Логистическая регрессия и SVM

Повторение

Классификация

- $Y = \{-1, +1\}$
- -1 – отрицательный класс
- +1 – положительный класс
- Алгоритм $a(x)$ должен возвращать одно из двух чисел

Линейный классификатор

$a(x) = sign(w_0 + \sum_{j=1}^{d} w_j x_j)$ - выдает знак вещественного числа

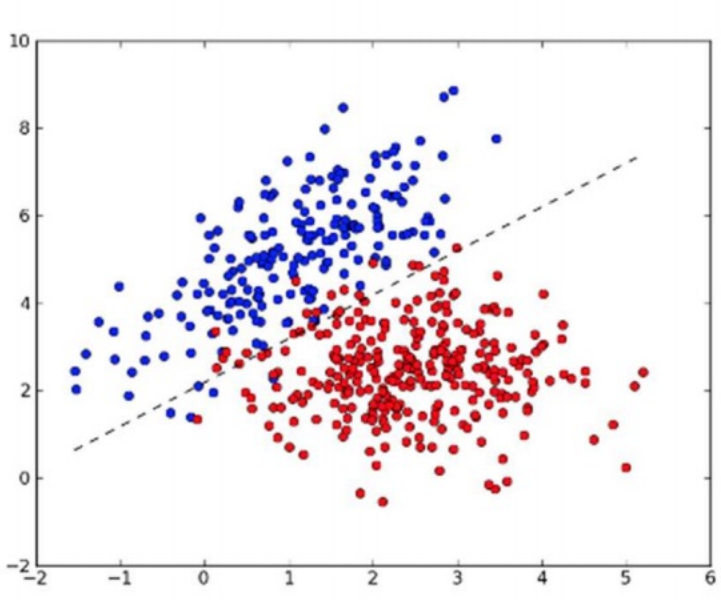
Свободный коэффицент

веса

признаки

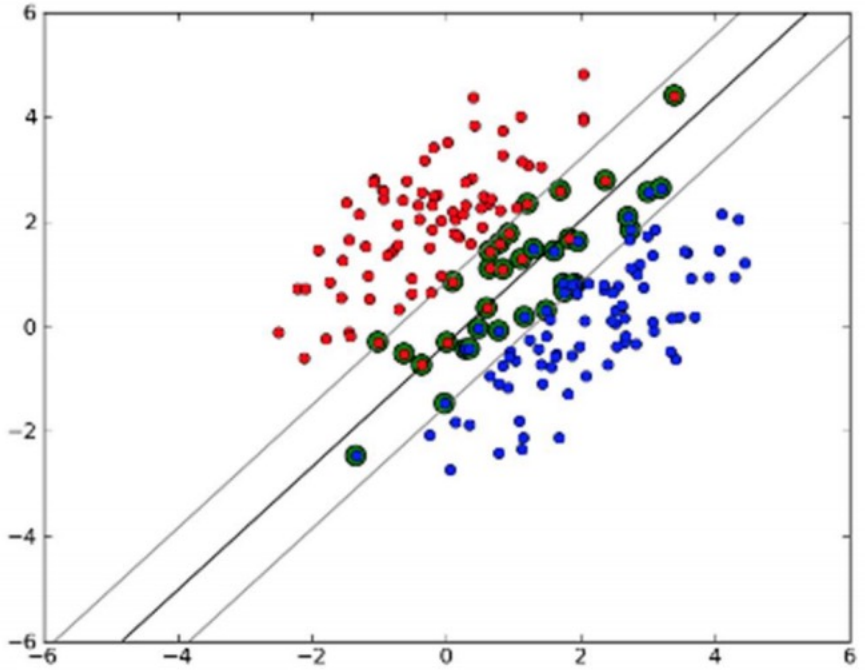
Геометрия

- Линейный классификатор проводит гиперплоскость
- $< w, x > < 0$ – обьект слева от нее
- $< w, x > > 0$ – обьект справа от нее



Отступ

- $M_i = y_i < w, x_i >$
- $M_i > 0$ - классификатор дает верный ответ
- $M_i < 0$ - классификатор ошибается
- Чем дальше отступ от нуля, тем больше уверенности



Линейный классификатор

- Линейный классификатор разделяет два класса гиперплоскостью
- Чем больше отступ по модулю, тем дальше объект от гиперплоскости
- Знак отступа говорит о корректности предсказания

Функция потерь в классификации

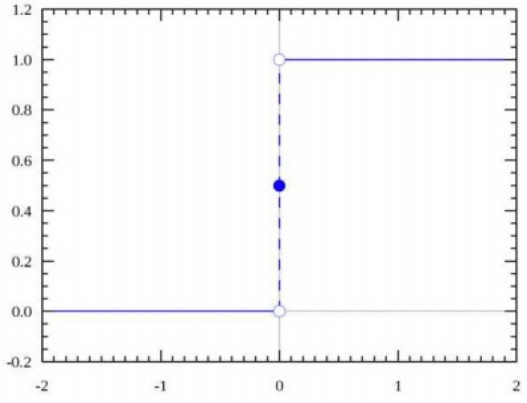
- Частый выбор – бинарная функция потерь
  $L(y, a) = [a \neq y]$
- Функционал ошибки – доля ошибок (error rate)
  $\mathrm{Q}(y, a) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) \neq y_i]$
- Нередко измеряют долю верных ответов (accuracy):
  $\mathrm{Q}(y, a) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) = y_i]$

Доля ошибок для линейного классификатора

- Функционал ошибки:

$$Q(w, X) = \frac{1}{l} \sum_{i=1}^{l} [sign(< w, x_i >) \neq y_i]$$

- Индикатор – недифференцируемая функция



Отступы для линейного классификатора

- Функционал ошибки:

$$Q(w, X) = \frac{1}{l}\sum_{i=1}^{l}[sign(< w, x_i >) \neq y_i]$$

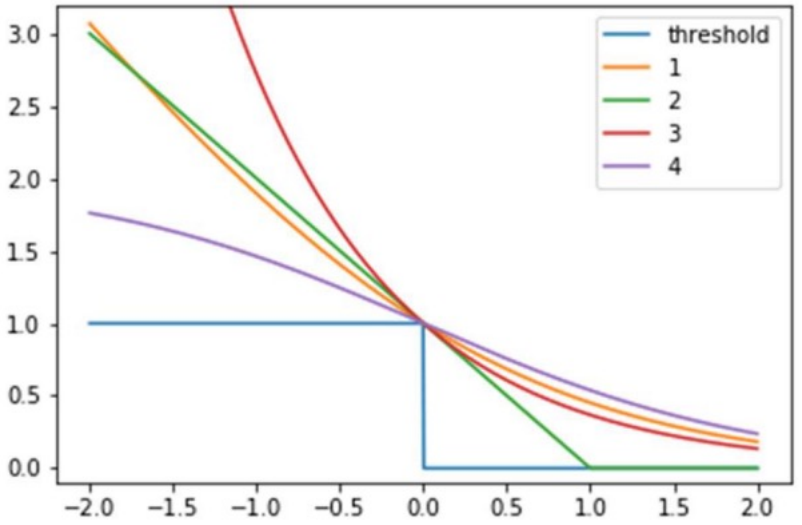
- Альтернативная запись:

$$Q(w, X) = \frac{1}{l}\sum_{i=1}^{l}[y_i < w, x_i >< 0]$$

$y_i < w, x_i > = M_i$ - отступ

Примеры верхних оценок

- $\tilde{L}(M) = \log(1 + e^{-M}) -$ логистическая
- $\tilde{L}(M) = \max(0, 1 - M) -$ кусочно $-$ линейная
- $\tilde{L}(M) = e^{-M} -$ экспоненциальная
- $\tilde{L}(M) = \frac{2}{1+e^{M}}$ - сигмоидная



Качество классификации

- Доля неправильных ответов:

$$\frac{1}{l}\sum_{i=1}^{l}[a(x_i) \neq y_i]$$

- Доля правильных ответов:

$$\frac{1}{l}\sum_{i=1}^{l}[a(x_i) = y_i]$$

Матрица ошибок

	Y = 1	Y = -1
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN0

Точность (precision)

Можно ли доверять классификатору при $a(x) = 1$?

$$precision(a, X) = \frac{TP}{TP + FP}$$

Полнота (recall)

Как много положительных обьектов находит классификатор?

$$recall(a, X) = \frac{TP}{TP + FN}$$

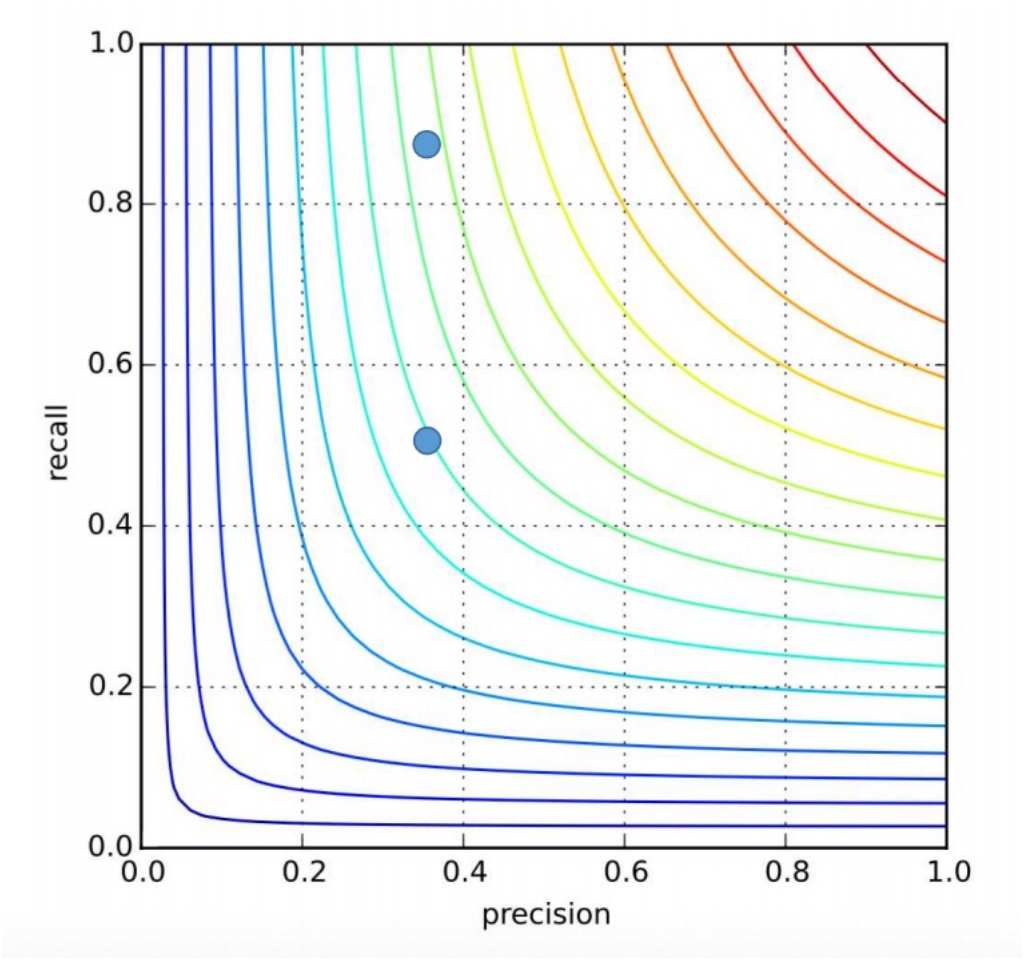
F-мера

$$F = \frac{2 * precision * recall}{precision + recall}$$

- $precision = 0.4, recall = 0.5$
- $F = 0.44$

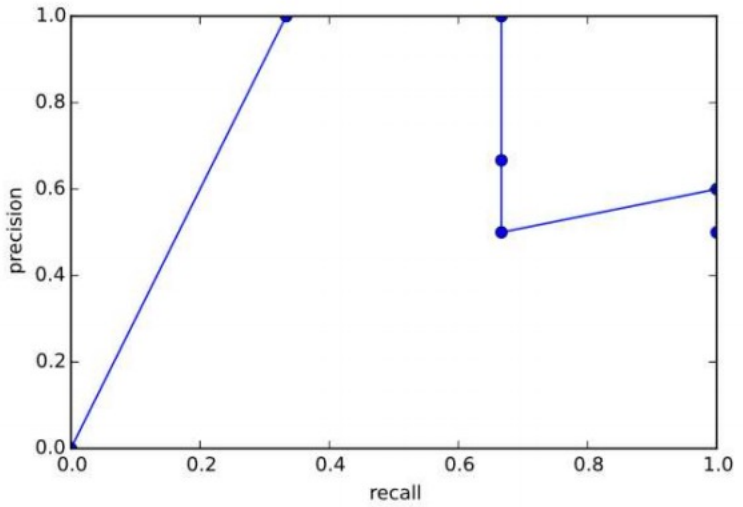
- $precision = 0.4, recall = 0.9$
- $F = 0.55$

Работает!



PR-кривая

- Кривая точности-полноты
- Ось X — полнота
- Ось Y — точность
- Точки — значения точности и
- полноты при последовательных порогах



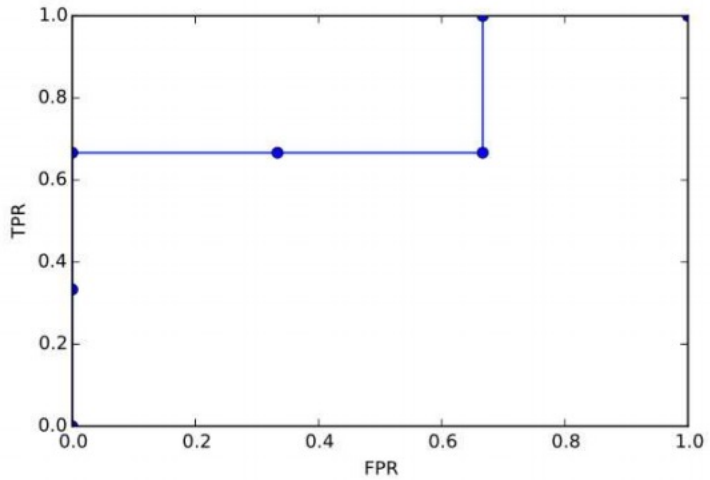
ROC-кривая

- Receiver Operating Characteristic
- Ось X — False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

- Ось Y - True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$



ROC-кривая

AUC (Area Under Curve) - площадь под ROC-кривой.

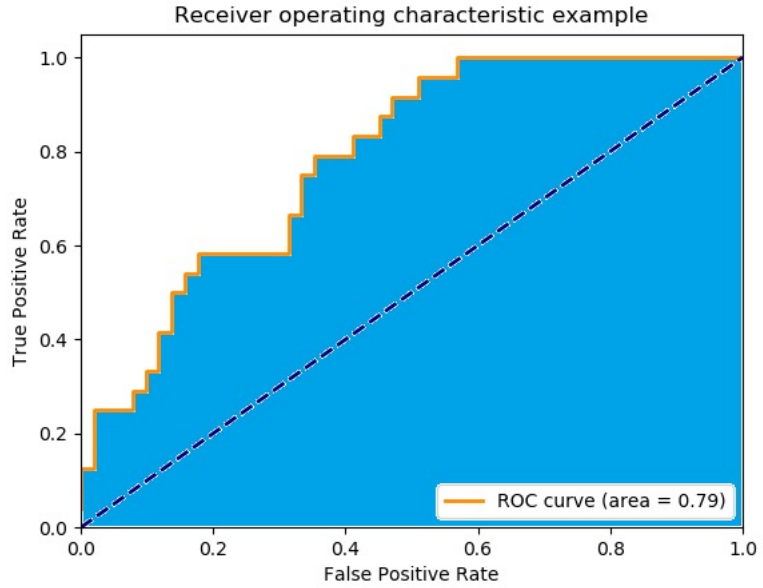
$$AUC \in [0; 1].$$

- $AUC = 1$ -

идеальная классификация

- $AUC = 0.5$ -

случайная классификация



ROC-кривая

- Receiver Operating Characteristic

- Ось X — False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

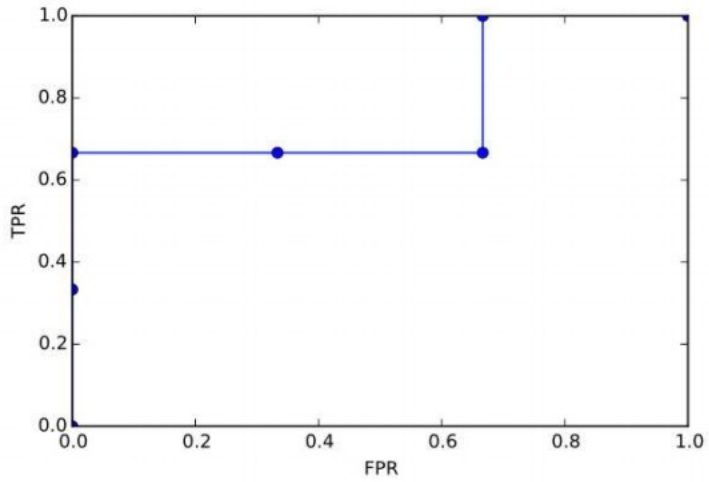
$FP + TN$ – число отрицательных обьектов

- Ось Y - True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

T$P + FN$ – число положительных обьектов

=



AUC-ROC

$$FPR = \frac{FP}{FP + TN} \qquad TPR = \frac{TP}{TP + FN}$$

FP и TP нормируются на размеры классов

- AUC-ROC не поменяется при изменении баланса классов
- Учитывает True Negatives
- Идеальный алгоритм: $AUC - ROC = 1$
- Худший алгоритм: $AUC - ROC \approx 0.5$

Логистическая регрессия

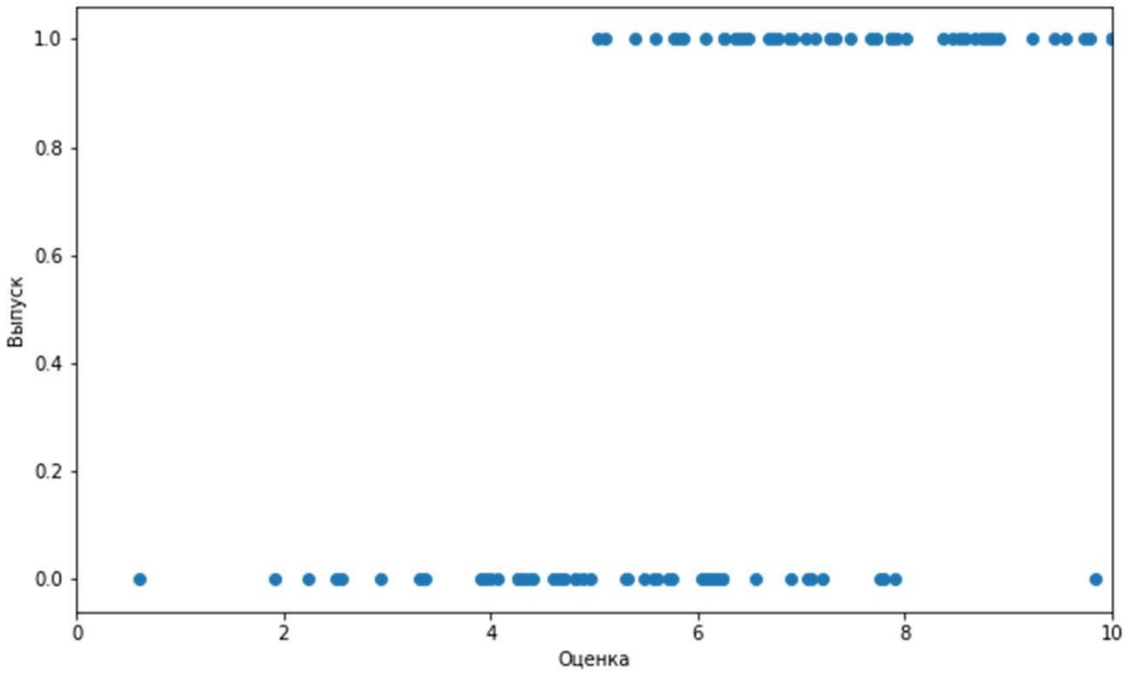
Логистическая регрессия

- Решаем задачу бинарной классификации: $\mathbb{Y} = \{-1, +1\}$

- Минимизация верхней оценки:

$$\frac{1}{\ell}\sum_{i=1}^{\ell} \log(1 + \exp(-y_i\langle w, x_i\rangle)) \to \min_w$$

Предсказание вероятностей



Предсказание вероятностей

Предсказание вероятностей

- Кредитный скоринг
- Стратегия: выдавать кредит только клиентам с $b(x) > 0.9$
- 10% невозвращённых кредитов — нормально

Предсказание вероятностей

- Баннерная реклама
- $b(x)$— вероятность, что пользователь кликнет по рекламе
- $c(x)$— прибыль в случае клика
- $c(x)b(x)$— хотим оптимизировать

Предсказание вероятностей

- Прогнозирование оттока клиентов
- Медицинская диагностика
- Поисковое ранжирование (насколько веб-страница соответствует запросу?)

Предсказание вероятностей

Будем говорить, что модель $b(x)$ предсказывает вероятности, среди объектов $b(x) = p$ если доля положительных равна p.

Логистическая регрессия

Логистическая регрессия - это линейный классификатор!

–

Предсказание вероятностей

Линейный классификатор

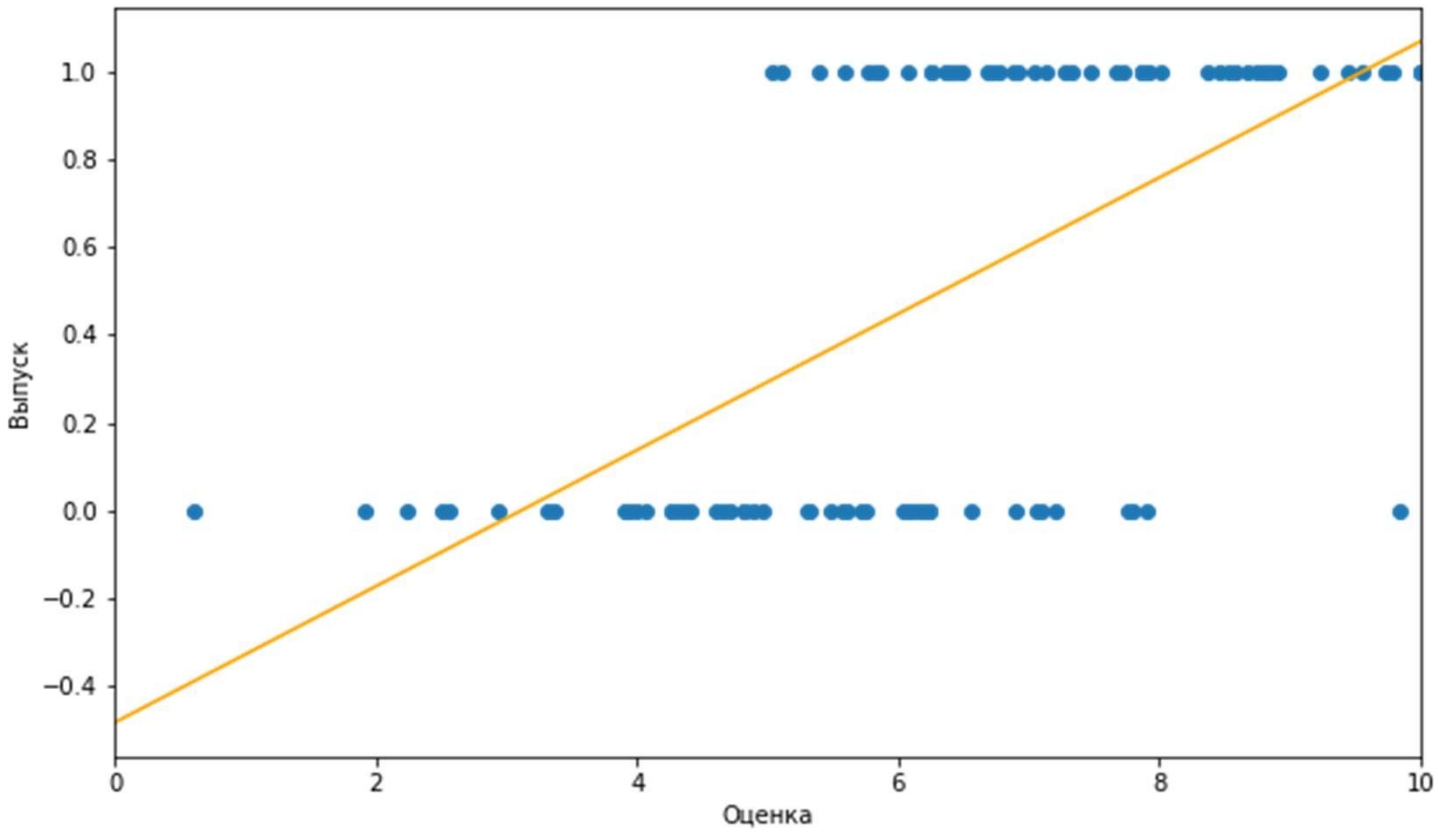
$$a(x) = \text{sign}\,\langle w, x\rangle$$

- Обучим как-нибудь — например, на логистическую функцию потерь:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log(1+\exp(-y_i\langle w, x_i\rangle)) \to \min_w$$

- Может, $\langle w, x\rangle$ сойдёт за оценку?

Предсказание вероятностей



$$b(x) = w_1 x + w_0$$

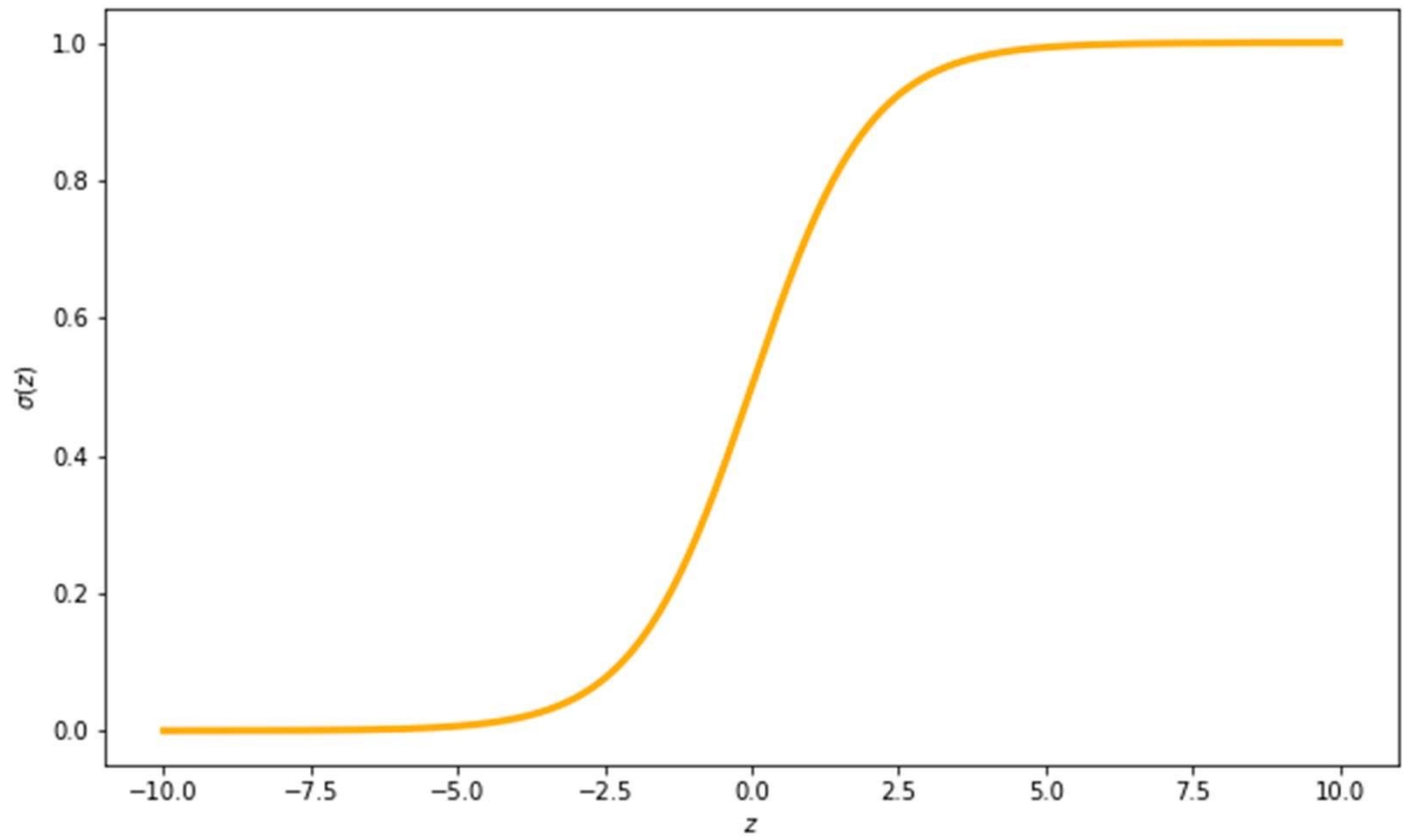
Линейный классификатор

- Переведём выход модели на отрезок [0, 1]
- Например, с помощью сигмоиды[1]:

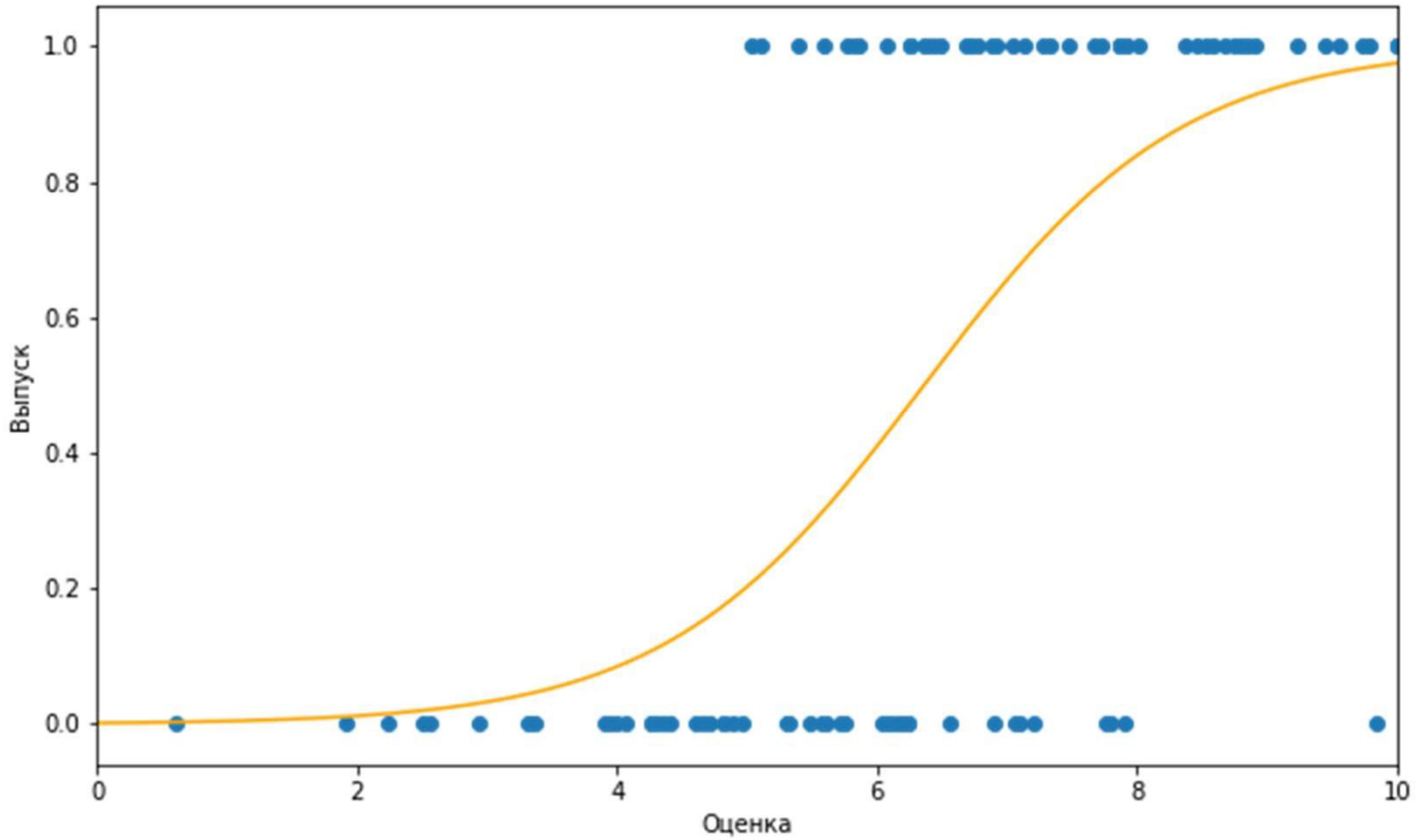
$$\sigma(\langle w, x \rangle) = \frac{1}{1 + \exp(-\langle w, x \rangle)}$$

[1] https://sebastianraschka.com/faq/docs/logistic-why-sigmoid.html

Сигмоида

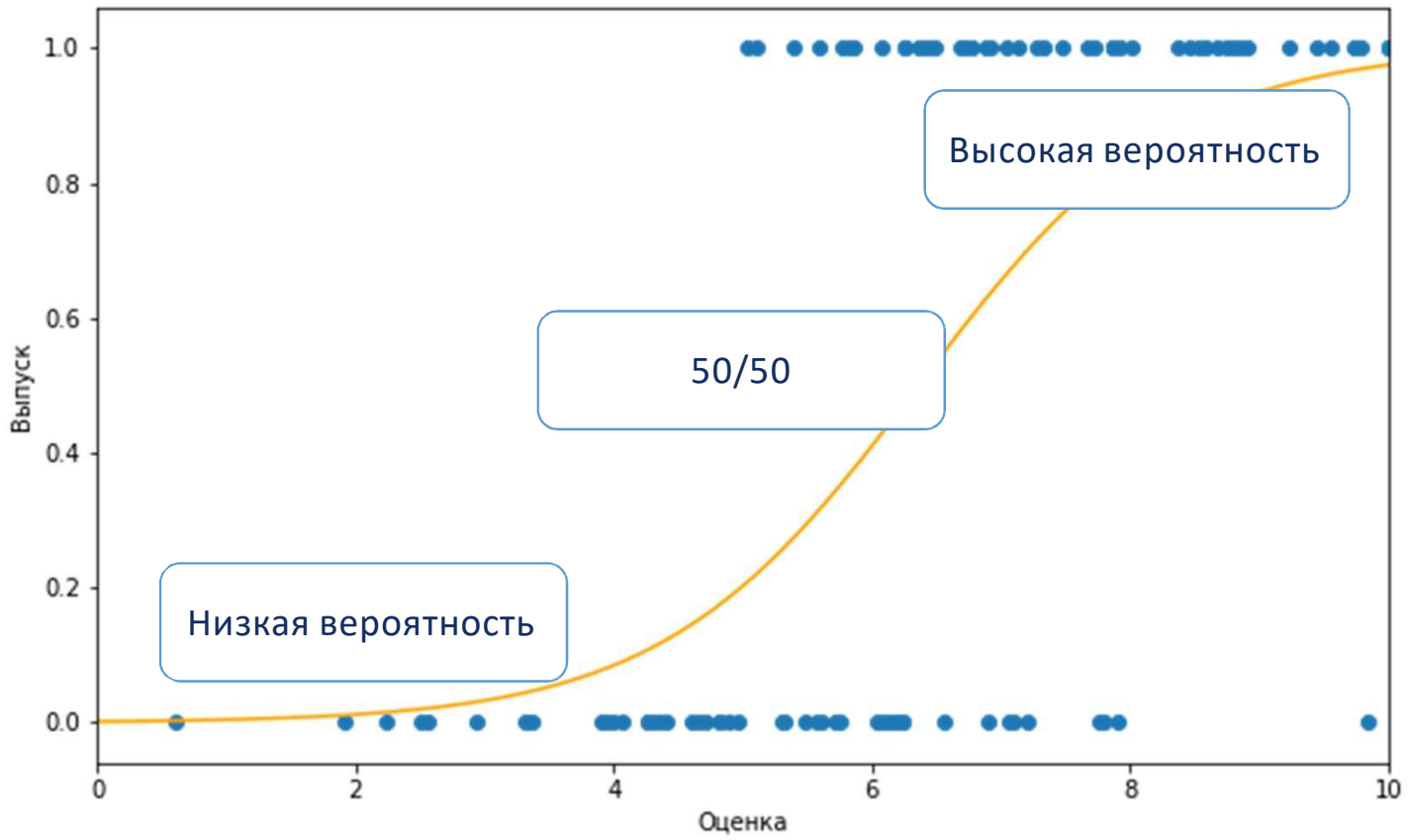


Предсказание вероятностей



$$b(x) = \sigma(w_1 x + w_0)$$

Предсказание вероятностей



Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?

Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$

Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$ или $\langle w, x_i \rangle \to +\infty$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$ или $\langle w, x_i \rangle \to -\infty$

Предсказание вероятностей

- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$ или $\langle w, x_i \rangle \to +\infty$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$ или $\langle w, x_i \rangle \to -\infty$
- То есть задача — сделать отступы на всех объектах максимальными

$$y_i \langle w, x_i \rangle \to \max_w$$

Предсказание вероятностей

- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$

$$-\sum_{i=1}^{\ell} \{[y_i = 1]\sigma(\langle w, x_i \rangle) + [y_i = -1](1 - \sigma(\langle w, x_i \rangle))\} \to \min_w$$

Предсказание вероятностей

$$-\sum_{i=1}^{\ell}\{[y_i = 1]\sigma(\langle w, x_i\rangle) + [y_i = -1]\big(1 - \sigma(\langle w, x_i\rangle)\big)\} \to \min_w$$

- Если $y_i = +1$ и $\sigma(\langle w, x_i\rangle) = 0$, то штраф равен 1
- Надо строже!

Предсказание вероятностей

$$-\sum_{i=1}^{\ell}\{[y_i = 1]\log\sigma(\langle w, x_i\rangle) + [y_i = -1]\log(1 - \sigma(\langle w, x_i\rangle))\} \to \min_w$$

- Если $y_i = +1$ и $\sigma(\langle w, x_i\rangle) = 0$, то штраф равен $-\log 0 = +\infty$
- Достаточно строго
- Функция потерь называется **log-loss**

$$L(y, z) = -[y = 1]\log z - [y = -1]\log(1 - z)$$

Логистическая регрессия

$$-\sum_{i=1}^{\ell} \{[y_i = 1] \log \sigma(\langle w, x_i \rangle) + [y_i = -1] \log (1 - \sigma(\langle w, x_i \rangle))\} =$$

$$-\sum_{i=1}^{\ell} \left\{[y_i = 1] \log \frac{1}{1 + \exp(-\langle w, x \rangle)} + [y_i = -1] \log \left(1 - \frac{1}{1 + \exp(-\langle w, x \rangle)}\right)\right\} =$$

$$-\sum_{i=1}^{\ell} \left\{[y_i = 1] \log \frac{1}{1 + \exp(-\langle w, x \rangle)} + [y_i = -1] \log \left(\frac{1}{1 + \exp(\langle w, x \rangle)}\right)\right\} =$$

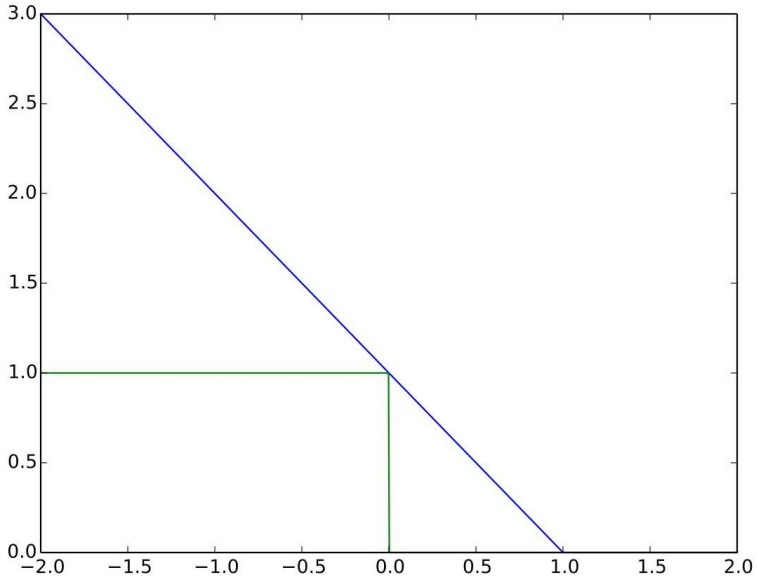
$$\sum_{i=1}^{\ell} \{[y_i = 1] \log(1 + \exp(-\langle w, x \rangle)) + [y_i = -1] \log (1 + \exp(\langle w, x \rangle))\} =$$

$$\sum_{i=1}^{\ell} \log(1 + \exp(-y_i \langle w, x_i \rangle))$$

Метод опорных векторов

Hinge loss

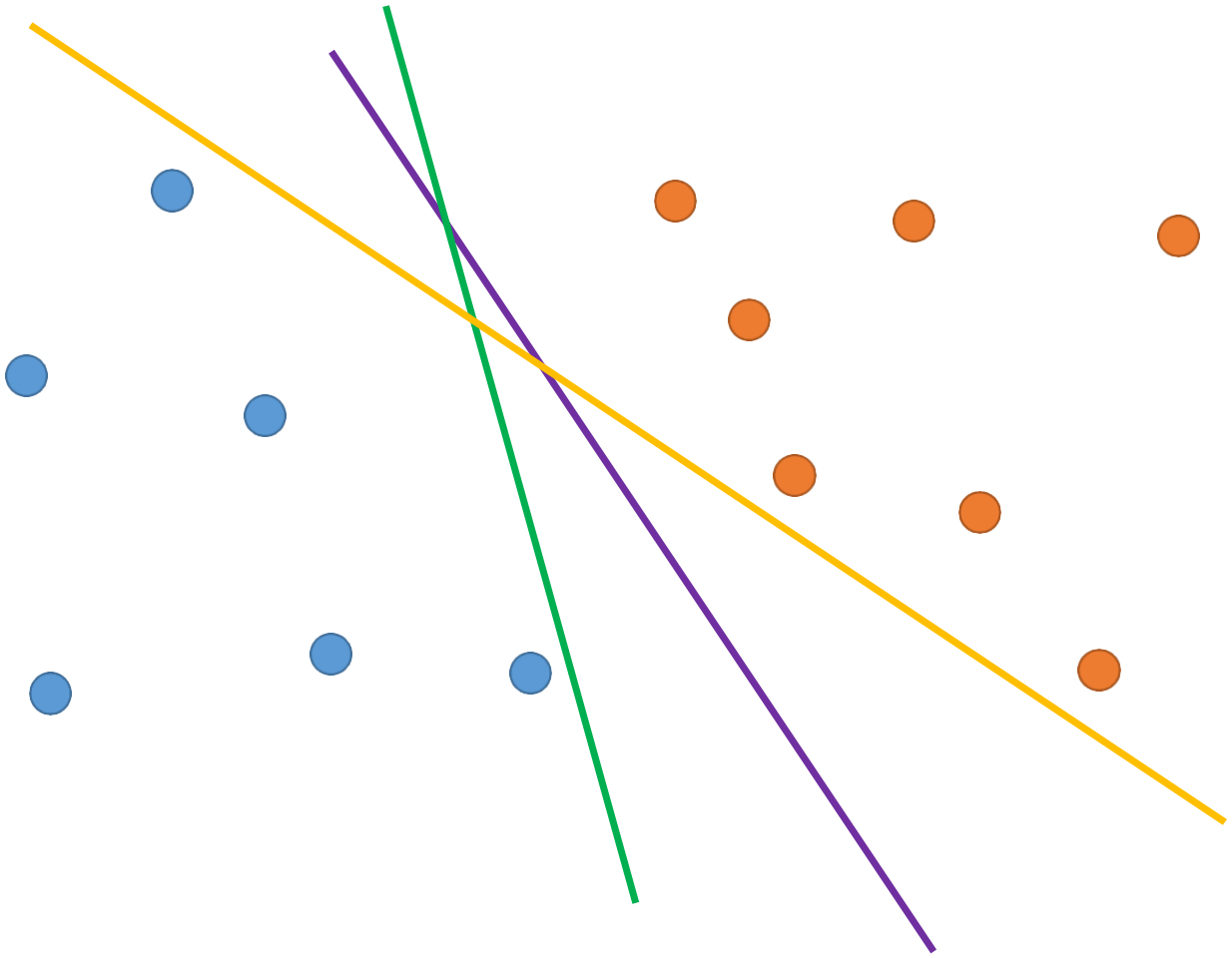
- Бинарная классификация: $\mathbb{Y} = \{-1, +1\}$



- Минимизация верхней оценки:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\max(0, 1 - y_i\langle w, x_i\rangle) \to \min_w$$

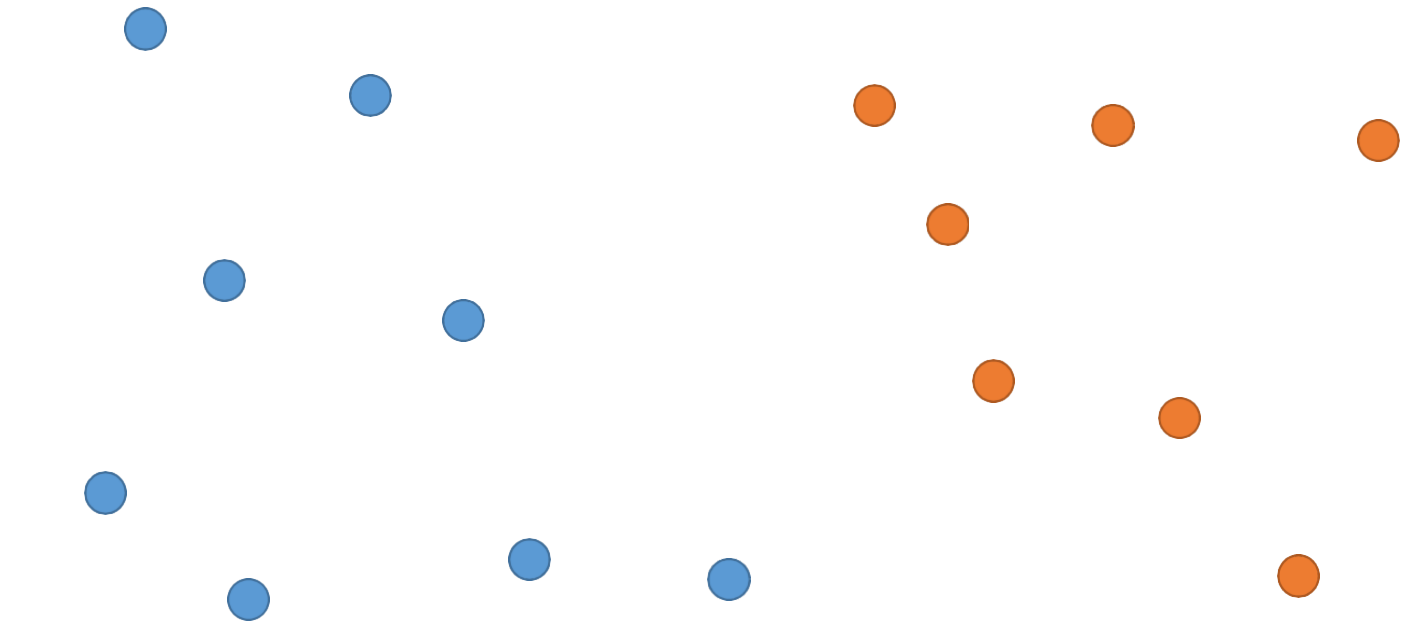
Какой классификатор лучше?



Отступ классификатора

- Будем максимизировать отступ классификатора — расстояние от гиперплоскости до ближайшего объекта

Отступ классификатора



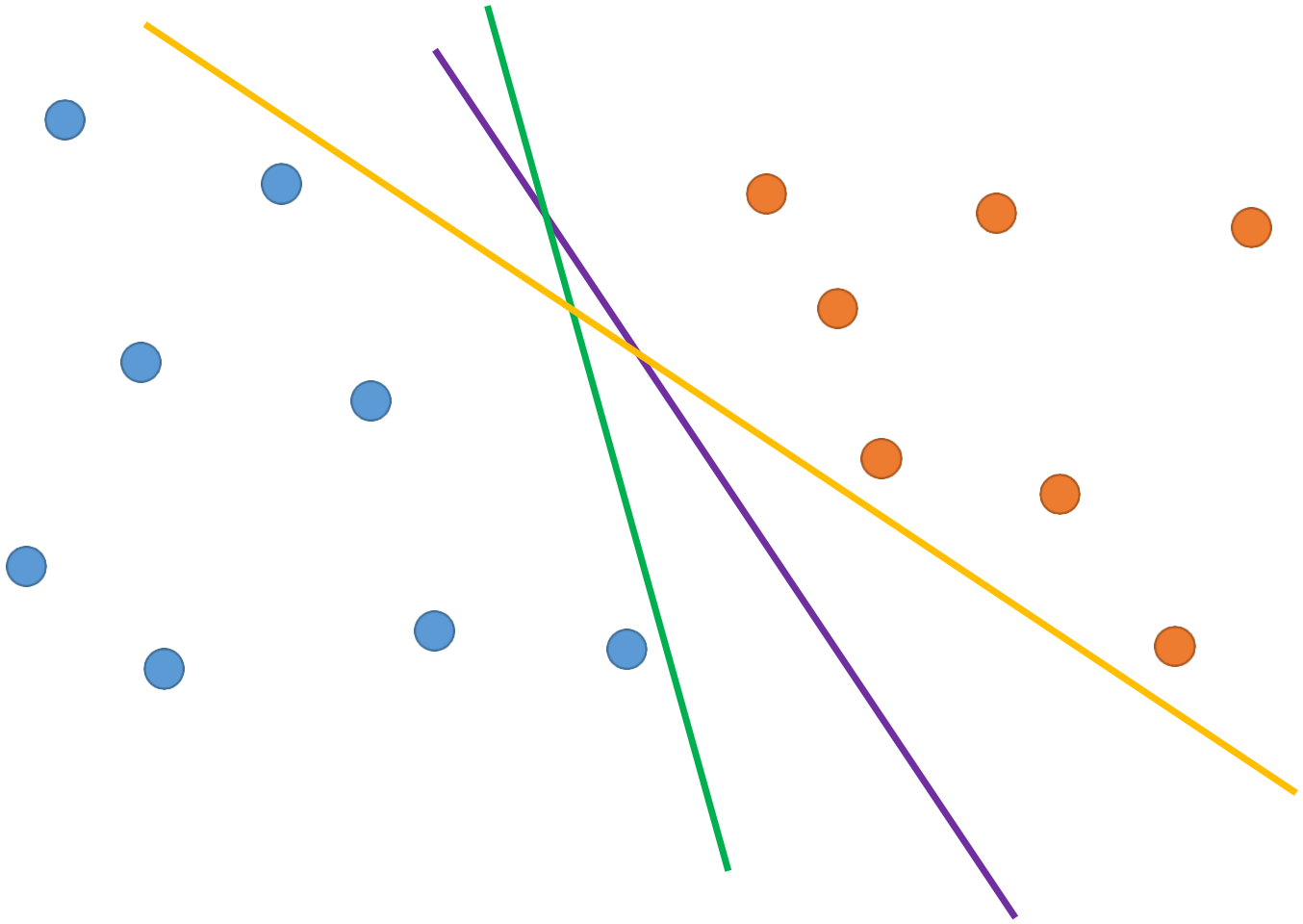
Отступ классификатора

- Будем максимизировать отступ классификатора — расстояние от гиперплоскости до ближайшего объекта
- При этом будет стараться сделать поменьше ошибок
- По сути, делаем как можно меньше предположений о модели, и верим, что это понизит вероятность переобучения

Простой случай

- Будем считать, что выборка линейно разделима
- Существует линейный классификатор, не допускающий ни одной ошибки

Линейно разделимый случай



Линейно разделимый случай

- **Требование 1**: $y_i(\langle w, x_i \rangle + w_0) > 0$ для всех $i = 1, \dots, \ell$
- **Требование 2**: максимальный отступ классификатора

Отступ классификатора

- Расстояние от точки до гиперплоскости $\langle w, x \rangle + w_0 = 0$:

$$\frac{|\langle w, x \rangle + w_0|}{\|w\|}$$

- Отступ классификатора:

$$\min_{i=1,\ldots,\ell} \frac{|\langle w, x_i \rangle + w_0|}{\|w\|}$$

Небольшое предположение

- Линейный классификатор:

$$a(x) = \text{sign}\left(\langle w, x_i \rangle + w_0\right)$$

- Если мы поделим w и w_0 на число $\mathrm{k} > 0$, то выходы классификатора никак не поменяются:

$$a(x) = \text{sign}\left(\frac{\langle w, x_i \rangle + w_0}{k}\right) = \text{sign}\left(\langle w, x_i \rangle + w_0\right)$$

Небольшое предположение

- Поделим w и w_0 на $\min\limits_{i=1,\dots,\ell} |\langle w, x_i \rangle + w_0| > 0$, после этого будет выполнено

$$\min_{i=1,\dots,\ell} |\langle w, x_i \rangle + w_0| = 1$$

Отступ классификатора

- Расстояние от точки до гиперплоскости $\langle w, x \rangle + w_0 = 0$:

$$\frac{|\langle w, x \rangle + w_0|}{\|w\|}$$

- Отступ классификатора:

$$\min_{i=1,\ldots,\ell} \frac{|\langle w, x_i \rangle + w_0|}{\|w\|} = \frac{\min\limits_{i=1,\ldots,\ell} |\langle w, x_i \rangle + w_0|}{\|w\|} = \frac{1}{\|w\|}$$

Линейно разделимый случай

- **Требование 1**: $y_i(\langle w, x_i \rangle + w_0) > 0$ для всех $i = 1, \ldots, \ell$
- **Требование 2**: максимальный отступ классификатора

$$\frac{1}{\|w\|} \to \max_w$$

Линейно разделимый случай

- **Требование 1**: $y_i(\langle w, x_i \rangle + w_0) > 0$ для всех $i = 1, \dots, \ell$
- **Требование 2**: максимальный отступ классификатора

$$\frac{1}{\|w\|} \to \max_w$$

- При условии, что $\min_{i-1,\dots,\ell} |\langle w, x_i \rangle + w_0| = 1$

Линейно разделимый случай

- **Требование 1**: $y_i(\langle w, x_i \rangle + w_0) > 0$ для всех $i = 1, \dots, \ell$
- **Требование 2**: максимальный отступ классификатора

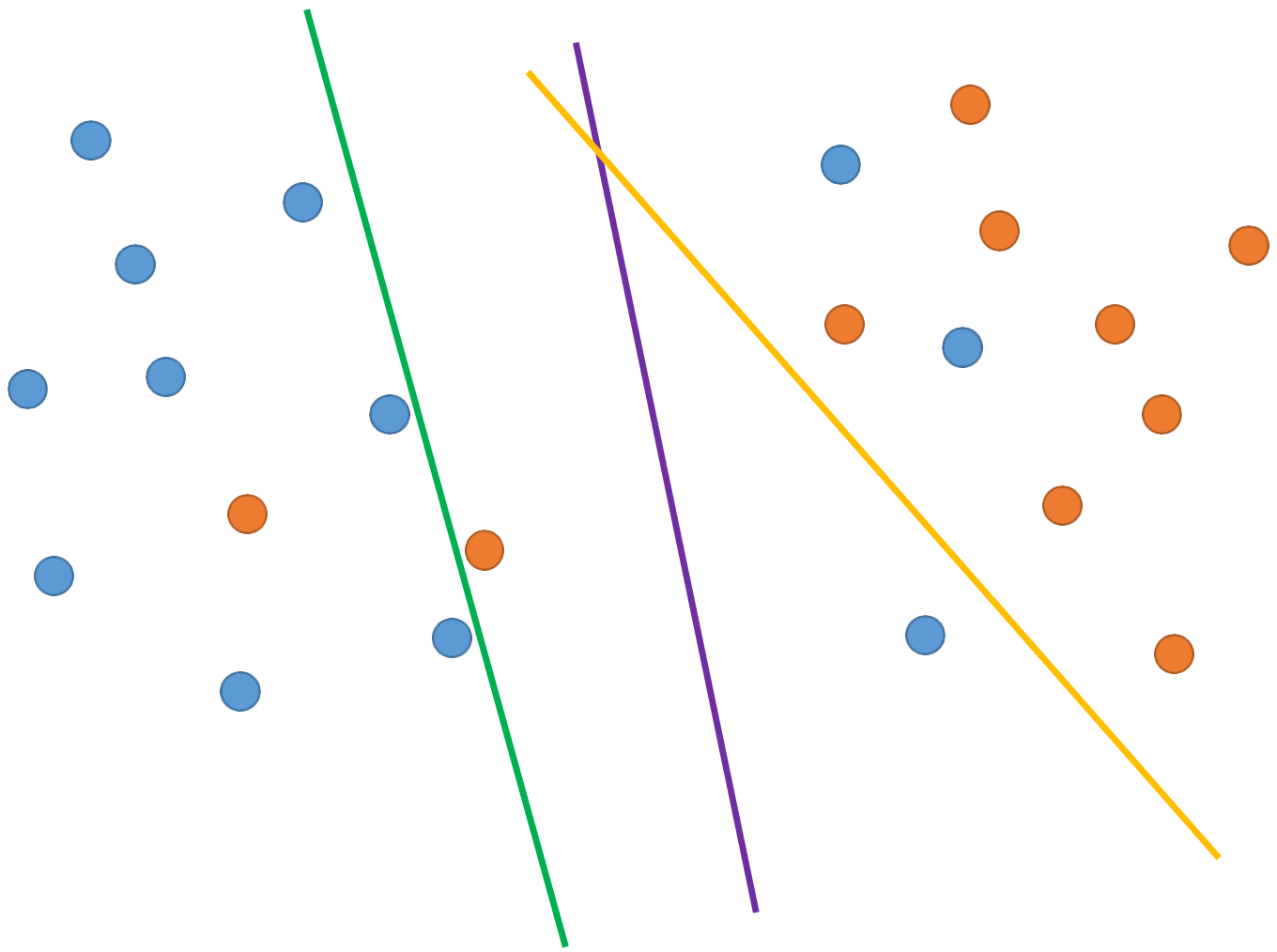
$$\frac{1}{\|w\|} \to \max_w$$

- При условии, что $|\langle w, x_i \rangle + w_0| \geq 1$
- И мы минимизируем $\|w\|$ — тогда где-то модуль отступа будет равен 1

Метод опорных векторов (SVM)

$$\begin{cases} \|w\|^2 \to \min\limits_{w,w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 \end{cases}$$

Линейно неразделимый случай



Линейно неразделимый случай

- Любой линейный классификатор допускает хотя бы одну ошибку

$$\begin{cases} \|w\|^2 \to \min\limits_{w,w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 \end{cases}$$

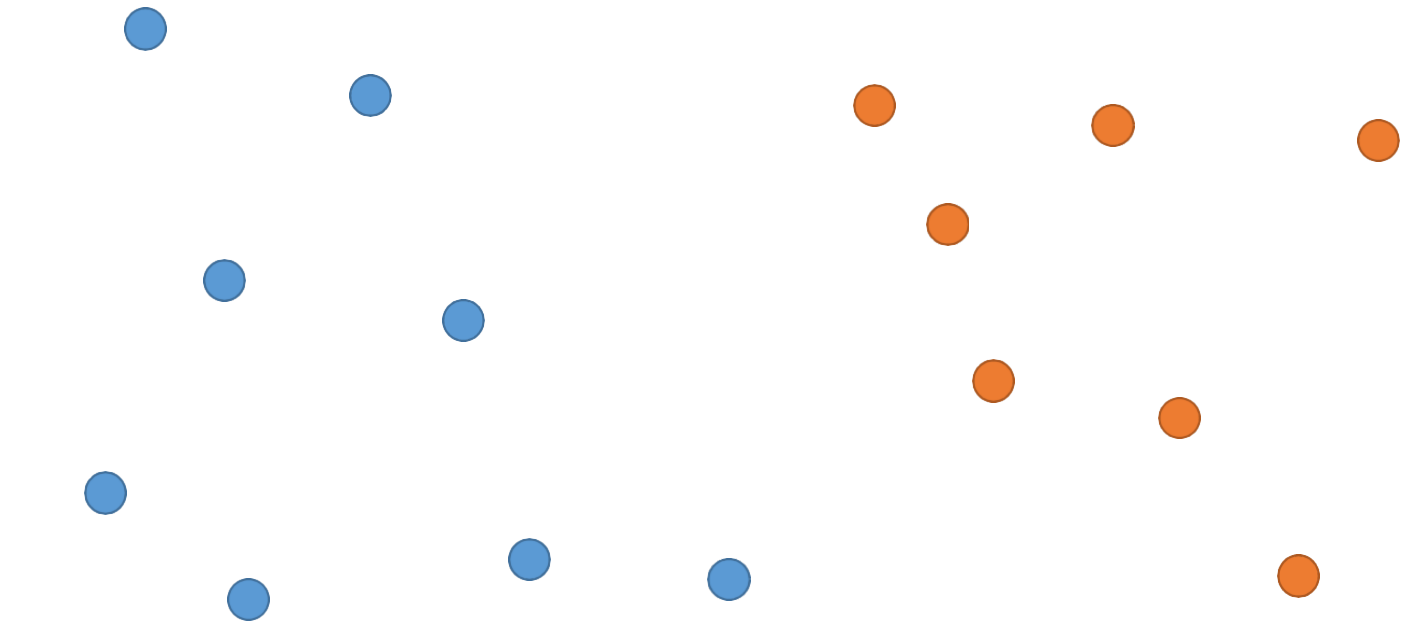
Линейно неразделимый случай

$$\begin{cases} \|w\|^2 \to \min\limits_{w,w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Линейно неразделимый случай

$$\begin{cases} \|w\|^2 \to \min\limits_{w,w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - 10^{1000} \end{cases}$$

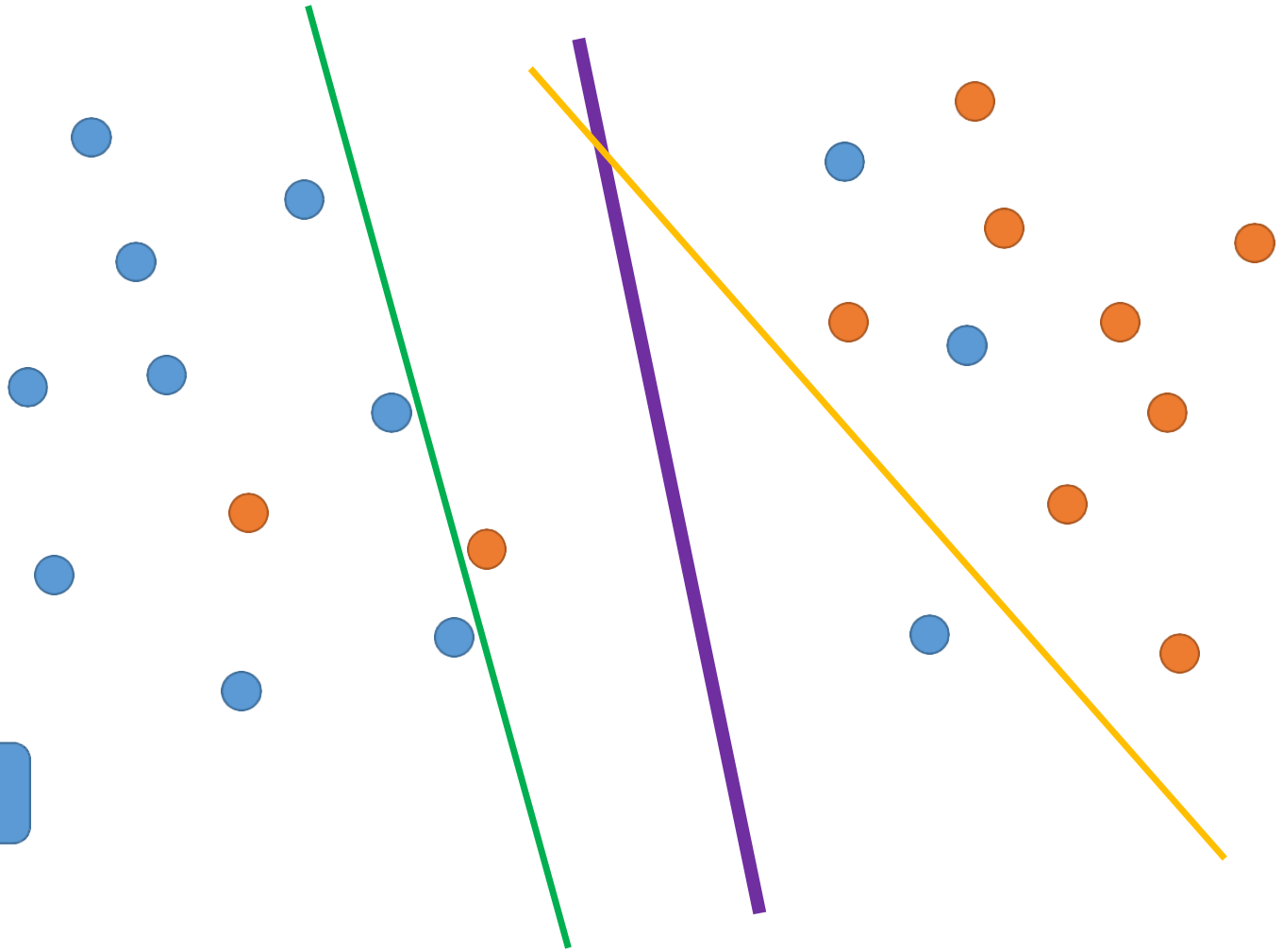
Отступ классификатора



Метод опорных векторов

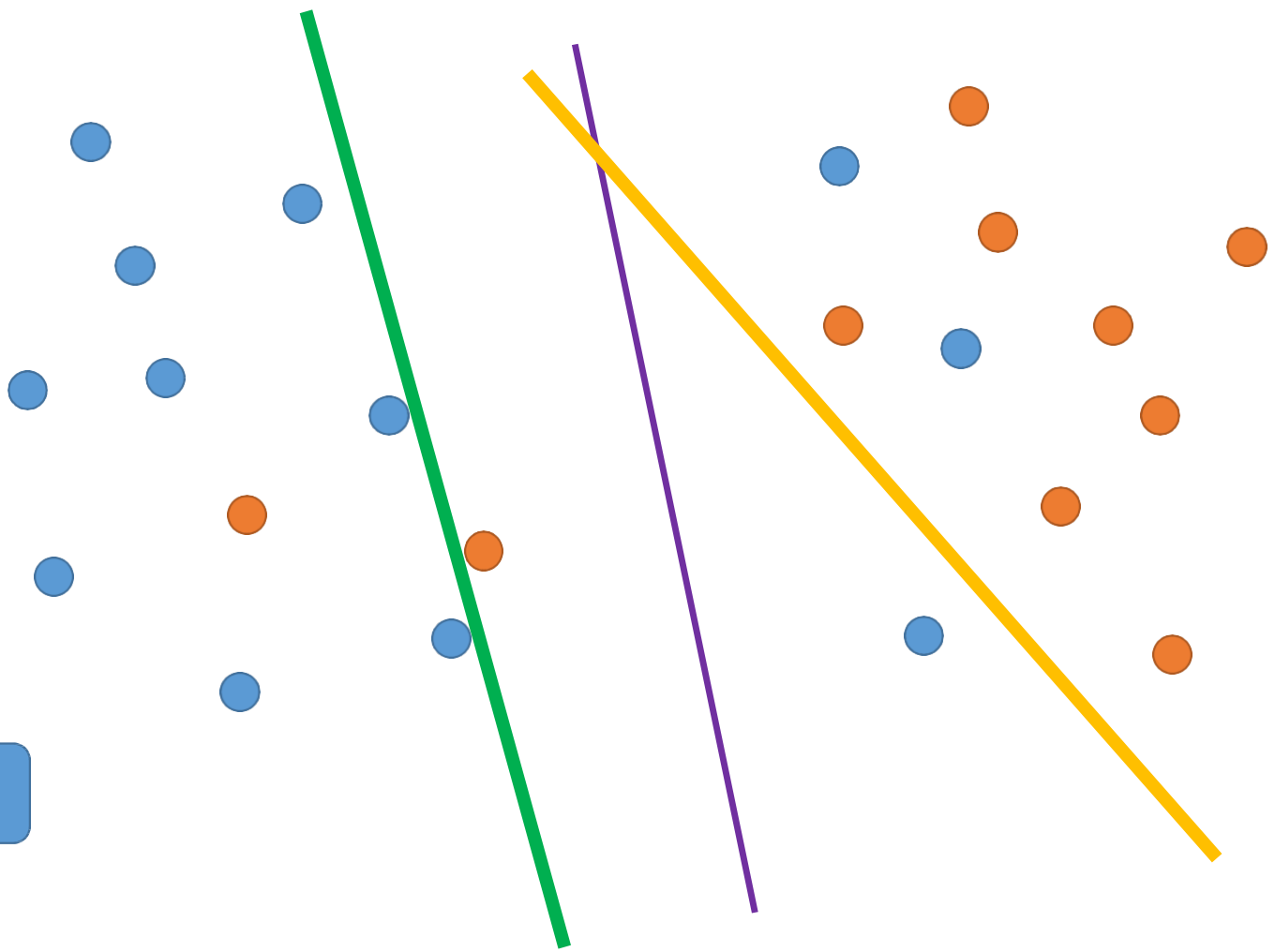
$$\begin{cases} \|w\|^2 + C\sum_{i=1}^{\ell}\xi_i \to \min\limits_{w,w_0,\xi_i} \\ y_i(\langle w, x_i\rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Линейно неразделимый случай



Маленькое C

Линейно неразделимый случай



Большое C

Метод опорных векторов

$$\begin{cases} \|w\|^2 + C\sum_{i=1}^{\ell} \xi_i \to \min\limits_{w, w_0, \xi_i} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

- Объединим ограничения:

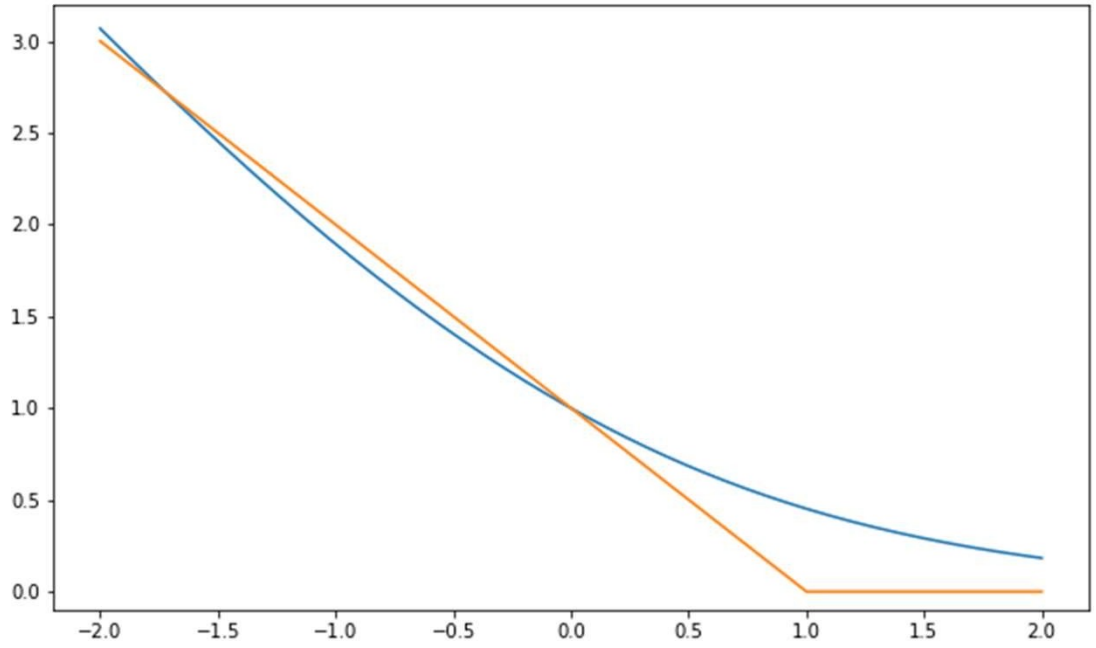
$$\xi_i \geq \max\big(0, 1 - y_i(\langle w, x_i \rangle + w_0)\big)$$

Метод опорных векторов

$$C\sum_{i=1}^{\ell}\max\left(0, 1 - y_i(\langle w, x_i\rangle + w_0)\right) + \|w\|^2 \to \min_{w,w_0}$$

- Функция потерь (hinge loss) + регуляризация

Сравнение логистической регрессии и SVM



Резюме

- Логистическая регрессия минимизирует логистические потери
- Метод опорных векторов основан на идее максимизации отступа классификатора

Метод опорных векторов

Спасибо за внимание!

Ildar Safilo

@Ildar_Saf
irsafilo@gmail.com
https://www.linkedin.com/in/isafilo/