Мистическая Оптимальность Наивной Байесовской Модели: Оценка Вероятности в Системе Классификаторов

Купервассер Олег Юрьевич

ООО «ТРАНЗИСТ ВИДЕО», резидент Технопарка Сколково

В статье доказана оптимальность наивной байесовской модели.

Глава 2.1.1 Введение

Байесовы Классификаторы широко в настоящее время используются для распознания, идентификации и получения знания. Области применения - например, обработка изображения, медицина, химия (QSAR) [1] на Рис. 1.

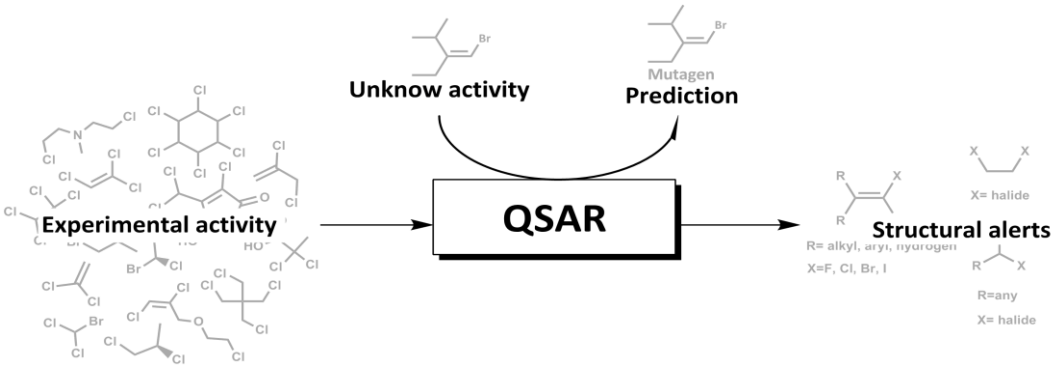

Рис. 1 QSAR

Позвольте нам дать пример использования методов QSAR в статьях [2,3]: «Молекулярное распознавание и связывание, выполняемого белками, являются фоном всех биохимических процессов в живой клетке. В частности, обычный механизм действия лечебного препарата - эффективное закрепление и ингибирование деятельности целевого белка. Прямое моделирование молекулярных взаимодействий в комплексах ингибитор- белок - основа современного метода компьютерной разработки лекарств, однако такое моделирование является чрезвычайно сложной проблемой … В предлагаемом методе, подобие для связываемого участка осуществляется распознаванием существования химически и пространственно аналогичных областей по отношению к области связывания. Мы представляем новый метод определения местного подобия для участка связывания, основанный на анализе полной окружающей среды белка для фрагментов лиганда. Сравнение области связывания изучаемого белка (цель) с трехмерной структурой другого белка (аналога) в комплексе с лигандом позволяет фрагментам лиганда от аналогичного комплекса быть соотнесенным с их возможным положением в целевом участке, так, чтобы полная окружающая среда белка для фрагмента и его аналога были подобны. Соответствующие окружения - подобные области и фрагменты, примыкающие к целевому участку, рассматриваются как типовые образцы связывания. Набор таких целевых образцов связывания, полученных из базы данных аналоговых комплексов, формирует подобную облаку структуру (облако фрагментов), которая является сильным инструментом для компьютерной разработки лекарств»

Особое значение Байесовские Классификаторы имею в медицинской диагностике и биоинформатике. Очень хороший пример может быть найден в статье [4].

Однако, таинственным путем Наивный Байесовский Классификатор обычно дает очень хорошее представление распознания. Более сложные модели Байесового Классификатора не могут улучшить его значительно [1]. Мы даем простое доказательство оптимальности Наивного Байесовского Классификатора, которое может объяснить этот интересный факт.

В статье [5] авторы объясняют это замечательное свойство. Однако, они используют некоторое предположение (Zero-One Loss), которое уменьшает универсальность и общность этого рассмотрения. Мы даем более общее доказательство оптимальности Наивного Байесовского Классификатора [1]. Последующее интересное развитие задачи было сделано в [6] (2004), [7] (2006). Однако, к сожалению эта статья не включают анализа предыдущей [1].

Позвольте нам сформулировать коротко основную задачу, которую мы пытаемся решить. Предположиv, что у нас есть набор некоторых объектов и набор переменных, которые характеризуют эти объекты. Для каждого объекта мы знаем вероятностное распределение для каждой переменной. Однако, у нас нет никакой информации о корреляциях переменных. Теперь, предположим, что мы знаем значения переменных для некоторого объекта из набора объектов. Какова вероятность, что эта выборка соответствует некоторому объекту? Это - типичная задача распознания при условии неполноты информации.

Позвольте нам рассмотреть самый простой случай, когда никакие корреляции не существуют между переменными. В этом случае, Наивная Байесовская модель - точное решение задачи. Мы докажем, что для случая, когда мы ничего не знаем о корреляциях - Наивная Байесовская модель не точное, но оптимальное решение в некотором смысле. Более детально, мы докажем, что Наивная Байесовская модель дает минимальную среднюю ошибку по всем возможным моделям корреляций. Мы предполагаем в этом доказательстве, что у всех моделей корреляций есть равная вероятность. Мы думаем, что этот результат может объяснить описанную выше таинственную оптимальность Наивной Байесовской модели.

Материал организован следующим образом. В разделе 2.1.2 мы даем точное математическое определение задачи для двух переменных и двух объектов. В разделе 2.1.3 мы определяем обозначения. В разделе 2.1.4 мы даем общую форму условной вероятности для всех возможных корреляций наших переменных. В разделе 2.1.5 мы находим ограничения функций, описывающих корреляции. В разделе 2.1.6 мы даем определение расстоянию между двумя вероятностями (корреляционными) моделями. В разделе 2.1.7 мы находим ограничения для наших основных функций. В разделе 2.1.8 мы решаем свою главную задачу - мы доказываем оптимальность Наивной Байесовской модели для равномерного распределения всех возможных корреляций. В разделе 2.1.9 мы находим среднюю ошибку между Наивной Байесовской моделью и фактической моделью при равномерном распределении всех возможных корреляций. В разделе 2.1.10 мы расматриваем случай более чем двух переменных и объектов. Последний раздел - заключение.

Глава 2.1.2 Постановка задачи

Пусть A - случайная величина, со значениями $0,1$. Предположим, что априорная вероятность $P(A) = P(A = 1)$ известна и обозначим ее θ. Пусть X_1, X_2 - две случайные величины, со значениями в некотором множестве, например $]-\infty;+\infty[$. Мы имеем следующую информацию: $X_1 = x_1$ и $X_2 = x_2$ (полученную из измерения). Кроме того, у нас есть две системы - "классификаторы", которые для данных x_1 и x_2 дают:

$$P(A = 1/X_1 = x_1) = P(A/x_1) \doteq \alpha ,$$

$$P(A = 1/X_2 = x_2) = P(A/x_2) \doteq \beta .$$

Мы хотим оценить вероятности $P(A = 1/X_1 = x_1, X_2 = x_2) = P(A/x_1, x_2)$ в терминах α, β и θ. Более определенно, мы хотим найти функцию $\Gamma_{opt}(\alpha, \beta, \theta)$, которая в среднем является наилучшим приближением для $P(A/x_1, x_2)$ в смысле, определенным явно в продолжении (см. рис. 2).

Глава 2.1.3 Предварительные обозначения

$\rho_{X_1,X_2}(x_1, x_2)$ - совместная ФПВ (функция плотности вероятности) X_1 и X_2.

$\rho_{X_1,X_2/A}(x_1, x_2) \doteq h(x_1, x_2)$ - совместная ФПВ X_1 и X_2 для $A = 1$. В терминах $h(x_1, x_2)$ и θ мы можем написать $P(A/x_1, x_2)$ следующим образом:

$$P(A/x_1, x_2) = \frac{\theta h(x_1, x_2)}{\theta h(x_1, x_2) + (1-\theta)\bar{h}(x_1, x_2)}, \quad (1)$$

Где

$\bar{h}(x_1, x_2) \doteq \rho_{X_1,X_2/\bar{A}}(x_1, x_2)$ - совместная ФПВ X_1 и X_2 для $A = 0$.

Мы имеем:

$$\rho_{X_1}(x_1) = \int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1, x_2)dx_2,$$

$$\rho_{X_2}(x_2) = \int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1, x_2)dx_1,$$

$$h_1(x_1) \doteq \rho_{X_1/A}(x_1) = \int_{-\infty}^{+\infty} h(x_1, x_2)dx_2,$$

$$h_2(x_2) \doteq \rho_{X_2/A}(x_2) = \int_{-\infty}^{+\infty} h(x_1, x_2)dx_1,$$

$$\bar{h}_1(x_1) \doteq \rho_{X_1/\bar{A}}(x_1) = \int_{-\infty}^{+\infty} \bar{h}(x_1, x_2)dx_2,$$

$$\bar{h}_2(x_2) \doteq \rho_{X_2/\bar{A}}(x_2) = \int_{-\infty}^{+\infty} \bar{h}(x_1, x_2)dx_1.$$

Глава 2.1.4 Общая форма $P(A/x_1, x_2)$

Определим функции $g(x_1, x_2)$ и $\bar{g}(x_1, x_2)$

$$g(x_1,x_2)\doteq\frac{h(x_1,x_2)}{h_1(x_1)h_2(x_2)},$$

$$\bar{g}(x_1,x_2)\doteq\frac{\bar{h}(x_1,x_2)}{\bar{h}_1(x_1)\bar{h}_2(x_2)}.$$

Заметим, что если X_1 и X_2 являются условно независимыми, т.е.

$$h(x_1,x_2)=\rho_{X_1X_2/A}(x_1,x_2)=\rho_{X_1/A}(x_1)\rho_{X_2/A}(x_2)=h_1(x_1)h_2(x_2),$$

тогда

$$g(x_1,x_2)=\bar{g}(x_1,x_2)=1.$$

Определим следующую монотонно неубывающую функцию плотности вероятности:

$$H_1(x_1)\doteq\int_{-\infty}^{x_1}h_1(z)dz,$$

$$H_2(x_2)\doteq\int_{-\infty}^{x_2}h_2(z)dz,$$

$$\bar{H}_1(x_1)\doteq\int_{-\infty}^{x_1}\bar{h}_1(z)dz,$$

$$\bar{H}_2(x_2)\doteq\int_{-\infty}^{x_2}\bar{h}_2(z)dz.$$

Отметим, что, так как $H_1(x_1),H_2(x_2),\bar{H}_1(x_1)$ и $\bar{H}_2(x_2)$ являются монотонными (здесь, можно предположить, что $h_1(x_1),h_2(x_2),\bar{h}_1(x_1),\bar{h}_2(x_2)>0$, так, что $H_1(x_1),H_2(x_2),\bar{H}_1(x_1)$ и $\bar{H}_2(x_2)$ монотонно увеличиваются. Это ограничение, как будет показно в продолжении, является лишним), то существуют обратные функции $H_1^{-1}(x_1),H_2^{-1}(x_2),\bar{H}_1^{-1}(x_1)$ и $\bar{H}_2^{-1}(x_2)$. Мы можем поэтому определить:

$$J(a,b)\doteq g(H_1^{-1}(a),H_2^{-1}(b)),$$

$$\bar{J}(a,b)\doteq\bar{g}(\bar{H}_1^{-1}(a),\bar{H}_2^{-1}(b)),$$

Для полной ясности, мы с этого момента обознчим

$$J \doteq J(H_1(x_1), H_2(x_2)) = g(H_1^{-1}(H_1(x_1)), H_2^{-1}(H_2(x_2))) = g(x_1, x_2)\,,$$

$$\bar{J} \doteq \bar{J}(\overline{H}_1(x_1), \overline{H}_2(x_2)) = \bar{g}(\overline{H}_1^{-1}(\overline{H}_1(x_1)), \overline{H}_2^{-1}(\overline{H}_2(x_2))) = \bar{g}(x_1, x_2)\,.$$

По определению

$$h(x_1, x_2) = Jh_1(x_1)h_2(x_2)\,, \tag{2}$$

$$\bar{h}(x_1, x_2) = \bar{J}\bar{h}_1(x_1)\bar{h}_2(x_2)\,. \tag{3}$$

Мы теперь имеем:

$$h_1(x_1) \doteq \rho_{X_1/A}(x_1) = \rho_{X_1}(x_1)P(A/x_1)P(A) = \frac{\alpha\rho_{X_1}(x_1)}{\theta}\,, \tag{4}$$

$$h_2(x_2) \doteq \rho_{X_2/A}(x_2) = \rho_{X_2}(x_2)P(A/x_2)P(A) = \frac{\beta\rho_{X_2}(x_2)}{\theta}\,, \tag{5}$$

$$\bar{h}_1(x_1) \doteq \rho_{X_1/\overline{A}}(x_1) = \rho_{X_1}(x_1)P(\overline{A}/x_1)P(\overline{A}) = \frac{(1-\alpha)\rho_{X_1}(x_1)}{1-\theta}\,, \tag{6}$$

$$\bar{h}_2(x_2) \doteq \rho_{X_2/\overline{A}}(x_2) = \rho_{X_2}(x_2)P(\overline{A}/x_2)P(\overline{A}) = \frac{(1-\alpha)\rho_{X_2}(x_2)}{1-\theta}\,. \tag{7}$$

Следовательно, из (2),(3)

$$h(x_1, x_2) = J\,\frac{\alpha\beta\rho_{X_1}(x_1)\rho_{X_2}(x_2)}{\theta^2}\,,$$

$$\bar{h}(x_1, x_2) = \bar{J}\,\frac{(1-\alpha)(1-\beta)\rho_{X_1}(x_1)\rho_{X_2}(x_2)}{(1-\theta)^2}\,.$$

Сейчас из (1)

$$P(A/x_1,x_2)=\frac{\frac{J}{\theta}\alpha\beta\rho_{X_1}(x_1)\rho_{X_2}(x_2)}{\frac{J}{\theta}\alpha\beta\rho_{X_1}(x_1)\rho_{X_2}(x_2)+\frac{\bar{J}}{(1-\theta)}(1-\alpha)(1-\beta)\rho_{X_1}(x_1)\rho_{X_2}(x_2)}$$

$$=\frac{\alpha\beta}{\alpha\beta+\frac{\bar{J}}{J}\frac{\theta}{1-\theta}(1-\alpha)(1-\beta)}.$$

(8)

Заметим, что в случае условной независимости $J=\bar{J}=1$ и (8) становится точным решением $\Gamma(\alpha,\beta,\theta)=P(A/x_1,x_2)$.

Глава 2.1.5 Ограничения на функции $J(a,b)$ и $\bar{J}(a,b)$

Мы имеем

$$h_1(x_1)=\int_{-\infty}^{+\infty}J(H_1(x_1),H_2(x_2))h_1(x_1)h_2(x_2)dx_2. \tag{9}$$

Следовательно

$$1=\int_{-\infty}^{+\infty}J(H_1(x_1),H_2(x_2))h_2(x_2)dx_2=\int_0^1 J(H_1(x_1),H_2(x_2))dH_2(x_2). \tag{10}$$

Таким образом, мы имеем следующее условие

$$\int_0^1 J(a,b)db=1, \tag{11}$$

И аналогично

$$\int_0^1 J(a,b)da=1. \tag{12}$$

Аналогично мы имеем

$$\int_0^{1-} J(a,b)da = 1$$
$$\int_0^{1-} J(a,b)db = 1. \qquad (13)$$

Очевидно

$$J(a,b), \bar{J}(a,b) \geq 0, \qquad (14)$$

$$\int_0^1 \int_0^1 J(a,b)dadb = \int_0^1 \int_0^{1-} J(a,b)dadb = 1. \qquad (15)$$

Множество всех решений (11), (12), (13), (14), (15) вместе с (8) определяет множество всей возможной реализаций $P(A/x_1, x_2)$

Дадим пример решения (11), (12) и (14), (15):

Пусть $\rho(x)$ функция такая что $\rho(x) \geq 0$ и $\int_0^1 \rho(x)dx = 1$

Тогда

$$J(a,b) = \begin{cases} \rho(a-b) & ,a \geq b \\ \rho(a-b+1) & ,a < b \end{cases},$$

удовлетворяет (11),(12) и (14),(15).

Глава 2.1.6 Определение расстояния

Мы определяем расстояние между предложенной апроксимацией функции $P(A/x_1, x_2)$ - $\Gamma(\alpha, \beta, \theta)$ и настоящей функцией $P(A/x_1, x_2)$ следующим образом:

$$\| \Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2) \| \doteq \iint_{-\infty}^{+\infty} \rho_{X_1X_2}(x_1,x_2)[\Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2)]^2 dx_1 dx_2 .$$

Сейчас мы имеем из (2),(3) и (4),(5), (6),(7)

$$\rho_{X_1X_2}(x_1,x_2) = \theta h(x_1,x_2) + (1-\theta)\bar{h}(x_1,x_2) = \theta J h_1(x_1)h_2(x_2) + (1-\theta)\bar{J}\bar{h}_1(x_1)\bar{h}_2(x_2) =$$
$$[J\alpha\beta\theta + \bar{J}(1-\alpha)(1-\beta)(1-\theta)]\rho_{X_1}(x_1)\rho_{X_2}(x_2),$$

$$\| \Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2) \| =$$
$$\iint_{-\infty}^{+\infty} \rho_{X_1}(x_1)\rho_{X_2}(x_2)[J\alpha\beta\theta + \bar{J}(1-\alpha)(1-\beta)(1-\theta)](\Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2))^2 dx_1 dx_2 \qquad (16)$$
$$= \int_0^1\int_0^1 [\frac{J\alpha\beta}{\theta} + \frac{\bar{J}(1-\alpha)(1-\beta)}{(1-\theta)}](\Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2))^2 dF_1(x_1)dF_2(x_2).$$

где

$$F_1(x_1) = \int_{-\infty}^{x_1} \rho_{X_1}(z)dz,$$

$$F_2(x_2) = \int_{-\infty}^{x_2} \rho_{X_2}(z)dz.$$

Глава 2.1.7 Ограничение на основные функции

Мы рассмотрим в дальнейшем все функции с аргументами $1 \geq F1, F2 \geq 0$, а не x_1, x_2. У нас есть шесть функций от $F1, F2$, которые определяют (16): $J, \bar{J}, H_1, H_2, \alpha, \beta$. Позвольте нам, написать другие функции помощью этих функций и найти ограничения для этих функций.

(i)

$$\alpha = P(A/x_1) = \theta h_1(x_1)/\rho_{X_1}(x_1) = \theta \frac{dH_1}{dx_1}\frac{dF_1}{dx_1} = \theta \frac{dH_1}{dF_1}.$$

Таким же образом

$$\beta = \theta \frac{dH_2}{dF_2}.$$

Мы знаем, что функции H_1, F_1, H_2, F_2 являются суммирующими функциями для функций распределения x_1, x_2, соответственно. Эти функции - монотонно неубывающие функции и изменяются от 0 до 1 по определению функции распределения. Поэтому, мы можем получить

следующие ограничения для функций H_1, H_2, поскольку функции F_1, F_2 существуют:

$H_1(1) = H_2(1) = 1$.

$H_1(0) = H_2(0) = 0$,

$$0 \le \alpha = \theta \frac{dH_1}{dF_1}, \beta = \theta \frac{dH_2}{dF_2} \le 1,$$

$$0 \le \theta \le 1,$$

(ii)

$$\overline{H}_1(x_1) = \int_{-\infty}^{x_1} \overline{h}_1(x_1) = \int_{-\infty}^{x_1} \frac{(1-\alpha)\rho_{X_1}(x_1)}{1-\theta} dx_1 = \frac{1}{1-\theta}\int_{-\infty}^{x_1} - \frac{\theta}{1-\theta}\int_{-\infty}^{x_1} \alpha\rho_{X_1}(x_1)\theta dx_1$$

$$= \frac{F_1}{1-\theta} - \frac{\theta}{1-\theta} H_1(x_1).$$

Таким же образом

$$\overline{H}_2(x_2) = \frac{F_2}{1-\theta} - \frac{\theta}{1-\theta} H_2(x_2),$$

(iii)

$J(H_1(F_1), H_2(F_2))$:

$J(H_1(F_1), H_2(F_2)) \ge 0$

$\int_0^1 J(a,b)db = 1$

$\int_0^1 J(a,b)da = 1$,

$\bar{J}(\overline{H}_1(F_1), \overline{H}_2(F_2))$:

$\bar{J}(\overline{H}_1(F_1), \overline{H}_2(F_2)) \ge 0$

$\int_0^1 \bar{J}(a,b)db = 1$

$\int_0^1 \bar{J}(a,b)da = 1$,

(iv)

$$P(A/x_1, x_2) = \frac{\frac{J\alpha\beta}{\theta}}{\frac{J\alpha\beta}{\theta} + \frac{\bar{J}(1-\alpha)(1-\beta)}{1-\theta}}. \qquad (17)$$

Глава 2.1.8 Оптимизация

Мы находим наилучшее приближение $\Gamma(\alpha, \beta, \theta)$ следующим образом

$$\min_{\Gamma(\alpha,\beta,\theta)} E[\| \Gamma(\alpha, \beta, \theta) - P(A/x_1, x_2) \|] \to \Gamma(\alpha, \beta, \theta),$$

где ожидаемое значение (или ожидание, или математическое ожидание, или среднее, или первый момент) $E[...]$ взято относительно совместной ФПВ и для возможной реализации:
$J, \bar{J}, \alpha, \beta, H_1, H_2$ для данных F_1 и F_2.

Ради краткости мы обозначаем:

$$C \doteq \frac{J\alpha\beta}{\theta} + \frac{\bar{J}(1-\alpha)(1-\beta)}{(1-\theta)},$$

$$D \doteq \frac{J\alpha\beta}{\theta}.$$

Затем из (17) и (16)

$$\| \Gamma(\alpha, \beta, \theta) - P(A/x_1, x_2) \| = \int_0^1 \int_0^1 C(\Gamma(\alpha, \beta, \theta) - D/C)^2 dF_1 dF_2$$
$$= \int_0^1 \int_0^1 dF_1 dF_2 [D^2 C + \Gamma^2(\alpha, \beta, \theta) C - 2\Gamma(\alpha, \beta, \theta) D]. \qquad (18)$$

Таким образом

$$\min_{\Gamma(\alpha,\beta,\theta)} E[\| \Gamma(\alpha,\beta,\theta) - P(A/x_1, x_2) \|] =$$

$$\min_{\Gamma(\alpha,\beta,\theta)} E[\int_0^1\int_0^1 dF_1 dF_2 [D^2 C + \Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D]] =$$

$$\min_{\Gamma(\alpha,\beta,\theta)} E[\int_0^1\int_0^1 dF_1 dF_2 [D^2 C]] + \min_{\Gamma(\alpha,\beta,\theta)} E[\int_0^1\int_0^1 dF_1 dF_2 [\Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D]] =$$

$$Const + \min_{\Gamma(\alpha,\beta,\theta)} E[\int_0^1\int_0^1 dF_1 dF_2 [\Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D]] .$$

(19)

Остается рассчитать ожидаемую величину в (19)

Мы делаем следующее очевидное предположение

$$\rho_{J,\bar{J},\alpha,\beta,H_1,H_2/F_1,F_2}(J,\bar{J},\alpha,\beta,H_1,H_2/F_1,F_2) =$$
$$\rho_{J/H_1,H_1}(J/H_1,H_2)\rho_{\bar{J}/\bar{H}_1,H_1}(\bar{J}/\bar{H}_1,\bar{H}_2)\rho_{\alpha/F_1}(\alpha/F_1)\rho_{H_1/\alpha,F_1}(H_1/\alpha,F_1)\rho_{\beta/F_2}(\beta/F_2)\rho_{H_2/\beta,F_2}(H_2/\beta,F_2) . \quad (20)$$

8.1 Лемма 1

$$E[J(a,b)] = \int_0^{+\infty} \rho_{J(a,b)/a,b}(J(a,b)/a,b)J(a,b)dJ = 1 ,$$

$$E[\bar{J}(a,b)] = \int_0^{+\infty} \rho_{\bar{J}(a,b)/a,b}(\bar{J}(a,b)/a,b)\bar{J}(a,b)d\bar{J} = 1 .$$

Доказательство:

Позвольте нам рассматреть функцию: $\rho_{J(a,b)/a,b}$. Функция $J(a,b)$ определена на квадрате $0 \le a,b \le 1$. Позвольте нам сделать выборку функции J на этом квадрате его делением его на небольших квадраты (i,j) и определять значение функции J_{ij} на каждом квадрата i,j. Тогда следующие ограничения на функцию $J(***)$ могут быть написаны:

$$J_{ij} \ge 0 ,$$

$$\frac{1}{N}\sum_{i=1}^{N} J_{ij} = 1 ,$$

$$\frac{1}{N}\sum_{j=1}^{N} J_{ij} = 1 .$$

Здесь $i = 1,...,N$, $j = 1,...,N$

Все матрицы (J_{ij}), которые удовлетворяют эти условия, имеют равную вероятность. Определим функцию плотности вероятности

$\rho(J_{11},...,J_{ij},...,J_{NN})$.

Эта функция плотности вероятности должна быть симметричной относительно перестановки строк и столбцов в матрице (J_{ij}), потому что у плотности распределения имеет равную вероятность для всех матриц (J_{ij}), которые удовлетворяют вышеупомянутым условиям. Действительно, эти условия также симметричны относительно перестановки строк и столбцов в матрице (J_{ij}). Из условий симметрии, которые определяют эту функцию (ρ) относительно перестановки строк и столбцов в матрице (J_{ij}), мы можем заключить, что эта функция (ρ) также не преобразовывается относительно таких перестановок.

Рассмотрим функцию $\rho_{u/ij}(u/ij)$, которая является дискретной версией $\rho_{J(a,b)/a,b}(J(a,b)/a,b)$:

$$\rho_{u/ij}(u/ij) = \int...\int_0^{+\infty} \rho(J_{11},...,J_{nk},...,J_{ij} = u,...,J_{NN}) \prod_{(lm)\neq(ij)} dJ_{lm} .$$

Позвольте нам переставлять строки и столбцы (J_{ij}) таким способом, что элемент J_{ij} будет заменен элементом J_{nk}, и функция, $\rho(J_{11},...)$ не будет меняться после этого. Тогда из предыдущего уравнения мы получаем:

$$\rho_{u/ij}(u/ij) = \int...\int_0^{+\infty} \rho(J_{11},...,J_{nk} = u,...,J_{ij},...,J_{NN}) \prod_{(lm)\neq(nk)} dJ_{lm} = \rho_{u/nk}(u/nk) .$$

$$\rho_{u/ij}(u/ij) = \int...\int_0^{+\infty} \rho(J_{11},...,J_{nk} = u,...,J_{ij},...,J_{NN}) \prod_{(lm)\neq(nk)} dJ_{lm} = \rho_{u/nk}(u/nk) .$$

Из этого уравнения мы можем заключить, что $\rho_{u/ij}(u/ij)$ не зависит от ij поэтому $\rho_{J/ab}(J/ab)$ не зависит от ab и

$\rho_{J/ab}(J/ab) = \rho_J(J)$,

и

$$E[J(a,b)] = \int_0^{+\infty} \rho_J(J) J dJ = Const , \qquad (21)$$

из

$\int_0^1 \int_0^1 J(a,b) da db = 1$,

Мы можем заключить, что

$$\int_0^1\int_0^1 E[J(a,b)]dadb = 1.$$

Таким образом, мы получаем, что $\mathrm{Const} = 1$ в уравнении (21).

8.2 Лемма 2

Функции плотности вероятности α и β не зависят от F_1 и F_2.

$$\rho_{\alpha/F_1}(\alpha/F_1) = \rho_\alpha(\alpha),$$

$$\rho_{\beta/F_2}(\beta/F_2) = \rho_\beta(\beta).$$

Доказательство:

Позвольте нам сделать выборку функции $\alpha(F_1)$, деля область определения этой функции $F_1,[0,1]$ на интервалах $1/N, N\gg1$. Тогда имеем следующие ограничения для $\alpha_k, k = 1,..., N\ \ 0 \le \alpha_k \le 1$,

$$\frac{1}{N}\sum_{k=1}^{N}\alpha_k = \int_0^1 \theta dH_1(F_1)dF_1dF_1 = \theta.$$

Все столбцы (α_k), которые удовлетворяют этим условиям имеют равную вероятность. Позвольте рассмотреть соответствующую функцию $\rho(\alpha_1,...,\alpha_k,...,\alpha_1,...,\alpha_N)$. Из условий симметрии, которые определяют эту функцию относительно перестановок $\alpha_k \to \alpha_1$, функция $\rho(\alpha_1,...,\alpha_k,...,\alpha_1,...,\alpha_N)$ также не преобразовывается относительно таких перестановок. Таким образом, мы можем написать

$$\rho_k(u) = \int_0^1 \rho(\alpha_1,...,\alpha_k = u,...,\alpha_1,...,\alpha_N)\prod_{n\neq k} d\alpha_n = \int_0^1 \rho(\alpha_1,...,\alpha_k,...,\alpha_1 = u,...,\alpha_N)\prod_{n\neq l} d\alpha_n = \rho_1(u).$$

Из этих уравнений, мы можем заключить, что функция $\rho_{\alpha/F_1}(\alpha/F_1)$ не зависит от F_1.

$$\rho_{\alpha/F_1}(\alpha/F_1) = \rho_\alpha(\alpha).$$

Из (20) мы получаем

$$\mathrm{E}[\Gamma^2(\alpha,\beta,\theta)\mathrm{C}-2\Gamma(\alpha,\beta,\theta)\mathrm{D}]=$$
$$\int_0^1\int_0^1\rho_\alpha(\alpha)\rho_\beta(\beta)\mathrm{d}\alpha\mathrm{d}\beta\int_0^1\int_0^1\rho_{\mathrm{H}_1/\alpha,\mathrm{F}_1}(\mathrm{H}_1/\alpha,\mathrm{F}_1)\rho_{\mathrm{H}_2/\beta,\mathrm{F}_2}(\mathrm{H}_2/\beta,\mathrm{F}_2)\mathrm{d}\mathrm{H}_1\mathrm{d}\mathrm{H}_2$$
$$\int_0^\infty\int_0^\infty\rho_{\mathrm{J}}(\mathrm{J})\rho_{\bar{\mathrm{J}}}(\bar{\mathrm{J}})[\Gamma^2(\alpha,\beta,\theta)[\frac{\mathrm{J}\alpha\beta}{\theta}+\frac{\bar{\mathrm{J}}(1-\alpha)(1-\beta)}{1-\theta}]-2\Gamma(\alpha,\beta,\theta)\frac{\mathrm{J}\alpha\beta}{\theta}]\mathrm{d}\mathrm{J}\mathrm{d}\bar{\mathrm{J}}$$
$$=\int_0^1\int_0^1\rho_\alpha(\alpha)\rho_\beta(\beta)\mathrm{d}\alpha\mathrm{d}\beta[\Gamma^2(\alpha,\beta,\theta)[\frac{\mathrm{E}[\mathrm{J}]\alpha\beta}{\theta}+\frac{\mathrm{E}[\bar{\mathrm{J}}](1-\alpha)(1-\beta)}{1-\theta}]-2\Gamma(\alpha,\beta,\theta)\frac{\mathrm{E}[\mathrm{J}]\alpha\beta}{\theta}].$$

Позвольте нам определить

$$\bar{\mathrm{C}}=\frac{\alpha\beta}{\theta}+\frac{(1-\alpha)(1-\beta)}{1-\theta},$$
$$\bar{\mathrm{D}}=\frac{\alpha\beta}{\theta}.$$

Согласно Лемме 1, $\mathrm{E}[\mathrm{J}]=\mathrm{E}[\bar{\mathrm{J}}]=1$. Отсюда

$$\mathrm{E}[\Gamma^2(\alpha,\beta,\theta)\mathrm{C}-2\Gamma(\alpha,\beta,\theta)\mathrm{D}]=\int_0^1\int_0^1[\Gamma^2(\alpha,\beta,\theta)\bar{\mathrm{C}}$$
$$-2\Gamma(\alpha,\beta,\theta)\bar{\mathrm{D}}]\rho_\alpha(\alpha)\rho_\beta(\beta)\mathrm{d}\alpha\mathrm{d}\beta.$$

Остается найти

$$\min_{\Gamma(\alpha,\beta,\theta)}\int_0^1\int_0^1\mathrm{d}\mathrm{F}_1\mathrm{d}\mathrm{F}_2\int_0^1\int_0^1\mathrm{d}\alpha\mathrm{d}\beta\rho_\alpha(\alpha)\rho_\beta(\beta)[\Gamma^2(\alpha,\beta,\theta)\bar{\mathrm{C}}-2\Gamma(\alpha,\beta,\theta)\bar{\mathrm{D}}]. \quad (22)$$

Так как

$\rho_\alpha(\alpha)\rho_\beta(\beta)\ge 0$, если выражение в квадратных скобках минимизируется в каждой точк, тогда весь интеграл в (22) минимизируется. Таким образом, мы можем продолжить следующим образом

$$\frac{\partial}{\partial\Gamma}[\Gamma^2(\alpha,\beta,\theta)\bar{\mathrm{C}}-2\Gamma(\alpha,\beta,\theta)\bar{\mathrm{D}}]=2\Gamma(\alpha,\beta,\theta)\bar{\mathrm{C}}-2\bar{\mathrm{D}}=0.$$

Где оптимум $\Gamma(\alpha,\beta,\theta)$ дается следующим выражением

$$\Gamma_{opt}(\alpha,\beta,\theta) = \frac{\overline{D}}{\overline{C}} = \frac{\frac{\alpha\beta}{\theta}}{\frac{\alpha\beta}{\theta} + \frac{(1-\alpha)(1-\beta)}{1-\theta}}.$$

Глава 2.1.9 Среднее расстояние между предложенойапроксимацией функции $P(A/x_1, x_2)$ - $\Gamma(\alpha,\beta,\theta)$ и реальной функцией $P(A/x_1, x_2)$

Среднее расстояние из (18) следующее

$$DIS = E[\| \Gamma(\alpha,\beta,\theta) - P(A/x_1, x_2) \|] =$$
$$\int_0^1\int_0^1 \rho_\alpha(\alpha)\rho_\beta(\beta)d\alpha d\beta[\Gamma^2(\alpha,\beta,\theta)\overline{C} - 2\Gamma(\alpha,\beta,\theta)\overline{D}] + Const,$$

где $Const$ в этом уравнении определена следующим выражением

$$Const = E[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1,x_2)[P(A/x_1,x_2)]^2 dx_1 dx_2].$$

Из этого уравнения мы можем найти границы $Const$. Из $0 \le P(A/x_1, x_2) \le 1$ мы можем заключить:

$$Const \le E[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1,x_2)P(A/x_1,x_2)dx_1 dx_2] = E[\theta] = \theta.$$

Второе условие следующее

$$0 \le E[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1,x_2)[P(A/x_1,x_2) - \theta]^2 dx_1 dx_2] =$$
$$E[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1,x_2)[P(A/x_1,x_2)^2 + \theta^2 - 2P(A/x_1,x_2)\theta]dx_1 dx_2] =$$
$$E[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1,x_2)[P(A/x_1,x_2)]^2 dx_1 dx_2] - \theta^2.$$

Таким образом, из этих двуж уравнений мы можем заключить

$\theta^2 \le \text{Const} \le \theta$.

На следующем шаге мы хотели бы найти функцию $\rho_\alpha(\alpha)$ ($\rho_\beta(\beta)$)в уравнении для DIS.

Ограничения на функцию $\alpha(\mathrm{F}_1), 0 \le \mathrm{F}_1 \le 1$ следующие:

(i)

$$\int_0^1 \alpha(\mathrm{F}_1)\mathrm{dF}_1 = \theta,$$

(ii)

$$0 \le \alpha(\mathrm{F}_1) \le 1.$$

In discrete form (for $\mathrm{N} \to \infty$) we can rewrite $\alpha_{\text{set}} = \{\alpha_1, \alpha_2, ..., \alpha_{\mathrm{N}}\}$

(i)

$$\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_{\mathrm{i}} = \theta,$$

(ii)

$$0 \le \alpha_{\mathrm{i}} \le 1, \mathrm{i} = 1,2,..., \mathrm{N}.$$

Определим функцию $\mathrm{U}(\alpha_{\text{set}})$ следующим образом

$$\mathrm{U}(\alpha_{\text{set}}) = \begin{cases} \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_{\mathrm{i}} & \text{for} 0 \le \alpha_{\mathrm{i}} \le 1 \quad , \mathrm{i} = 1,2,..., \mathrm{N} \\ +\infty & \text{otherwise} \end{cases},$$

$$\mathrm{U}(\alpha_{\text{set}}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{U}_{\mathrm{i}}(\alpha_{\mathrm{i}}),$$

$$\mathrm{U}_{\mathrm{i}}(\alpha_{\mathrm{i}}) = \begin{cases} \alpha_{\mathrm{i}} & \text{for} 0 \le \alpha_{\mathrm{i}} \le 1 \\ +\infty & \text{otherwise} \end{cases}.$$

Тогда функция, которая является, однородной функцией плотности, учитывая ограничения (i),(ii), будет следующая

$$\rho_{\alpha_{\text{set}}}(\alpha_{\text{set}}) = \frac{1}{\mathrm{C}}\delta(\mathrm{U}(\alpha_{\text{set}}) - \mathrm{N}\theta)\,, \qquad (23)$$

где δ -делта-функция Дирака.

Константа C определяется таким образом:

$$\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}\rho_{\alpha_{\text{set}}}(\alpha_{\text{set}})\mathrm{d}\alpha_1\ldots\mathrm{d}\alpha_{\mathrm{N}} = 1\,.$$

Можно доказать, (см. курс "Statistical mechanics" [8]; преообразование от микроканонического к каноническому распределению) что для $\mathrm{N} \mapsto \infty$ распределение (23) равно следующему распределению:

$$\rho_{\alpha_{\text{set}}}(\alpha_{\text{set}}) = \frac{1}{\mathrm{Z}}\mathrm{e}^{-\mathrm{KU}(\alpha_{\text{set}})}\,,$$

где Z и K могут быть найдены из уравнений:

$$\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}\rho_{\alpha_{\text{set}}}(\alpha_{\text{set}})\mathrm{d}\alpha_1\ldots\mathrm{d}\alpha_{\mathrm{N}} = 1, \qquad (24)$$

$$\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}\mathrm{U}(\alpha_{\text{set}})\rho_{\alpha_{\text{set}}}(\alpha_{\text{set}})\mathrm{d}\alpha_1\ldots\mathrm{d}\alpha_{\mathrm{N}} = \mathrm{N}\theta\,. \qquad (25)$$

Функция $\rho_\alpha(\alpha)$ может быть найдена следующим образом

$$\rho_\alpha(\alpha) = \int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}\rho_{\alpha_{\text{set}}}(\alpha_1,\ldots,\alpha_{\mathrm{j}} = \alpha,\ldots,\alpha_{\mathrm{N}}) \prod_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\mathrm{N}}\mathrm{d}\alpha_{\mathrm{i}} = \frac{1}{\mathrm{D}}\mathrm{e}^{-\mathrm{KU}_{\mathrm{j}}(\alpha_{\mathrm{j}}=\alpha)}\,, \qquad (26)$$

где

$$\mathrm{D}^{\mathrm{N}} = \mathrm{Z}\,. \qquad (27)$$

Из уравнений (24),(25) мы можем найти

$$\frac{1}{Z} = \left(\frac{K}{1-e^{-K}}\right)^{N}, \tag{28}$$

$$\theta = \Lambda(K), \tag{29}$$

где $\Lambda(K)$ -убывающая функция

$$\Lambda(K) = \begin{cases} 1 & forK = -\infty \\ 0 & forK = +\infty \\ 1/2 & forK = 0 \\ \frac{1}{K} - \frac{1}{e^{K}-1} & otherwise \end{cases}.$$

Если K является корнем уравнения (29) мы можем записать из уравнений (26),(27),(28),(29) для функции $\rho_{\alpha}(\alpha)$:

$$\rho_{\alpha}(\alpha) = \begin{cases} \begin{cases} For & K = 0 \\ 1 & for 0 \le \alpha \le 1 \\ 0 & otherwise \end{cases} \\ \begin{cases} For & K = +\infty \\ 2\delta(\alpha) & 0 \le \alpha \le 1 \\ 0 & otherwise \end{cases} \\ \begin{cases} For & K = -\infty \\ 2\delta(\alpha-1) & 0 \le \alpha \le 1 \\ 0 & otherwise \end{cases} \\ \begin{cases} For & otherwise\ K \\ \frac{1}{D}e^{-K\alpha} & 0 \le \alpha \le 1 \\ 0 & otherwise \end{cases} \end{cases},$$

где $2\int_0^1 \delta(\alpha-1) = 2\int_0^1 \delta(\alpha) = 1$ и

$$\frac{1}{D} = \frac{K}{1-e^{-K}}.$$

Глава 2.1.10 Случай более чем двух переменных A и X.

Пусть A - случайная величина, со значениями на множестве $0,1,...,L$. Предположим, что априорная вероятность $P(A = i)$ известна и обозначим ее θ_i, здесь $i = 1,...,L$. Пусть $X_1,...,X_K$ случайные величины, со значениями на некотором множестве, например $]-\infty;+\infty[$. Мы имеем следующую информацию : $X_1 = x_1$,..., $X_K = x_K$ (полученную из измерений). Кроме того, у нас есть системы - "классификаторы", которые дают $x_1,...,\ x_K$:

$$P(A = i/X_j = x_j) \doteq \alpha_{ij} .$$

Мы хотим оценить вероятности $P(A = i/X_1 = x_1,...,X_K = x_K)$ в терминах α_{ij} и θ_i. Более определенно мы желаем найти функцию $\Gamma_{opt,M}(\alpha_{ij},\theta_i)$, которая в среднем является наилучшим приближением для $P(A = M/x_1,...,x_K)$. Тем же самым путем, что в случае двух переменных мы можем найти что $\Gamma_{opt,M}(\alpha_{ij},\theta_i)$ определенный уравнением:

$$\Gamma_{opt,M}(\alpha_{ij},\theta_i) = \frac{(\prod_{j=1}^{K}\alpha_{Mj})/\theta_M^{K-1}}{\sum_{i=1}^{L}(\prod_{j=1}^{K}\alpha_{ij})/\theta_i^{K-1}} .$$

Мы имеем очевидное ограничение на α_{ij}, θ_i

$$0 \le \alpha_{ij} \le 1$$

$$\sum_{i=1}^{L}\alpha_{ij} = 1,$$

$$0 \le \theta_i \le 1$$

$$\sum_{i=1}^{L}\theta_i = 1.$$

Глава 2.1.11 Выводы

Мы доказали, что Наивная Байесовская модель дает минимальную среднюю ошибку при равномерном распределении всех возможных корреляций между характерными переменными. Этот результат может объяснить описанную выше мистическую оптимальности Наивной Байесовской модели. Мы также нашли среднюю ошибку, которую Наивная Байесовская модель дает для равномерного распределения всех возможных корреляций.

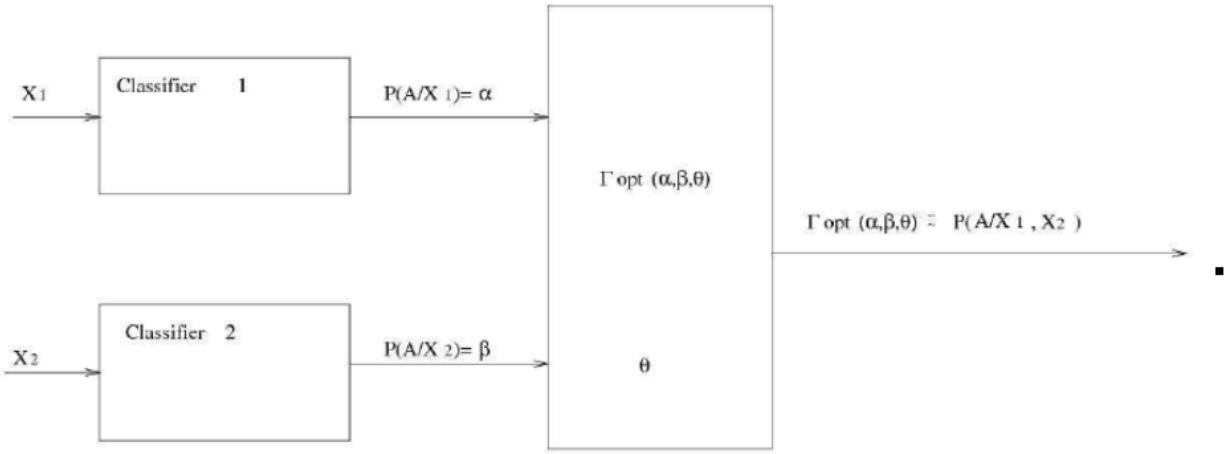

Рис 2 Функция $\Gamma(\alpha, \beta, \theta) : [0,1]^3 \mapsto [0,1]$

Библиография

[1] Kupervasser O., The mysterious optimality of Naive Bayes: Estimation of the probability in the system of "classifiers", Pattern Recognition and Image Analysis, 24 (1), pp. 1-10 (2014), arXiv:cs/0202020v1, (2002)
http://arxiv.org/abs/cs/0202020v1

[2] V. Ramensky, A. Sobol, N. Zaitseva, A. Rubinov, V. Zosimov, A novel approach to local similarity of protein binding sites substantially improves computational drug design results, Proteins: Structure, Function, and Bioinformatics, , 69(2), pp 349–357 (2007)

[3] S. Nikitin1, N. Zaitseva, O. Demina, V. Solovieva, E. Mazin, S. Mikhalev, M. Smolov, A. Rubinov, P. Vlasov, D. Lepikhin, D. Khachko, V. Fokin, C. Queen,V. Zosimov, A very large diversity space of synthetically accessible compounds for use with drug design programs, Journal of Computer-Aided Molecular Design, 19, pp 47–63 (2005)

[4] Raymer M. L., Doom T. E., Kuhn L. A., Punch W. F., "Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm," IEEE Transactions on Systems, Man, and Cybernetics, 33B, 802 (2003)

[5] Domingos, P., and Pazzani, M., On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning 29, 103 (1997)

[6] Zhang H., The Optimality of Naive Bayes, In FLAIRS Conference (2004)
http://www.cs.unb.ca/profs/hzhang/publications/FLAIRS04ZhangH.pdf

[7] Kuncheva L.I., On the optimality of Naive Bayes with dependent binary features, Pattern Recognition Letters, 27,830 (2006)

[8] Landau L.D., Lifshitz E.M., Statistical Physics, 5,Elsevier Science Technology, United Kingdom, (1996)