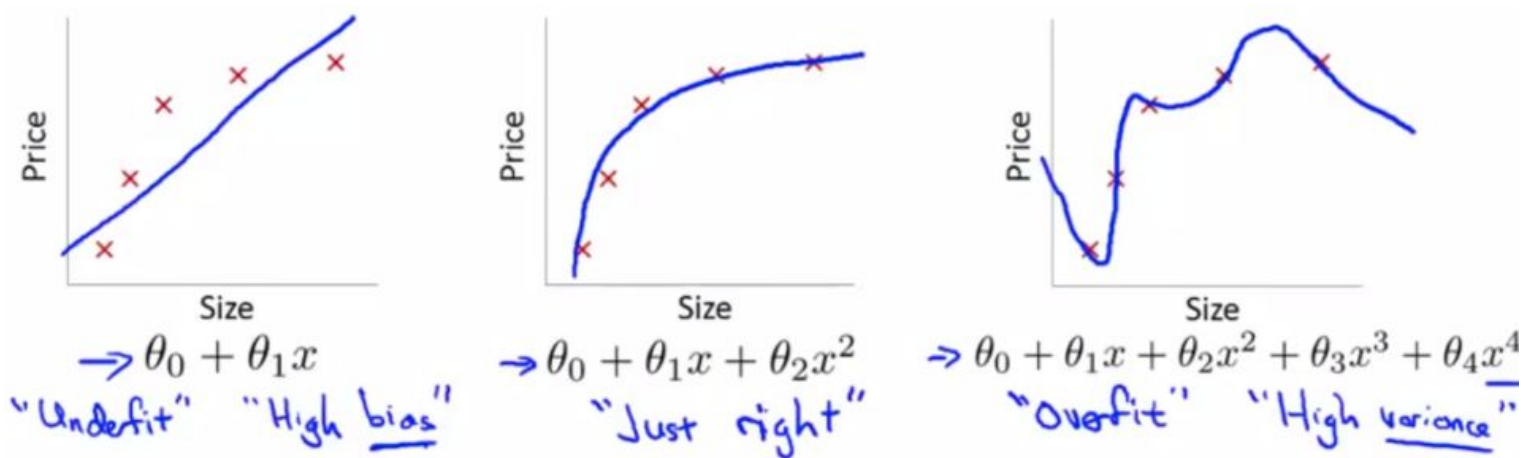
NOT SURE IF GOOD MODEL...



...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:

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

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.

Регрессия

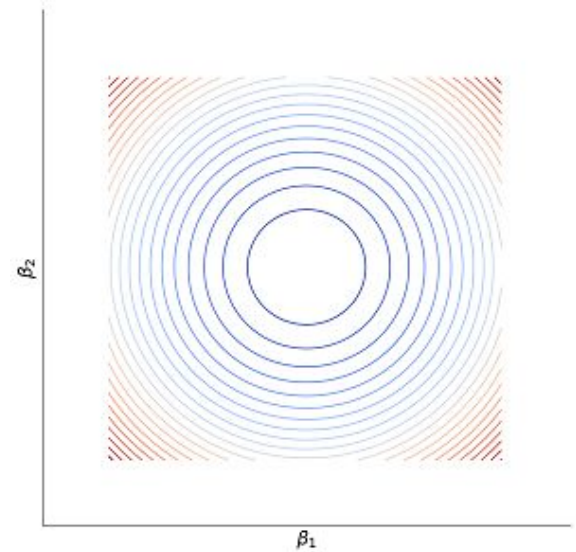
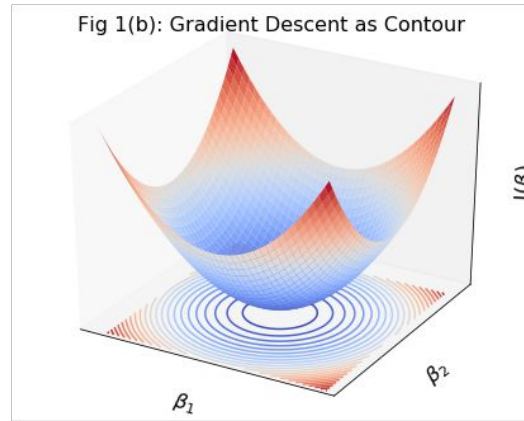
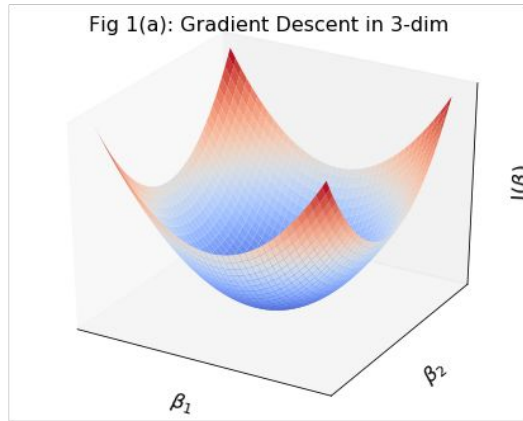


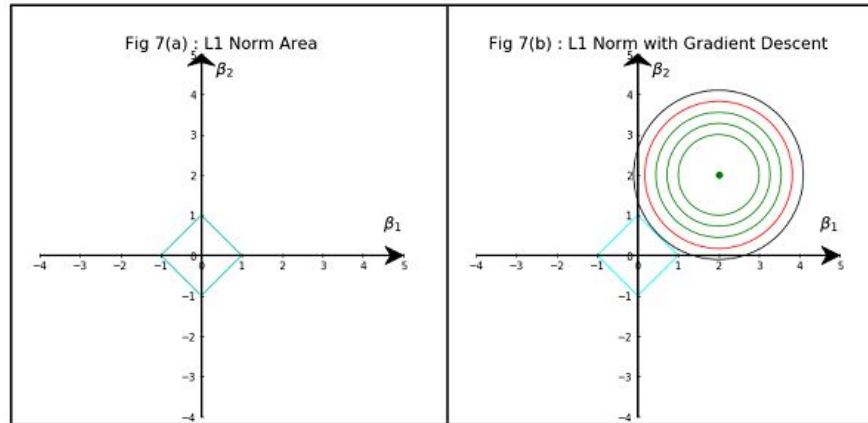
Fig 2: Gradient Descent on axes of β_1 and β_2

L1 Norm or Lasso Regression

L1 Norm is of the form $|\beta_1| + |\beta_2|$.

Modified Cost function for L1 Regularization is as follows:

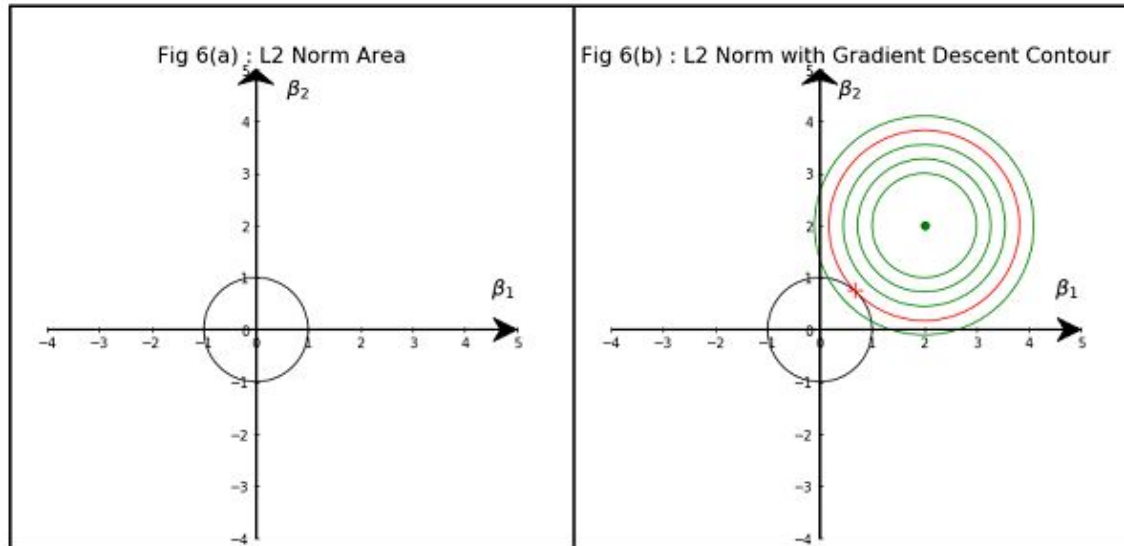
$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$



L2 Norm or Ridge Regression

L2 Norm is Euclidean distance norm of the form $|\beta_1|^2 + |\beta_2|^2$.

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



Метод к 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:

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

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

KNN NEIGHBORHOOD
SIZE

Small
↓
 $K = \text{Low Bias, High Variance}$

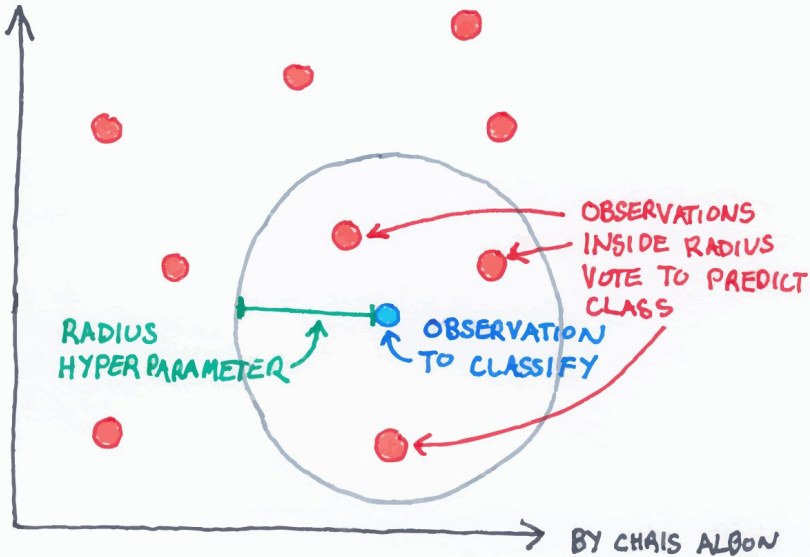
↑
LARGE
 $K = \text{High Bias, Low Variance}$

BY CHRIS ALBON

RADIUS-BASED

NEAREST NEIGHBOR CLASSIFIER

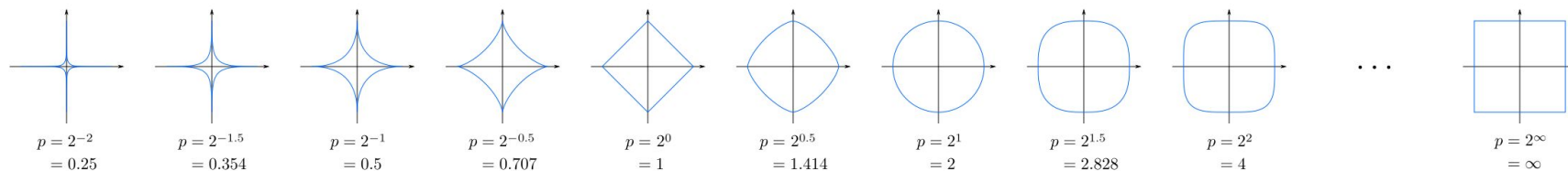
An alternative to
k-nearest neighbor
wherein the nearest
neighbor is determined
by a radius hyper-
parameter.



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-c6e030bfadd>
2. <https://github.com/esokolov/ml-course-hse/>
3. <https://chrisalbon.com/>
4. https://github.com/Slinkolgor/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>