Правило Байеса, Байесовский классификатор и комбинирование свидетельств

Итак, у нас есть три класса образов-слов, и одно загаданное слово. Открытые каждой буквы слова для нас является затратной операцией. Вероятности появления каждого слова равны 1/10.

Класс C_1	Класс C_2	Класс C_3
GRAF	ITAK	AVAX
SCAD	SORT	
PORT	MRAK	
TORT	IVAN	
	BORN	

$$P("??A?") = 6/10$$
$$P("S???") = 2/10$$
$$P("S???"|"??A?") = 1/6$$

Вот именно тот факт, что $P("S???") \neq P("S???"|"??A?")$, и дает нам условную зависимость "S???" от "??A?" (и наоборот).

Давайте пример с классами разберем, чтоб совсем под формулу из «Википедии» подогнать:

$$P("S???"|C_2) = 1/5$$
$$P("S???"|C_2, "??A?") = 0$$

Не равны, условной независимости нет. Кстати, именно это и имелось в виду, когда говорилось, что данные первого теста можно использовать для выбора второго теста, то есть значение какого признака определять следующим (хотя это и трудно с вычислительной точки зрения в большинстве задач, но в медицине, к примеру, именно так и поступают).

В «Википедии» приведен пример со спамом – анализ электронной почты. Пример тоже спорный, однако уже ближе к теме наивного классификатора. Если рассматривать некоторое произвольное письмо, то определить зависимость 31-го слова от 2-го весьма сложно (очевидно, если такая зависимость и есть, то очень слабая). Хотя если известно, что 2-е слово «*лукоморья*», то вероятность того, что 31-е слово – «*леший*», гораздо выше, чем у слов «*квазитопологический*» или «*уважаемый*». Вот не знали бы мы 2-го слова, считали бы априорно, что

P(«*уважаемый*») > P(«*леший*») > P(«*квазитопологический*»)

Возиться с такими зависимостями очень сложно, практически нереально, вот и считаем, что вероятность слов в письмах не зависит от других слов, уже найденных в письме, тем более, что анализируются не все слова, а только определенные слова-маркеры. В этом случае метод комбинирования свидетельств с предположением об их независимости – лучшее, что у нас есть, но при этом все-таки приближение.

Итак, есть письмо электронное, нужно определить вероятности, с которыми оно относится к классам «спам» и «не спам». Пусть у нас есть два слова (*i*-е и *j*-е), к примеру, «*таймшер*» и «*кино*». Для этих слов известно, что с вероятностью 0.01 «*таймшер*» встречается в нормальных письмах, и с вероятностью 0.08 – в спаме. Сумма не 1, потому что в большинстве писем (спам и не спам) это слово не встречается. Вот вероятность того, что письмо спам, если в нем встретилось слово «*таймшер*», может быть 0.9, а вероятность того, что не спам – 0.1, сумма тут должна быть 1. Аналогичные вероятности должны быть заданы и для слова «*кино*», а также нам потребуется значение априорной вероятности того, что письмо – спам, например, среди всех писем 80% – спам, это и будем использовать.

Вводим классы *Spam* и *NotSpam*, исходные данные:

$$P(\text{«таймшер»}|Spam) = 0.08 \qquad P(\text{«таймшер»}|NotSpam) = 0.01$$

$$P(\text{«кино»}|Spam) = 0.02 \qquad P(\text{«кино»}|NotSpam) = 0.05$$

$$P(Spam) = 0.08 \qquad P(NotSpam) = 0.02$$

По формулам считаем отношение правдоподобия (ну, или можно логарифм правдоподобия посчитать, смотря для чего):

$$\frac{P(Spam|\text{«таймшер»\&«кино»})}{P(NotSpam|\text{«таймшер»\&«кино»})} = \frac{P(Spam)}{P(NotSpam)} \prod_i \frac{P(\omega_i|Spam)}{P(\omega_i|NotSpam)}$$

$$P\big(Spam\big|\,\text{«таймшер»\&«кино»}\big) + P\big(NotSpam\big|\,\text{«таймшер»\&«кино»}\big) = 1$$

Подставляем значения, решаем систему уравнений, получаем результат:

$$P\big(Spam\big|\,\text{«таймшер»\&«кино»}\big) = .9275362319$$

$$P(NotSpam\,|\,\text{«таймшер»\&«кино»}) = .07246376812$$

Получили вероятности, можно делать вывод.

Теперь, собственно, по поводу применимости первого или второго варианта. Первый вариант, строго говоря, более предпочтителен, так как позволяет учесть связь между свидетельствами. К сожалению, такой подход далеко не всегда работает – для реальных примеров нужно хранить большой объем информации, сопоставимый с полным совместным распределением. У нас вообще были явно заданы все возможные образы. На практике такое обычно не применимо. Если к первому

примеру применить второй способ, получим значения неверные (отличающиеся), как раз из-за неверного предположения о независимости свидетельств (метод потому и называется «наивный»). Однако второй подход гораздо проще в реализации. Очевидно, что первый способ применить к фильтрации спама невозможно. Второй же гораздо проще в реализации, на реальных задачах показывает неплохие результаты, и работает.