Вероятностные алгоритмы Наивный Байес

Лекция № 9

Лектор: Шевляков Артём

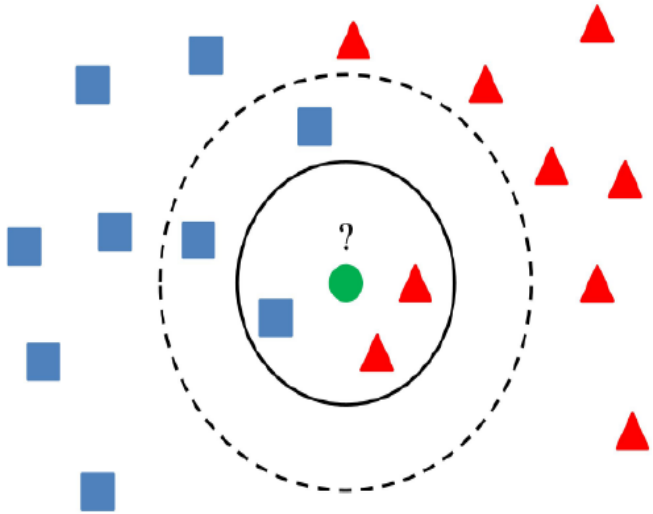
Есть алгоритмы, которые выдают не метку класса, а вероятность принадлежности классам

Пример предсказания

Объект	Рост	Вес	Пол (0-ж, 1-м) (предсказанный)
A	180	70	0.75
B	150	45	0.1

Примеры таких алгоритмов

1. наивный Байес
2. логистическая регрессия (см. ниже)
3. модификация kNN (подумайте, какая?)



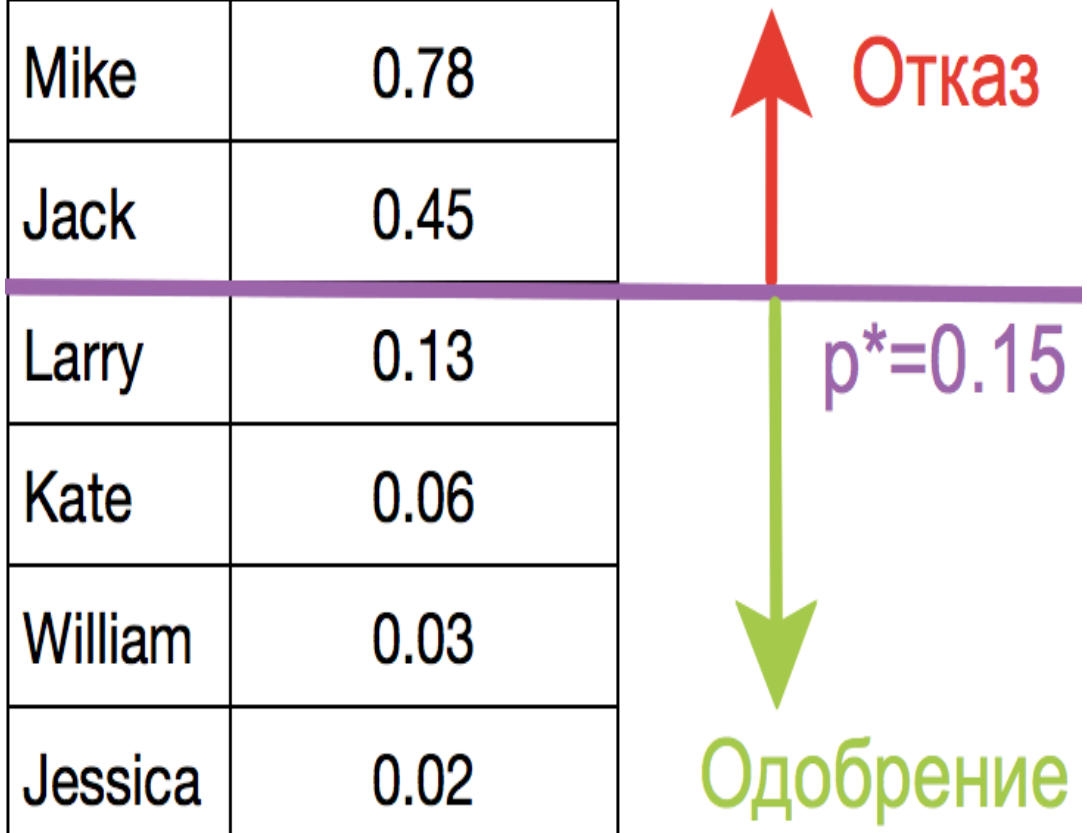
4. модификация решающего дерева дерева (подумайте, какая?)
5. …

Зачем нужны такие алгоритмы предсказания?

1. Когда данные будут подаваться на вход другим алгоритмам (вероятность несёт в себе дополнительную информацию: нашу уверенность в классификации).
2. Возможность переложить ответственность на заказчика)))))
3. Помогают избежать обвинений в богохульстве))))

Классический пример (кредитный скоринг)

Клиент	Вероятность невозврата
Mike	0.78
Jack	0.45
Larry	0.13
Kate	0.06
William	0.03
Jessica	0.02



Повторим школьную теорию вероятностей

Вероятность события

Пусть A – событие. Например,
A={монета выпадет орлом}
A={докладчик помрёт во время лекции}
A={на игральном кубике выпадет 6}
Для событий можно найти их вероятность Pr(A) следующим образом:
Пусть $A_1, A_2, \ldots, A_n$ – все элементарные исходы эксперимента, тогда
Pr(A)={число исходов при которых наблюдается A}/n

Пример 1

Эксперимент: кидаем кубик.
Элементарные исходы: «выпадет 1»,…, «выпадет 6»
Для краткости:

1,2,3,4,5,6

Найдите вероятности следующих событий:
A={выпадет четное число}
A={выпадет число, делящиеся на 3}
A={выпадет делитель числа 6}
A={выпадет отрицательное число}
A={выпадет число меньше 10}

Пример 2

Эксперимент: заводим двоих детей.
Элементарные исходы (М – мальчик, Д- девочка):

ММ,МД,ДМ,ДД

Найдите вероятности следующих событий:
А={дети будут разного пола}
А={будет хотя бы одна девочка}
А={старшим ребенком будет мальчик}
А={число девочек не равно числу мальчиков}

Условные вероятности

Pr(A|B) (читается как «вероятность A при условии B»)
Условная вероятность считается по правилу:
1. Мысленно представьте, что событие B уже произошло.
2. Удалите элементарные исходы, которые НЕ удовлетворяют событию B.
3. Найдите вероятность события A для нового множества элементарных исходов, используя формулу с предыдущих слайдов.

Пример 1*

Эксперимент: кидаем кубик.
Элементарные исходы:

1,2,3,4,5,6

Найдите вероятность Pr(A|B) для :
1) A={выпадет четное число}, B={выпадет число, делящиеся на 3}
2) A={выпадет 6}, B={выпадет делитель числа 6}
3) A={выпадет 1}, B={выпадет четное число}

Пример 1*

Эксперимент: кидаем кубик.
Элементарные исходы:

1,2,3,4,5,6

Найдите вероятность Pr(A|B) для :
1) A={выпадет четное число}, B={выпадет число, делящиеся на 3}
2) A={выпадет 6}, B={выпадет делитель числа 6}
3) A={выпадет 1}, B={выпадет четное число}

А теперь подсчитайте Pr(B|A) для пп.1-3

Пример 2*

Эксперимент: заводим двоих детей.
Элементарные исходы (М – мальчик, Д- девочка):

ММ,МД,ДМ,ДД

Найдите вероятность Pr(A|B) для :
1) A={дети будут разного пола} B={среди детей есть девочка}
2) A={дети будут разного пола} B={старший ребенок - девочка}

Пример 2*

Эксперимент: заводим двоих детей.
Элементарные исходы (М – мальчик, Д- девочка):

ММ,МД,ДМ,ДД

Найдите вероятность Pr(A|B) для :
1) A={дети будут разного пола} B={среди детей есть девочка}
2) A={дети будут разного пола} B={старший ребенок - девочка}

А теперь подсчитайте Pr(B|A) для пп.1-2

Задачка более близкая к проблеме классификации

В выборке 60% женщин и 40% мужчин. Известно, что среди них курит 10% женщин и 70% мужчин. Про человека NN известно, что он курит. С какой вероятностью NN является женщиной (мужчиной)?

Если вы решите эту задачу, то вы фактически осуществите классификацию NN (предскажете значение его пола).

Решение

В выборке 60% женщин и 40% мужчин. Известно, что среди них курит 10% женщин и 70% мужчин. Про человека NN известно, что он курит. С какой вероятностью NN является женщиной (мужчиной)?

Обозначим:

М={человек является мужчиной}

Ж={человек является женщиной}

К={человек курит}

Из условия можно найти

Pr(М), Pr(Ж), Pr(К|М), Pr(К|Ж).

Решение

Имеем:
Pr(М)=0.4, Pr(Ж)=0.6, Pr(К|М)=0.7, Pr(К|Ж)=0.1
Из этих чисел можно даже найти Pr(К):
Pr(К)= Pr(М)*Pr(К|М)+Pr(Ж)*Pr(К|Ж)=0.4*0.7+0.6*0.1 =0.34=34%

Но нас интересуют вероятности несколько других событий (каких?):

Решение

Имеем:
Pr(М)=0.4, Pr(Ж)=0.6, Pr(К|М)=0.7, Pr(К|Ж)=0.1
Из этих чисел можно даже найти Pr(К):
Pr(К)= Pr(М)*Pr(К|М)+Pr(Ж)*Pr(К|Ж)=0.4*0.7+0.6*0.1
=0.34=34%

Но нас интересуют вероятности несколько других событий:
Pr(М|К)=?
Pr(Ж|К)=?
А как их найти?

Есть формула Байеса!

$$Pr(B|A)=Pr(A|B)*Pr(B)/Pr(A)$$

Для нашей задачи получаем:
$Pr(М|К)=Pr(К|М)*Pr(М)/Pr(К)=0.7*0.4/0.34=28/34$
$Pr(Ж|К)= Pr(К|Ж)*Pr(Ж)/Pr(К)=0.1*0.6/0.34=6/34$

- фактически мы получили оценки принадлежности «классу мужчин» и «классу женщин»

Что делать с вероятностью при классификации?

На прошлом слайде были получены вероятности принадлежности объекта «классу мужчин» и «классу женщин» (28/34 и 8/34). Что с ними делать, чтобы получить точный (а не вероятностный) ответ?

1. Выбрать класс с наибольшей вероятностью (в данном примере выбрать «мужчин»).

2. С распределением 28/34 и 8/34 сгенерировать случайную величину. Ее значение и будет окончательным ответом.

Что делать с вероятностью при классификации?

2. Установить пороговое значение π для вероятности. Если вероятность принадлежности «классу мужчин» больше π, то объект классифицируется как мужчина. Иначе – как женщина.

Такой подход оправдан, когда ущерб от ошибки классификации различен для первого и второго класса (см. пример на след. слайде).

Террорист или нет?

Допустим мы оцениваем не вероятности принадлежности классам «мужчин» и «женщин», а к классам «террористов» и «честных людей». Допустим, что для человека А были получены вероятностные оценки на принадлежности к классу террористов 6/34 и к классу «честных людей» 28/34.

Его «честность» в 4.5 раза выше его терроризма и в соответствии с методами пп1-2 он (скорей всего) будет окончательно отнесен к классу «честных».

Террорист или нет?

Но тут над понимать, что цена ошибки «террориста приняли за честного» гораздо выше цены ошибки «честного приняли за террориста». А для объекта А принадлежность террористам не такая уж и маленькая (6/34=18%) – заведомо выше доли террористов в населении страны.

Здесь оправдано установить маленький порог π (например, π=10%) и людей, у которых вероятность принадлежности «классу террористов» больше π, отправлять на дополнительную проверку.

Наивный Байес

А как получить вероятности классов, когда признаков >1?

В задаче про курение нецелевой признак был один («курит или нет»). А если нецелевых признаков больше?
Пусть

	Курит	Любит кошек	Пол (Y)
A	0	1	1
B	0	1	0
C	1	0	1
D	1	0	0
E	1	0	1

	Курит	Любит кошек	Пол (Y)
F	0	0	?

Вычисления для объекта F

Если применять формулу Байеса для объекта F, то нужно будет знать вероятности:
Pr(К'*Л' | М)
Pr(К'*Л' | Ж)
где
К={курит}, К'={не курит},
Л={любит кошек}, Л'= {не любит кошек},
*= союз «и».
А как вычислить эти вероятности?

Теорема о произведении вероятности

$$Pr(A*B)=Pr(A)*Pr(B),$$

если события A и B не зависят друг от друга.

Решая нашу задачу, сделаем допущение, что курение НЕ ЗАВИСИТ от любви к кошкам и наоборот (а в жизни это действительно так?). Тогда мы получаем:

$Pr(К'*Л' \mid М)= Pr(К' \mid М)* Pr(Л' \mid М)$,

$Pr(К'*Л' \mid Ж)= Pr(К' \mid Ж)*Pr(Л' \mid Ж)$,

а вероятности из правых частей равенств можно найти по таблице.

Решение примера

Из таблицы получаем вероятности:

$Pr(Ж)=2/5$ $Pr(М)=3/5$

$Pr(К|Ж)=1/2$ $Pr(К'|Ж)=1/2$

$Pr(К|М)=2/3$ $Pr(К'|М)=1/3$

$Pr(Л|Ж)=1/2$ $Pr(Л'|Ж)=1/2$

$Pr(Л|М)=1/3$ $Pr(Л'|М)=2/3$

	Курит	Любит кошек	Пол (Y)
A	0	1	1
B	0	1	0
C	1	0	1
D	1	0	0
E	1	0	1

Решение примера

По формуле Байеса для объекта F получаем вероятности принадлежности классам:

Pr(Ж|К’Л’)= Pr(К’Л’|Ж)*Pr(Ж)/Pr(К’Л’)
Pr(М|К’Л’)= Pr(К’Л’|М)*Pr(М)/Pr(К’Л’)

Используем предположение о независимости курения от любви к кошкам:

Pr(Ж|К’Л’)= Pr(К’|Ж)* Pr(Л’|Ж)* Pr(Ж)/Pr(К’Л’)
Pr(М|К’Л’)= Pr(К’|М)* Pr(Л’|М)* Pr(М)/Pr(К’Л’)

Решение примера

Подставляем числа:

$Pr(Ж|К'Л') = (1/2 * 1/3 * 2/5)/Pr(К'Л') = (1/15)/Pr(К'Л')$
$Pr(М|К'Л') = (2/3 * 1/3 * 3/5)/Pr(К'Л') = (2/15)/Pr(К'Л')$

Кстати, величину $Pr(К'Л')$ можно не вычислять (она нам не нужна), а воспользоваться равенством

$$Pr(Ж|К'Л') + Pr(М|К'Л') = 1$$

И получить

$Pr(Ж|К'Л') = 1/3$ $Pr(М|К'Л') = 2/3$

Решение примера

Таким образом, принадлежность классу мужчин в 2 раза выше принадлежности классу женщин.

$Pr(Ж|К'Л') = 1/3$ $Pr(М|К'Л') = 2/3$

Вот таким макаром и происходит классификация с помощью формулы Байеса.

Аналогично обрабатываются таблицы с бОльшим числом признаков.

Допущение о независимости признаков

Работа классификатора существенно использует предположение о независимости признаков друг от друга.

Но независимость признаков (когда коэфф. корр=0) не часто наблюдается на практике.

Поэтому наш классификатор называют наивным байесовским классификатором (НБК).

Когда между признаками существует сильная зависимость НБК может сильно ошибаться.

Работа НБК при зависимых признаках

В этой таблице признаки коррелируют (более того: они совпадают).
Проведем расчет принадлежности классам объекта F
1) без использования второго признака,
2) с двумя признаками.

	Курит	Любит кошек	Пол (Y)
A	0	0	1
B	0	0	0
C	1	1	1
D	1	1	0
E	1	1	1

	Курит	Любит кошек	Пол (Y)
F	0	0	?

Используем только один признак

Имеем: Pr(Ж)=2/5, Pr(К’|Ж)=1/2, Pr(М)=3/5, Pr(К’|М)=1/3

Pr(Ж|К’)= Pr(К’|Ж)* Pr(Ж)/Pr(К’)=(1/5)/Pr(К’)
Pr(М|К’)= Pr(К’|М)* Pr(М)/Pr(К’)=(1/5)/Pr(К’)

Из равенства Pr(Ж|К’)+Pr(М|К’)=1 находим
Pr(Ж|К’)=1/2, Pr(М|К’)=1/2

	Курит	Любит кошек	Пол (Y)
A	0	0	1
B	0	0	0
C	1	1	1
D	1	1	0
E	1	1	1

Используем оба признака

Имеем: $Pr(Ж)=2/5$, $Pr(К'|Ж)=1/2$, $Pr(Л'|Ж)=1/2$,
$Pr(М)=3/5$, $Pr(К'|М)=1/3$, $Pr(Л'|М)=1/3$

$Pr(Ж|К'Л')=Pr(К'|Ж)*Pr(Л'|Ж)*Pr(Ж)/Pr(К'Л')=(1/10)/Pr(К'Л')$
$Pr(М|К'Л')=Pr(К'|М)*Pr(Л'|М)*Pr(М)/Pr(К'Л')=(1/15)/Pr(К'Л')$

Из равенства $Pr(Ж|К'Л')+Pr(М|К'Л')=1$
находим
$Pr(Ж|К'Л')=3/5$, $Pr(М|К'Л')=2/5$

	Курит	Любит кошек	Пол (Y)
A	0	0	1
B	0	0	0
C	1	1	1
D	1	1	0
E	1	1	1

Ответ-то изменился!!!

… хотя никакой принципиально новой информации мы в таблицу не добавляли.

Вот к чему приводит наличие зависимостей между признаками!!!

Зависимые признаки нужно удалять (или как-то модифицировать) – об этом в теме «Отбор признаков»

Плюсы НБК:

1. Простота и быстрая работа. Используется в системах массового обслуживания. Например, при фильтрации спама.
2. В некоторых случаях НБК хорошо работает даже для коррелированных признаков (так что стОит рискнуть)))). Например, при фильтрации спама признаки, как правило, не являются независимыми – этот эффект и рассмотрим на след. слайдах.

Фильтрация спама (общая постановка)

Задача: есть текст письма, определить является ли оно спамом (класс 1) или нет (класс 0).

Какие признаки есть у письма?
Самый простой способ (count vectorizer): для каждого слова из словаря завести свой бинарный признак (он будет равен 1, если это слово встречается в рассматриваемом письме).

	лекарство	запой	геморрой	заработать	миллион	Путин	Спам?
A	1	1	1	0	0	0	1
B	1	1	0	0	0	0	1
C	1	0	1	0	0	0	1
D	0	0	0	1	1	0	1
E	1	0	0	0	1	0	1
F	0	0	0	0	1	1	0

Результат работы НБК

Был загружен «**SMS Spam Collection Data Set**». Нормальные сообщения там помечены словом «ham», а плохие «spam».

ham What you doing?how are you?

ham Ok lar... Joking wif u oni...

ham dun say so early hor... U c already then say...

ham MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*

ham Siva is in hostel aha:-.

ham Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor. Then he started guessing who i was wif n he finally guessed darren lor.

spam FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop

spam Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of Australia? Text MQUIZ to 82277. B

spam URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

Результат работы НБК

kNN для фильтрации спама показал precision=1 recall=0.61

НБК для фильтрации спама показал precision=0.98 recall=0.92 при пороге π=0.5

И это несмотря на то, что в данных много зависимых признаков.

Имейте в виду: в задаче про спам очень важно правильно установить порог π классификации.

Напомню смысл порога π: если

$$\Pr(\text{Спам} \mid \text{текст письма}) > \pi,$$

то письмо относим к классу «Спам».

В этой задаче порог должен быть большим (а почему?).

Вообще говоря…

… там много зависимых признаков (к тому же объем выборки небольшой).

Но НБК не обязательно на них начнет косячить (проверено на практике)!

Имейте в виду: в задаче про спам очень важно правильно установить порог π классификации.

Напомню смысл порога π: если

$$\Pr(\text{Спам} \mid \text{текст письма}) > \pi,$$

то письмо относим к классу «Спам».

В этой задаче порог должен быть большим (а почему?).

Показатели качества алгоритма, выдающего вероятности

Проблема:

Как оценить качество работы алгоритма, выдающего вероятности принадлежности к классам.
Самый простой способ: округлить вероятности до целых значений и получить обычный алгоритм классификации.

Объект	Рост	Вес	Пол (0-ж, 1-м) (предсказанный)	Ответ (округл)
A	180	70	0.75	1
B	150	45	0.1	0

А потом считать для округленных значений считать precision, recall и т.д.

Но:

такой способ не является оптимальным, особенно когда

1) классы не сбалансированы (число объектов одного класса много больше объектов другого класса);
2) цены ошибок за неправильные классификации объектов класса 1 и класса 0 различны.

Для вероятностей используют свои показатели качества

1. log-loss
2. ROC-AUC
3. …

На каких идеях основаны эти величины?
Смысл формулы для log-loss такой:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

модель штрафуется сильнее тогда, когда она сильно уверена в неправильном ответе.

На последующих слайдах мы рассмотрим ROC-AUC.

Качество классификации в первом приближении

Пусть алгоритм выдал оценки, как показано в табл. 1. Упорядочим строки табл. 1 по убыванию ответов алгоритма – получим табл. 2. Ясно, что в идеале её столбец «класс» тоже станет упорядочен (сначала идут 1, потом 0); в случае «слепого угадывания» будет случайное распределение 0 и 1.

Так вот: число нарушений порядка в табл. 2 определяет качество алгоритма.

id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

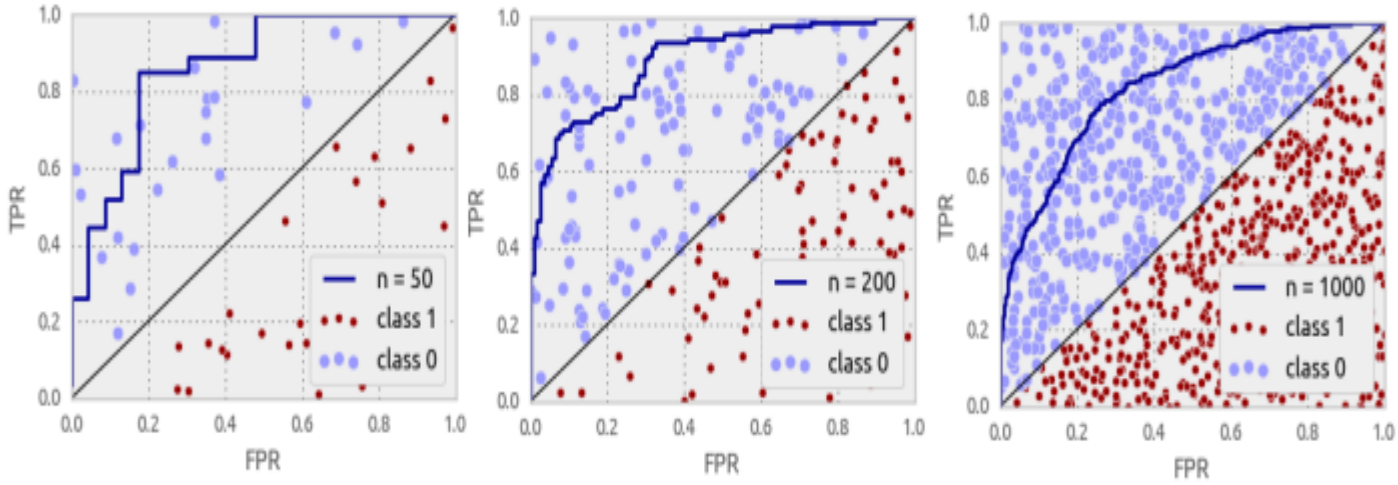
Табл. 1

id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2

ROC (receiver operating characteristic) - кривая

Выглядит примерно так (синяя синяя):

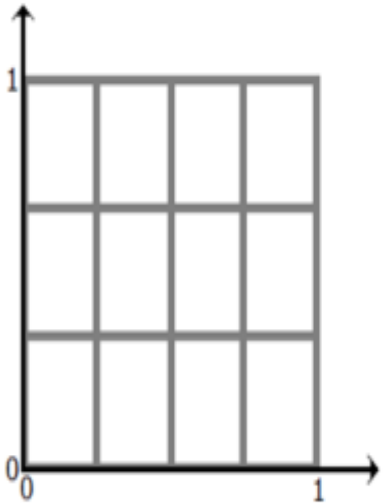


Чем больше площадь (AUC=area under curve) под ней – тем лучше.

А как строить ROC-кривую?

Строим ROC-кривую

Как и все показатели качества ROC-кривая строится по тестовой выборке.
Пусть n –число нулей в тестовой выборке, m – число единиц. Надо взять единичный квадрат на координатной плоскости, разбить его на m равных частей горизонтальными линиями и на n – вертикальными

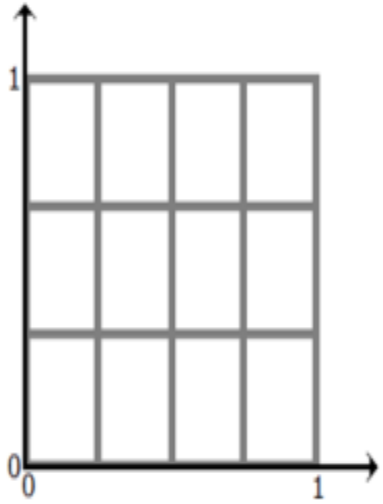


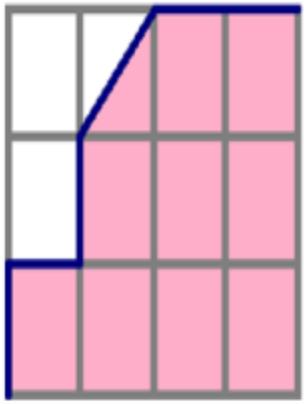
id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2

Строим ROC-кривую

Теперь будем просматривать строки в таблице сверху вниз и прорисовывать на сетке линии, переходя их одного узла в другой. Стартуем из точки (0, 0). Если значение метки класса в просматриваемой строке 1, то делаем шаг вверх; если 0, то делаем шаг вправо.



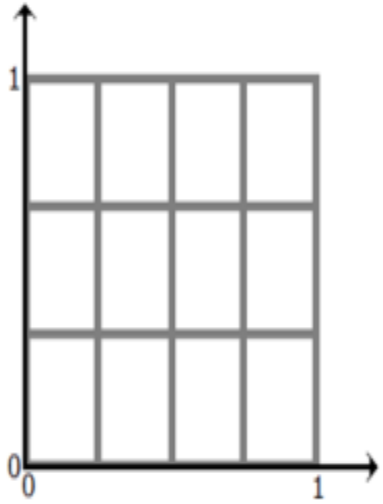


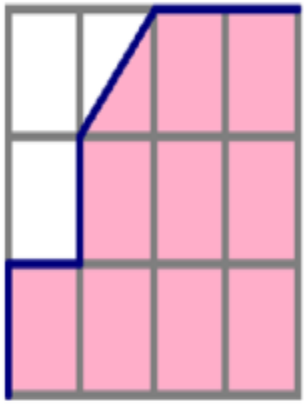
id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2

Важное правило:

если у нескольких объектов значения меток равны, то мы делаем сразу шаг в точку, которая на *a* блоков выше и *b* блоков правее,
где *a* – число единиц в группе объектов с одним значением метки, *b* – число нулей в ней.



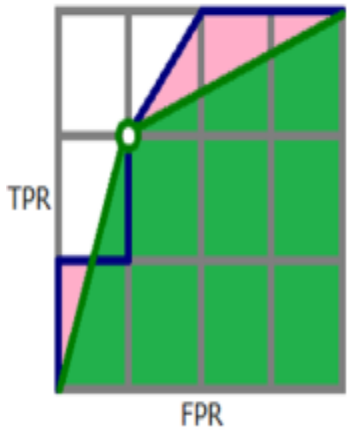


id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2

Как выбрать порог *p* для округления с помощью ROC-кривой?

Смысл порога *p*: если вероятность для объекта выше *p*, то объект относят к классу 1.
Для этого нужно взять точку с кривой. Порог p для нее будет равен вероятности объекта, который использовался при построении кривой в эту точку.
Точка на рисунке соотв. порогу 0.3

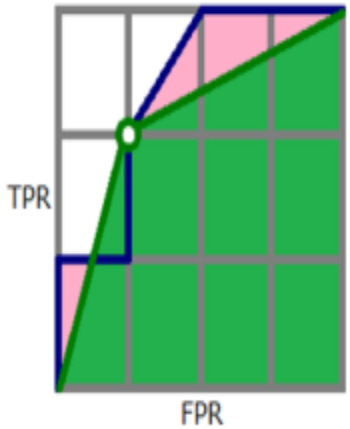


id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2

А что значат координаты точки на ROC-прямой?

Значение по горизонтали: FPR (false positive rate), доля объектов класса 0, которые неверно классифицированы нашим алгоритмом
Значение по вертикали: TPR (true positive rate, recall), доля объектов класса 1, которые верно классифицированы нашим алгоритмом.



id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2

Использованная литература

1. Т.Сегаран «Программируем коллективный разум» (фильтрация спама)
2. https://alexanderdyakonov.wordpress.com/2017/07/28/auc-roc-%D0%BF%D0%BB%D0%BE%D1%89%D0%B0%D0%B4%D1%8C-%D0%BF%D0%BE%D0%B4-%D0%BA%D1%80%D0%B8%D0%B2%D0%BE%D0%B9-%D0%BE%D1%88%D0%B8%D0%B1%D0%BE%D0%BA/ (про показатели качества) Оттуда же взяты и картинки

Приложение: логистическая регрессия

Формально:

Логистическую регрессию (ЛР) можно рассматривать как один из видов линейного классификатора, результаты которого преобразуются в вероятности.

Ниже я напомню, что делают линейные классификаторы.

Правило классификации

Если объект А описывается признаками $(x_1,x_2,\ldots x_n)$, то он принадлежит классу 1, если

$w_1x_1+w_2x_2+\ldots+w_nx_n>0,$

и объект А принадлежит классу -1, если

$w_1x_1+w_2x_2+\ldots+w_nx_n<0.$

Числа w_i являются параметрами модели и настраиваются по тренировочной выборке.

(Как вы уже заметили, в теории линейной классификации классы удобнее обозначать через 1 и -1).

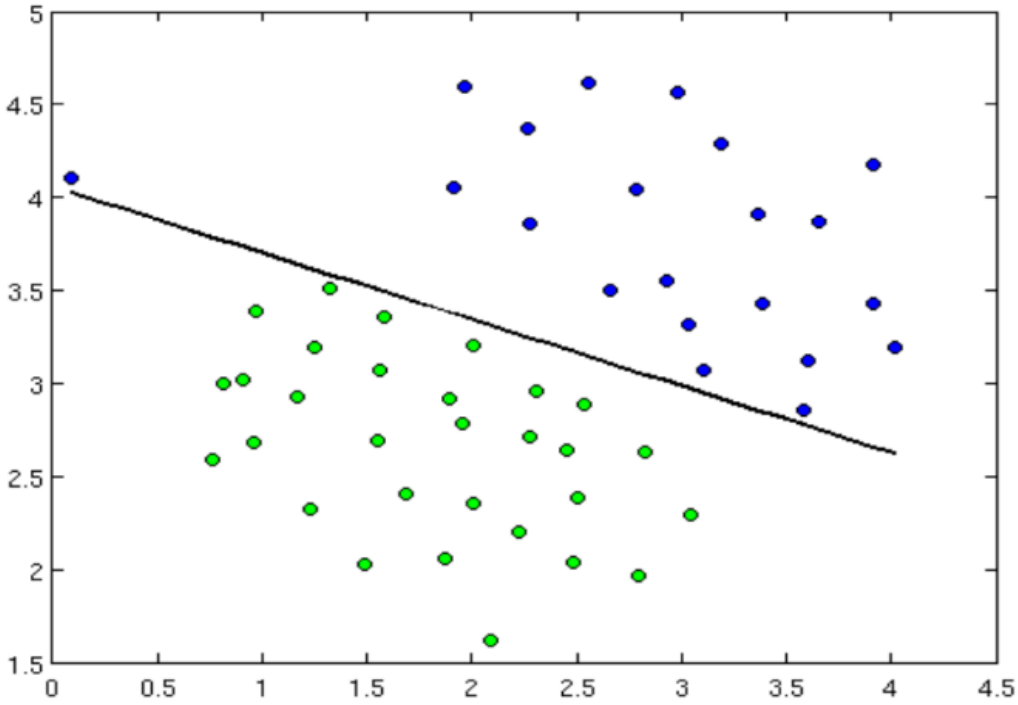
Геометрическая интерпретация

Фактически строится разделяющая гиперплоскость, разделяющая классы.

Например, если признаков 2,

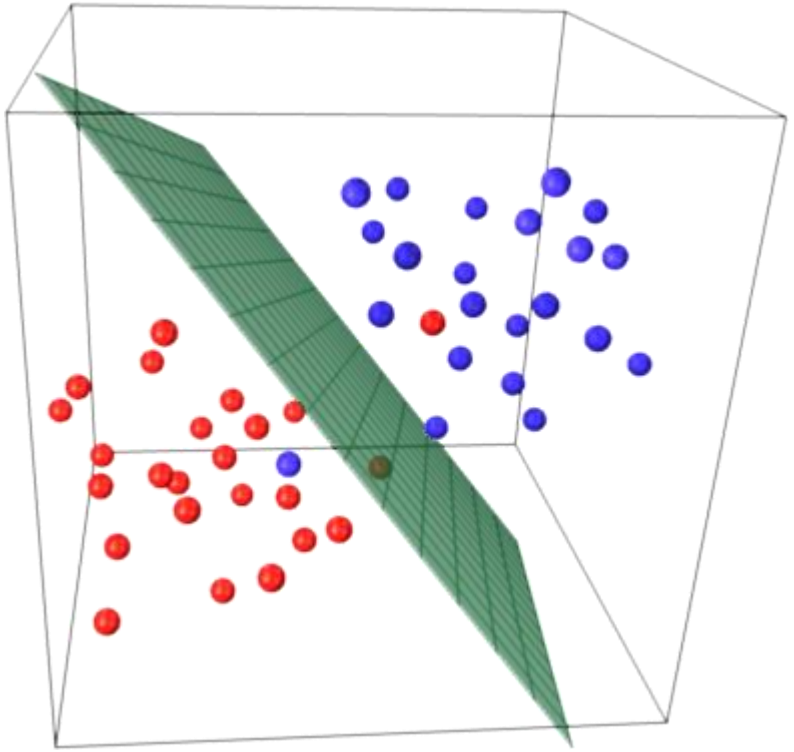
то строится

прямая.



https://habrahabr.ru/company/iticapital/blog/279963/

если признаков 3 штуки, то строится гиперплоскость в пространстве.



https://medium.com/data-science-group-iitr/support-vector-machines-svm-unraveled-e0e7e3ccd49b

Когда линейный классификатор ошибается?

Вспомним правило классификации:

если $w_1x_1+w_2x_2+\ldots+w_nx_n+w_0>0$, то объект относим к классу 1 (иначе к классу -1).

Когда модель допускает ошибку на тренировочной выборке?
Это происходит в 2-х случаях
1) когда $w_1x_1+w_2x_2+\ldots+w_nx_n>0$, но объект из класса -1
2) когда $w_1x_1+w_2x_2+\ldots+w_nx_n<0$ но объект из класса 1

Когда модель ошибается?

Иными словами, модель ошибается на объекте тренировочной выборки, если

$$M = y(w_1x_1 + w_2x_2 + \ldots + w_nx_n) < 0$$

здесь y – значение целевого признака объекта.

Величина M называется отступом.

Что минимизировать?

Введем обозначение

$$[M < 0] = \begin{cases} 1, \text{если } M < 0 \\ 0, \text{если } M > 0 \end{cases}$$

тогда для объектов тренировочной выборки нужно минимизировать функцию

$$\sum_{i=1}^{m} [M_i < 0] = \sum_{i=1}^{m} [y_i(w_1 x_1 + w_2 x_2 + \cdots + w_n x_n)]$$

где M_i – отступ i-го объекта тренировочной выборки, y_i – истинная метка класса i-го объекта тренировочной выборки.

Что минимизировать?

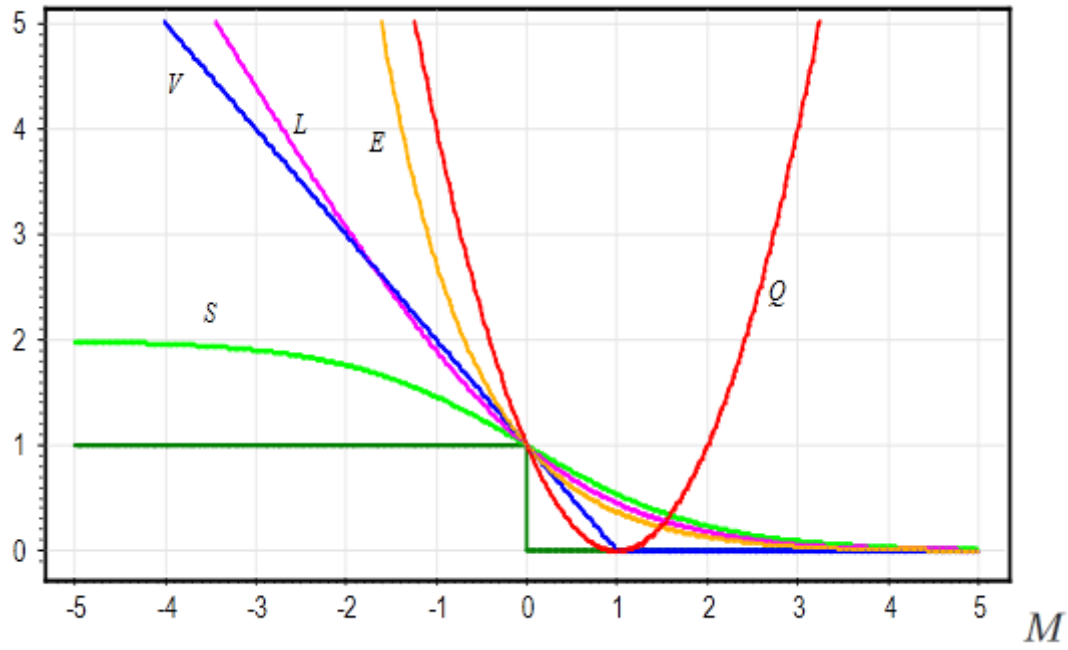
Есть проблема: эта функция

$$\sum_{i=1}^{m} [M_i < 0] = \sum_{i=1}^{m} [y_i(w_1 x_1 + w_2 x_2 + \cdots + w_n x_n)]$$

не дифференцируемая (нельзя взять производную). Искать минимум таких функций – это мазохизм.

Идея: найти дифференцируемую функцию *f*, которая приближенно равна [M<0]. Желательно также, чтобы выполнялось неравенство [M<0]<f (мажорирование).

И таких функций несколько



Все эти функции мажорируют функцию [M<0] (темно-зеленого цвета)

$$Q(M) = (1-M)^2 \quad \text{— квадратичная;}$$
$$V(M) = (1-M)_+ \quad \text{— кусочно-линейная;}$$
$$S(M) = 2(1+e^M)^{-1} \quad \text{— сигмоидная;}$$
$$L(M) = \log_2(1+e^{-M}) \quad \text{— логистическая;}$$
$$E(M) = e^{-M} \quad \text{— экспоненциальная.}$$

http://ru.learnmachinelearning.wikia.com/wiki

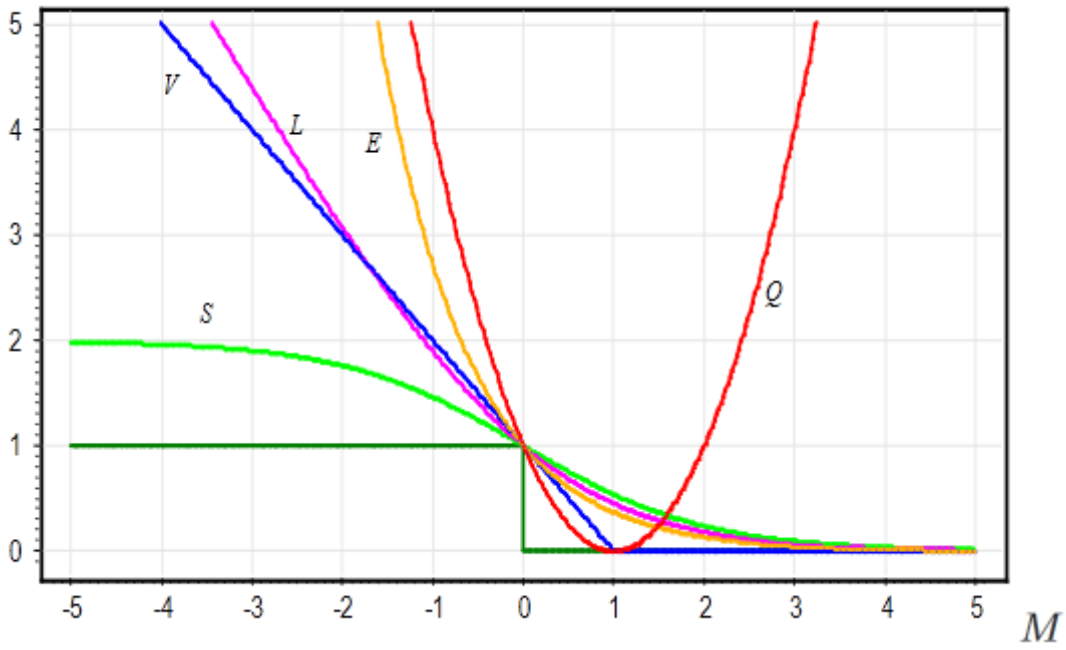
Схема построения линейного классификатора

1. Выбрать мажорирующую [M<0] функцию.
2. Составить выражение для минимизации.
3. Найти точку минимума выражения. Она даст оптимальные значения весов w_i.

ЛР укладывается в эту схему

Для мажорирования выражения нужно выбрать функцию

$$L(M)=\ln(1+e^{-M})$$



А дальше…

http://turboportal.ru/pics/vkpm/1132-vse-idet-po-planu-egor-letov.html

То есть

Нужно взять сумму

$$L(M)=ln(1+e^{-M})$$

по всем объектам тренировочной выборки (выражение будет зависеть от весов w_i) и найти точку минимума.

Например, для данных

Объект	X1	X2	Y
A	1	2	1
B	3	4	-1

Нужно минимизировать функцию

$$\ln\left(1+e^{-1(w_1 1+w_2 2+w_0)}\right)+\ln\left(1+e^{1(w_1 3+w_2 4+w_0)}\right)$$

Переход к вероятностям

Линейный классификатор действует так:

если $w_1x_1+w_2x_2+\ldots+w_nx_n+w_0>0$, то объект относим к классу 1 (иначе к классу -1).

А в ЛР делаем так:

вероятность принадлежности классу 1 равна $\sigma(w_1x_1+w_2x_2+\ldots+w_nx_n+w_0)$,
где $\sigma(z)$ – сигмоидная функция (сигмоид)

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