Ближайшие соседи: чей это цветок?

У цветка потерялась этикетка. Рядом есть измерения знакомых ирисов. Можно ли восстановить вид по похожим соседям?

На столе лежит ирис без этикетки. В справочнике есть измерения других цветов и названия их видов. Самое доступное решение: измерить неизвестный цветок и поискать похожие записи. Если у похожего цветка подписан вид, можно предположить такой же у нашего.

Звучит разумно, пока два похожих цветка не оказываются разных видов. Кому верить? Разберём способ, который позволяет проследить каждый голос. Для этого достаточно читать таблицу, сравнивать длины и считать до пяти. Все вычисления на карте выполнит интерактив.

Ранее

Классификация выбирает категорию. В уроке о спам-фильтре мы сравнивали один признак с порогом. Здесь сравним новый объект с уже известными примерами.

← Классификация: как фильтр узнаёт спам

1. Как цветок становится точкой

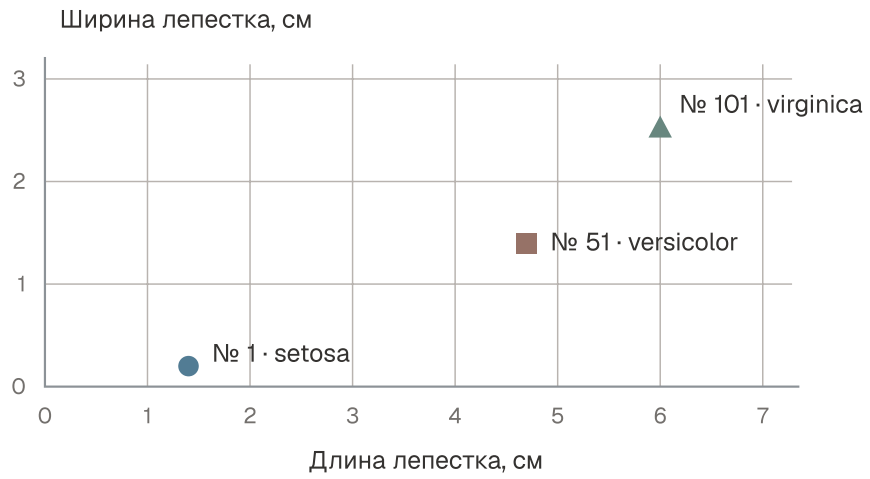
Возьмём два **признака**: длину лепестка и его ширину. Это две измеримые характеристики одного цветка. В таблице каждая строка описывает отдельный экземпляр. Последний столбец хранит известный вид: эти подписи понадобятся для голосования. У проверочного цветка вид сначала скроем, а после прогноза откроем для сравнения.

Ссылки

Набор Iris в UCI содержит измерения 150 ирисов трёх видов. Здесь используем длину и ширину лепестка из файла `bezdekIris.data`; исходные единицы — сантиметры.

Цветок в исходном файле	Длина, см	Ширина, см	Вид
№ 1	1,4	0,2	setosa
№ 51	4,7	1,4	versicolor
№ 101	6,0	2,5	virginica

На карте длина растёт вправо, ширина вверх. Цветок с длинным узким лепестком окажется справа внизу. Точка обозначает весь цветок; она не показывает форму его лепестка. Вид обозначим цветом и формой значка, чтобы различать группы даже без цвета.



Каждая строка превращается в одну точку. Положение кодирует два измерения; значок показывает известный вид. Масштаб одинаков: один сантиметр по обеим осям занимает одну клетку.

Карта сохраняет только выбранные измерения. Запах, цвет и место, где рос ирис, в этот пример не попали. Два цветка могут оказаться рядом на карте, хотя выглядят по-разному. Сходство всегда относится к тем признакам, которые мы решили сравнивать.

Пример
В таблице квартир соседями могут стать квартиры в разных городах, если сравнивать только площадь и число комнат. Модель не узнает про город, пока мы не включим его в описание.

> На карте цветов длина лепестка растёт вправо, ширина — вверх. У нового цветка лепесток длиннее, но уже, чем у отмеченного. Куда относительно отмеченной точки поставить новую?
>
> ○ Вправо и вверх
> ○ В то же место: вид ещё неизвестен
> ○ Влево и вверх
> ○ Вправо и вниз

2. Один сосед даёт один ответ

Теперь добавим неизвестный цветок. Найдём ближайшую точку справочника и возьмём её вид как прогноз. **Метод ближайшего соседа** (*nearest neighbor*) использует близость как правило: похожие измерения могут означать одинаковый вид.

В простом примере известны только длины: у цветка А лепесток 4 см, у цветка Б — 6 см. У нашего 4,4 см. До А разница 0,4 см, до Б — 1,6 см. Ближе А. Если есть ещё ширина, сравниваем расстояние на плоскости; можно представить линейку между точками. Формула для работы с картой не нужна.

Заметка
Известные примеры называют обучающими. У этого метода обучение в простейшем варианте означает сохранить их признаки вместе с видами. При новом запросе он заново ищет соседей.

Такое правило иногда ошибается. У ближайшего цветка могли неверно записать вид. Или это редкий экземпляр на границе двух групп. Тогда один неудачный пример определит весь ответ.

Тот же способ на пингвинах

Попробуем перенести правило на настоящие измерения Palmer Penguins. Для маленького справочника специально выберем всего три птицы из полного набора. Сравнивать будем один признак — длину ласта: чем меньше разница в миллиметрах, тем ближе сосед.

№ записи	Длина ласта, мм	Известный вид
1	181	Adelie — Адели
153	211	Gentoo — папуанский
277	192	Chinstrap — антарктический

У проверочной птицы № 3 ласт длиной 195 мм. Её самой в справочнике нет. Кто ближе: птица с ластом 181, 211 или 192 мм? Назовите прогноз, затем откройте настоящий вид.

Какого пингвина мы узнали?
Разницы равны 14, 16 и 3 мм. Ближе всего запись № 277, поэтому прогноз — Chinstrap. Но в исходном файле проверочная птица № 3 относится к Adelie. Расчёт выполнен правильно, а прогноз ошибочен: виды могут иметь похожие размеры, и наш справочник очень мал. Добавление других измерений и примеров стоит проверить, но оно не обещает безошибочного ответа.

Мы не строили лучший определитель пингвинов: три записи выбраны для разбора одного решения. Полный CSV позволяет проверить измерения и найти другие примеры.

Посмотрим, что произойдёт, если спросить несколько соседей.

3. Дадим соседям проголосовать

Берём три ближайшие точки. Две подписаны versicolor, одна virginica. Каждый цветок получает один голос, поэтому прогноз будет versicolor. Число выбранных соседей принято обозначать буквой **k**; название **k-NN** означает «k ближайших соседей». При k = 3 голосуют три цветка, даже если в справочнике их сотня. При равных голосах наш интерактив сообщает «ничья». Если расстояния совпали, раньше выбирается строка с меньшим номером: это заранее принятое правило.

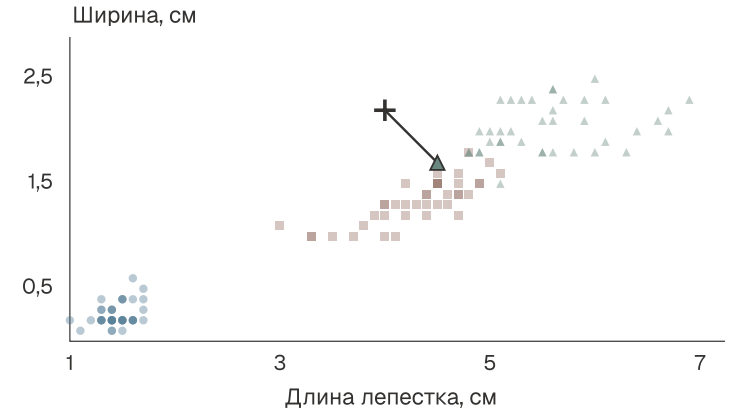


Из трёх ближайших примеров два голосуют за versicolor. Более далёкие точки в этом решении не участвуют.

Перед экспериментом предположите: обязательно ли пять соседей выберут тот же вид, что один? На карте уже стоит учебная точка: длина 4 см, ширина 2,2 см. Сначала посмотрите на ближайшего, назовите вид и проверьте подсчёт. Затем выберите пять соседей. Кнопка «Следующий голос» позволяет открывать их по одному; рядом меняются счёт и текущий лидер.

Кто ближе к вашему цветку?

Один сосед: первый прогноз



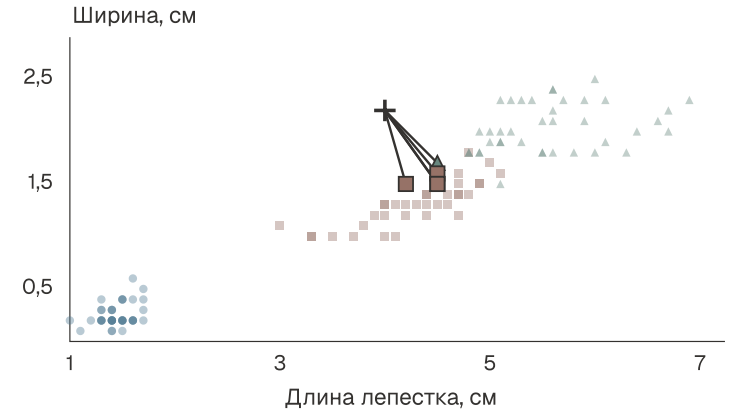
○ setosa · □ versicolor · △ virginica · + ваш цветок

Прогноз: virginica. Голоса: setosa 0 · versicolor 0 · virginica 1.

Учебная точка: 4,0 × 2,2 см. Настоящий вид неизвестен.

1. № 107: virginica · 4.5 × 1.7 см

Пять соседей: большинство меняет прогноз



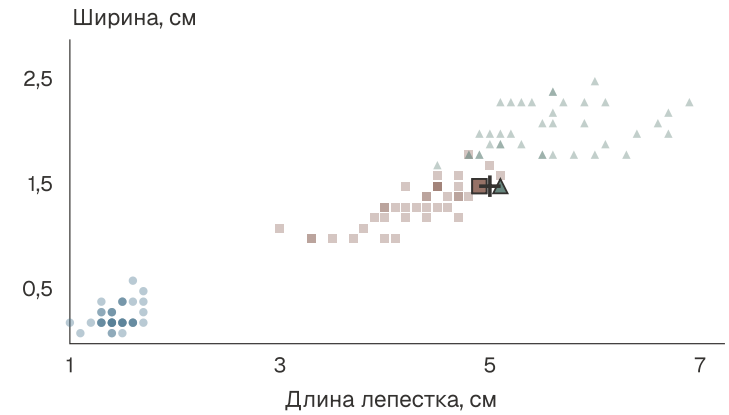
○ setosa · □ versicolor · △ virginica · + ваш цветок

Прогноз: versicolor. Голоса: setosa 0 · versicolor 4 · virginica 1.

Учебная точка: 4,0 × 2,2 см. Настоящий вид неизвестен.

1. № 107: virginica · 4.5 × 1.7 см
2. № 62: versicolor · 4.2 × 1.5 см
3. № 86: versicolor · 4.5 × 1.6 см
4. № 52: versicolor · 4.5 × 1.5 см
5. № 67: versicolor · 4.5 × 1.5 см

Цветок № 120: три соседа, обе длины в сантиметрах



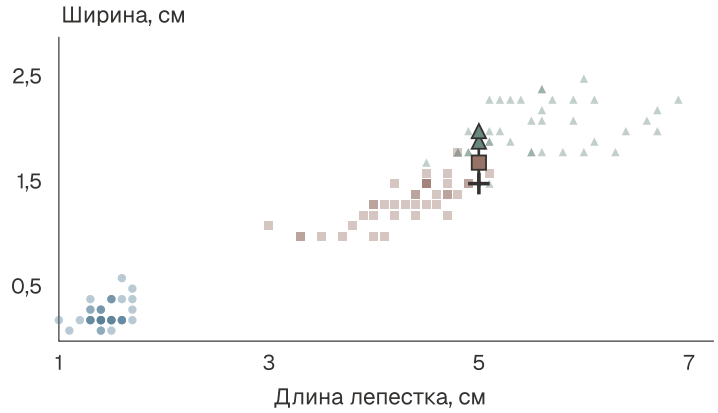
○ setosa · □ versicolor · △ virginica · + ваш цветок

Прогноз: versicolor. Голоса: setosa 0 · versicolor 2 · virginica 1.

Цветок № 120: 5,0 × 1,5 см. Он исключён из соседей. Его вид: virginica. Прогноз ошибочен.

1. № 53: versicolor · 4.9 × 1.5 см
2. № 73: versicolor · 4.9 × 1.5 см
3. № 134: virginica · 5.1 × 1.5 см

Тот же цветок: длина в миллиметрах, ширина в сантиметрах



○ setosa · □ versicolor · △ virginica · + ваш цветок

Прогноз: virginica. Голоса: setosa 0 · versicolor 1 · virginica 2.

Цветок № 120: 5,0 × 1,5 см. Он исключён из соседей. Его вид: virginica. Прогноз совпал. В расчёте длина ×10; на карте обе оси в см.

1. № 78: versicolor · 5.0 × 1.7 см
2. № 147: virginica · 5.0 × 1.9 см
3. № 114: virginica · 5.0 × 2.0 см

Начните с одного соседа, затем выберите пять. Двигайте измерения ползунками. Для проверки масштаба установите цветок № 120 и три соседа; сравните оба варианта единиц. Карта содержит 120 обучающих ирисов; 30 отложены и не голосуют.

Интерактивный эксперимент — в электронной версии урока.

Один сосед выбирает virginica, пять выбирают versicolor. Первую точку мы придумали, поэтому правильный вид неизвестен: здесь наблюдаем смену решения. Цветок № 120 ниже позволит проверить прогноз по настоящей метке.

> Среди семи ближайших цветков 4 относятся к versicolor, остальные к virginica. Сколько голосов получит virginica?

История

Способ сравнивать с ближайшими примерами исследовался задолго до современных нейросетей. Статья Cover и Hart «Nearest Neighbor Pattern Classification» вышла в 1967 году; ссылка на публикацию.

4. Почему миллиметры меняют ответ

Цветок не меняется, когда длину записывают в миллиметрах. Но число становится в десять раз больше. Если программа сравнивает эти числа как раньше, различия в длине начинают сильнее влиять на выбор соседей.

Представьте два известных цветка относительно нового. У А длина отличается на 0,2 см, ширина на 0,8 см. У Б длина отличается на 0,5 см, ширина совпадает. На карте в сантиметрах Б ближе: до него 0,5 см, до А больше 0,8 см. Теперь запишем только длину в миллиметрах: различия станут 2 и 5. При прежнем способе считать расстояние ближе окажется А.

В интерактиве выберите цветок № 120, три соседа и вариант «длина в мм, ширина в см». Сравните список соседей с исходным. Оси для удобства остаются в сантиметрах; меняется расчёт расстояния. Поэтому выбранные соседи могут выглядеть дальше на прежней картинке.

Подумай

Если перевести в миллиметры обе оси, все расстояния увеличатся одинаково. Порядок соседей сохранится. Проблему создаёт изменение относительного веса признаков.

> Метод ближайшего соседа определяет вид цветка по евклидову расстоянию между измерениями лепестков. У всех цветов, включая новый, длину перевели из сантиметров в миллиметры, ширину оставили в сантиметрах. Расстояние считают по полученным числам без нормировки. Почему прогноз может измениться?
>
> ○ Миллиметры всегда точнее сантиметров
>
> ○ У цветка изменился настоящий вид
>
> ○ Стало больше обучающих цветов
>
> ○ Различия по этой оси получили больший вес

Масштаб нужно выбрать осмысленно. Часто признаки приводят к сопоставимым диапазонам по обучающим данным. Подходящий способ проверяют на отдельных примерах; универсального масштаба для любых задач нет.

Профи. **Как вычислить расстояние**

Для двух признаков используем евклидово расстояние: $d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$. Буквы a и b обозначают два цветка, индексы 1 и 2 — длину и ширину. Для сортировки можно сравнивать квадраты расстояний: квадратный корень сохраняет порядок неотрицательных чисел.

Для А из примера квадрат расстояния равен $0{,}2^2 + 0{,}8^2 = 0{,}68$, для Б — $0{,}5^2 = 0{,}25$. После перевода длины: $2^2 + 0{,}8^2 = 4{,}64$ и $5^2 = 25$. Порядок поменялся.

Выберите трёх ближайших

Цель: Получите прогноз versicolor по большинству трёх соседей

```
from collections import Counter
rows = [(0.1, 'virginica'), (0.2, 'versicolor'),
        (0.3, 'versicolor'), (0.9, 'virginica')]
k = 1  # TODO: выберите число соседей
nearest = sorted(rows)[:k]
votes = Counter(name for distance, name in nearest)
prediction = votes.most_common(1)[0][0]
print(f'Прогноз: {prediction}; соседей: {len(nearest)}')
```

5. Экзамен на незнакомом цветке

Вернитесь к интерактиву в шаге 3. Выберите цветок № 120, три соседа и «Обе длины в сантиметрах». Назовите прогноз, проверьте подсчёт голосов и нажмите «Узнать настоящий вид».

Разбор после проверки
В исходном наборе он подписан virginica, а прогноз при этих настройках — versicolor: вот конкретная ошибка. Другие настройки могут исправить её, но это ещё не повод считать их лучшими для всех цветов.

Мы заранее отложили каждую пятую строку исходного файла: 30 цветов из 150. Среди соседей остаются 120. Если разрешить проверочному цветку голосовать за самого себя, он окажется на нулевом расстоянии. При одном соседе получится пересказ известной метки вместо проверки на новом объекте.

Больше соседей тоже не всегда лучше. Если спросить всех, побеждать начнёт самый частый вид справочника независимо от измерений запроса; при равных количествах будет ничья. А если спросить одного, ответ сильнее зависит от случайной точки. Подходящее число соседей выбирают по результатам проверки.

Заметка

Такое разделение выбрано для воспроизводимого учебного опыта. Это не соревнование моделей. Если постоянно менять настройки по этим 30 цветам, понадобится ещё одна независимая проверка.

Может ли большинство ошибиться?

Проверим сразу все 30 отложенных цветов. Для каждого посчитаем голоса за выбранный вид. Например, «4/5» означает, что четыре из пяти соседей поддержали прогноз. На графике такие цветы собраны в один столбик; его высота показывает их количество. При одном соседе все решения единогласны. Значит ли это, что все они верны?

Соседи согласны. А ответ верен?

Один голос: все решения единогласны



○ прогноз верен · × ошибка. Один значок — один цветок.

Верно 29 из 30. Цветок № 120: 1 из 1 за versicolor, но настоящий вид — virginica.

Ссылки

Идея сравнить голоса с настоящими ответами адаптирована из графика kNN в IDS Skoltech. Здесь используем знакомые ирисы, отдельные проверочные цветы и считаем каждый цветок явно.

Пять голосов: большинство тоже может ошибиться



○ прогноз верен · × ошибка. Один значок — один цветок.

Верно 29 из 30. Цветок № 120: 4 из 5 за versicolor, но настоящий вид — virginica.

Откройте ошибки, затем выберите пять соседей. Даже четыре голоса из пяти могут поддержать неверный вид. Согласие соседей не равно гарантии правильного ответа.

Интерактивный эксперимент — в электронной версии урока.

> Метод одного соседа копирует вид ближайшего цветка из справочника. При проверке ему подали те же цветы, что использовали для обучения, и получили все ответы верно. Что проверить прежде всего?
>
> ○ Достаточно ли яркие цвета на карте
> ○ Не слишком ли короткие названия видов
> ○ Не выбирает ли каждый цветок самого себя
> ○ Все ли измерения записаны целыми числами

Эпик. **Сколько чужих меток изменят большинство?**

Среди семи ближайших соседей пять имеют метку А и два метку Б. Расстояния фиксированы. Какое минимальное число меток нужно заменить, чтобы победил Б? Обобщите для нечётного k и текущего большинства m голосов за А. Докажите минимальность.

▾*Решение*

Одна замена А на Б уменьшает отрыв А сразу на два голоса.
В примере после одной замены счёт 4:3, после двух 3:4. Минимум две.
Для нечётного k нужно оставить у А не более $(k-1)/2$ голосов: заменить $m-(k-1)/2$. Меньшее число оставляет А строго больше половины голосов. Это рассуждение предполагает два класса и неизменный состав соседей; для трёх классов условия победы другие.

Следующий способ

Дерево решений задаёт последовательность вопросов о признаках. Вместо списка соседей мы сможем проследить путь от первого вопроса до ответа.

Механизм: найти ближайшие известные примеры и посчитать голоса.

Настройки: число соседей и масштаб признаков способны изменить решение.

Проверка: новый цветок не должен подсказывать собственный вид.

Домашняя работа

1. Для новой учебной точки с длиной лепестка 4 см и шириной 2,2 см сравните прогнозы по одному и пяти ближайшим соседям. В таблице даны пять ближайших известных цветов, от самого близкого к самому дальнему. Каждый выбранный сосед даёт один голос за свой вид. Запишите голоса и объясните смену решения.

№	Длина, см	Ширина, см	Вид
107	4,5	1,7	virginica
62	4,2	1,5	versicolor
86	4,5	1,6	versicolor
52	4,5	1,5	versicolor
67	4,5	1,5	versicolor

2. Цветок № 120: длина лепестка 5 см, ширина 1,5 см. При сантиметрах по обеим осям выбраны соседи № 53, 73, 134. Если длину перевести в миллиметры, оставив ширину в сантиметрах, выбраны № 78, 147, 114. По таблице сравните прогнозы трёх соседей и объясните изменение.

№	Длина, см	Ширина, см	Вид
53	4,9	1,5	versicolor
73	4,9	1,5	versicolor
134	5,1	1,5	virginica
78	5	1,7	versicolor
147	5	1,9	virginica
114	5	2	virginica

3. Со звёздочкой: придумайте пять известных точек двух классов и одну новую так, чтобы прогноз менялся при переходе от одного соседа к трём.

Источники

- Palmer Penguins: описание измерений (https://allisonhorst.github.io/palmerpenguins/reference/penguins.html) — данные Kristen Gorman и Palmer Station LTER, CC0; справочник из записей 1, 153, 277 и проверочная запись 3 файла penguins.csv.

- Iris, UCI (https://archive.ics.uci.edu/dataset/53/iris): источник измерений, описание признаков, CC BY 4.0.
- IDS Skoltech: прогнозы kNN (https://ids.skoltech.ru/seminars/seminar02/02-ml-basics.html#now-its-time-to-build-a-model): источник идеи сравнения голосов с настоящими ответами.
- Файлы Iris (https://archive.ics.uci.edu/static/public/53/iris.zip): используется `bezdekIris.data`, номера строк отсчитываются от 1; длина и ширина лепестка — третий и четвёртый столбцы. Каждая пятая строка исключена из обучения.
- Методы ближайших соседей, scikit-learn (https://scikit-learn.org/stable/modules/neighbors.html): алгоритм, голосование, выбор расстояния.
- Cover и Hart, 1967 (https://doi.org/10.1109/TIT.1967.1053964): исследование классификации по ближайшему соседу.