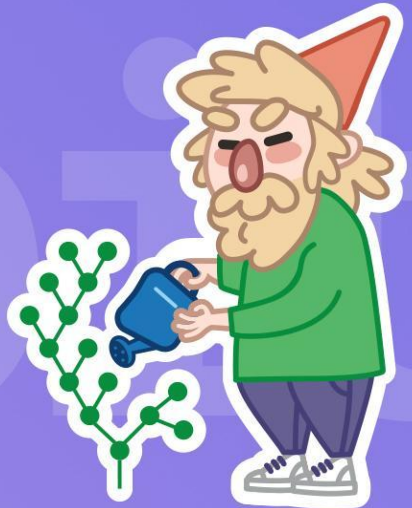
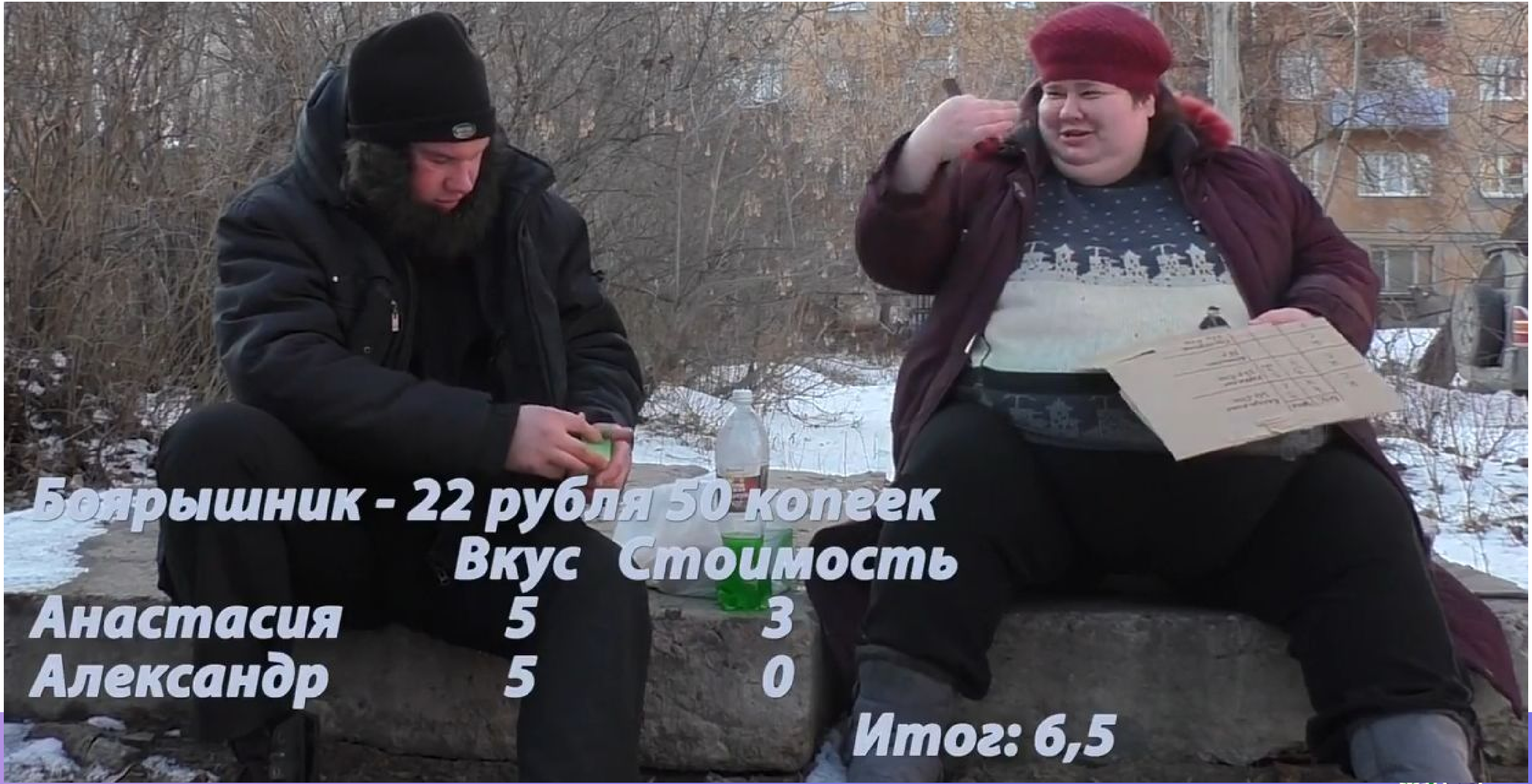
Наивный Байес и вероятностный подход к классификации

Лекция 6



Пример задачи регрессии (омская версия)



Ликбез: вероятность и формула Байеса

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 A1,A2,…,An – 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):

МM,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 (каких?):

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:

$Pr(\text{М}|\text{К})=Pr(\text{К}|\text{М})*Pr(\text{М})/Pr(\text{К})=0.7*0.4/0.34=28/34$

$Pr(\text{Ж}|\text{К})= Pr(\text{К}|\text{Ж})*Pr(\text{Ж})/Pr(\text{К})=0.1*0.6/0.34=6/34$

- 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 п для вероятности. Если вероятность принадлежности «классу мужчин» больше п, 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 п (например, п=10%) и людей, у которых вероятность принадлежности «классу террористов» больше п, 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 А и В 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' | М)= Pr(К' | М)* Pr(Л' | М),

Pr(К'*Л' | Ж)= Pr(К' | Ж)*Pr(Л' | 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')

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)=2/3

Pr(Ж|К’)= Pr(К’|Ж)* Pr(Ж)/Pr(К’)=(1/5)/Pr(К’)
Pr(М|К’)= Pr(К’|М)* Pr(М)/Pr(К’)=(2/5)/Pr(К’)

Из равенства Pr(Ж|К’)+Pr(М|К’)=1 находим
Pr(Ж|К’)=1/3, Pr(М|К’)=2/3

	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)=2/3, Pr(Л′|М)=2/3

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′)=(4/15)/Pr(К′Л′)

Из равенства Pr(Ж|К′Л′)+Pr(М|К′Л′)=1 находим
Pr(Ж|К′)=3/11, Pr(М|К′)=8/11

	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.

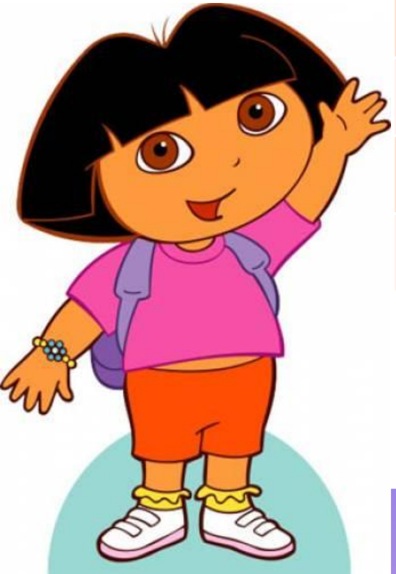
Фильтрация спама (общая постановка)

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

Какие признаки есть у письма?
Самый простой способ: для каждого слова из словаря завести свой бинарный признак (он будет равен 1, если это слово встречается в рассматриваемом письме).
Получится примерно как на след. слайде…

Помогите Даше найти зависимые признаки

	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



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 п классификации.

Напомню смысл порога п: если

Pr(Спам | текст письма)> п,

то письмо относим к классу «Спам».

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

Использованная литература

1. Т.Сегаран «Программируем коллективный разум» (фильтрация спама)