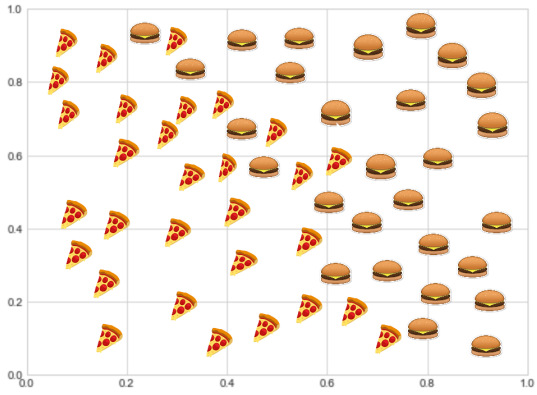
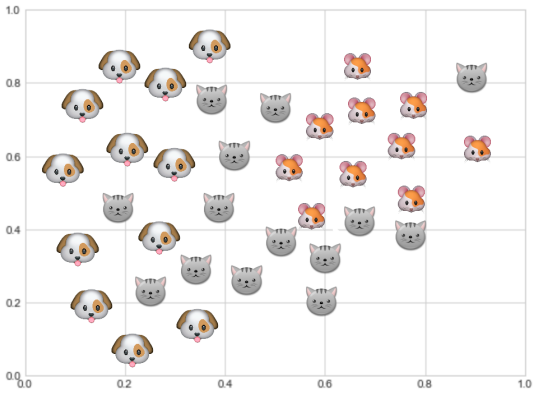
Семинар 8: Соседи, деревья, кросс-валидация

Задача 1 (классификация в картинках)

Нам нужно научиться отделять пиццу от бургеров, а также котиков от пёсиков и от мышек. Проведите на картинках линии, которые отделят одни классы от других. Да, это и есть машинное обучение. Но обычно кривые рисуем не мы, а компуктер.

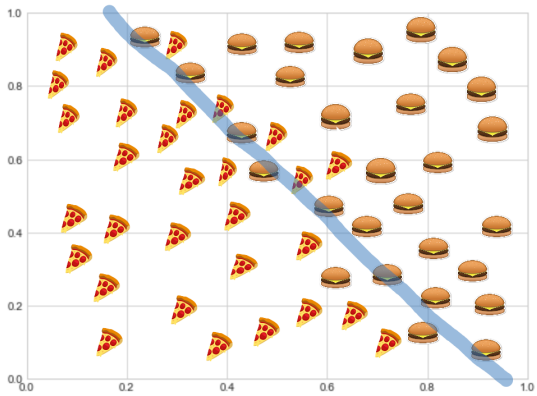


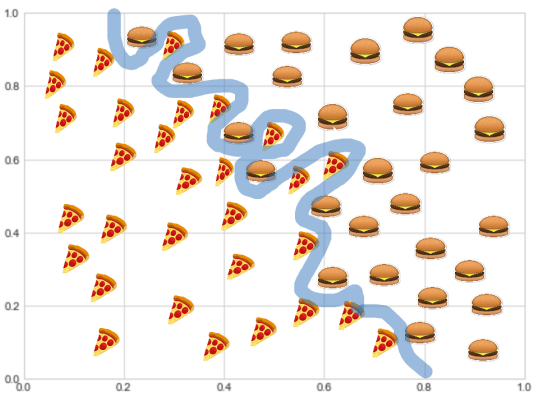


Почему нельзя провести между пиццей и бургерами слишком подробную и извилистую границу? В чём проблема самого правого верхнего котика? Что такое переобучение? Как понять переобучились ли мы?

Решение:

Сначала обсудим бургеры и пиццу. Первый вариант: провести между ними прямую. Тогда мы в части случаев ошибёмся и признаем некотрые бургеры пиццей, а некоторые пиццы бургерами. Второй вариант: провести извилистую разделительную линию, которая чётко разграничит бургеры и пиццу. Вопрос: какой из этих двух вариантов лучше?





Если у нас в выборке оказались все пиццы и бургеры мира, и других быть не может, вторая граница нам подойдёт. Мы подстроимся под все особенности нашей гениральной пицце-бургерной совокупности и будем всегда чётко и безошибочно отличать одно от другого.

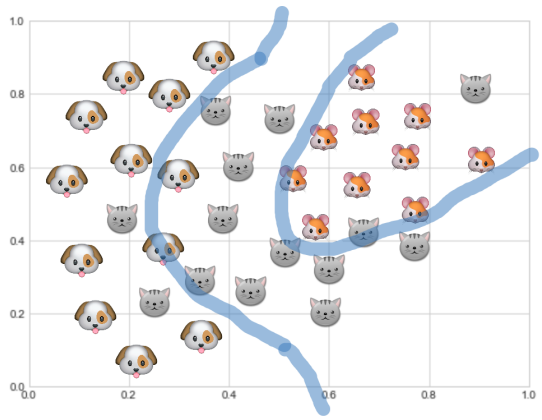
НО в нашем распряжении обычно находится не вся генеральная совокупность, а лишь какая-то её часть. Мы в выборке видим не все возможные варианты, и хотим обучить наш классифи-

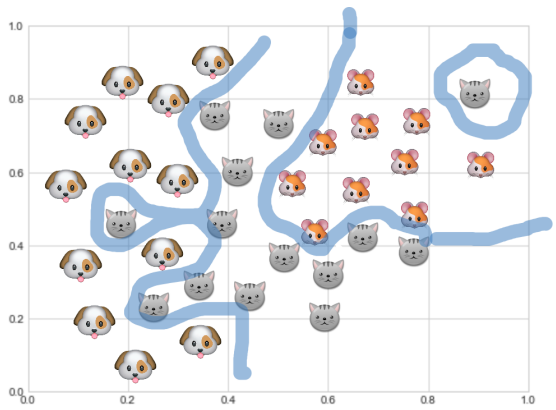
катор обощать. Если к нам попадает новая пицуля или бургер, классификатор должен адекватно сработать на них.

Скорее всего, пиццы, проникшие на территорию бургеров, обладают какими-то аномальными особенностями, на детекцию которых затачивать классификатор нет никакого смысла. Если мы попробуем сделать это, мы влезем на территорию бургеров, и на новых объектах, которые оказались обычными бургерами, будем делать ошибки, подумав, что это аномальные пиццы. Из-за этого лучше разграничить бургеры и пиццы простой линией, которая изображена на первой картинке.

Ещё раз, ещё раз. Если мы проведём подробную границу, мы заточим классификатор под особенности выборки, вместо того, чтобы научить его отличать пиццу от бургера в общем случае. Такие ситуации называются переобучением. И это главная головная боль людей, занимающихся машинным обучением. С переобучением у них идёт вечная борьба.

Теперь посмотрим на котиков, пёсиков и мышек. Снова мы можем провести границы между ними разными способами.





Снова мы можем провести более-менее простую границу и иногда ошибаться. Ну знаете, есть такие собаки мелкие, похожие на кошек. Или даже на мышек. И, если мы будем специфицировать границу под этих собак, мы начнём ошибаться на котах, так как подобные аномалии встречаются редко.

Основная проблема верхнего котика в том, что он аномальный. Каким-то образом он попал на территорию мышек. Выделять для него свою зону будет плохой идеей, так как в таком случае мы будем переобучать классификатор под конкретный выброс.

Осталось обсудить главный вопрос: как понять а не переобучились ли мы. Для этого обычно дробят выборку на две части: тренировочную и тестовую. На тренировочной учат алгоритм (в данном случае границу между классами), а на тестовой проверяют насколько хорошо он работает. Насколько часто алгоритм на тестовой части делает ошибку.

Если получается, что на обучающей выборке качество высокое, а на тестовой низкое — мы переобучились и вместо того, чтобы научить модель обобщать закономерности, существующие в данных, обучили его под особенности конкретной выборки. Если на тестовой выборке качество сравнимо с обучающей, значит мы научились извлекать какие-то реальные закономерности.

Бьюсь об заклад, что для простых линий, качество на тесте для бургеров и мышек будет выше, чем для сложных. Конечно же, простые границы оказываются хороши не всегда, но всегда имеет смысл сначала построить простую модель, а после сравнивать с ней сложные.

Задача 2 (KNN, кросс-валидация)

На плоскости расположены колонии рыжих и чёрных муравьёв. Рыжих колоний три и они имеют координаты $(-1,-1)$, $(1,1)$ и $(3,3)$. Чёрных колоний тоже три и они имеют координаты $(2,2)$, $(4,4)$ и $(6,6)$.

а) Чем KNN отличается от K-means?

б) Поделите плоскость на «зоны влияния» рыжих и чёрных используя метод одного ближайшего соседа.

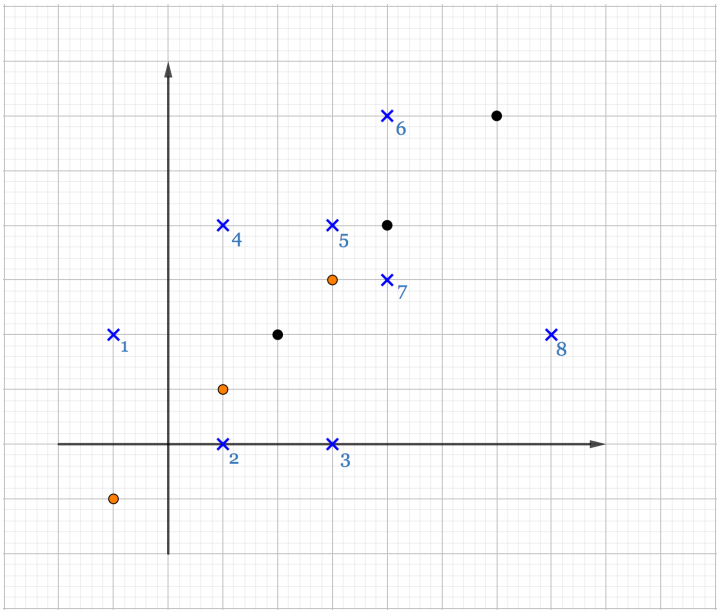
в) Поделите плоскость на «зоны влияния» рыжих и чёрных используя метод трёх ближайших соседей.

г) С помощью кросс-валидации с выкидыванием отдельных наблюдений выберите оптимальное число соседей k перебрав $k \in \{1,3,5\}$. Целевой функцией является количество несовпадающих прогнозов.

Решение:

а) KNN — это метод классификации. Для неё мы знаем ответы для каждого из объектов, и учим алгоритм отличать одни ответы от других. K-means — это метод кластеризации. Для неё мы не знаем ответов ни на одном из объектов. Мы учим алгоритм выделять области похожих объектов.

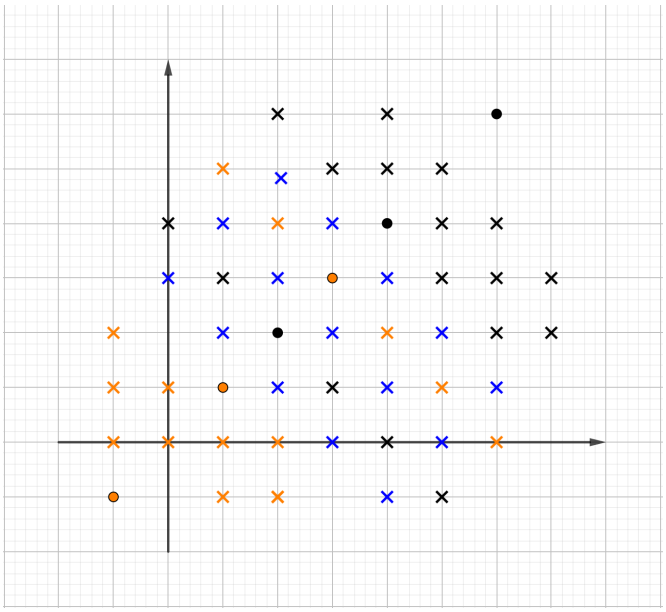
б) Будем ради удобства измерять расстояние между муравейниками в метрах. Давайте отметимна плоскости несколько рандомных точек и посмотрим к чьей зоне влияния они относятся.



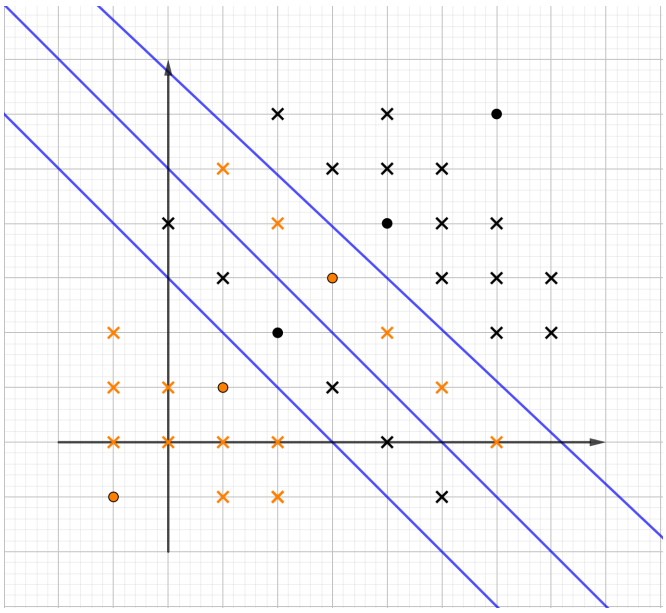
Точка номер один явно будет в зоне влияния рыжих муравьёв. До ближайшего рыжего муравейника нужно пройти $\sqrt{5}$ метров, до ближайшего чёрного 3 метра. Точка два тоже рыжая.

По аналогии точки восемь и шесть оказывается чёрными. С оставшимися точками возникают проблемы. Например, от точки номер пять одинаковое расстояние как до чёрного, так и до рыжего муравейников. Она является спорной. Судя по всему, именно через неё пройдёт граница. Давайте попробуем нащупать побольше подобных пограничных точек.

Если точка принадлежит рыжим муравьям, будем помечать её рыжим крестом. Если чёрным, то чёрным. Если это спорная точка, то синим.

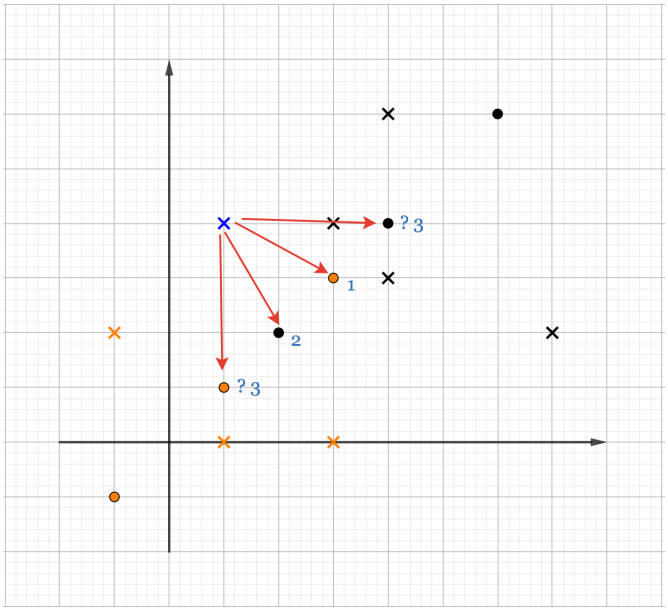


Кажется, что мы нащупали границы, вдоль которых находятся спорные территории. Осталось только прочертить их.

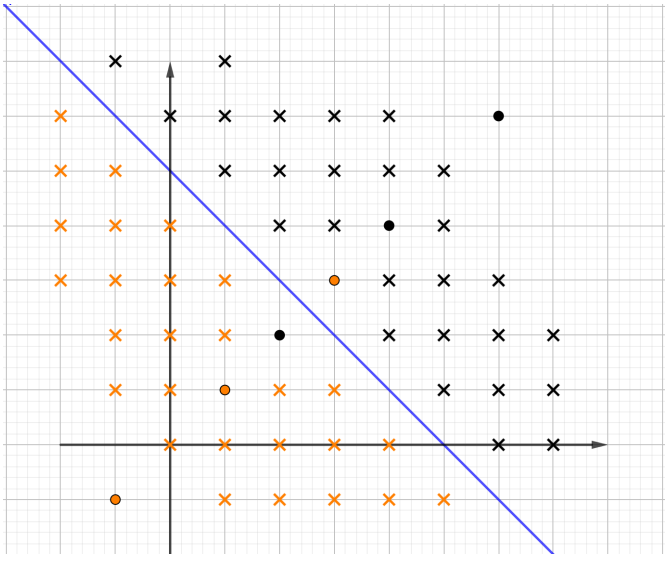


в) Теперь попробуем поделить плоскость на зоны влияния, используя метод трёх ближайших соседей. Посмотрим на самую первую картинку, где мы нанесли на плоскость рандомные точки, и попробуем порассуждать в чьей зоне влияния оказывается какая точка.

Для первой точки две из трёх ближайших — рыжие. Она находится в рыжей зоне влияния. По аналогии происодит со второй и третьей точками. Пятая, шестая, седьмая и восьмая точки оказываются в зоне влияния чёрных муравьёв и окрашиваются в чёрные цвета. Проблемы возникают только с четвёртой точкой. Ближайшие к ней две точки — рыжая и чёрная. Решение надо принимать по третьему ближайшему соседу. Третью ближайшую точку найти не удаётся, так как рыжая и чёрная точка находятся от неё н одинаковых расстояниях. Выходит, что мы оказались на границе.



Попробуем нащупать ешё пограничных точек и провести пограничную линию.



И это граница? У нас же есть ошибки! Да, есть. Но давайте вспомним мораль, которую мы извлекли из первого упражнения: слишком детализированная граница между классами приводит к переобучению.

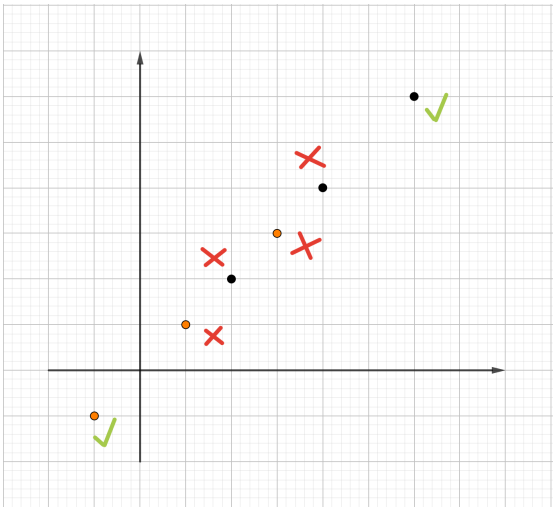
Порассуждаем в терминах джунглей. Есть поле, на нём селятся муравьи. Логично ли с их стороны селиться полосками? Конечно же нет. Намного логичнее было бы, что по историческим причинам на одной стороне поля живут рыжие муравьи, на второй чёрные. У нас

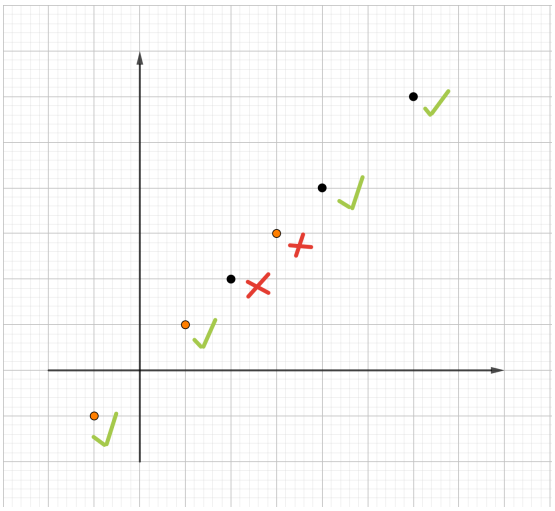
в выборке оказалось несколько примеров муравейников. И по ним мы попытались нащупать границу для зон влияния. На границе вполне может происходить такое, что муравьи проникают на территорию друг-друга.

Проводя излишние полосы, мы переходим от выуживания реальных закономерностей, существующих в джунглях, к излишнему фрагментированию обучающей выборки, то есть переобучаемся под её особенности.

г) Давайте убедимся в том, что алгоритм трёх ближайших соседей, проводящий одни разграничительную линию между муравьями, работает лучше, чем алгоритм одного ближайшего соседа. Для этого воспользуемся стратегией кросс-валидации.

Кросс-валидация состоит в следующем: давайте будем закрывать по очереди разные части выборки ладошкой. На оставшейся выборке будем обучать модель, а на скрытой проверять её качество. Будем делать так много раз и посмотрим на итоговое качество.





Закрываем ладошкой самую нижнюю точку. По оставшимся четырём расчерчиваем границы. Мы по методу одного ближайшего соседа относим эту точку к рыжим муравьям. Это оказывается правильным решением. Угадали.

Закроем ладошкой вторую снизу точку. Расчертим границы. Она окажется ближе всего к чёрным муравьям. Но на самом деле она рыжая. Ошибка… Также проделаем с остальными точками. В итоге получится, что мы совершаем целых 4 ошибки. По аналогии сделаем с методом трёх ближайших соседей и получим всего лишь 2 ошибки.

Чувствуете? Мы ошибаемся из-за излишней детализации, которую нам навязывает метод одного ближайшего соседа. Кросс-валидация позволяет это отследить. А что, если взять 5 ближайших соседей? Тогда мы ошибёмся абсолютно в каждой точке.

На самом деле k это гиперпараметр метода ближайших соседей. Мы можем подобрать его оптимальным образом с помощью кросс-валидации. В данном примере оптимально будет выбрать $k = 3$.

Задача 3 (древо для классификации)

Машка пять дней подряд гадала на ромашке, а затем выкладывала очередную фотку «Машка с ромашкой» в инстаграмчик. Результат гадания — переменная y_i, количество лайков у фотки —

переменная x_i. Постройте классификационное дерево для прогнозирования y_i с помощью x_i на обучающей выборке:

y_i	x_i
плюнет	10
поцелует	11
поцелует	12
к сердцу прижмёт	13
к сердцу прижмёт	14

Дерево строится до идеальной классификации. Критерий деления узла на два — минимизация числа допущенных ошибок[1]. Правило прогнозирования в каждой вершине: в качестве прогноза выдаем тот класс, представителей которого в вершине больше. Предположим, что под фоткой стоит 15 лайков, каков будет результат гадания?

Решение:

Мы должны обучить дерево, которое будет по переменной x, число лайков от парня, прогнозировать переменную y, состояние отношений Маши. Обычно деревья учат по-жадному. Будем смотреть, какое разбиение по переменной x сильнее всего уменьшает ошибку, и выбирать его.

Ошибку мы договорились считать как долю неверных ответов. Обычно на практике при разбиении вершины на две используют не такой критерий, но мы для простоты используем его.

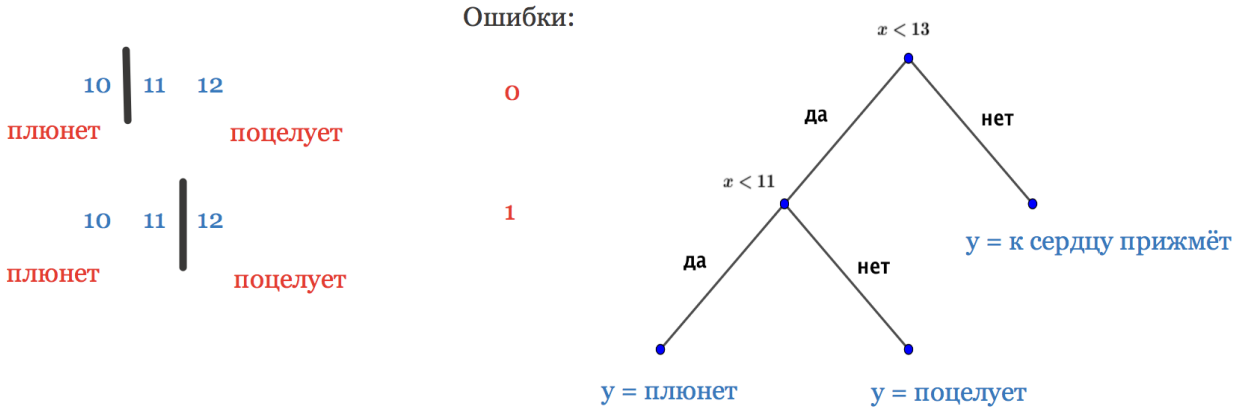
При делении вершины на две между 10 и 11 лайкаи, слева у нас окажется плюнет. Именно его мы и будем там прогнозировать. Справа окажется два поцелует и два к сердцу прижмёт. Надо спрогнозировать в этой вершине класс, представителей которого тут большинство, чтобы сделать поменьше ошибок. Так как у нас оба класса представлены в одинаковом объёме, неважно что мы спрогнозируем. В любом случае получим две ошибки.

При дроблении вершины на две между 11 и 12 лайками, слева оказывается плюнет и поцелует. Одна ошибка. Справа оказывается два к сердцу прижмёт и одно поцелует. Спрогнозируем к сердцу прижмёт, так как их большинство, и получи одну ошибку. В сумме у нас две ошибки.

[1]На самом деле на практике так не делают. Обычно для разбиения узла при строительстве классификационных деревьев используют энтропию. О том, что это такое, можно погуглить.

Рассуждая аналогичном образом приходим к выводу, что самое классное разбиение между 12 и 13. При нём мы совершаем только одну ошибку. В дереве, мы будем задавать вопрос: «А количество лайков меньше 13?» Если да, будем идти налево и прогнозировать, что нас поцелуют. Если нет, будем идти направо и прогнозировать, что нас прижмут к сердцу.

Справа в листе дерева у нас оказались объекты одного класса. Слева в листе дерева содержутся объекты разных классов. Можно сделать ещё одно разбиение.



В итоге в нашем дереве окажется три листа, в каждом из которых мы будем делать прогноз. Обратите внимание, что дерево запомнило выборку. Деревья постоянно так делают. В этом их существенный минус. Чтобы победить его, деревья стригут. Либо используют как части более сложных моделей. Например, как часть случайного леса.

Предположим, что под фоточками Маши от Паши накопилось 15 лайков. Что ждёт её отношения? Начинаем идти по решающему дереву, чтобы сделать прогноз. Число лайков меньше 13? Нет. Идём направо. Кажется, Машу прижмут к сердцу. Это наш прогноз.

Задача 4 (древо для регрессии)

Миша работает в маленькой кофейне. Харио Малабар Монсун является фирменным напитком этой кофейни. Мише интересно узнать как именно ведёт себя спрос на напиток y_i в зависимости от температуры за окном t_i. Четыре дня Миша записывал свои наблюдения:

t_i	y_i
21	1
19	2
12	8
8	8

Сегодня он решил обучить регрессионное дерево. В качестве функции потерь он использует

$$\sum (y_i - \hat{y}_i)^2.$$

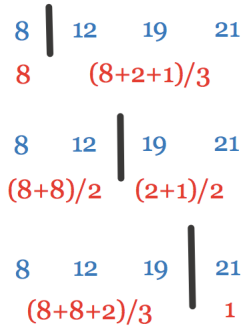
а) Обучите регрессионное дерево.

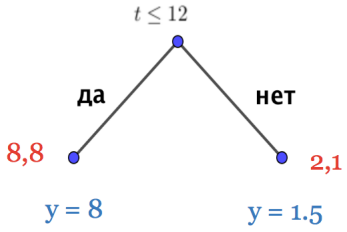
б) Какой прогноз на сегодня сделает дерево Миши, если за окном 13 градусов?

Решение:

На этом семинаре мы сажаем дерево. Будем ли мы на следущем строить дом и рожать ребёнка — большой вопрос. Мы должны по переменной t спрогнозировать переменную y. Для этого нужно обучить дерево. Учить мы его будем по-жадному. Будем смотреть какое разбиение по переменной t сильнее всего уменьшает ошибку, и выбирать его.

На первом шаге у нас есть три способа сделать разбиение по переменной t:





- Мы можем отправть в левую вершину все ситуации, где температура меньше либо равна 8 градусам. В таком случае, когда мы идём по дереву налево, мы будем прогнозировать, что потребители выпьют 8 чашек кофе. Когда мы идём вправую вершину, мы будем прогнозировать, что потребители выпьют 3.6 чашек кофе. Это среднее всех y, попавших в правую вершину. Давайте посчитаем ошибку, которую при этом будет допускать дерево.

$$(8-8)^2+(8-3.6)^2+(2-3.6)^2+(1-3.6)^2=28.68.$$

- Мы можем отправить в левую вершину все ситуации, где температур меньше либо равна 12. В таком случае слева прогноз будет 8, а справа 1.5. Найдём ошибку:

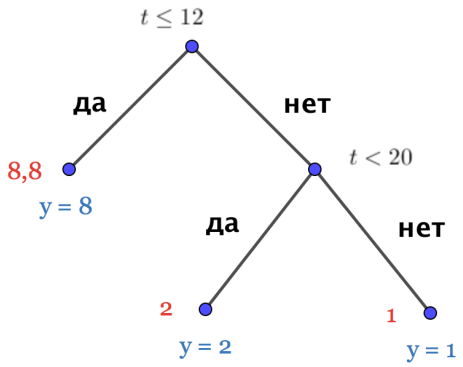
$$(8-8)^2+(8-8)^2+(2-1.5)^2+(1-1.5)^2=0.5.$$

- В третьей ситуации получаем, что

$$(8-6)^2+(8-6)^2+(2-6)^2+(1-1)^2=24.$$

Оптимальным для разбиения оказывается второй вариант. Он сильнее всего уменьшает ошибку. Выбрав его, мы отправляем влевую вершину две восьмёрки и получаем в ней нулевую ошибку. Вправую вершину мы отправляем двойку и единицу.

В правой вершине нужно сделать ещё одну итерацию, чтобы отделить двойку от единицы. Тогда обучение дерева будет окончено. Итоговое дерево будет иметь вид:



Сделаем прогноз для 13 градусов. Для этого пройдёмся по дереву от корня к одному из листов. На улице меньше или равно 12 градусов? нет. Идём направо. На улице меньше 20 градусов? Да. Идём налево. В кофейне купят 2 чашки.

Обратите внимание, что дерево идеально запомнило обучающую выборку. Оно слишком сильно фрагментировало её. Это является переобучением. Чтобы деревья не переобучались и не вылизывали выборку, обычно останавливают обучение деревьев досрочно:

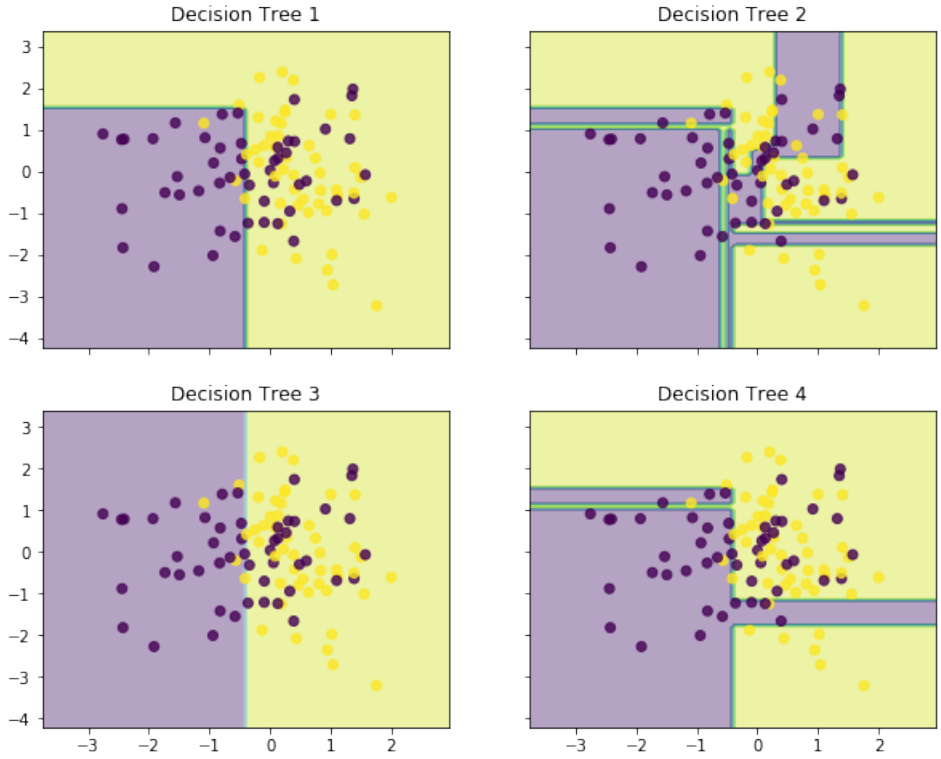
- Когда в вершини оказалось не менее 10 объектов
- Когда дерево построилось до 20 листьев.
- Когда глубина дерева оказалась равна 5.

Конечно же конкретные цифры здесь для пример. Они являются гиперпараметрами и подбираются также как мы подбирали k в методе ближайших соседей на предыдущем семинаре.

Другой путь — применять сразу много деревьев. Примером такой модели является случайный лес.

Задача 5

Ниже изображены разделяющие поверхности для задачи бинарной классификации, соответствующие решающим деревьям разной глубины. Какое из изображений соответствует наиболее глубокому дереву? Какой примерной глубине дерева соотвествует каждая из картинок?



Решение:

Чем глубже дерево, тем сильнее оно фрагментирует нашу выборку, и тем сильнее оно выделяет в ней самые микроскопические кусочки. Сильнее всего выборка фрагментирована на верхней правой картинке, значит это разбиение плосткости на части соотвествует самому глубокому дереву.

На третьей картинке плоскость дробиться на части один раз. Значит в дереве есть один сплит. Его глубина равна единице. На первой картинке появляется ещё одно дополнительное разбиение по оси y, глубина дерева увеличивается до двух.

На картинке номер 4 мы делаем два дополнительных разбиения правой части и два левой. Глубина дерева уже не менее трёх. На второй картинке всё становится ещё глубже.

1. Ещё задачи

Тут лежит ещё несколько задач для самостоятельного решения. Возможно, похожие будут в самостоятельной работе...

Задача 6

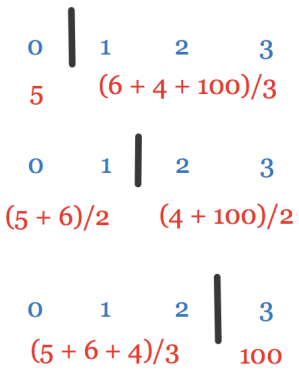
Выращиваем регрессионное дерево в домашних условиях! Вот вам выборка для этого:

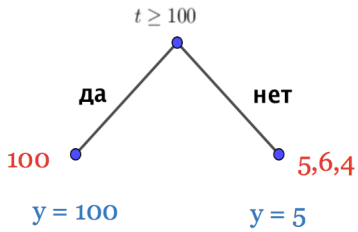
x_i	y_i
0	5
1	6
2	4
3	100

Критерий деления вершины — минимизация квадратичной функции потерь. Критерий остановки — три листа. Зачем нужен критерий остановки? Как дерево ведёт себя с выбросами?

Решение:

У нас есть три способа раздробить по x дерево.



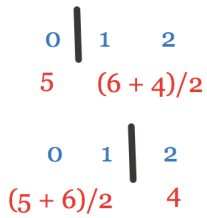


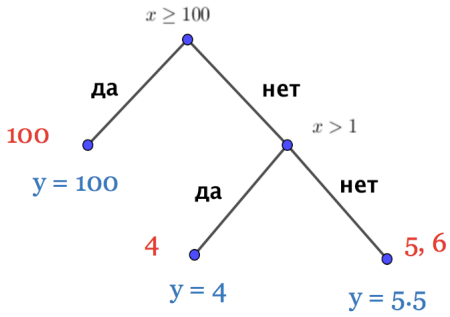
Посчитаем для каждого способа квадратичную ошибку:

- $(5-5)^2 + (6-36.6)^2 + (4-36.6)^2 + (100-36.6)^2 = 6018.68$
- $(5-5.5)^2 + (6-5.5)^2 + (4-52)^2 + (100-52)^2 = 4608.5$
- $(5-5)^2 + (6-5)^2 + (4-5)^2 + (100-100)^2 = 2$

Выгоднее всего оказывается обособить первым же отсечением выброс. Это нормальная ситуация. На практике так происходит регулярно. Деревья изолируют выбросы в отдельные вершины, и они никак не портят работу с основной выборкой. Такое свойство называется нечувствительностью к выбросам или робастностью к выбросам. В следущем упражнении, мы с вами встретимся с ещё одной моделью, которая устойчива к выбросам.

Сделаем второй шаг разбиения.





Посчитаем для каждого способа квадратичную ошибку:

- $(5-5)^2 + (6-5)^2 + (4-5)^2 = 2$
- $(5-5.5)^2 + (6-5.5)^2 + (4-4)^2 = 0.5$

Понятно, что дробить нужно, обосабливая четвёрку. После этого нужно остановиться. По условию задачи критерий остановки — три листа у дерева. Ошибка бы продолжила убывать для тренировочной выборки. На тестовой она бы возрастала. Обычно подобные критерии ранней остановки помогают избежать переобучения.

Кстати говоря, именно благодаря тому, что деревья на первом же шаге изолируют выбросы, случайный лес можно из прогнозной модели модернизировать в модель, которая неплохо справляется с поиском аномалий. Подумайте на досуге как именно можно сделать это.

Задача 7

Пятачок собрал данные о визитах Винни-Пуха в гости к Кролику. Здесь x_i - количество съеденного мёда в горшках, а y_i - бинарная переменная, отражающая застревание Винни-Пуха при входе

y_i	x_i
0	1
1	4
1	2
0	3
1	3
0	1

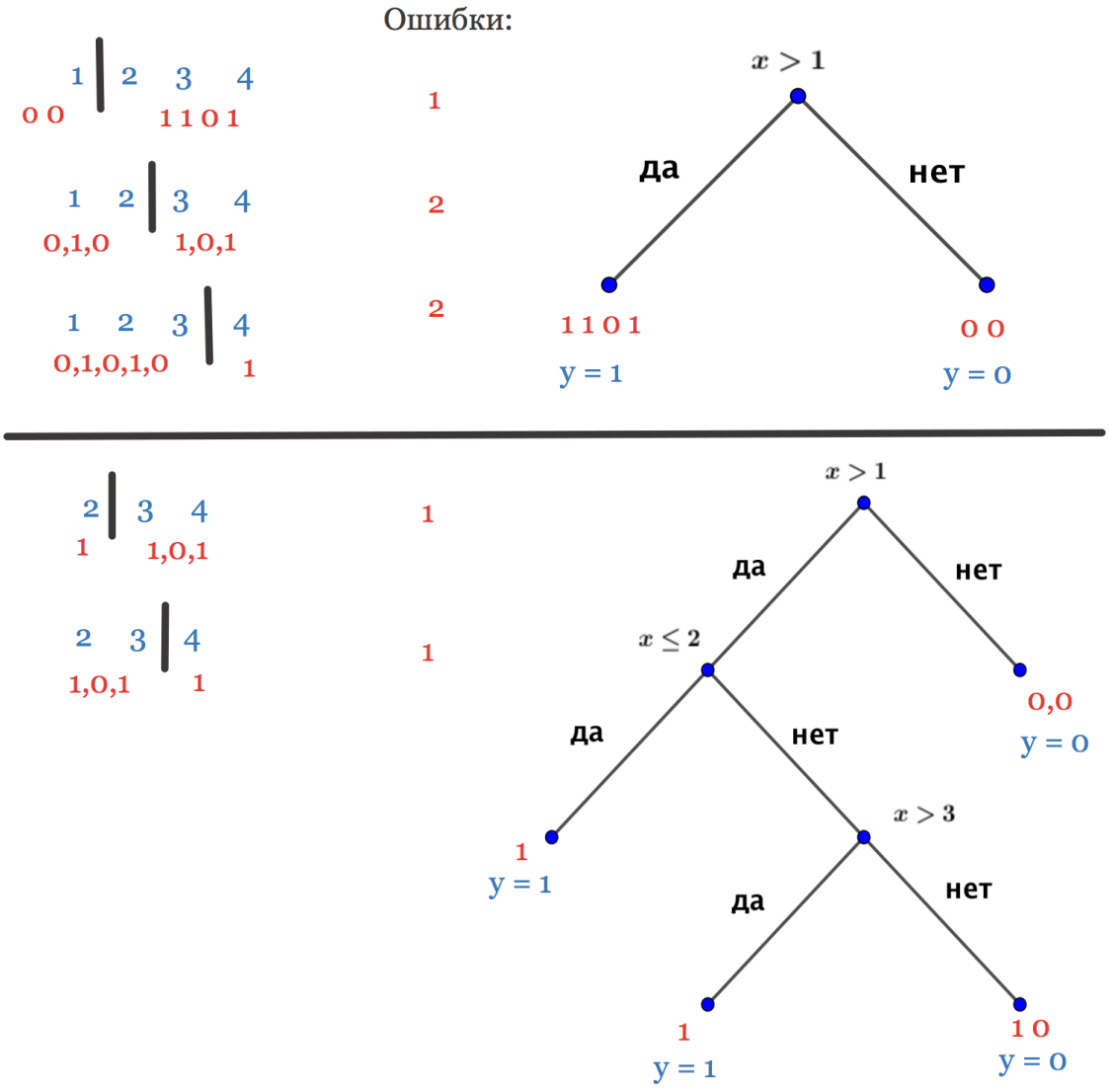
a) Пятачок собирается оценить дерево по всей выборке. Помогите очень маленькому существу сделать это.

б) Пятачок узнал у Иа-Иа, что оказывается выборку надо делить на тренировочную и тестовую Поэтому он отложил последние два наблюдения для теста. Оцените дерево по первым четырём наблюдениям и проверьте его работоспособность по последним двум.

в) Пятачок поговорил с Совой и узнал, что деревья часто переобучаются. Она рассказала ему, что над деревьями надо строить ансамбли. Например, случайный лес. Пятачок решил построить лес из двух деревьев. Первое дерево он строит на наблюдениях с первого по третье, второе на наблюдениях со второго по четвёртое. Третье дерево на наблюдениях $1, 2, 4$. Помогите пяточку построить лес и оценить качество его работы на тестовой выборке.

Решение:

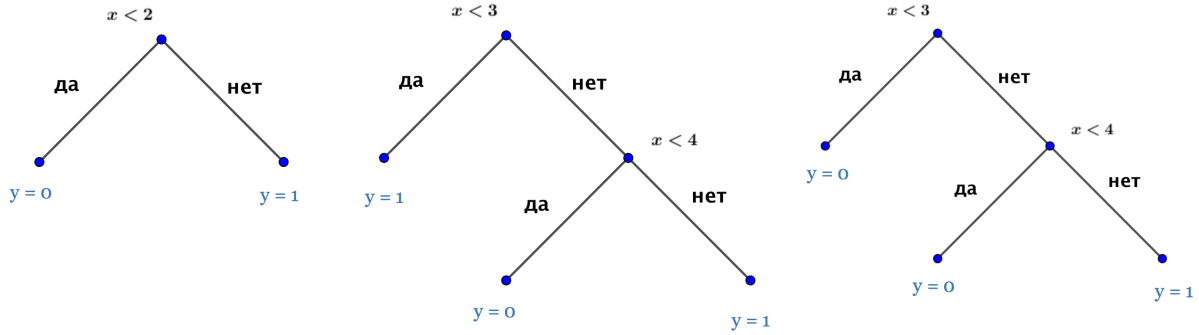
а) Бедный малютка Пяточок! Он даже не понимал, на какие муки он себя обрекает, когда собирался строить свою модель для прогнозирования того, что произойдёт с Винни! Как же хорошо, что мы оказались рядом и подставили маленькому существу своё большое дружеское плечо. Для начала построим дерево сразу на всей выборке.



Обратите внимание, что это дерево ошибается из-за того, что при $x = 3$ у нас есть как факт застревания медведя в норе, так и факт его прохождения сквозь нору.

б) Когда мы строим дерево на первых четырёх наблюдениях, первое разбиение можно сделать либо по единице, либо по четвёрке. В обеих ситуациях совершается одна ошибка. Для удобства выберем первый случай. Дальше снова неважно где делать разбиение. Сделаем его в двойке. В итоге получим дерево из пункта а). На тестовой выборке дерево делает одну ошибку при $x = 3$.

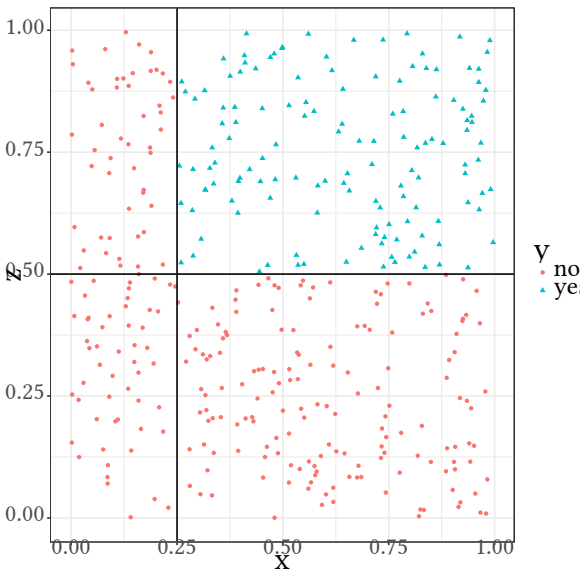
в) Лес, который должен получиться в ходе обучения, изображён на кратинке:



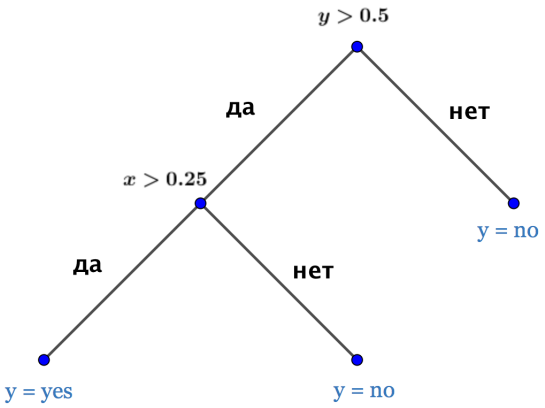
Для $x = 3$ первое дерево прогнозирует 1, второе 0, третье 0. Большая часть говорит, что 0, его и берём. Это ошибка. Для $x = 1$ первое первое и третье деревья прогнозируют 0, второе 1, берём ноль и не ошибаемся.

Задача 8

По данной диаграмме рассеяния постройте классификационное дерево для зависимой переменной y:



Решение:



Если мы дробим сначала по оси y, то мы сразу же довольно сильно уменьшаем неопределённость и ошибаемся только на верхнем левом прямоугольнике, прогнозируя, что он синего цвета.

Если мы дробим сначала по переменной x, то мы будем ошибаться на нижнем правом прямоугольнике. Там ошибка намного страшнее. Значит, сначала произойдёт разбиение по y, затем по x. Именно в этом состоит жадная процедура обучения дерева: уменьшить ошибку при каждом разбиении как можно сильнее.

Задача 9

Рассмотрим обучающую выборку для прогнозирования y с помощью x и z:

y_i	x_i	z_i
y_1	1	2
y_2	1	2
y_3	2	2
y_4	2	1
y_5	2	1
y_6	2	1
y_7	2	1

Будем называть деревья разными, если они выдают разные прогнозы на обучающей выборке. Сколько существует разных классификационных деревьев для данного набора данных?

Решение: Либо мы сначала дробим по x, потом по z. Либо наоборот.