GeekBrains

Классификация и использование логистической регрессии в задачах классификации

Урок 4
На этой лекции вы найдете ответы на такие вопросы как:

- Какие задачи решает классификация
- Что такое логистическая регрессия
- Метрики качества классификации

Булгакова Татьяна

Преподаватель в GeekBrains, Нетология,Skillfactory

С 2010 года занимаюсь DataScience и NN. Фрилансер

- Участвовала в разработке программы по настройке оборудования для исследования пространственного слуха китообразных НИИ ИПЭЭ РАН
- Участвую в разработке рекомендательных систем по настройке нейростимуляторов для медицинских центров
- Работаю над курсом по нейронным сетям

План курса

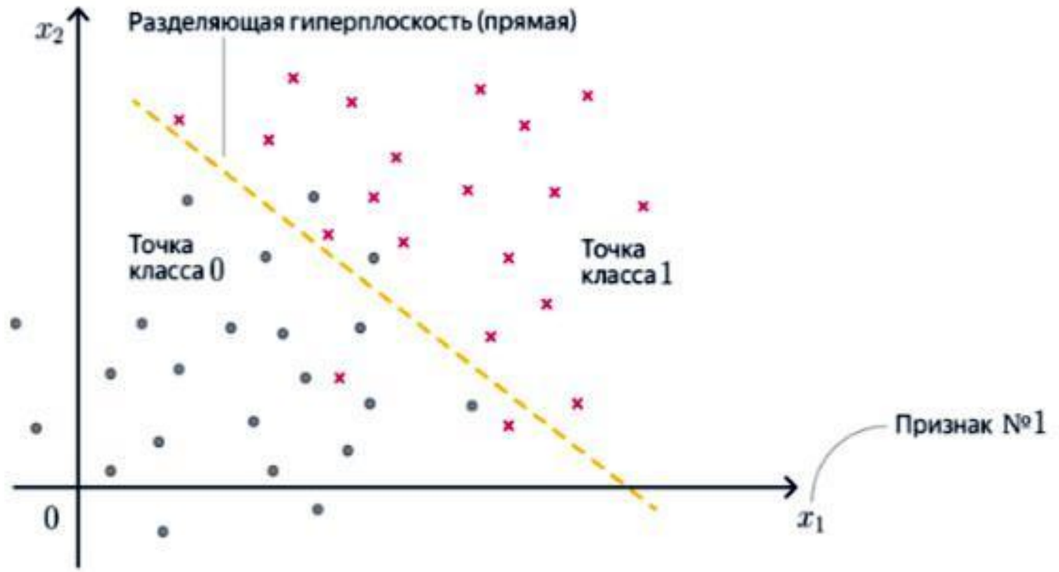
1. Первичный и визуальный анализ данных
2. Описательные статистики в контексте EDA. Корреляция и корреляционный анализ
3. Регрессия и использование библиотеки Scikit-learn в задачах обучения с учителем
4. Классификация и использование логистической регрессии в задачах классификации
5. Функционалы ошибки и поиск оптимальных параметров в задачах машинного обучения
6. Проблема переобучения и недообучения модели. Кросс-валидация и регуляризация.
7. Ансамблирование и использование деревьев решений в задачах машинного обучения
8. Генерация признаков. Методы отбора признаков. Подбор гиперпараметров.
9. Обучение без учителя. Понижение размерности. Алгоритмы понижения размерности
10. Кластеризация и решение задачи группировки данных в машинном обучении

Что будет на уроке сегодня

- 📌 Какие задачи решает классификация
- 📌 Что такое многоклассовая классификация
- 📌 Что такое линейная регрессия
- 📌 Метрики задачи классификации

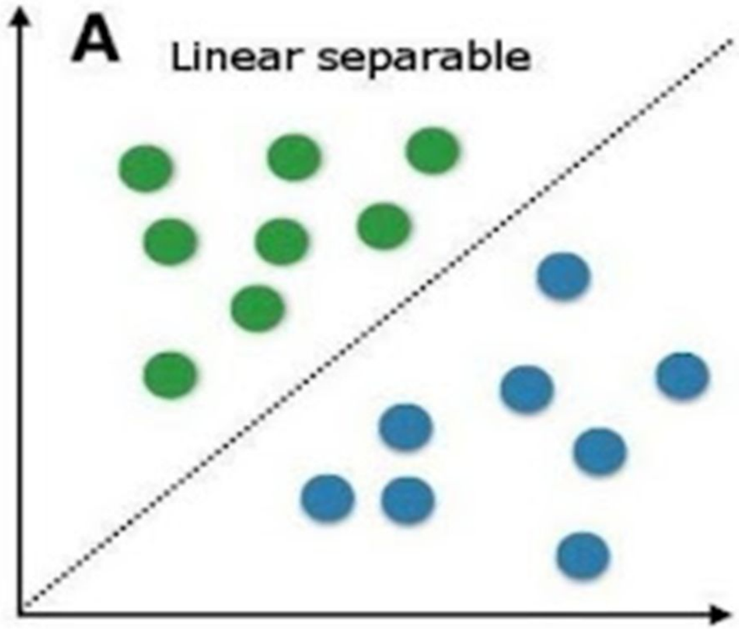
Классификация

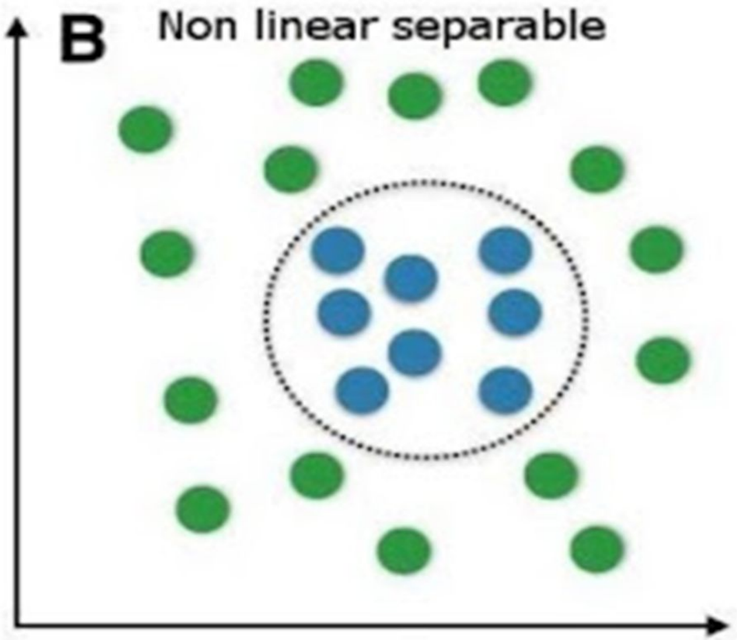
Задача классификации – это задача обучения с учителем (supervised learning), при которой модель должна классифицировать объекты



Классификация

Разделимость означает, что объекты могут быть разделены на два или более класса таким образом, чтобы объекты внутри каждого класса были похожи между собой, но отличались от объектов в другом классе.





Разделимость данных

Для определения разделимости данных часто используются статистические методы.

- Метод главных компонент (PCA Principal Component Analysis)
- Метод опорных векторов (SVM Support Vector Machine)
- Алгоритм кластеризации
- Алгоритмы машинного обучения
- Визуализация данных

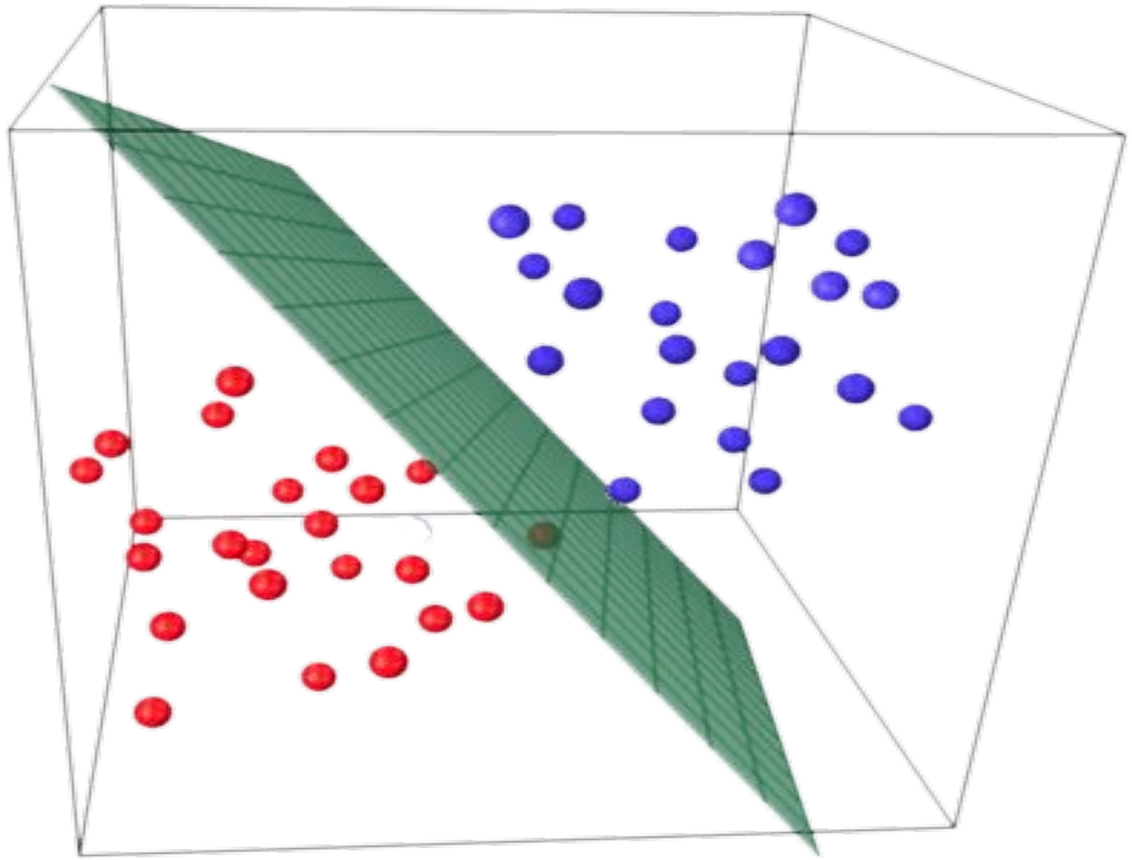
Разделимость данных

👉 Разделимость данных позволяет определить возможность классификации объектов в заданных наборах данных.

👉 Для определения разделимости данных нужно использовать различные методы

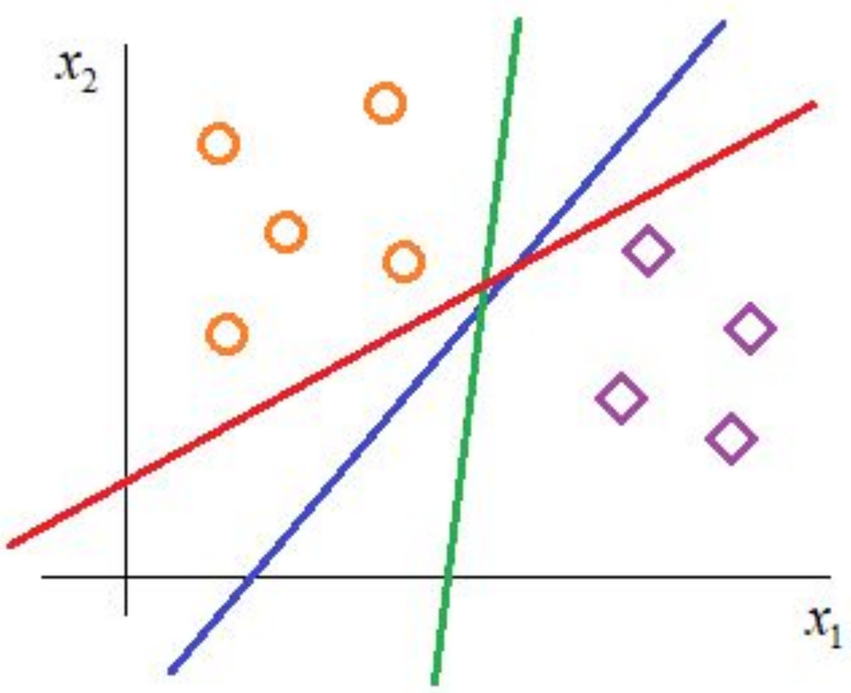
👉 Выбор метода будет зависеть от структуры и количества данных и требуемой точности решения.

Пример реализации линейной классификации



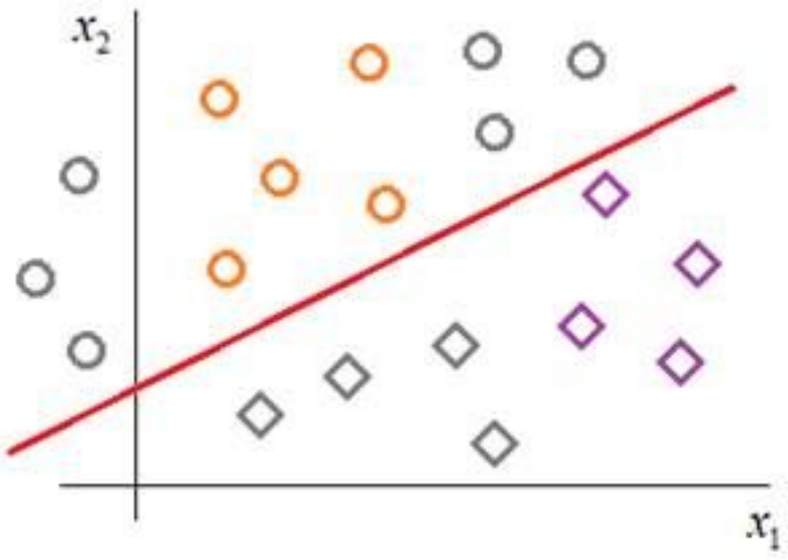
Пример реализации линейной классификации

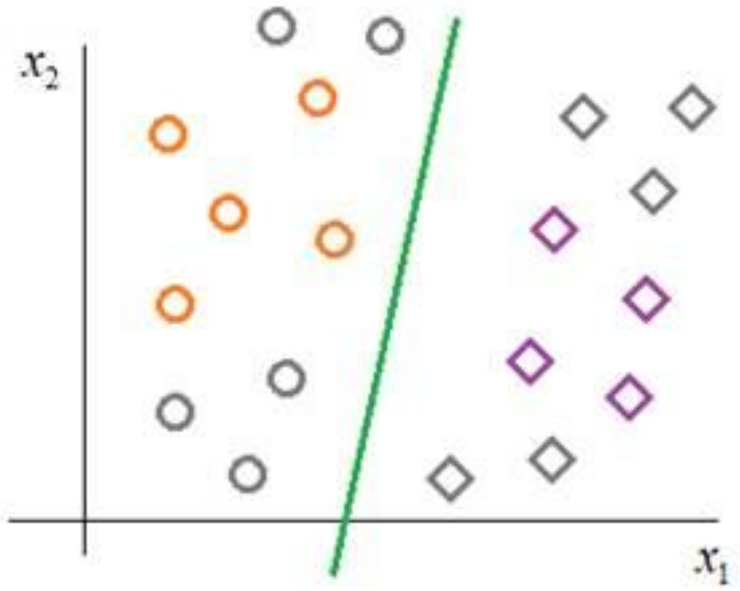
Для простоты предположим двумерное пространство признаков. Каждый объект, принадлежащее классу, может быть представлено как точка на плоскости.



Пример реализации линейной классификации

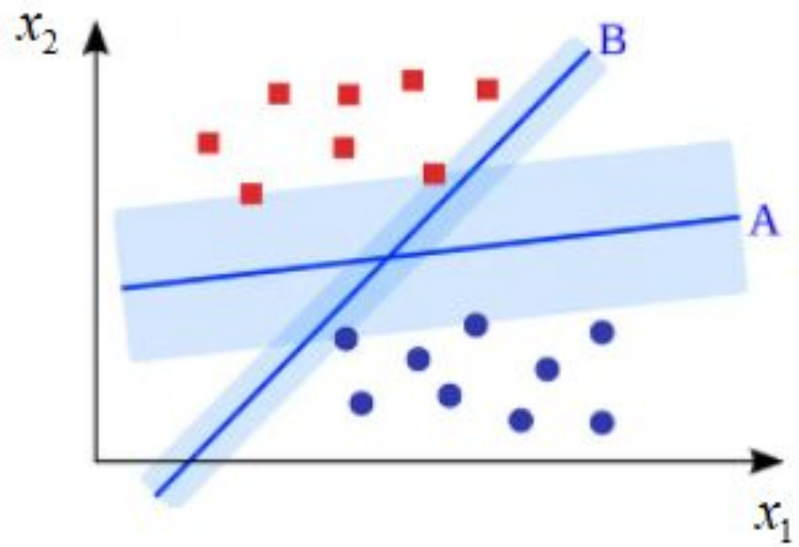
Возникает вопрос, какое разделение лучше. В машинном обучении предполагается, что модель, обученная на выборке, должна хорошо работать на любом другом наборе из того же распределения.





Модель опорных векторов - SVM

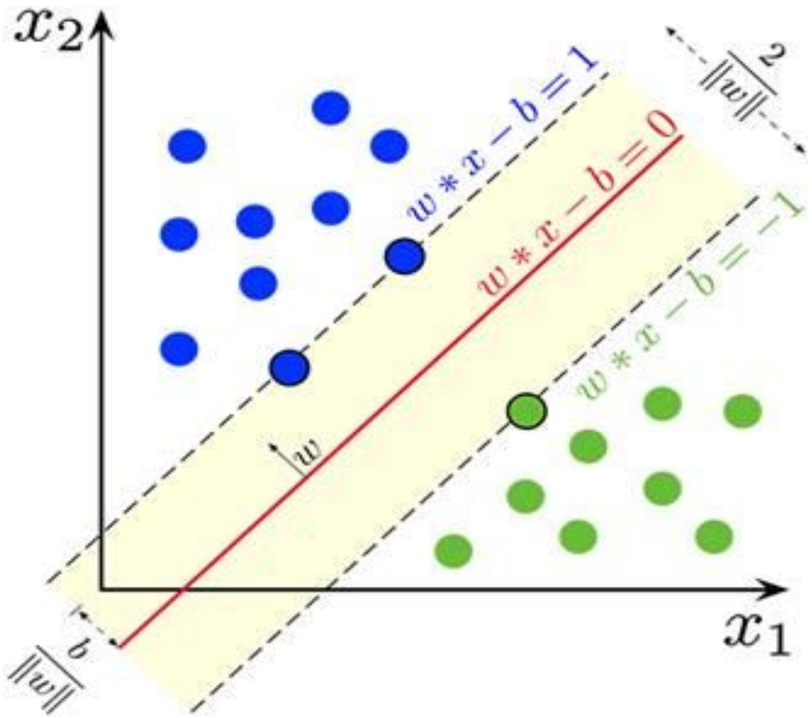
С точки зрения SVM, оптимальная разделительная гиперплоскость - это гиперплоскость, которая образует самую широкую полосу между двумя классами объектов. Сама разделительная гиперплоскость располагается в центре этой полосы

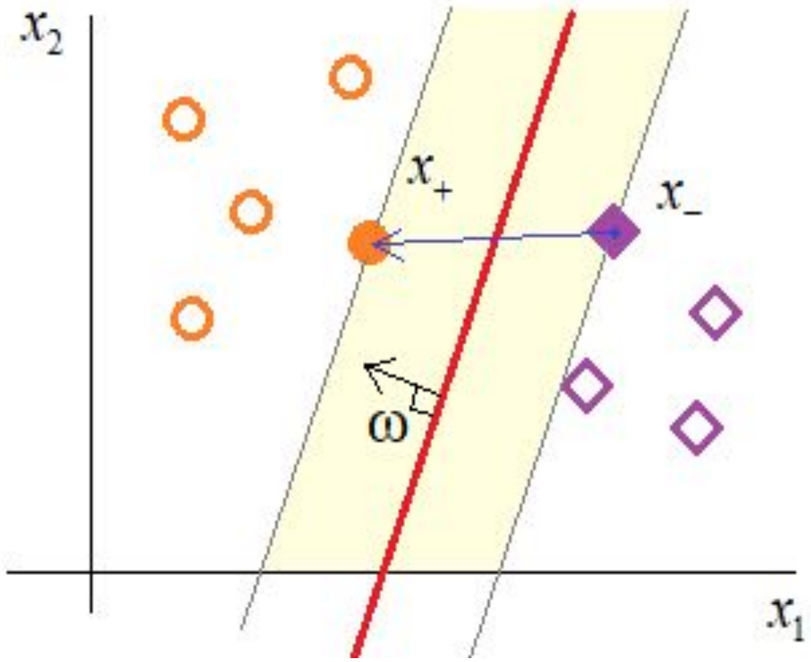


Модель опорных векторов - SVM

$$a(x) = sign(w^T x - b) = sign(\langle w, x \rangle - b) = sign(w * x - b)$$

$$a(x) \in \{-1; +1\}$$





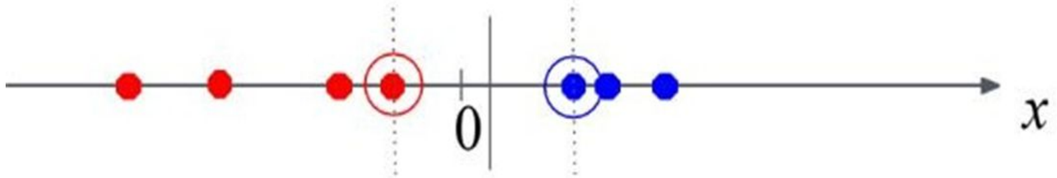
Модель опорных векторов - SVM

В итоге, модель SVM может быть использована для классификации новых точек данных на основе изученных свойств гиперплоскости и положения этих точек относительно нее.

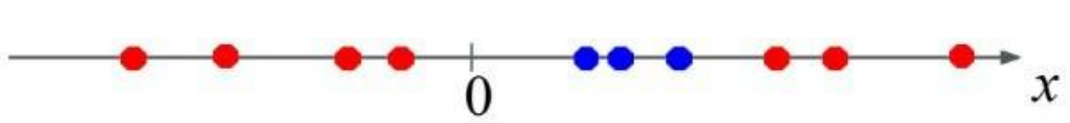
Ядро распределения или еще раз о разделимости данных

Ядро распределения данных в машинном обучении — это функция, которая измеряет сходство между двумя данными или объектами в пространстве признаков

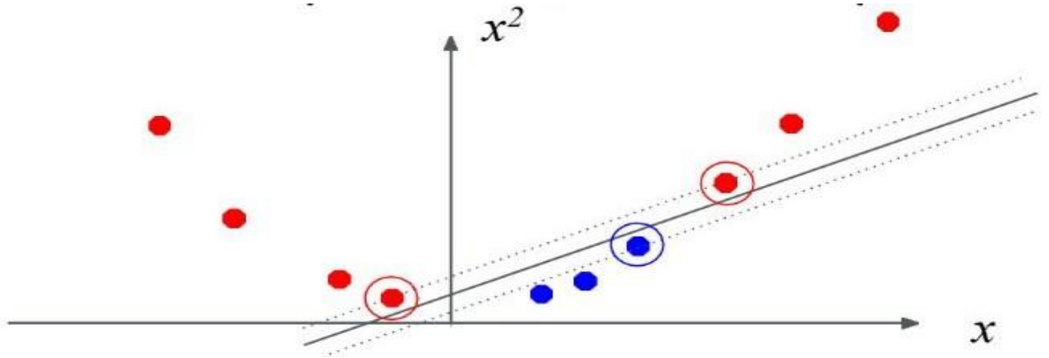
Линейно разделимые датасеты хорошо классифицируются



Но что делать, если они не линейно разделимы?



Можно попробовать отобразить данные в пространство более высокой размерности



Ядро распределения или еще раз о разделимости данных

Ядерная функция принимает на вход два вектора и возвращает их скалярное произведение в пространстве более высокой размерности.

SVM зависит от скалярного произведения

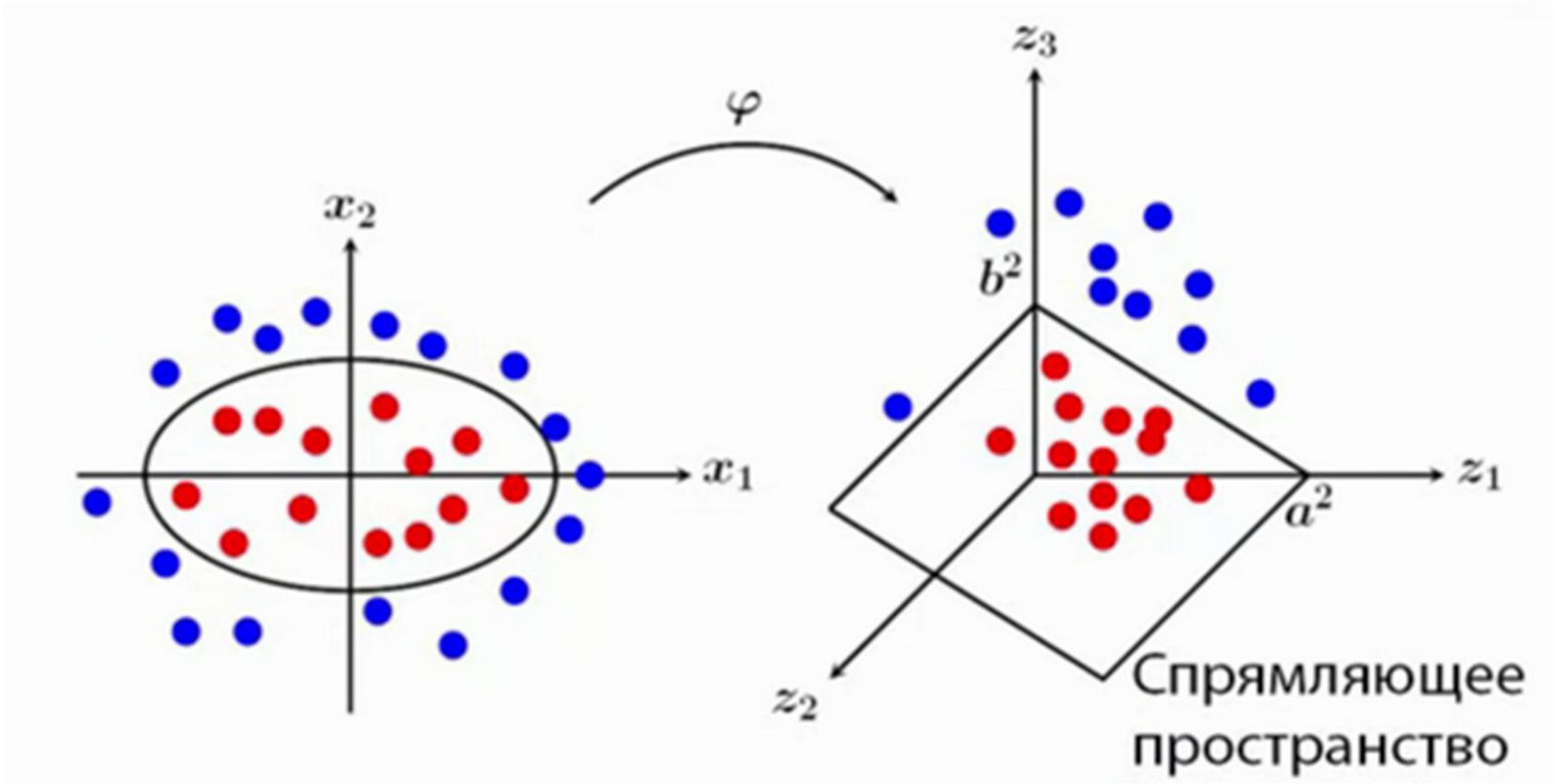
$$K(x_i, x_j) = x_i^T x_j$$

Если каждая точка отображается в пространство более высокой размерности при помощи Ф: x → φ(x), тогда скалярное произведение становится:

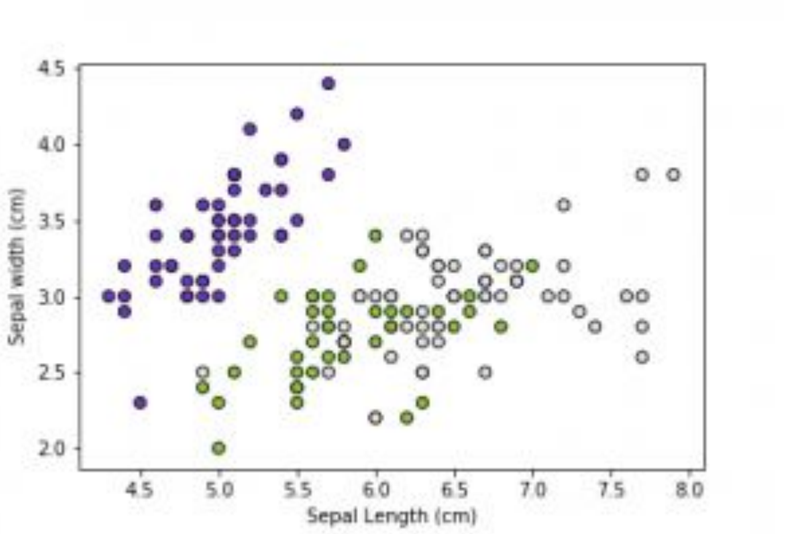
$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

Функция ядра - это функция, соответствующая скалярному произведению в производстве более высокой размерности

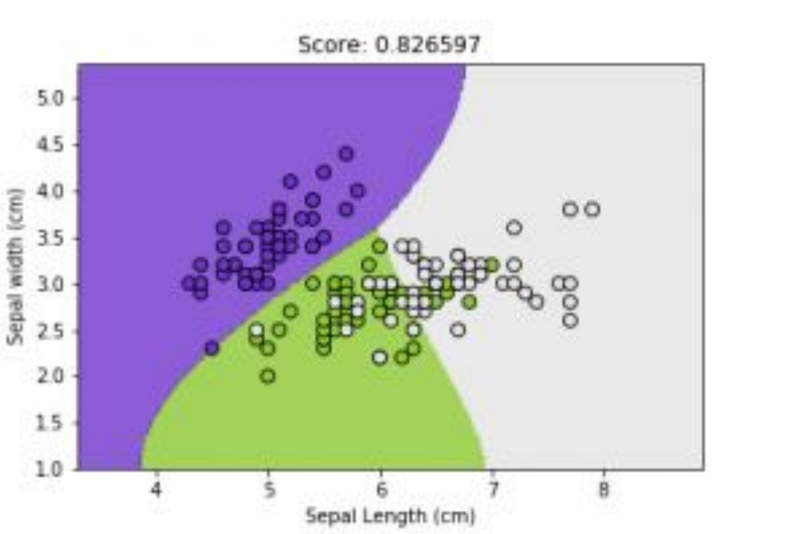
Ядро распределения или еще раз о разделимости данных



Ядро распределения или еще раз о разделимости данных

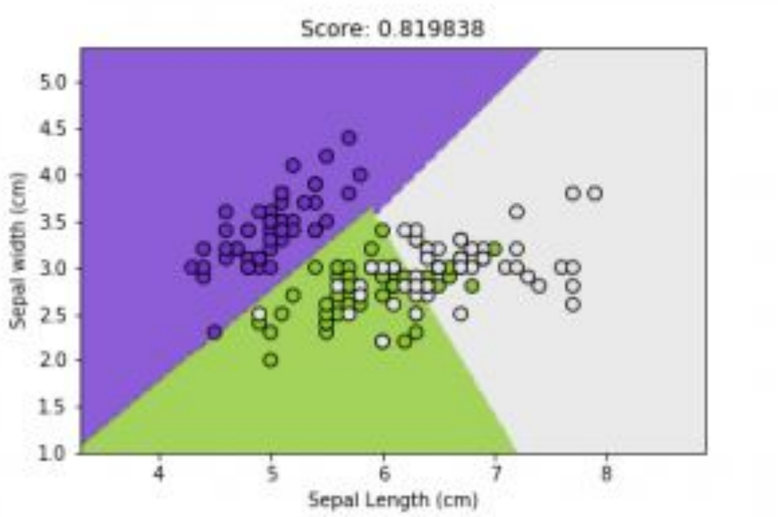


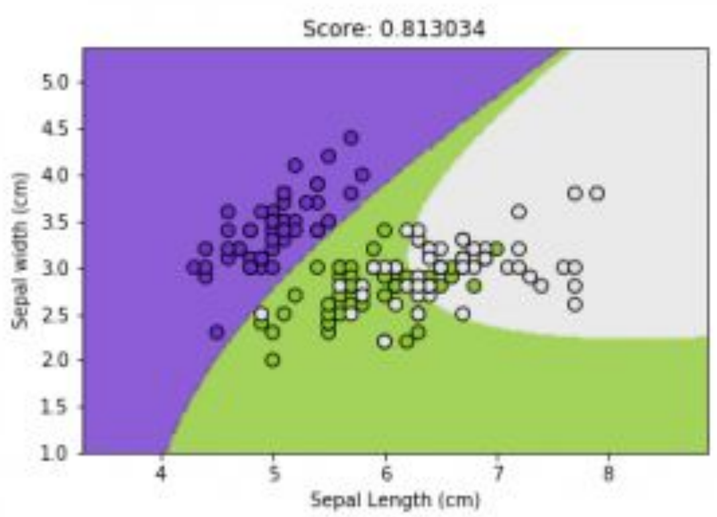
1. Линейное ядро



2. Полиномиальное ядро

Ядро распределения или еще раз о разделимости данных





3. Радиальная базисная функция (RBF) ядро

$$K(x, y) = \exp(-\gamma \| x - y \|^2)$$

4. Сигмоидное ядро

$$K(x, y) = \tanh(\alpha x \cdot y + c)$$

Логистическая регрессия

Пока нам неизвестно, как обучается модель линейной классификации, но уже ясно, что окончательные прогнозы можно рассчитать с помощью следующей формулы:

$$y = sign \langle w, x_i \rangle$$

Задание: Можно ли решить данную задачу, как задачу регрессии?

Логистическая регрессия

Ответ:

Если мы попробуем предсказать числа -1 и 1, минимизируя MSE, учитывая знак, но результаты получатся плохими. Во-первых, регрессия дает очень маленькую ошибку для объектов, которые находятся близко к *плоскости разделения*, но не с той стороны. Во-вторых, предсказание, например, 5 вместо 1 - это ошибка. Однако, если знак правильный, то модуль числа не имеет значения.

То есть, нам нужна прямая, которая разделяет эти точки, а не проходит через них!

Логистическая регрессия

Это линейный классификатор, позволяющий оценивать вероятности принадлежности объектов классам

Проблема в том, что вероятность по определению является значением между 0 и 1, и нет простого способа обучить линейную модель, чтобы она удовлетворяла этому ограничению.

Способ преодоления этой проблемы заключается в обучении линейной модели правильному прогнозированию объектов, связанных с вероятностями в диапазоне (-∞, ∞)

Логистическая регрессия

Затем ответ модели преобразуется в вероятность. Одним из таких объектов является logit

$$\langle w, x_i \rangle = \log\left(\frac{p}{1-p}\right)$$

$$e^{\langle w, x_i \rangle} = \frac{p}{1-p}$$

$$p = \frac{1}{1+e^{-\langle w, x_i \rangle}}$$

Функция в правой части называется **сигмоидой** и обозначается

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Таким образом, $p = \sigma(\langle w, x_i \rangle)$

Логистическая регрессия

Применим метод максимума правдоподобия. Правдоподобие позволяет понять, насколько вероятно получить данные значения таргета *y* при данных *X* и весах *w*.

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

Оптимизировать произведение неудобно

$$\ell(w, X, y) = \sum_i \big(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\big) = \sum_i \big(y_i \log(\sigma(\langle w, x_i \rangle)) + (1 - y_i) \log(1 - \sigma(\langle w, x_i \rangle))\big)$$

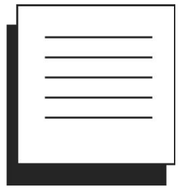
Логистическая регрессия

Учитывая, что $$\sigma(-z) = \frac{1}{1+e^{z}} = \frac{e^{-z}}{e^{-z}+1} = 1 - \sigma(z),$$

Нас интересует w, которое максимизирует вероятность. Умножим на минус один, чтобы получить минимизирующую функцию потерь:

$$L(w, X, y) = -\sum_{i} \left(y_i \log(\sigma(\langle w, x_i \rangle)) + (1 - y_i) \log(\sigma(-\langle w, x_i \rangle)) \right)$$

Логистическая регрессия



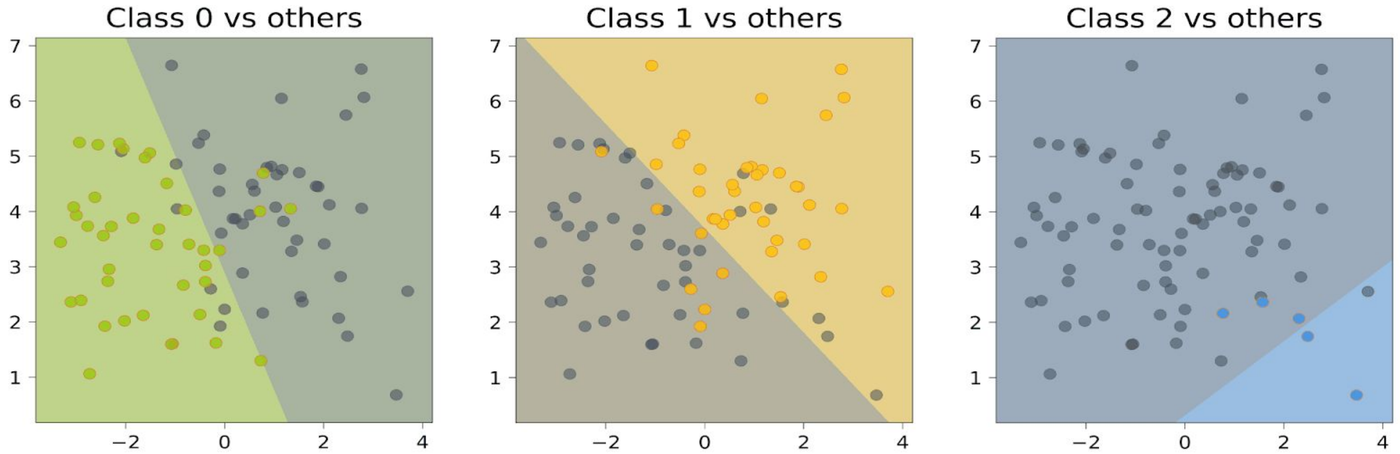
Это вероятность того, что класс положительный, и как нам перейти от этого к оценке самого класса?
Все предсказания положительны и находятся в диапазоне от 0 до 1.
Интуитивно, хотя и не совсем точно, можно ответить так: "Возьмите порог 0,5".
Более точным является индивидуальный выбор этого порога для уже построенных регрессий, которые минимизируют требуемые метрики на отложенной тестовой выборке.
Например, доли положительных и отрицательных классов должны примерно соответствовать истинным долям.

Многоклассовая классификация

Один против всех (one-versus-all)

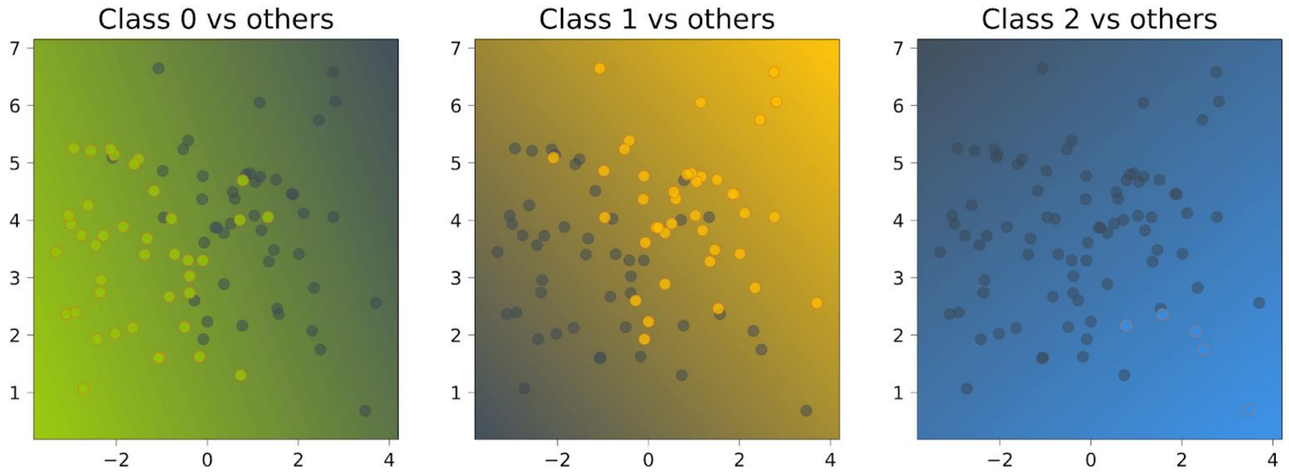
$$b_k(x) = \text{sgn}\left(\langle w_k, x\rangle + w_{0k}\right)$$

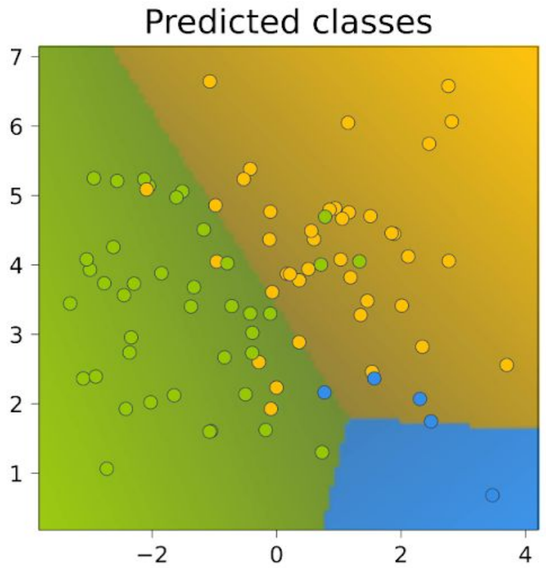
$$a(x) = \text{argmax}_k\left(\langle w_k, x\rangle + w_{0k}\right)$$



Многоклассовая классификация

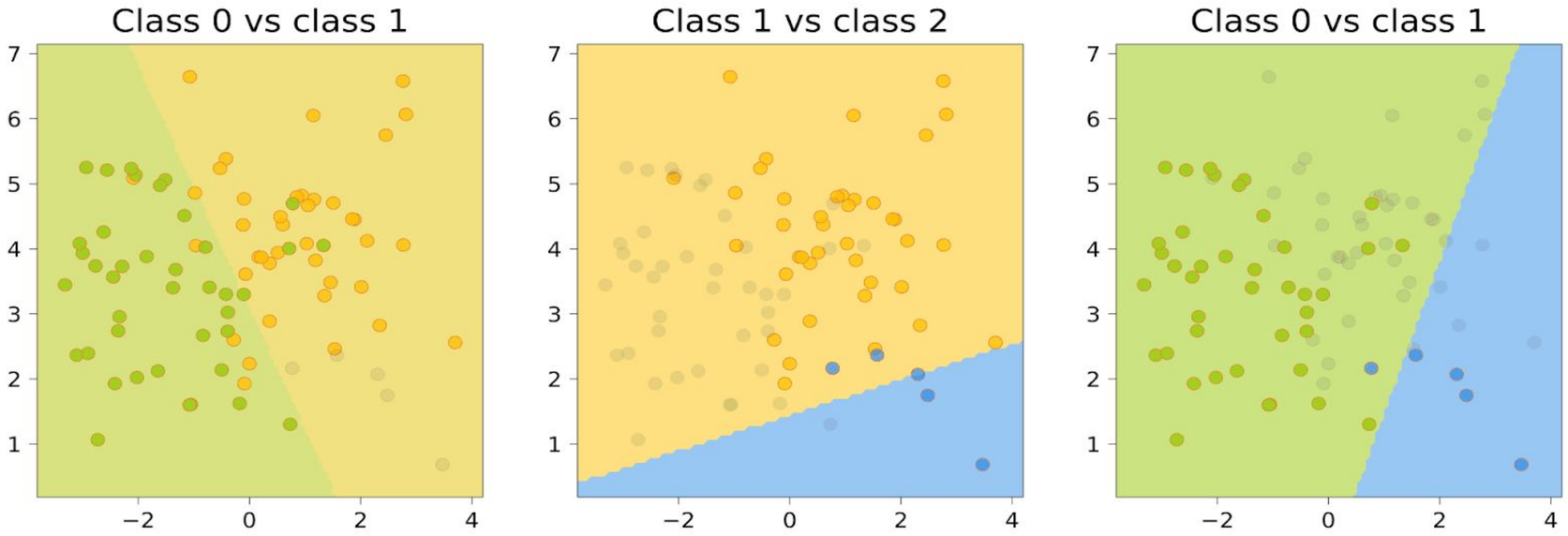
Один против всех (one-versus-all)





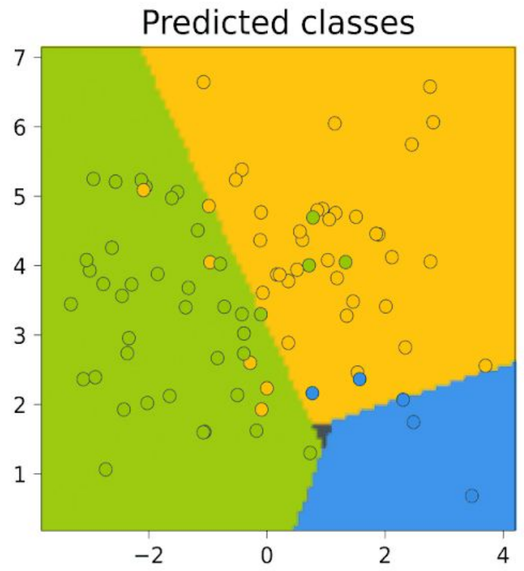
Многоклассовая классификация

Все против всех (all-versus-all)



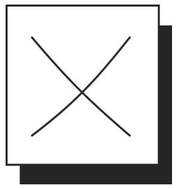
Многоклассовая классификация

Все против всех (all-versus-all)



Accuracy доля объектов, для которых мы правильно предсказали класс

$$\text{Accuracy}(y, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[y_i = f(x_i)]$$

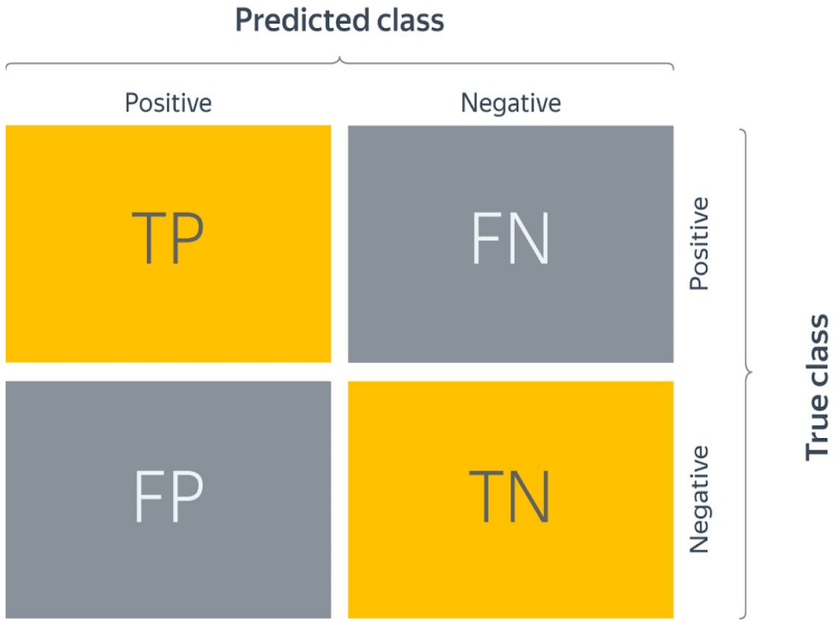


Не учитывает дисбаланс классов

Не учитывается стоимость ошибок для различных классов объектов

Confusion matrix (матрица ошибок)

Исторически проблема бинарной классификации - это проблема обнаружения необычных объектов в большом потоке объектов



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Точность (precision)

Учет доли правильно предсказанных положительных объектов среди всех объектов, предсказанных положительным классом

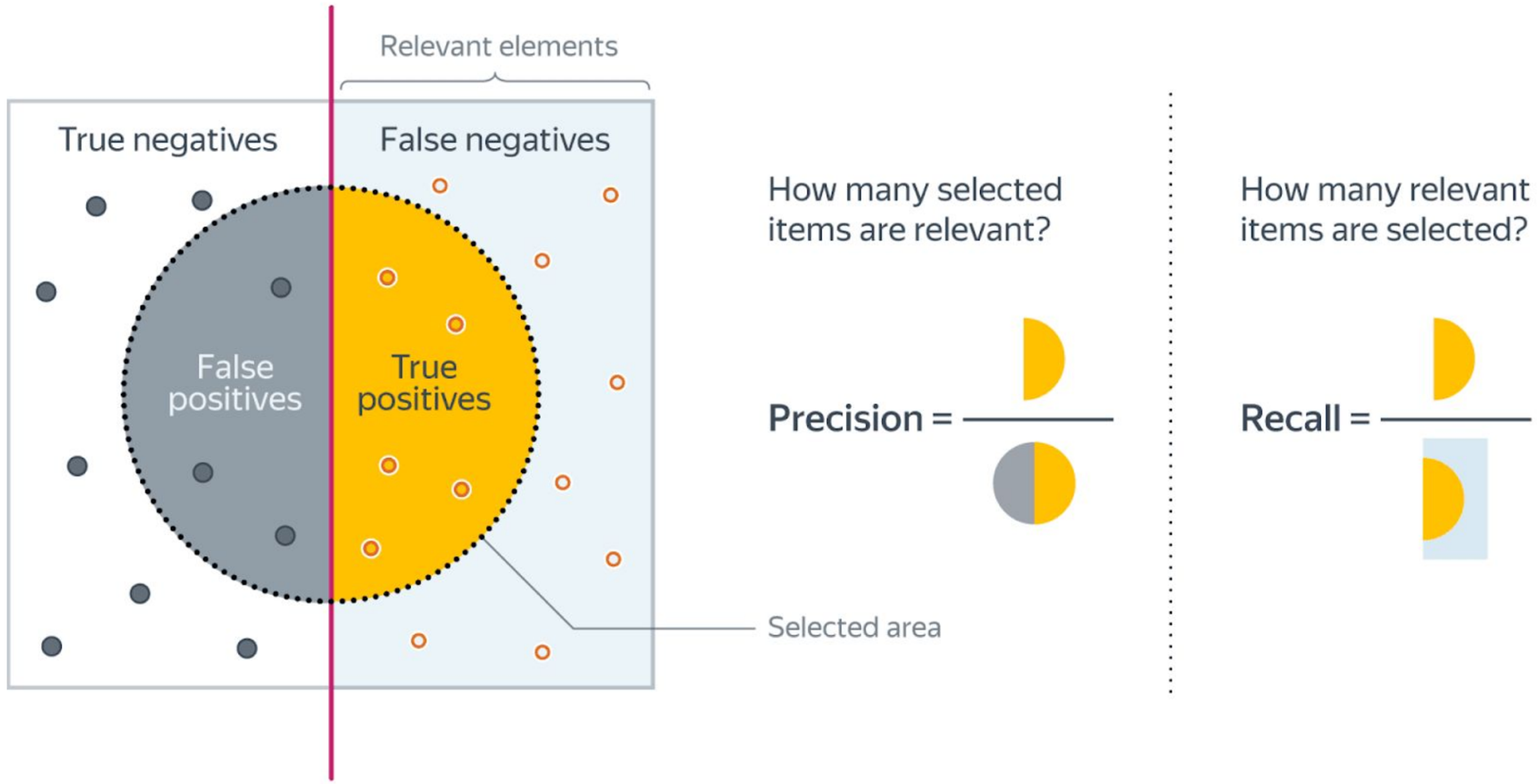
$$\text{Precision} = \frac{TP}{TP + FP}$$

Полнота (recall)

Учет доли правильно найденных положительных элементов среди всех элементов положительного класса

$$\text{Recall} = \frac{TP}{TP + FN}$$

Точность (precision) и Полнота (recall)



F-мера

среднее гармоническое Precision и Recall

$$F_\beta = (\beta^2 + 1)\frac{Recall \cdot Precision}{Recall + \beta^2 Precision}$$

F-мера объединяет точность и полноту в одну метрику, позволяя оценить их совместно. Коэффициент бетта в F-мере, обозначаемый как β, контролирует баланс между точностью и полнотой. Значение β определяет вес точности в соотношении к полноте.

В частности, когда β = 1, F-мера является сбалансированной метрикой, которая учитывает и точность, и полноту одинаково. Коэффициенты β < 1 делают F-меру более взвешенной в пользу точности, в то время как β > 1 делает ее более взвешенной в пользу полноты.

F-мера

F1-мерой частный случай F-меры при β = 1

$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} = 2\frac{Recall \cdot Precision}{Recall + Precision}$$

Выводы

1. Классификация - это задача отнесения объектов к заранее определенным классам на основе имеющихся данных

2. Модели SVM (Support Vector Machines) и логистической регрессии - две популярные модели для решения задач классификации.

3. Многоклассовая классификация - это задача классификации, в которой объекты могут быть отнесены к одному из нескольких классов.

4. Метрики качества классификации - это меры, используемые для оценки эффективности модели классификации.

5. Информация о задачах классификации полезна для понимания процесса решения задачи и выбора наиболее подходящей модели.

6. В результате лекции мы получили навыки по выбору модели классификации, обработке и анализу данных, а также понимание метрик качества классификации и их применения для оценки результатов моделей.

Итоги лекции

1. Умение выбирать подходящие методы и алгоритмы классификации: понимание задачи поможет определить, какие методы и алгоритмы классификации лучше всего подходят для решения данной задачи.

2. Навыки предобработки данных: понимание задачи помогает определить, какие признаки могут быть полезны для классификации и как обрабатывать или очищать данные, чтобы улучшить процесс классификации.

3. Умение работы с несбалансированными данными: задачи классификации могут столкнуться с проблемой несбалансированных данных, где классы имеют различное количество образцов.

4. Навыки оценки модели: понимание задачи поможет в выборе подходящих метрик оценки качества модели классификации.

Спасибо за внимание