Теория вероятностей, ДЗ №2

Мироненко Ольга

Заранее прошу прощения за длинное решение к задаче 3: мне было интересно разобраться во взаимосвязях различных параметров в ней, для себя, своего понимания.

Задание 1

Пусть событие X - пациент болен X, тогда событие $\overline{X}$ - пациент не болен X. Пусть событие F - пациент - женщина, тогда событие $\overline{F}$ - пациент - мужчина.

Из условия задачи известно, что $P(F) = P(\overline{F}) = 0.5$, $P(A|F) = 0.05$, $P(A|\overline{F})$ = 0.15.

Нужно найти вероятность того, что пациент, сдавший положительный анализ на заболевание X, - мужчина ($P(\overline{F}|A)$).

Из теоремы Байеса: $P(\overline{F}|A) = P(\overline{F})\frac{P(A|\overline{F})}{P(A)}$.

Чтобы найти распространённость болезни X в среднем по пополяции, воспользуемся формулой полной вероятности: $P(A) = P(AF) + P(A\overline{F}) = P(A|F) * P(F) + P(A|\overline{F}) * P(\overline{F}) = 0.05 * 0.5 + 0.15 * 0.5$ = 0.1.

Тогда $P(\overline{F}|A) = 0.5 * 0.15 /$ 0.1 = 0.75.

Таким образом, вероятность того, что сдавший положительный анализ на заболевание X, - мужчина, составляет 75%.

Задание 2

Пусть событие I - анализ выполнен на первом приборе, тогда событие $P(\overline{I})$ - анализ выполнен на втором приборе. Пусть также событие A - в результатах анализа получена абракадабра, тогда событие $\overline{A}$ - результаты без абракадабры.

Из условия задачи известно, что $P(I) = 0.9$, $P(\overline{I}) = 0.1$, $P(A|I) = 0.01$, $P(A|\overline{I}) = 0.1$.

Нужно найти вероятность того, что полученная абракадабра была выдана первым прибором ($P(I|A)$).

Из теоремы Байеса: $P(I|A) = P(I)\frac{P(A|I)}{P(A)}$.

Чтобы найти среднюю распространённость абракадабры в результатах работых обоих приборов, воспользуемся формулой полной вероятности: $P(A) = P(AI) + P(A\overline{I}) = P(A|I) * P(I) + P(A|\overline{I}) * P(\overline{I}) = 0.01 * 0.9 + 0.1 * 0.1$ = 0.019.

Тогда $P(I|A) = 0.9 * 0.01 /$ 0.019 = 0.47.

Таким образом, вероятность того, что абракадабра была получена именно на первом приборе, составляет 47%.

Задание 3

Для удобства я буду использовать обозначение H для события ``здоров'' и $\overline{H}$ - для события ``болен''.

В задаче необходимо оценить апостериорную вероятность того, что пациент, сдавший отрицательный анализ, на самом деле здоров. Используя формулу Байеса, мы можем это сделать с помощью ``апдейта'' априорной вероятности того, что пациент здоров, а именно:

$$P(H|-) = P(H)\frac{P(-|H)}{P(-)}$$

Частоту получения отрицательного теста, в среднем, по больным и здоровым пациентам, мы можем найти по формуле полной вероятности, которая складывается из вероятности ``встретить'' здорового пациента с отрицательным тестом и вероятности ``встретить'' больного пациента с отрицательным тестом в общей популяции, каждая из которых зависит от вероятности ``встретить'' здорового/больного пациента в общей популяции, соответственно, и вероятности ``встретить'' внутри этих групп пациентов с отрицательными тестами, а именно:

$$P(-) = P(H-) + P(\overline{H}-) = P(-|H) * P(H) + P(-|\overline{H}) * P(\overline{H})$$

Из определения, вероятность получения отрицательного теста для здорового пациента, $P(-|H)$, называется специфичностью (TNR), а вероятность получения положительного теста для больного пациента, $P(+|\overline{H})$, - чувствительностью (TPR). Соответственно, вероятность получения отрицательного теста для больного пациента составит $P(-|\overline{H}) = 1 - P(+|\overline{H}) = 1 - TPR$.

Таким образом, можем переписать формулу Байеса в следующем виде:

$$P(H|-) = P(H)\frac{TNR}{P(H) * TNR + P(\overline{H}) * (1 - TPR)}$$

Подставив исходные данные, получим: $P(H|-) = 0.95\frac{0.8}{0.95*0.8+0.05*(1-0.9)} = 0.99$

Для понимания, каким образом вероятность того, что пациент, сдавший отрицательный тест, действительно здоров, зависит от параметров, участвующих в оценке, на мой взгляд, лучше упростить полученную выше формулу следующим образом (при ненулевых вероятности ``встретить'' здорового пациента и специфичности теста):

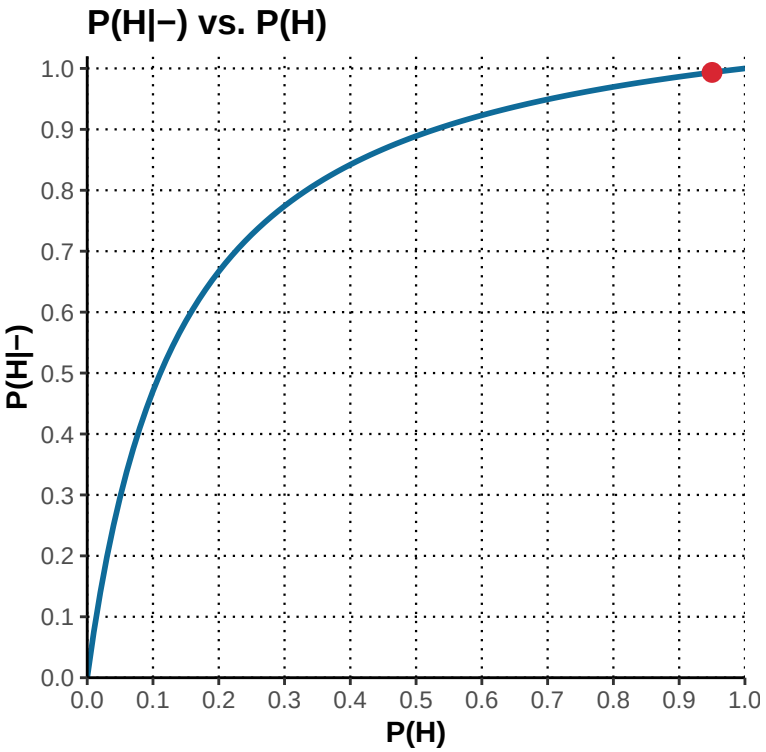
$$P(H|-) = \frac{1}{1 + \frac{P(\overline{H})}{P(H)} * \frac{1-TPR}{TNR}} = \frac{1}{1 + \frac{P(\overline{H})}{P(H)} * \frac{FNR}{TNR}} = \frac{1}{1 + \frac{1}{Odds_{H_0}} * \frac{FNR}{TNR}}$$

где $FNR = P(-|\overline{H})$ - это вероятность получения ложноотрицательного результата, $Odds_{H_0} = \frac{P(H)}{P(\overline{H})} = \frac{P(H)}{1-P(H)}$ - априорный шанс здоровья (отсутствия тестируемого заболевания), а соотношение $\frac{FNR}{TNR} = LR_-$ называется отношением правдоподобия для отрицательного теста и показывает, во сколько раз чаще встречаются ложноотрицательные результаты по сравнению с истинно отрицательными (чем оно больше, тем менее полезен такой тест в диагностике).

Получается, что **апостериорная вероятность здоровья прямо пропорциональна априорному шансу здоровья** (а поскольку шанс здоровья растёт, хотя и нелинейно, с увеличением априорной вероятности здоровья, то и от априорной вероятности здоровья) и **обратно пропорциональна отношению правдоподобия для отрицательного теста** (значит, с ухудшением качества этого

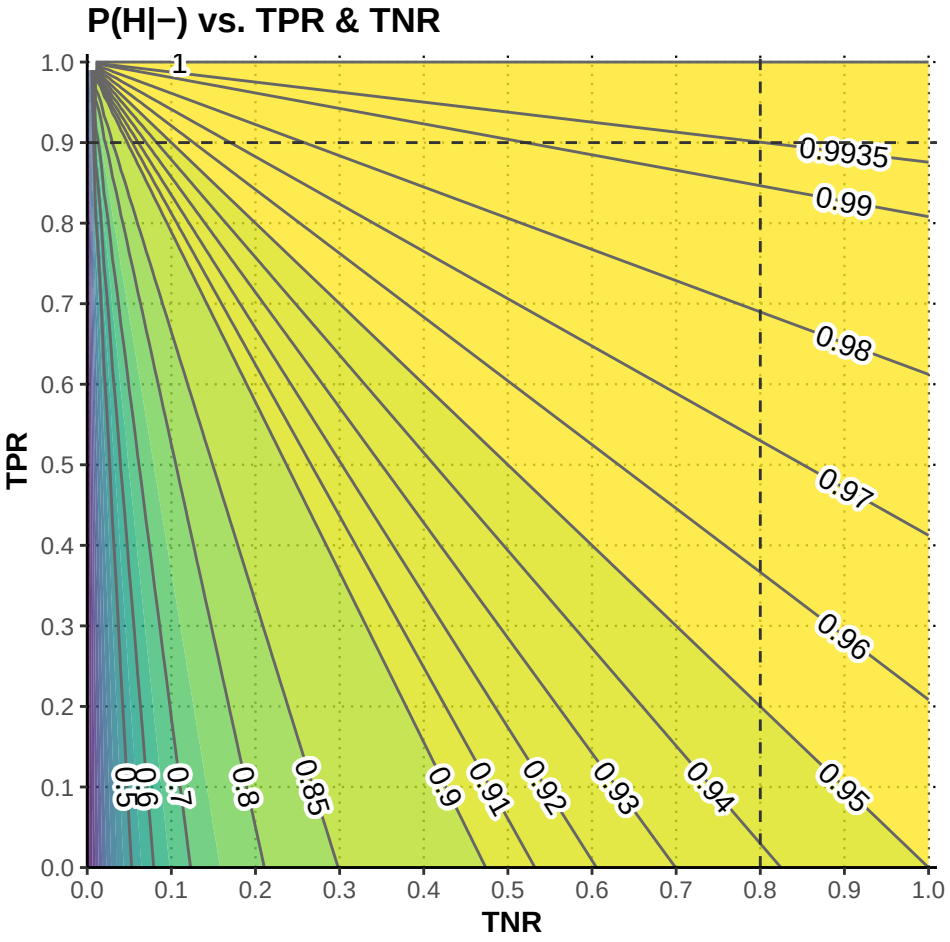
теста, оцененного с помощью данного показателя, апостериорная вероятность здоровья также уменьшается).

При данных в задаче чувствительности и специфичности теста изобразим, как апостериорная вероятность здоровья будет зависеть от априорной (красным показана точка со значениями, соответствующими исходным условиям задачи):



То есть **с уменьшением распространённости заболевания апостериорная вероятность здоровья увеличивается невозрастающими темпами, при фиксированных значениях чувствительности и специфичности**. Если последние обе будут равны по 0.5, то зависимость превратится в линейную. При прочих значениях кривая будет ближе или дальше от биссектрисы.

Что касается взаимосвязи апостериорной вероятности здоровья с чувствительностью и специфичностью теста, то при фиксированной чувствительности, она увеличивается с ростом специфичности, так же, как и при фиксированной специфичности, увеличивается с ростом чувствительности. Вместе с тем, одинаковый прирост того или другого даёт разный по размеру эффект в отношении апостериорной вероятности. Я постаралась изобразить это на графике ниже при данной в задаче распространённости заболевания (сплошными линиями показаны изолинии для апостериорных вероятностей от 0.5 до 1, в том числе для полученной в задаче), жирным пунктиром - линии для данных в задаче значений чувствительности и специфичности):

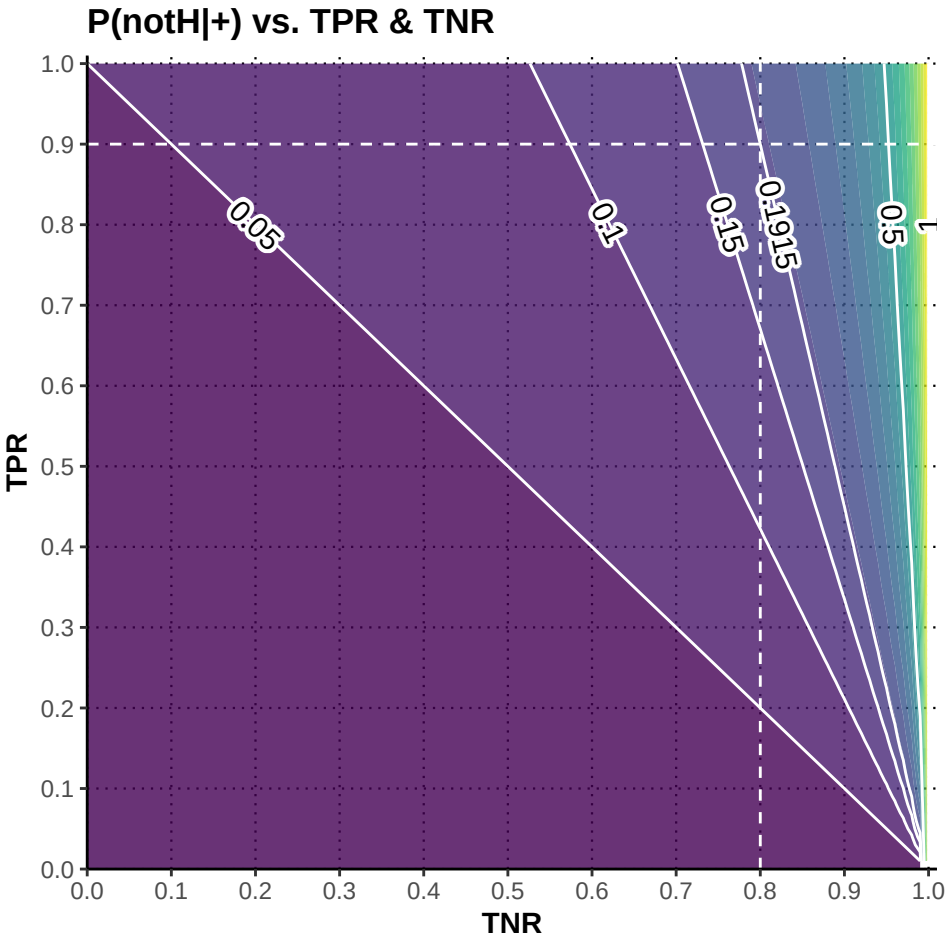


Получается, что **при данной распространённости заболевания, мы получим бОльшую, по сравнению с априорной, апостериорную вероятность здоровья при любых сочетаниях чувствительности и специфичности, попадающих в правый верхний угол квадрата, представленного на графике (над диагональю)**. Причём это будет верно для любой распространённости заболевания, не только для данной в задаче: все сочетания чувствительности и специфичности, лежащие в этой части графика, будут способствовать бОльшему апдейту априорной вероятности здоровья. Эти сочетания можно описать любым из соотношений:

- $TNR > 1 - TPR$, или $TNR > FNR$, или $LR_- < 1$)
- $TPR > 1 - TNR$, или $TPR > FPR$, или $LR+ > 1$, где $FPR = P(+|H)$ - вероятность получения ложноположительного результата, а $LR_+ = \frac{TPR}{FPR}$ - отношение правдоподобия для положительного теста.

Если у нас уже есть конкретный тест, с некоторой чувствительностью и специфичностью, то повысить свою уверенность в том, что пациент, сдавший отрицательный тест, действительно здоров, можно либо повышая и чувствительность, и специфичность теста, либо повышая одно, но понижая другое в определённых пределах (так, чтобы не опускаться ниже изолинии). В последнем случае если мы меняем специфичность на некоторую величину Δ_{TNR}, такую что $-TNR_0 < \Delta_{TNR} \leq 1 - TNR_0$ (TNR_0 - текущее значение специфичности), то для неуменьшения апостериорной вероятности здоровья чувствительность мы можем поменять соответствующим образом на величину, меньшую этой, и она будет тем меньше, чем меньше распространённость заболевания в популяции. Иными словами, **при прочих равных условиях, в плане увеличения апостериорной уверенности в здоровье пациента с отрицательным тестом по сравнению с априорной увеличение чувствительности теста на некоторую величину даст бОльший эффект, чем увеличение специфичности теста на эту же величину**.

При этом вариант теста, когда всем обследуемым ставится ``+" (чувствительность такого теста 1, а специфичность - 0), нам не подходит, поскольку при стремлении специфичности к нулю апостериорная вероятность здоровья также будет стремиться к нулю (хотя в точке $TNR = 0, TPR = 1$ она и будет не определена). Что касается тестов с близкой к 1 чувствительностью и минимальной специфичностью, то да, такой тест будет практически всегда выдавать ``+" здоровым пациентам, но уж если покажет ``-", то мы практически точно можем быть уверены, что пациент здоров :) С другой стороны, апостериорная вероятность болезни при положительном результате такого теста будет равна априорной, т.е. тест будет, по сути, неинформативным (тут мы уже возвращаемся к примеру, который был на лекции). Ниже приведу иллюстрацию для этого (здесь изолинии показывают апостериорную вероятность болезни при положительном тесте от 0.05 (априорная вероятность = распространённость болезни) до 1, в том числе полученную в задаче). По ней также наглядно видно, что **при прочих равных условиях, в плане увеличения апостериорной уверенности в болезни пациента с положительным тестом по сравнению с априорной увеличение специфичности теста на некоторую величину даст бОльший эффект, чем увеличение чувствительности теста на эту же величину**.



* * *

Также мне показалось интересным дальше преобразовать полученное выше соотношение для апостериорной вероятности здоровья, - и получить следующее:

$$\frac{P(H|-)}{1-P(H|-)} = \frac{Odds_{H_0}}{LR_-}$$

Левая часть этого соотношения - это апостериорный шанс здоровья (отсутствия тестируемого заболевания), т.е.:

$$Odds_{H_1} = \frac{Odds_{H_0}}{LR_-}$$

Таким образом, **апостериорный шанс здоровья прямо пропорционален априорному шансу здоровья и обратно пропорционален отношению правдоподобия для отрицательного теста**.

Аналогично можно было получить выражение для апостериорной вероятности болезни при положительном тесте:

$$Odds_{\overline{H}_1} = Odds_{\overline{H}_0} LR_+$$

Задание 4

Запишем формулу полной вероятности для события A:

$P(A) = P(AB) + P(A\overline{B}) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$

Если $P(A) = P(A|B)$, получим $P(A|B) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$

Отсюда: $P(A|B) - P(A|B)P(B) = P(A|\overline{B})P(\overline{B})$

$P(A|B)(1 - P(B)) = P(A|\overline{B})P(\overline{B})$

$P(A|B)P(\overline{B}) = P(A|\overline{B})P(\overline{B})$

При непустом множестве событий $\overline{B}$ получим $P(A|B) = P(A|\overline{B})$.

Таким образом, из $P(A) = P(A|B)$ следует, что $P(A) = P(A|\overline{B})$. Ч.т.д.

Задание 5

Пусть у нас есть некоторое воздействие X и исход Y, тогда, по определению, отношение рисков исхода в группе с воздействием X по сравнению с группой без него, можно будет оценить следующим образом: $RR = \frac{P(Y|X)}{P(Y|\overline{X})}$.

Опять же, из определения, события X и Y будут независимыми, если $P(Y|X) = P(Y)$. Исходя из предыдущего задания мы это также можем записать в виде $P(Y|\overline{X}) = P(Y)$. Подставив полученные выражения в выражение для отношения рисков, получим, что для независимых событий X и Y $RR = 1$. Ч.т.д.