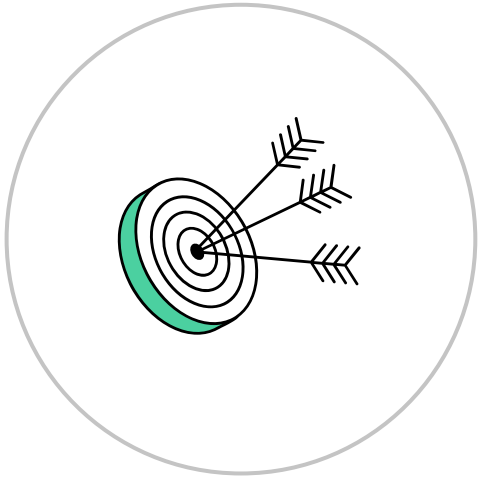
Классификация: логистическая регрессия и SVM

Цели занятия

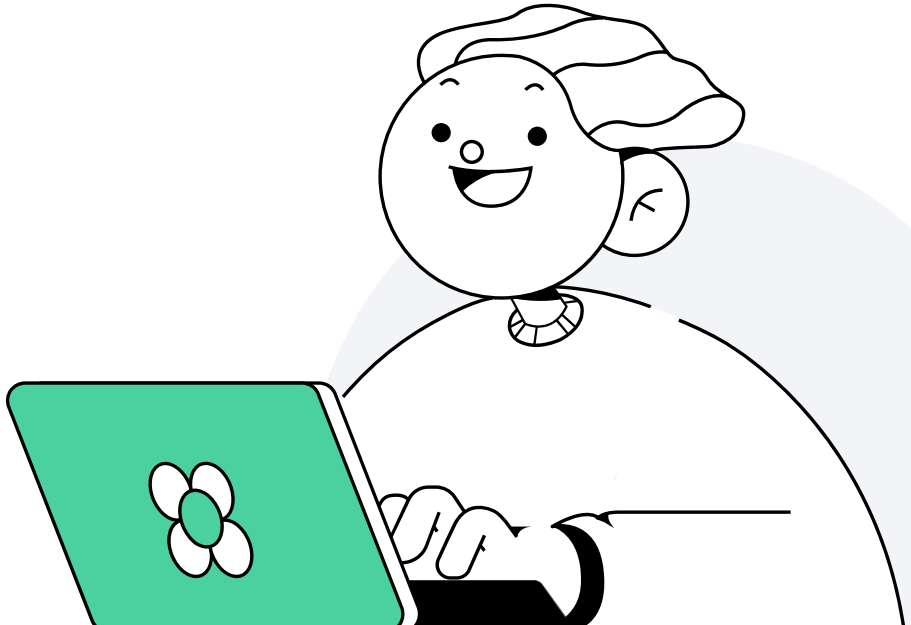
- Рассмотреть задачу классификации
- Познакомиться с линейным классификатором
- Узнать, как строится логистическая регрессия
- Усвоить метод построения модели Support Vector Machines (SVM)
- Решить задачи предсказания пола спортсмена и класса цветов



План занятия

1. Задача классификации
2. Логистическая регрессия
3. SVM
4. Практика
5. Итоги

*Нажми на нужный раздел для перехода

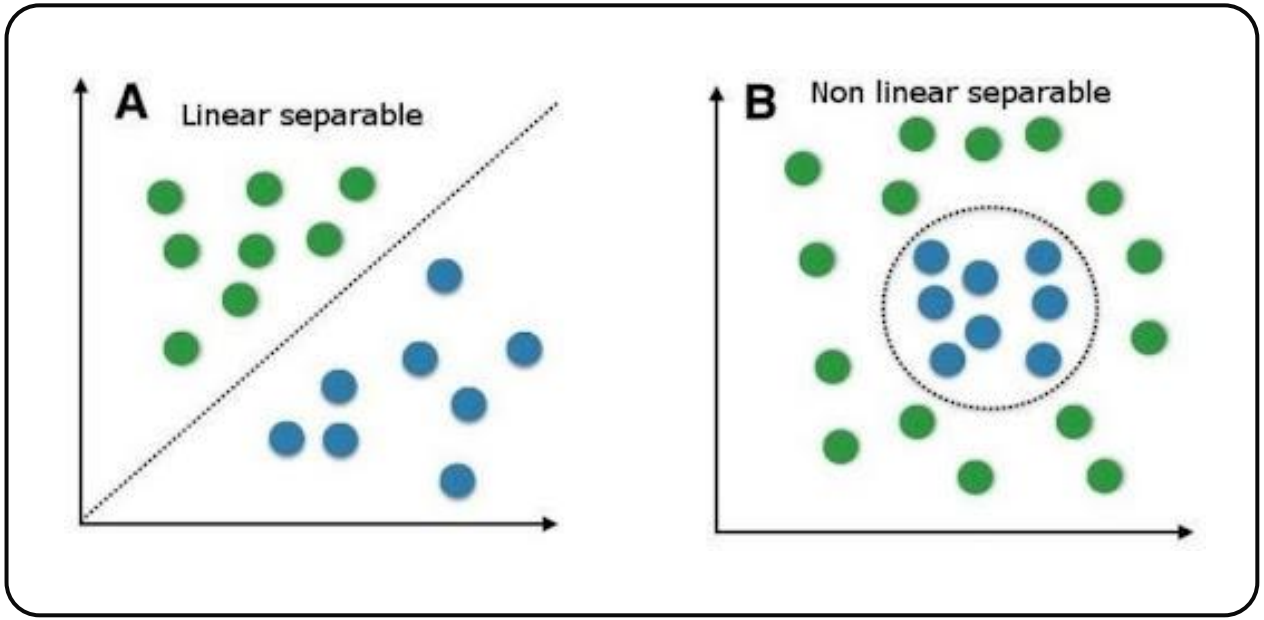


Задача классификации

Классификации — задача предсказания ответа из конечного множества вариантов

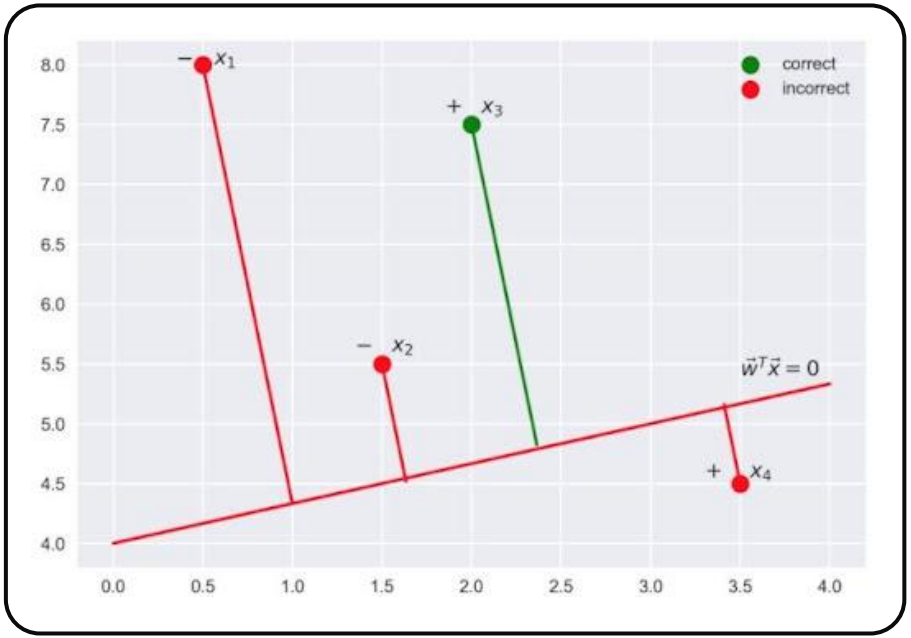
Линейный классификатор решает задачу разделения признакового пространства на две части, в каждом из которых находится свой класс

Линейная разделимость данных



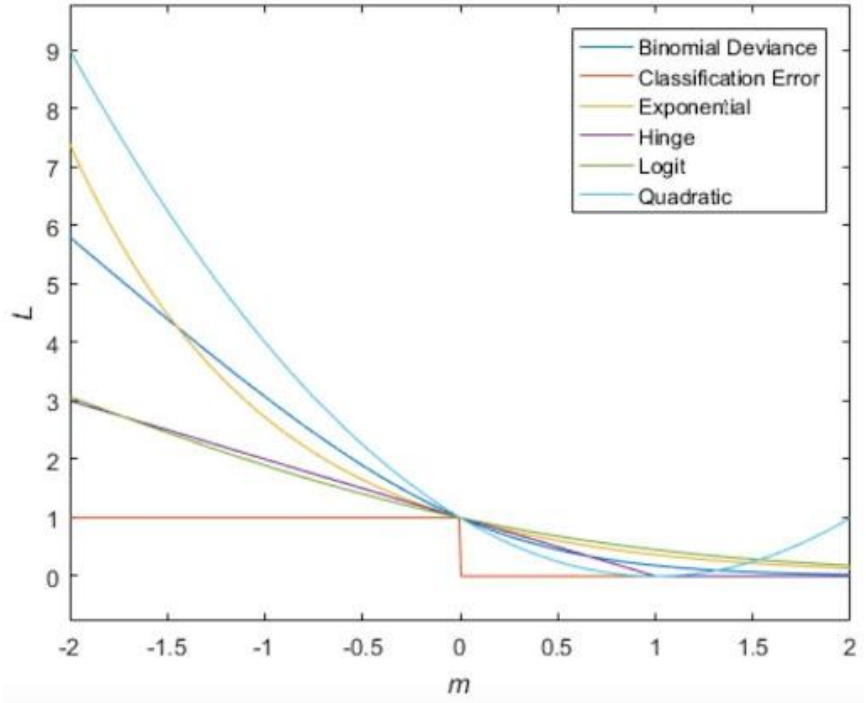
Отступ — расстояние от разделяющей гиперповерхности до объекта

Отступ можно понимать как степень погружённости объекта в свой класс



$$M(\overrightarrow{x_i}) = y_i \vec{w}^T \overrightarrow{x_i}$$

Функция потерь L как функция от отступа M



$$L = f(M)$$

Итоги раздела

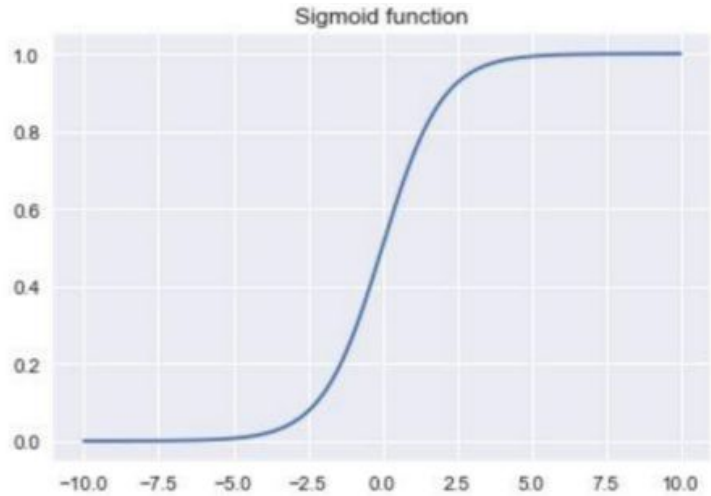
1. Рассмотрели линейную классификацию
2. Обсудили функции потерь для линейной классификации
3. Усвоили понятие отступа для классификации

Логистическая регрессия

2

Логистическая регрессия

Линейный классификатор, позволяющий оценивать вероятности принадлежности объектов классам



$$L = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

$$p = \frac{1}{1 + e^{-L}}$$

Функция потерь

$$p(y_i \mid x_i, w) = a_i^{y_i}(1 - a_i)^{1-y_i}$$

Модель предсказывает вероятность классов {0, +1}

$$p(y \mid X, w) = \prod_i p(y_i \mid x_i, w)$$

Максимизировать правдоподобие

$$\mathcal{L}_{log}(X, \vec{y}, \vec{w}) = \sum_i (-y_i \log a_i - (1 - y_i)\log(1 - a_i))$$

Функция потерь

Функция потерь

$$P\left(y = y_i \mid \overrightarrow{x_i}, \vec{w}\right) = \sigma(y_i \vec{w}^T \overrightarrow{x_i})$$

Модель предсказывает вероятность классов {-1, +1}

$$P\left(\vec{y} \mid X, \vec{w}\right) = \prod_{i=1}^{\ell} P\left(y = y_i \mid \overrightarrow{x_i}, \vec{w}\right)$$

$$\log P\left(\vec{y} \mid X, \vec{w}\right) = -\sum_{i=1}^{\ell} \log(1 + \exp^{-y_i \vec{w}^T \overrightarrow{x_i}})$$

Максимизировать правдоподобие

$$\mathcal{L}_{log}(X, \vec{y}, \vec{w}) = \sum_{i=1}^{\ell} \log(1 + \exp^{-y_i \vec{w}^T \overrightarrow{x_i}})$$

Функция потерь

Итоги раздела

1. Рассмотрели метод построения логистической регрессии
2. Познакомились с функцией потерь logloss

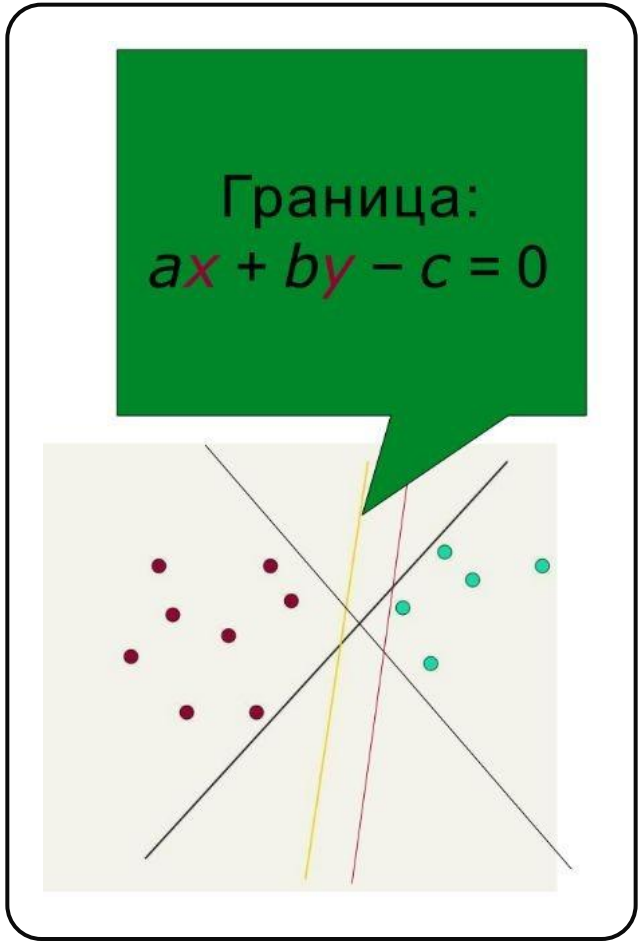
SVM

3

Множество гиперплоскостей

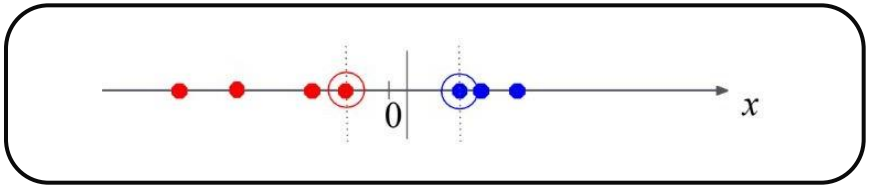
SVM находит оптимальную разделяющую поверхность

Максимизирует «зазор»

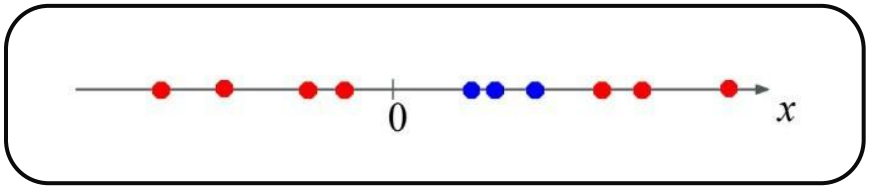


Non-linear SVMs

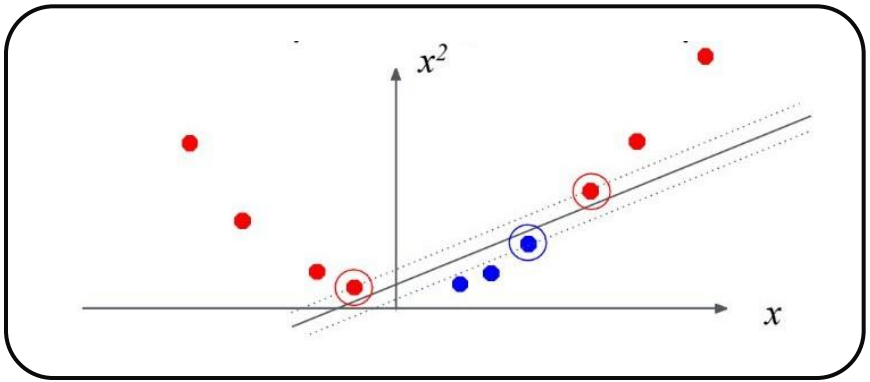
Линейно разделимые датасеты хорошо классифицируются.



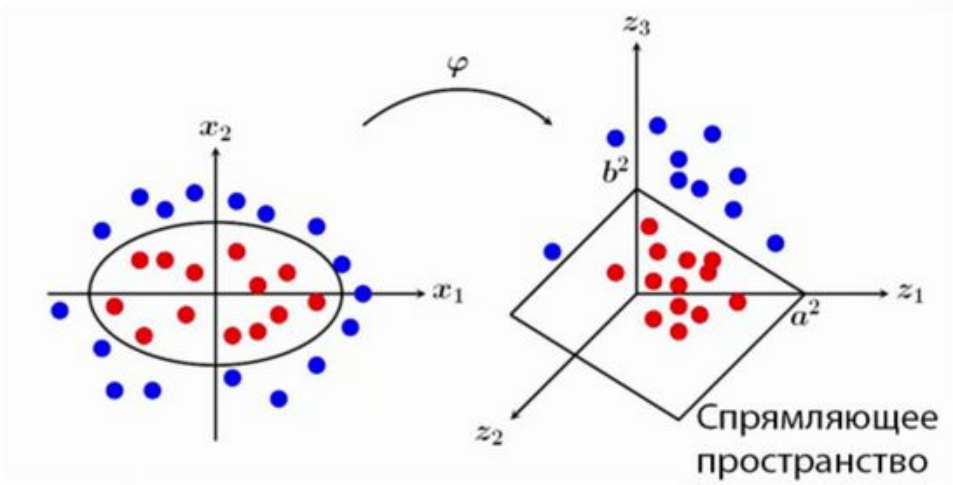
Но что делать, если они не линейно разделимы?



Можно попробовать отобразить данные в пространство более высокой размерности



The «Kernel Trick»



Kernels

- Полиномиальное

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}')^d$$

- Полиномиальное со смещением

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^d$$

Kernels

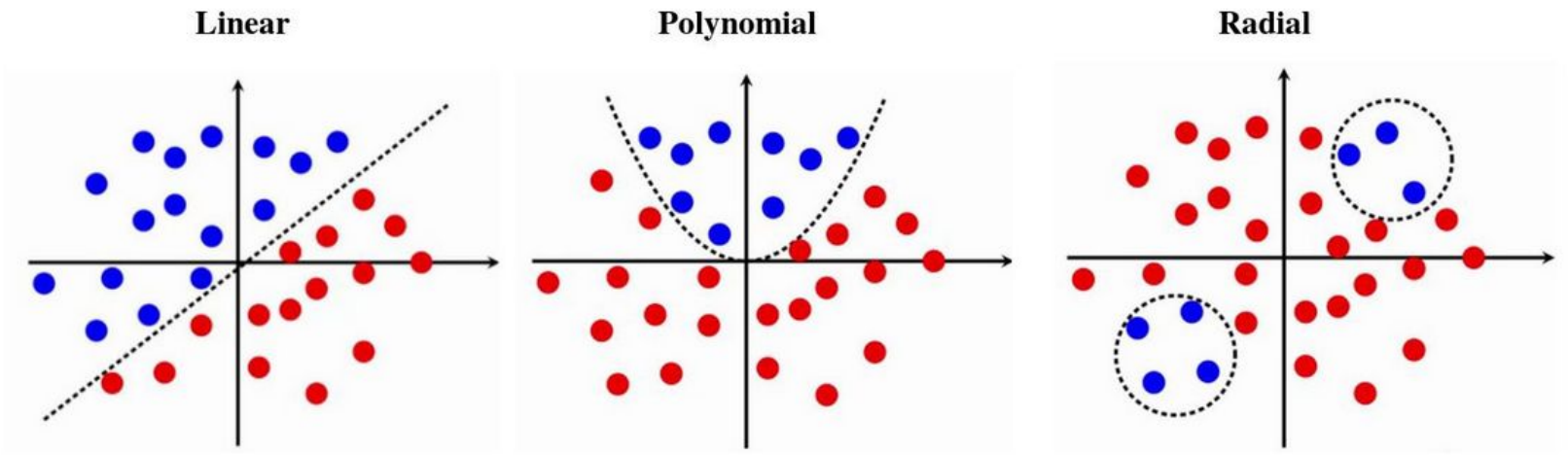
- Радиальная базисная функция

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2), \text{ для } \gamma > 0$$

- Радиальная базисная функция Гаусса

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

Kernels



Итоги раздела

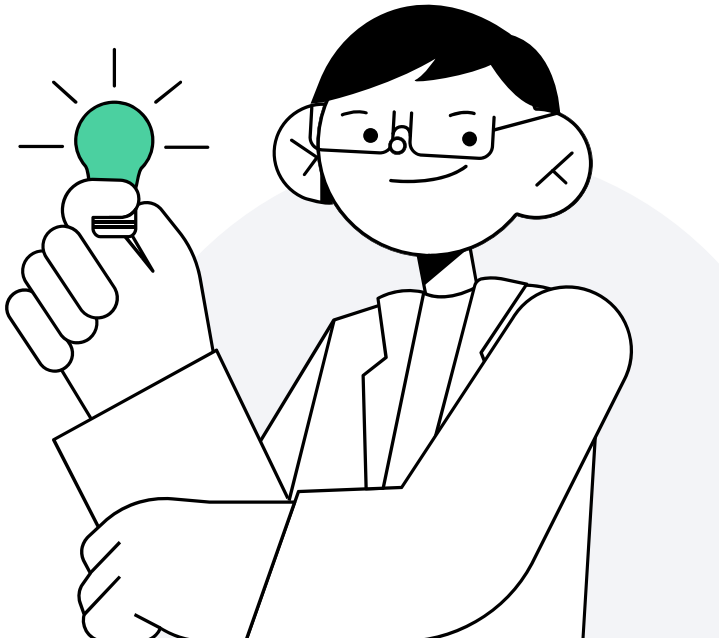
1. Познакомились с моделью Support Vector Machines (SVM)
2. Рассмотрели способы перевода линейно неразделимой выборки в линейно разделимую

Практика

4

Вы узнаете

- Как обучить логистическую регрессию через Sklearn
- Как оценивать качество модели для задачи классификации
- Как визуализировать разделяющие плоскости моделей
- Как менять ядра для модели SVM



Задача

Необходимо построить логистическую регрессию для предсказания пола спортсменов и классов цветов ириса

Необходимо:

- загрузить набор данных
- выбрать необходимые характеристики
- построить логистическую регрессию
- построить SVM
- визуализировать разделяющие плоскости классификатора
- оценить качество работы модели

Итоги занятия

1. Рассмотрели задачу классификации
2. Познакомились с линейным классификатором
3. Узнали, как строится логистическая регрессия
4. Усвоили метод построения модели SVM
5. Решили задачи предсказания пола спортсмена и класса цветов

Дополнительные материалы

- Линейные модели классификации и регрессии
- Курс «Основы статистики» на Stepik
- Пережёвывая логистическую регрессию
- Логистическая регрессия
- Реализация логистической регрессии

Классификация: логистическая регрессия и SVM