Регуляризация. SVM. Ансамбли. Визуализация.

Маша Шеянова, masha.shejanova@gmail.co

План

- ЧаВО (часто возникающие ошибки)
- регуляризация (l1, l2)
- SVM
- ансамбли
- визуализация

Типичные ошибки

no attribute *lower* / *lower* not found

Когда: при вызове fit_transform / transform в векторизаторе

Почему: что-то внутри функции ожидало строку, а получило не её

Давайте разберёмся. CountVectorizer и TfidfVectorizer принимают на вход iterable (массив, список, столбец pandas и т.д.).

Если сделать **train_test_split(df.text)**, в x_train и x_test лежит то, что им надо.

CountVectorizer и TfidfVectorizer выдают разреженную матрицу bag of words. Если сделать **train_test_split(baf_of_words)**, в x_train и x_test лежит то, что надо потом скормить классификатору, но никак не векторизатору.

warning: precision / recall is ill-defined

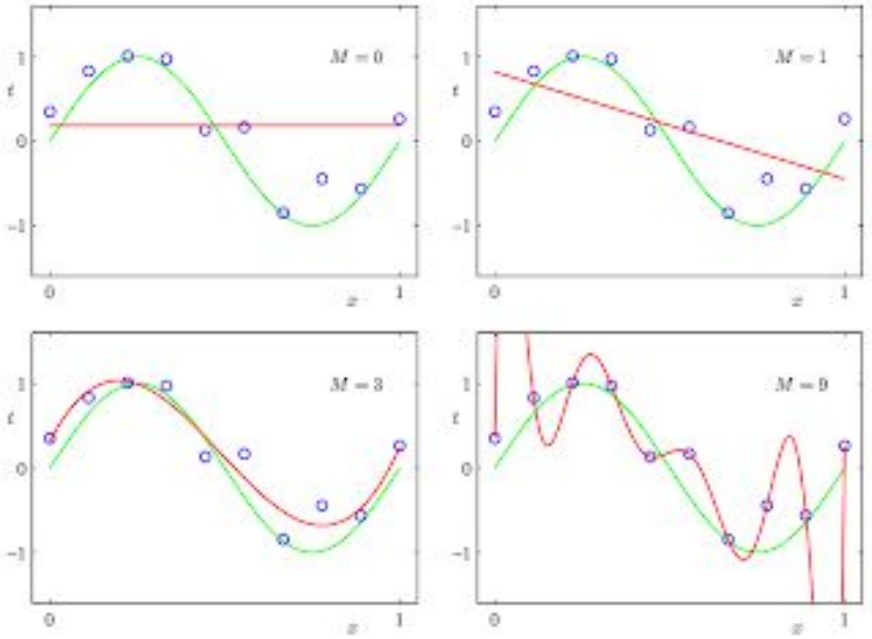
Когда при подсчёте точности или полноты снизу дроби получается 0, результат не определён. sklearn приравнивает его к нулю, чтобы хоть что-то показать. Как это может получиться?

- классы сильно несбалансированы, в test оказалось 0 примеров класса
- ошибка в коде — что-то перепутали при подсчёте + плохая модель

Эта ошибка была у меня в домашке. *classification_report* принимает ***первым*** аргументом вектор (массив, список) *true values*, а ***потом*** *predicted values*.

Регуляризация

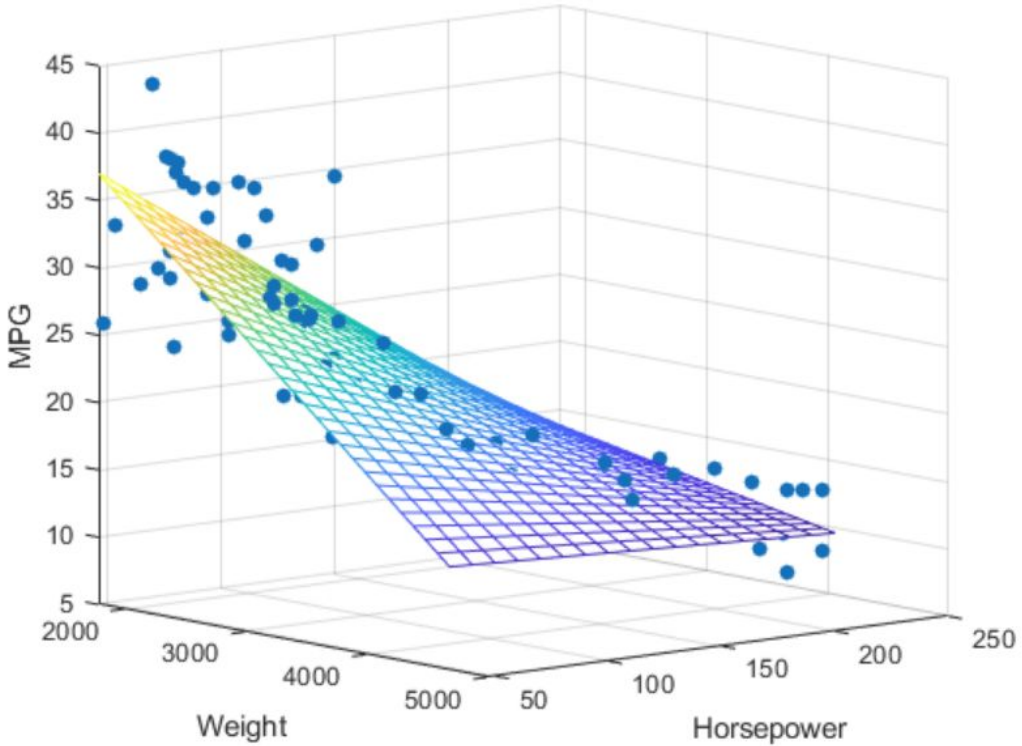
Недообучение и переобучение



Две верхние картинки — недообучение, нижняя слева — оптимальная аппроксимация, нижняя справа — переобучение.

(Источник: «Pattern Recognition and Machine Learning» (Figure 1.4) («Распознавание образов и машинное обучение» (рис. 1.4)).)

Неустранимая ошибка



Расход топлива зависит от двух признаков: вес машины, лошадиные силы.

Почему точки не лежат на одной прямой?

- неучтённые признаки
- шум в данных

Шум и неучтённые признаки — причина **неустранимой ошибки**.

Смещение и дисперсия (bias and variance)

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

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

Вообще, дисперсия в статистике — это про то, насколько случайная величина отклоняется от мат. ожидания.

$$D[X] = M\left[\left(X - M[X]\right)^2\right]$$

Разложение bias и variance

$$Err(x) = \left(E[\hat{f}\,(x)] - f(x)\right)^2 + E\left[\left(\hat{f}\,(x) - E[\hat{f}\,(x)]\right)^2\right] + \sigma_e^2$$

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

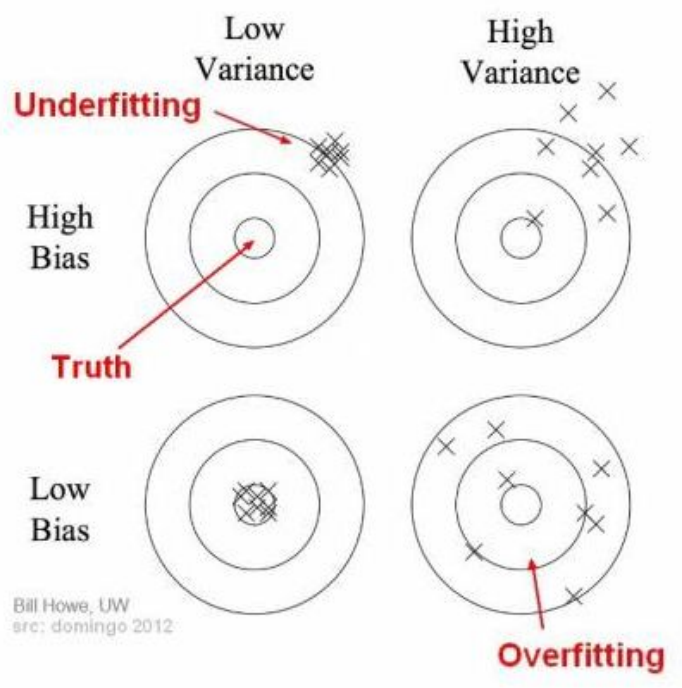
- квадрат **смещения** метода обучения можно рассматривать как ошибку, вызванную упрощением предположений, принятых в методе
- **дисперсия** метода обучения, или, интуитивно, как далеко метод обучения f^(x) уведёт от среднего значения
- **неустранимая ошибка** — то, что данные не могут нам дать

Bias-variance tradeoff

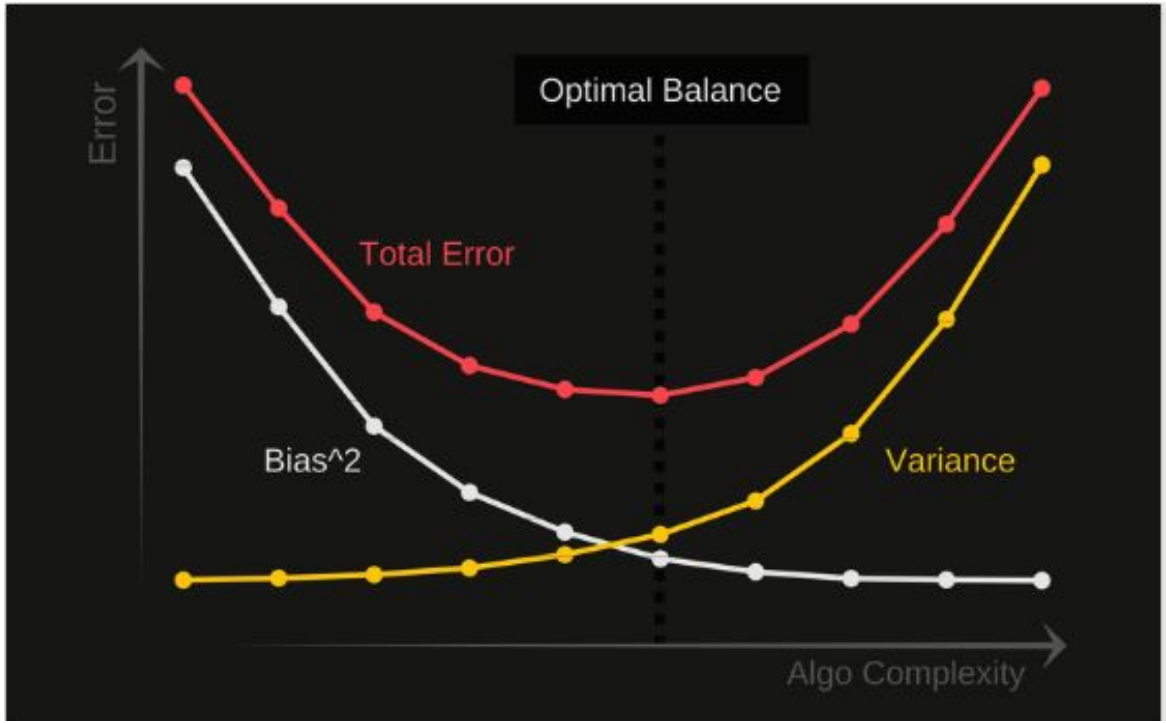
… или дилемма смещения–дисперсии.

Переусложнённые модели (что-то вроде $a * x^{39} + b * x^{35} + c * x^{31} + ...$) имеют высокую variance (их сильно кидает по точкам обучающих данных), и они будут склонны к переобучению. Простые модели (вроде $a * x + b$) имеют большой bias, и риск недообучения.

Источник картинки.



Bias-variance tradeoff



Источник картинки.

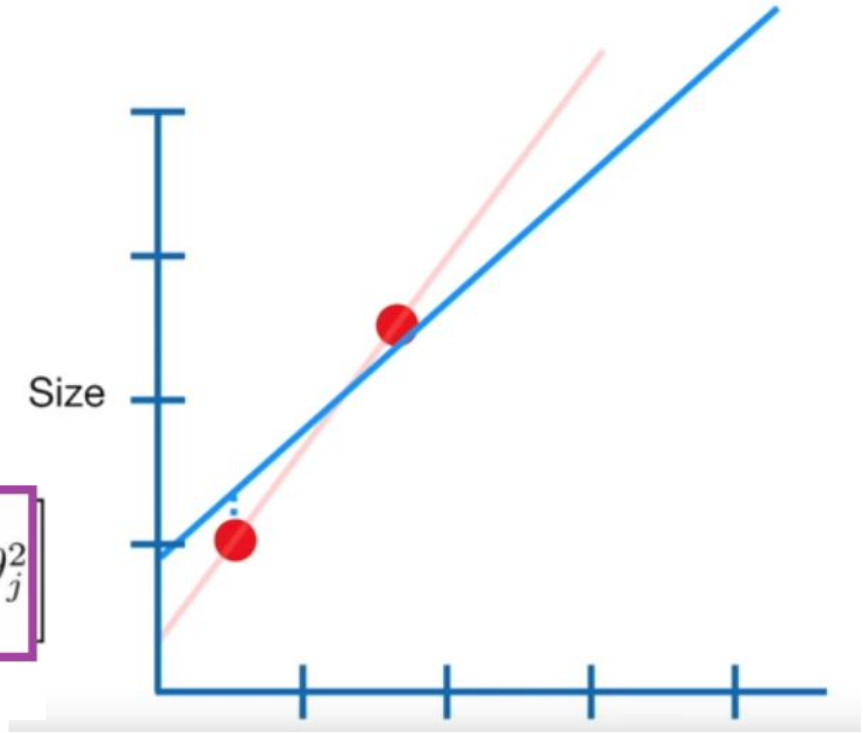
Наша цель — выбрать оптимальный баланс в сложности модели.

Регуляризация

Идея: вместо того, чтобы минимизировать просто функцию потерь J(X), будем минимизировать J(X) + λ * (какой-то из видов вид суммы коэффициентов).

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n}\theta_j^2\right]$$

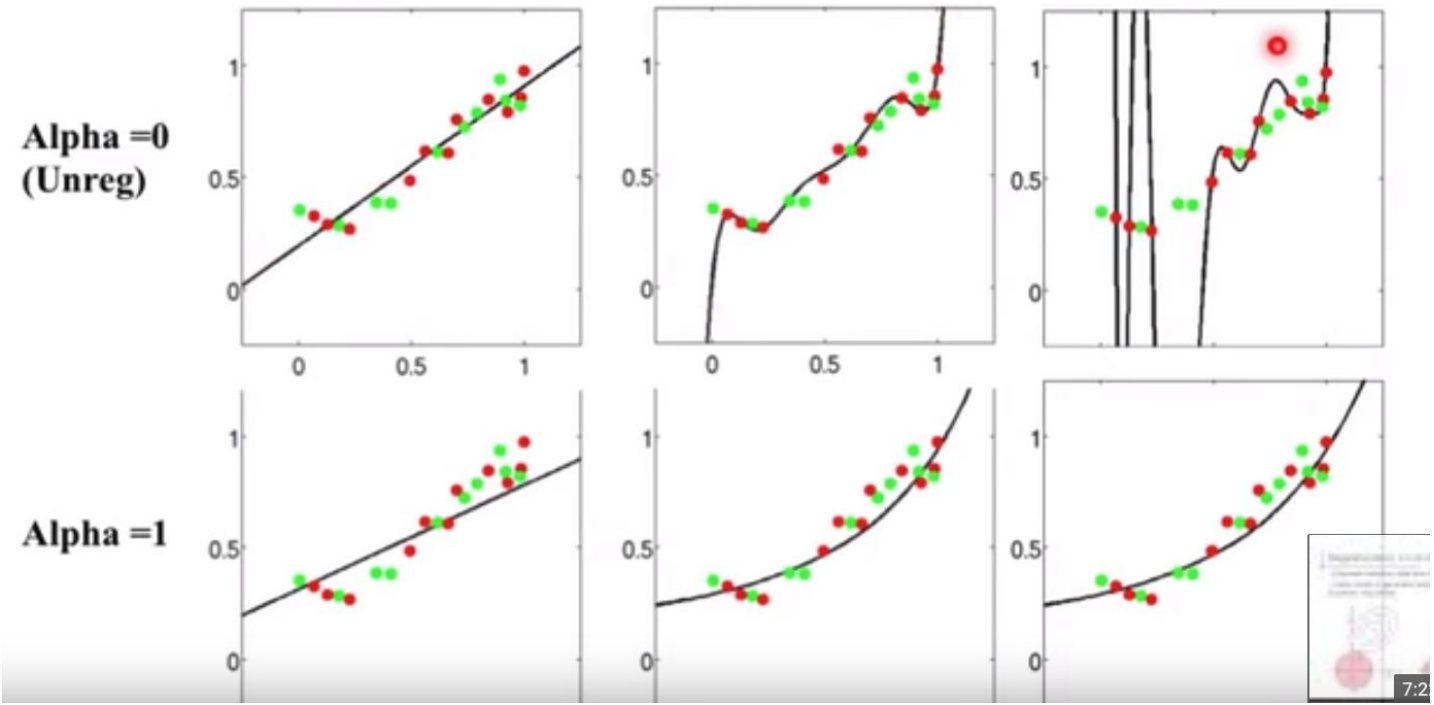
$$\min_\theta J(\theta)$$



альфа / лямбда

Чем больше эта величина, тем больше значения мы придаем штрафу за величину коэффициентов.

Источник.



L1 (Lasso)

L1 regularization (lasso regression) adds an L1 penalty equal to the absolute value of the magnitude of coefficients.

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Cost function

Просто складываем значения коэффициентов и умножаем на lambda.

L2 (Ridge Regression)

L2 regularization (Ridge regression) adds an L2 penalty equal to the square of the magnitude of coefficients. L2 will not yield sparse models and all coefficients are shrunk by the same factor (none are eliminated).

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Cost function

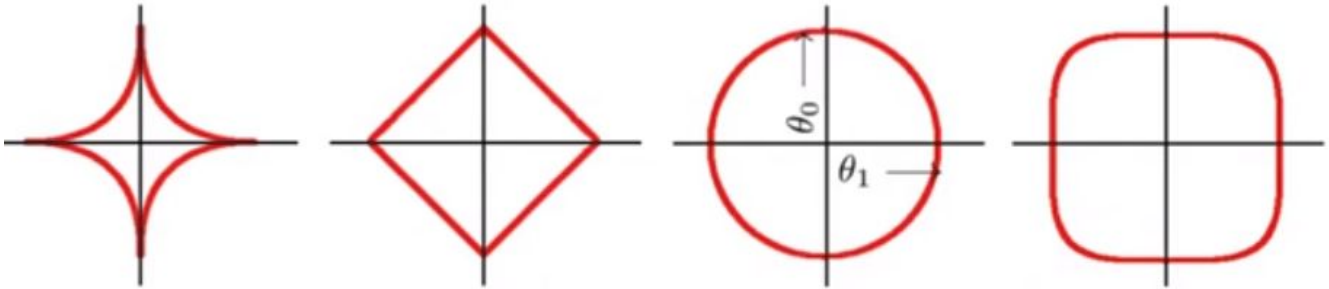
Возводим каждый коэффициент в квадрат и умножаем на lambda.

Разная регуляризация

Different regularization functions

- More generally, for the L_p regularizer: $\left(\sum_i |\theta_i|^p\right)^{\frac{1}{p}}$

Isosurfaces: $\|\theta\|_p = \text{constant}$



p=0.5

p=1

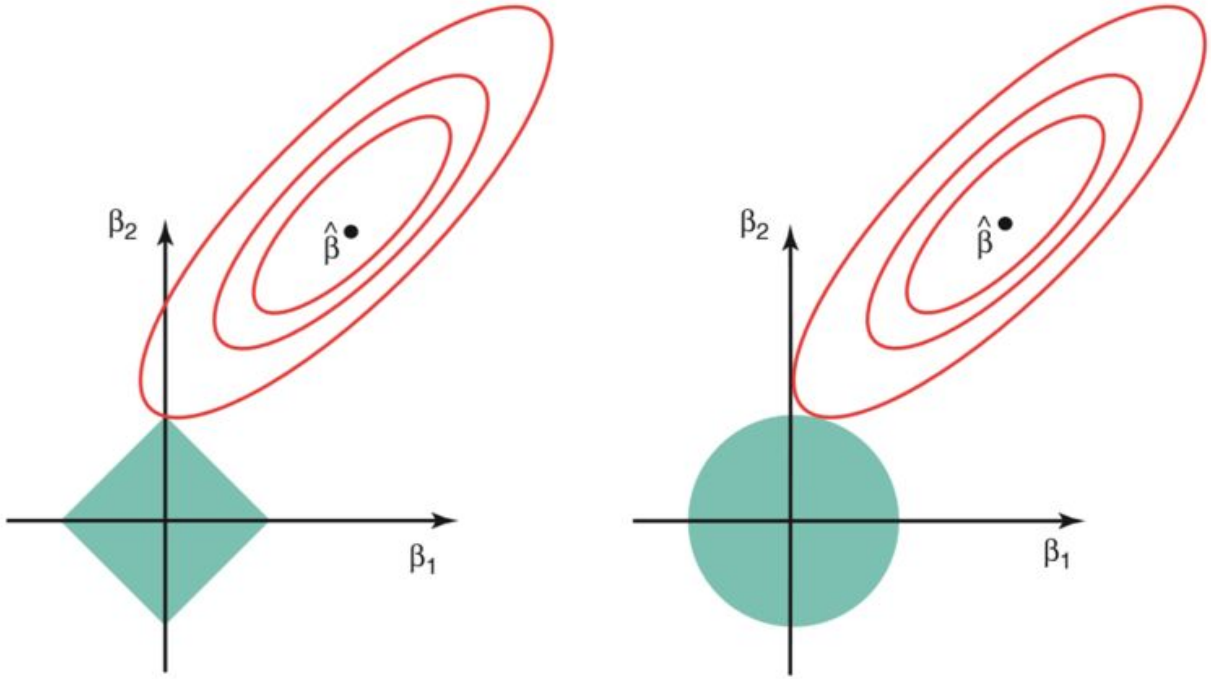
Lasso

p=2

Quadratic

p=4

L1 vs. L2



Комбинация двух фунцкий потерь.

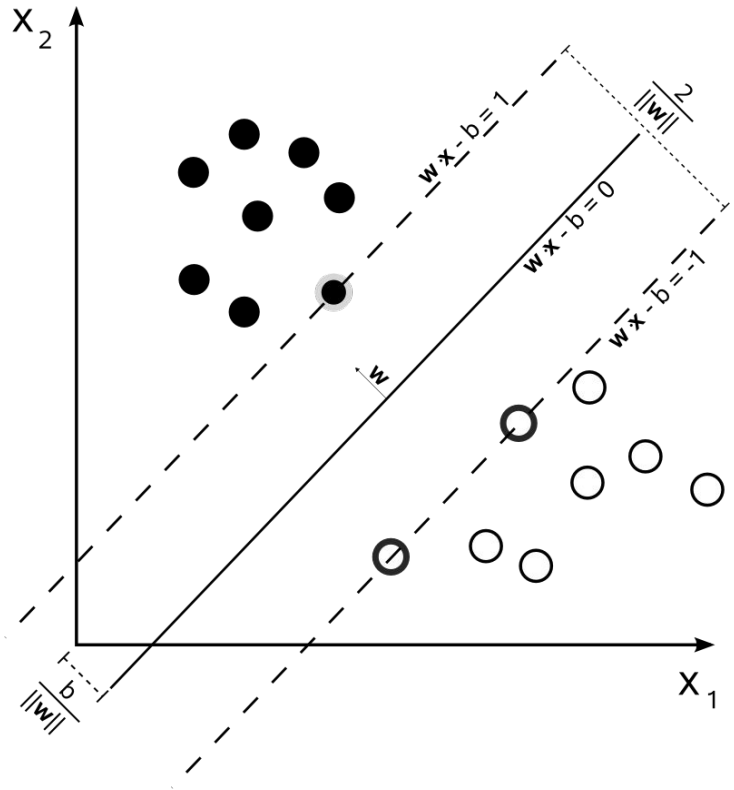
L1 будет занулять некоторые параметры, L2 не будет.

SVM

Идея

Есть данные, относящиеся к двум классам. Мы хотим построить разделяющую плоскость с наибольшим зазором между двумя классами.

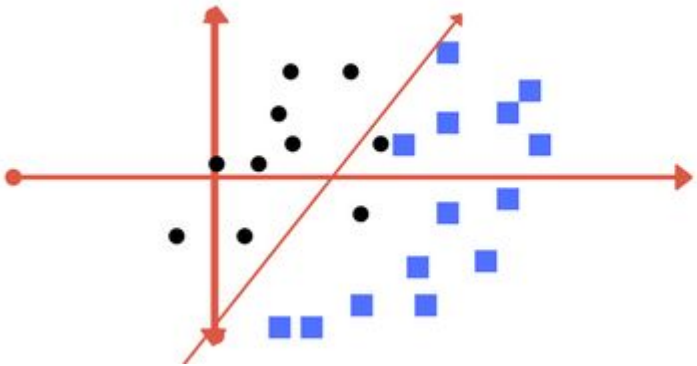
Подробнее SVM описан здесь.



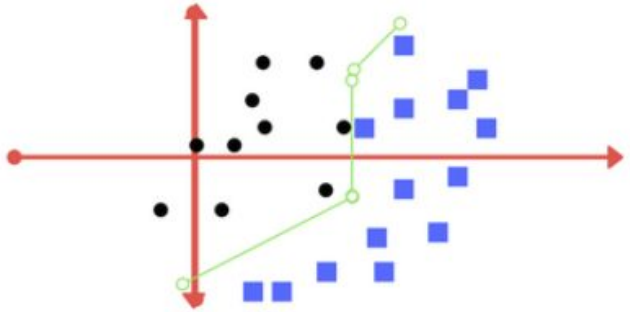
C parameter

For **large** values of C, the optimization will choose a **smaller-margin hyperplane**. Conversely, a very **small** value of C will cause the optimizer to look for a **larger-margin separating hyperplane**, even if that hyperplane misclassifies more points.

Маленький C:



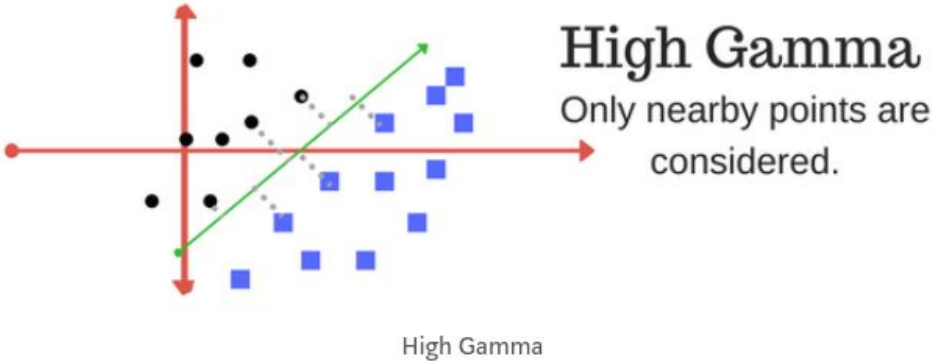
Большой C:



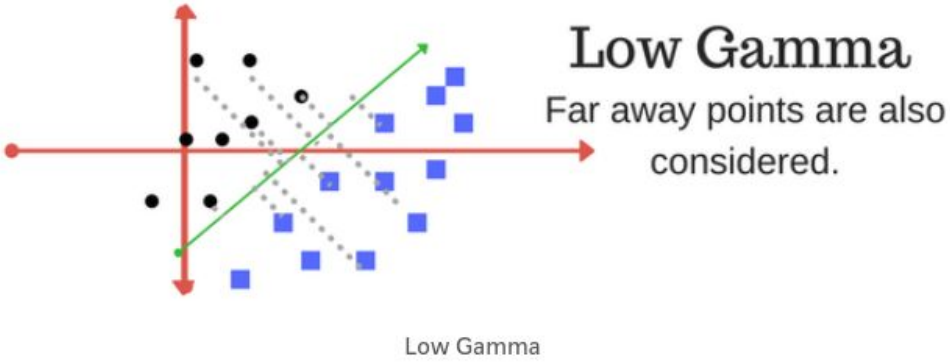
Gamma

With **low** gamma, points **far away** from plausible separation line **are considered** in calculation for the separation line.

High gamma means the points **close** to plausible line are considered in calculation.



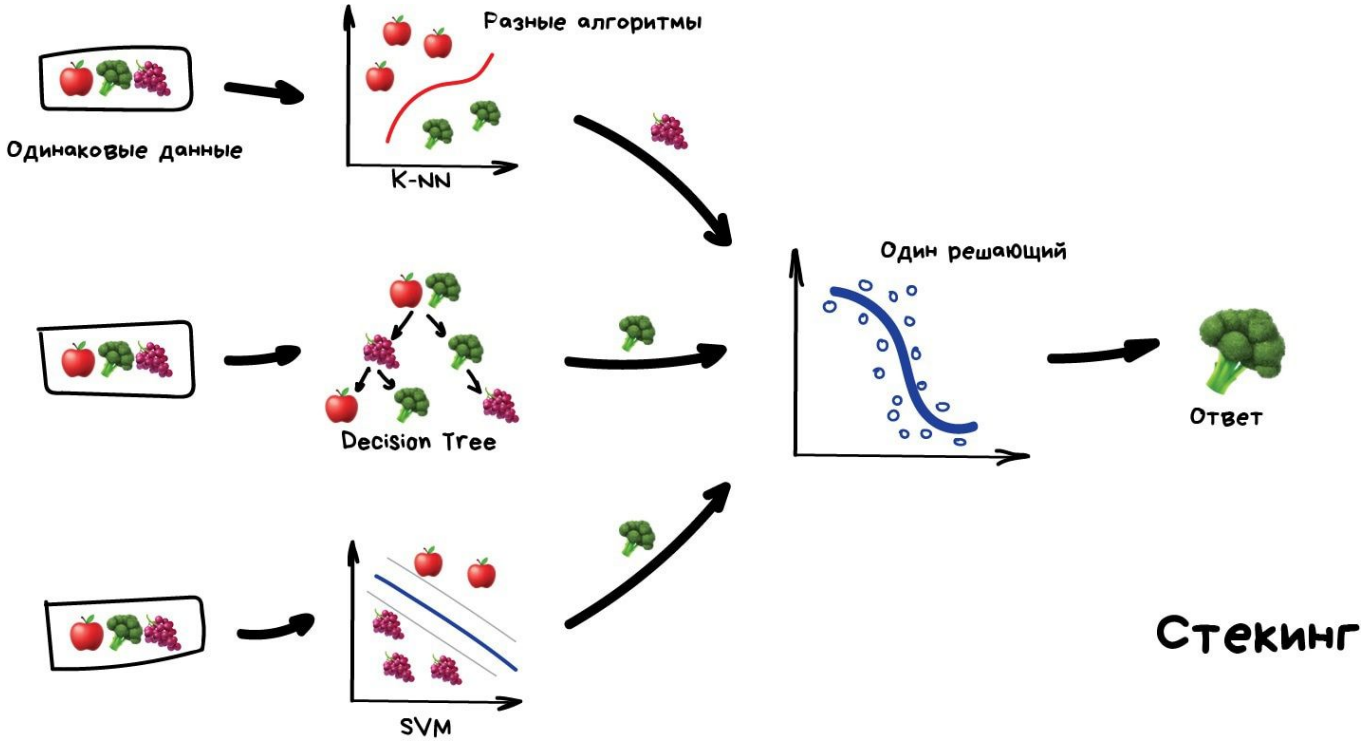
High Gamma



Low Gamma

Ансамбли
(Все картинки отсюда)

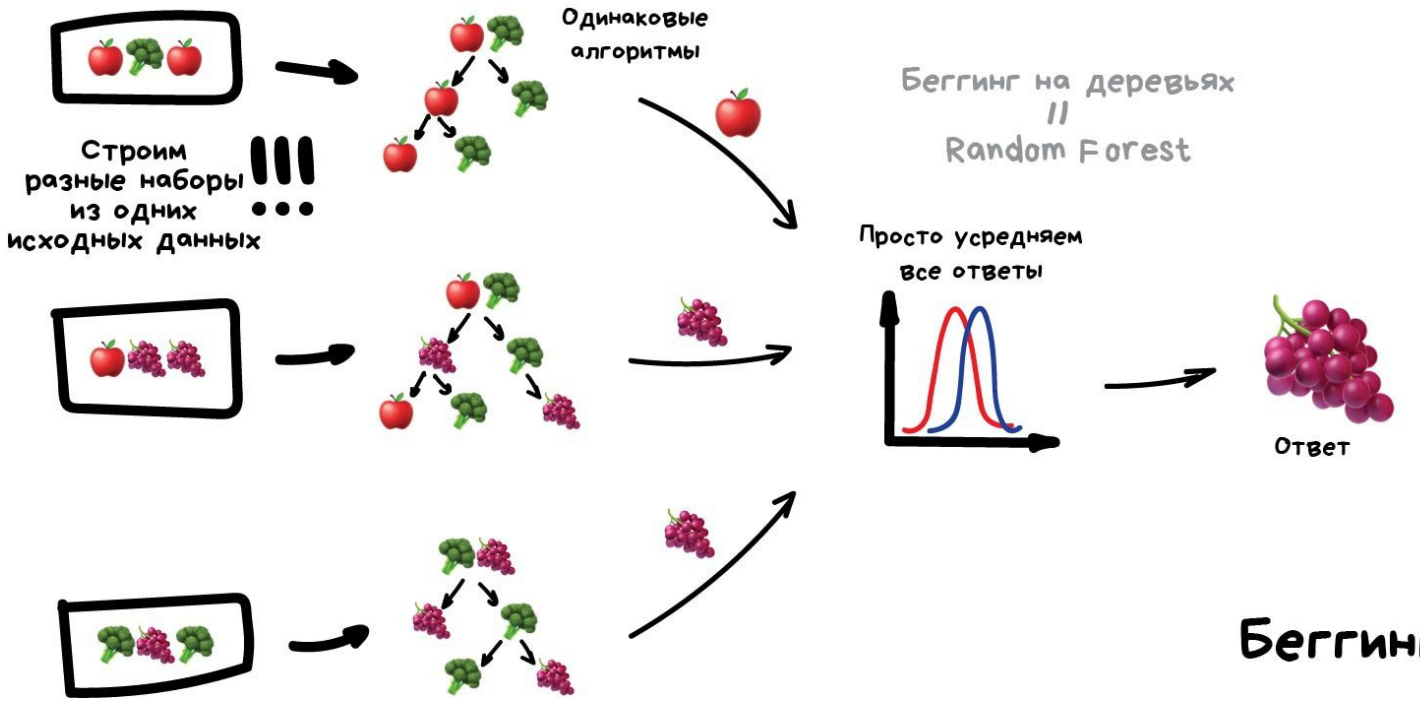
Stacking



Обучаем n разных алгоритмов и передаём их результаты на вход последнему. Ключевое слово — разных.

На практике применяется редко.

Bagging



Обучаем один алгоритм много раз на случайных выборках из данных. Потом усредняем ответы.

Данные в случайных выборках могут повторяться.

Random Forest (случайный лес)

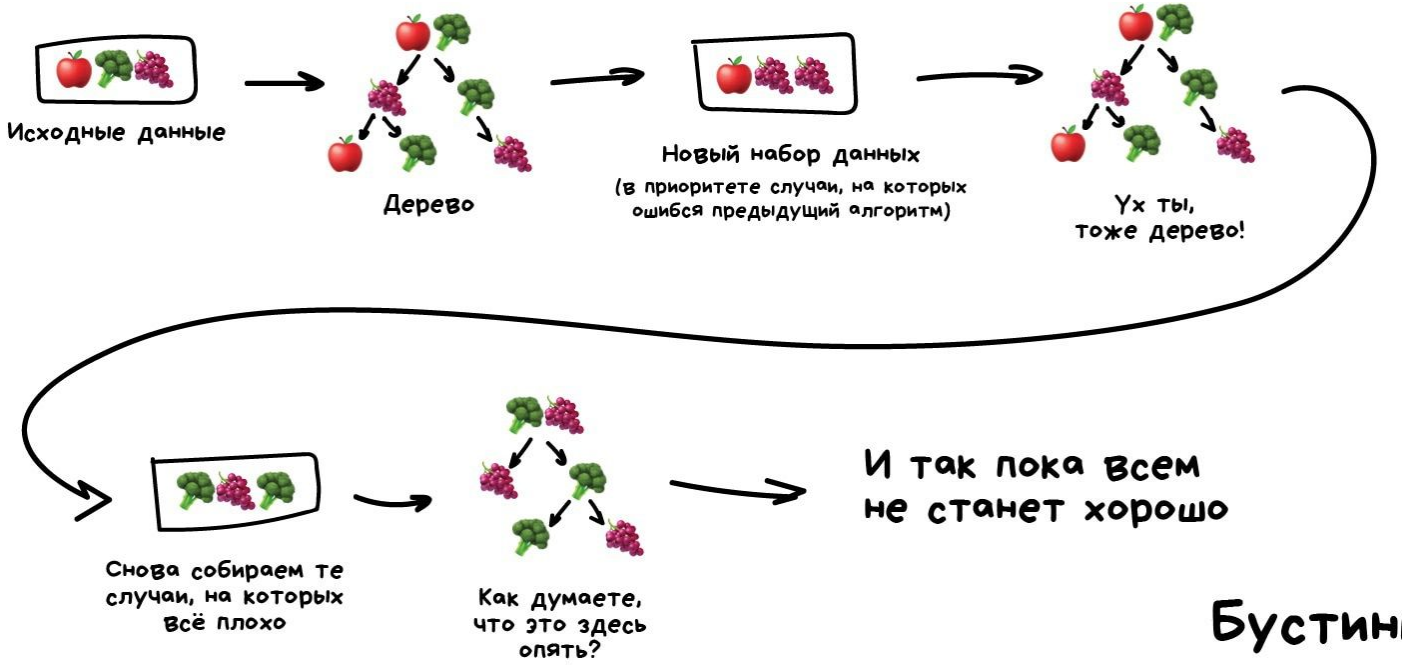
from sklearn.ensemble import RandomForestClassifier

Самый популярный вид бэггинга (потому что деревья склонны переобучаться).

Некоторые гиперпараметры:

- n_estimators — сколько деревьев
- параметры, используемые для деревьев (критерий, глубина дерева etc)

Boosting



Обучаем алгоритмы последовательно, каждый следующий уделяет особое внимание тем случаям, на которых ошибся предыдущий.

Бустинг

Boosting sklearn

Самые популярные алгоритмы для бустинга:

- XGBoost
- Light GBM
- CatBoost

Здесь разбираются их различия.

Визуализация и метрики

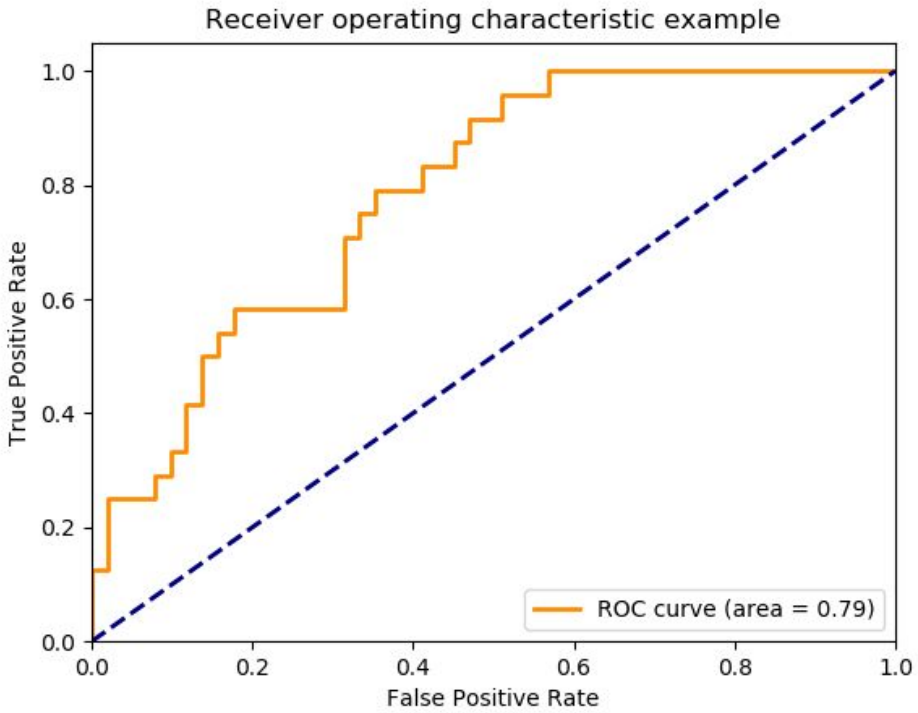
Receiver Operating Characteristic (ROC)

TPR — то же самое, что recall — TP / (TP + TN)

FPR — доля FP / (FP + TN)

Receiver Operating Characteristic — нахождение баланса между TPR и TFR при разных порогах принятия решения.

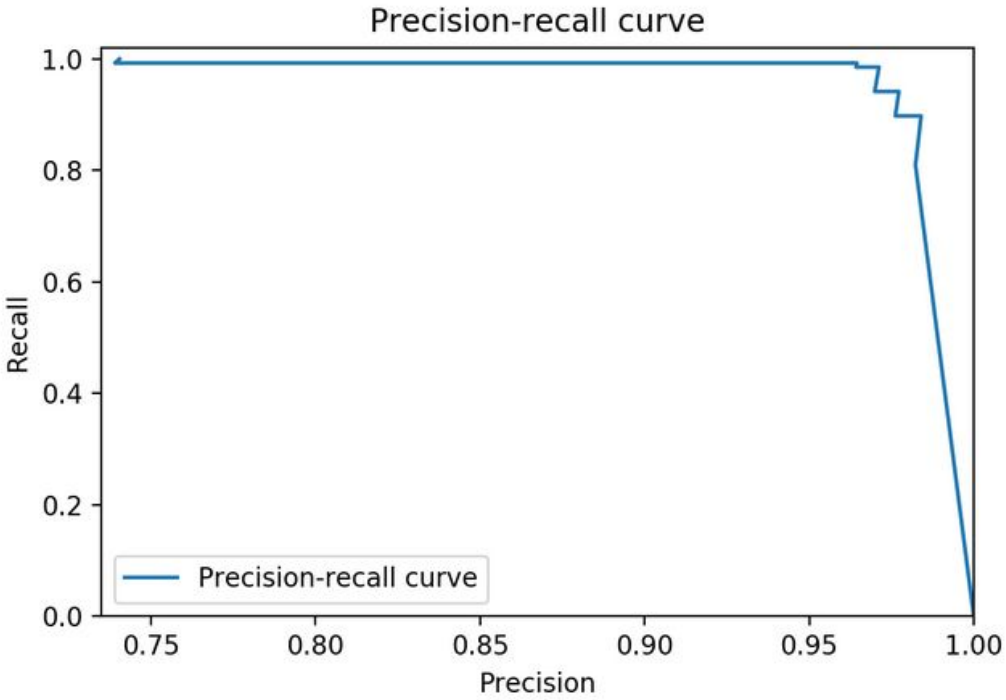
Receiver Operating Characteristic



Источник каринки. Там же показано, как построить её с помощью matplotlib.

Area under the ROC curve (AUC) is a good measure of the performance of the classification algorithm. If it is near 0.5, the classifier is not much better than random guessing, whereas **it gets better as the area gets close to 1**.

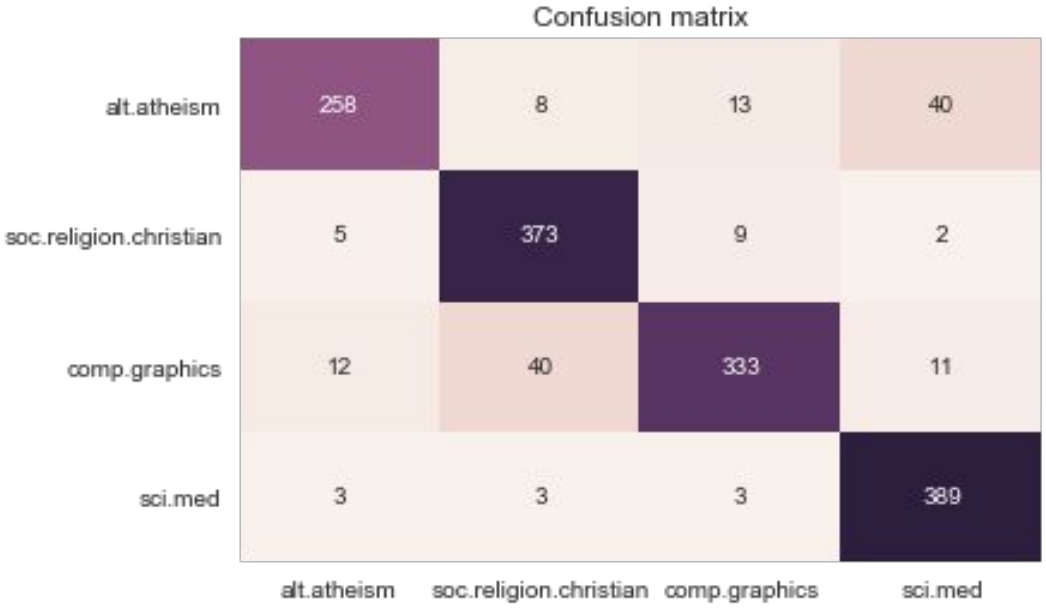
Precision-recall curve



With the precision-recall curve, the closer it is to the top-right corner, the better the algorithm.

And hence a larger **area under the curve** (AUC) indicates that the algorithm has higher recall and higher precision.

Confusion matrix



Раньше мы говорили о confusion matrix 2x2.

Но в задаче мультиклассовой классификации они могут быть и больше, и бывают очень полезны.

Источник картинки, данные про классификацию новостей.

Ресурсы

Почитать / посмотреть

Статьи:

- Дилемма смещения–дисперсии (рус)
- Understanding the Bias-Variance Tradeoff
- Regularization in Machine Learning
- introduction to data visualization in python
- SVM с картинками
- SVM с кодом
- Машинное обучение для людей, ансамбли (рус)
- объяснение про ROC curve

Видео про регуляризацию.