Задача фильтрации спама. Общая постановка: придумать алгоритм, который каждому электронному почтовому сообщению сопоставляет бинарную характеристику: -1, если это спам, и +1, если это обычное сообщение.

Обучать алгоритм будем по некоторому набору сообщений $M_1, M_2, \ldots, M_n$, для которых правильный «ответ» (спам это или нет) известен: $Y_1, Y_2, \ldots, Y_n$ ($Y_j \in \{-1;1\}$, $j = 1, \ldots, n$). В теории машинного обучения совокупность из n пар $\{(M_1, Y_1), \ldots, (M_n, Y_n)\}$ называется обучающей выборкой размера n. Упростим немного задачу, будем считать, что на вход алгоритму будет поступать не само сообщение M, а набор его числовых характеристик, то есть свяжем с каждым сообщением M некоторый числовой вектор $\vec{X} \in \mathbb{R}^d$. Какие это именно числовые характеристики, как они были получены – это отдельный вопрос, здесь мы его не будем обсуждать. Скажем лишь, что они достаточно хорошо различают классы (например, это может быть, длина всего сообщения, доля встречаемости символа $ или слова «выигрыш»). Тогда можно считать, что обучающая выборка представлена в виде $\{(\vec{X}_1, Y_1), \ldots, (\vec{X}_n, Y_n)\}$, где вектор $\vec{X}_j$ называют признаковым описанием j-го объекта (в данном случае сообщения). Построенный в ходе обучения алгоритм для каждого нового $\vec{X}$ возвращает ответ $a(\vec{X}) \in \{-1;1\}$.

Для формального решения задачи на языке математической статистики скажем, что задано распределение вероятности на пространстве признаков при фиксированном Y, то есть известны $P(\vec{X} \mid Y = 1)$ и $P(\vec{X} \mid Y = -1)$. Также известна априорная вероятность классов, то есть $P(Y = 1)$ и $P(Y = -1)$. Мы хотим вроде бы минимизировать вероятность ошибки нашего алгоритма: $R(a(\bullet)) = P(a(\vec{X}) \neq Y)$ или после применения формулы полной вероятности: $R(a(\bullet)) = P(a(\vec{X}) = -1 \mid Y = 1)P(Y = 1) + P(a(\vec{X}) = 1 \mid Y = -1)P(Y = -1)$. При этом стоит заметить, что ошибка отнесения хорошего сообщения к спаму для нас более грубая, чем, если мы не отфильтруем спам-сообщение. Поэтому в данной задаче уместно ввести разные штрафы ($C_\alpha > C_\beta$) за разного рода ошибки и минимизировать математическое ожидание штрафной функции: $R(a(\bullet)) = C_\alpha P(a(\vec{X}) = -1; Y = 1) + C_\beta P(a(\vec{X}) = 1; Y = -1)$. Выпишите формальное решение (Неймана-Пирсона) данной задачи.

На самом деле нам неизвестны $P(\vec{X} \mid Y = 1)$ и $P(\vec{X} \mid Y = -1)$, поэтому в реальности нужно делать какие-то предположения о том какому параметрическому классу распределений они относятся и оценивать по обучающей выборке неизвестные параметры. (Стоит подчеркнуть еще одну трудность – высокая размерность пространства признаков $d \approx 1000$). Можно пойти иным путем.

В теории машинного обучения подход к решению задачи следующий. Задаются некоторым семейством функций $S = (a(\vec{x}) : \mathbb{R}^d \to \{-1;1\})$, например, можно задать семейство разделяющих гиперплоскостей $S = (a_{\vec{\theta}, \vec{x}_0}(\vec{x}) = sign(\langle \vec{\theta}, \vec{x} - \vec{x}_0 \rangle), \quad \vec{x}_0, \vec{\theta} \in \mathbb{R}^d)$, и заменить средний риск эмпирическим

$\hat{a}(\bullet) = \arg\min_{a(\bullet) \in S} R_{\text{эмп}}(a(\bullet); \{(\vec{X}_j, Y_j)_{j=1}^n\})$, где

$R_{\text{эмп}}(a(\bullet); \{(\vec{X}_j, Y_j)_{j=1}^n\}) = \frac{1}{n}\sum_{i=1}^n C_\alpha I(a(\vec{X}_i) = -1; Y_i = 1) + C_\beta I(a(\vec{X}_i) = 1; Y_i = -1)$.

Основная надежда на то, что эмпирический риск достаточно хорошо приближает истинный. Ответ на вопрос: на сколько это справедливо – дает теория Вапника-Червоненкиса.

Обозначим $\Delta^S(x_1, ..., x_n)$ – число (бинарных) решающих правил класса S, по-разному классифицирующих объекты заданной выборки. Введем функцию роста $M^S(n) = \max \Delta^S(x_1, ..., x_n)$, где максимум берется по всем последовательностям $(x_1, ..., x_l)$ длины n. Покажите, что для семейства разделяющих гиперплоскостей $M^S(n) \leq 2^d C_n^d$.

Получите: $P\left\{\sup_{a(\bullet) \in S} \left| R(a(\bullet)) - R_{\text{эмп}}(a(\bullet); \{(\vec{X}_j, Y_j)_{j=1}^n\}) \right| > \varepsilon\right\} \leq 6M^S(2n)\exp\left[-\frac{\varepsilon^2(n-1)}{4}\right]$.