# Теорема про методы отбора признаков

## 1 Иллюстрация важности учета взаимосвязи между признаками.

Задача алгоритмов отбора признаков состоит в выборе подмножества признаков $\mathcal{A}$ исходного множества признаков $\{1, \ldots, n\}$ такого, что максимизируется некоторый критерий качества, который и определяет алгоритм отбора. При этом при наличии мультиколлинеарных признаков мультиколлинеарность устраняют путем удаления одного или нескольких мультиколлинеарных признаков из каждой группы мультиколлинеарных признаков, чтобы полученная матрица признаков $\mathbf{X}_{\mathcal{A}}$ не была плохо обусловленной и оценки параметров модели были потому устойчивыми. Покажем, что такой подход не является оптимальным и исключение мультиколлинеарных признаков может приводить к ухудшению качества прогноза.

## 2 Теорема

(Адуенко, 2016). Пусть имеется $l$ линейно независимых факторов и целевой вектор параметров $\mathbf{v} \in \mathbb{R}^l$. Обозначим $\mathbf{f}_i$ вектор значений факторов для $i$-го объекта. Пусть для каждого объекта вместо $\mathbf{f}_i$ наблюдается $\mathbf{x}_i = \mathbf{G}\mathbf{f}_i + \boldsymbol{\varepsilon}_i$, где $\mathbf{G}$ — матрица размера $n \times l, n \geqslant l$ полного ранга, а $\boldsymbol{\varepsilon}_i$ центрированный шум с невырожденной ковариационной матрицей $\boldsymbol{\Sigma}$.

Тогда оптимальной в терминах дисперсии шума $\mathbb{E}(\mathbf{w}^T\boldsymbol{\varepsilon}_i)^2$ оценкой $\mathbf{w}$, такой, что $\mathbb{E}\mathbf{w}^T\mathbf{x}_i = \mathbf{v}^T\mathbf{f}_i$, является оценка вида

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}\mathbf{G}(\mathbf{G}^T\boldsymbol{\Sigma}^{-1}\mathbf{G})^{-1}\mathbf{v}.$$

Доказательство. Рассмотрим сначала требование несмещенности $\mathbb{E}\mathbf{w}^T\mathbf{x}_i = \mathbf{v}^T\mathbf{f}_i$. Так как матрица $\mathbf{G}$ полного ранга, то значения всех факторов для любого объекта $\mathbf{f}_i$ при отсутствии шума можно точно восстановить по $\mathbf{x}_i$, причем при $n > l$ с избытком. При этом в силу отсутствия шума любой вектор $\mathbf{w}$, удовлетворяющий этому требованию, подходит и дает одинаковый результат. В частности можно оставить произвольные $l$ линейно независимых признаков и далее работать с ними. Из-за наличия шума ситуация изменяется и вектора $\mathbf{w}$, удовлетворяющие условию несмещенности $\mathbb{E}\mathbf{w}^T\mathbf{x}_i = \mathbf{v}^T\mathbf{f}_i$, перестают быть равноценными, так как для разных $\mathbf{w}$, $\mathbf{w}^T\mathbf{x}_i$ обладает разной дисперсией.

Рассмотрим сначала условие несмещенности. Из него получаем

$$\mathbb{E}\mathbf{w}^T\mathbf{x}_i = \mathbf{w}^T\mathbf{G}\mathbf{f}_i = \mathbf{v}^T\mathbf{f}_i,$$

откуда допустимые $\mathbf{w}$ удовлетворяют соотношению $\mathbf{G}^T\mathbf{w} = \mathbf{v}$. При этом

$$\mathbf{w}^T\mathbf{x}_i = \mathbf{w}^T(\mathbf{G}\mathbf{f}_i + \boldsymbol{\varepsilon}_i) = \mathbf{v}^T\mathbf{f}_i + \mathbf{w}_i^T\boldsymbol{\varepsilon}_i,$$

откуда условие оптимальности по дисперсии шума линейной комбинации принимает вид

$$\mathbb{E}(\mathbf{w}_i^T\boldsymbol{\varepsilon}_i)^2 = \mathbf{w}^T\mathbb{E}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^T)\mathbf{w} = \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \to \min_{\mathbf{w}},$$

откуда с учетом условия допустимости $\mathbf{w}$ получаем следующую задачу оптимизации

$$\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} \to \min_{\mathbf{w}},$$

$$\text{при условии } \mathbf{G}^T\mathbf{w} = \mathbf{v}$$

Решая задачу и вводя вектор множителей Лагранжа $\boldsymbol{\lambda} \in \mathbb{R}^l$, получаем условия оптимальности

$$2\mathbf{\Sigma}\mathbf{w} + \mathbf{G}\boldsymbol{\lambda} = \mathbf{0},$$
$$\mathbf{G}^T\mathbf{w} = \mathbf{v},$$

откуда

$$\mathbf{w} = \mathbf{\Sigma}^{-1}\mathbf{G}(\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathbf{G})^{-1}\mathbf{v}. \tag{1}$$

Проиллюстрируем доказанную теорему несколькими примерами

**2.1**

Случай совокупности $n$ мультиколлинеарных признаков, то есть $l = 1, \mathbf{G} = \mathbf{e} = [1, \dots, 1]^T, \mathbf{\Sigma}^{-1} = \mathrm{diag}(1/\sigma_1^2, \dots, 1/\sigma_n^2)$. В рассматриваемом случае имеется только один фактор, а признаковое описание представляет собой $n$ его зашумленных копий. Тогда все признаки являются мультиколлинеарными и с точки зрения отбора признаков из них стоит оставит только один. Однако в соответствии с теоремой такой подход не является оптимальным, а для достижения минимальной дисперсии шума требуется сложить признаки с некоторыми положительными весами. Получим выражение для этих весов в соответствии с (2.2).

$$\mathbf{w} = \begin{pmatrix} 1/\sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/\sigma_n^2 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \frac{v}{\frac{1}{\sigma_1^2} + \dots + \frac{1}{\sigma_n^2}} = v\left[\frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \dots + \frac{1}{\sigma_n^2}}, \dots, \frac{\frac{1}{\sigma_n^2}}{\frac{1}{\sigma_1^2} + \dots + \frac{1}{\sigma_n^2}}\right].$$

Таким образом, складывать признаки стоит в весами обратно пропорциональными дисперсии шума соответствующей копии актора. В частности, если $\sigma_1^2 = \dots = \sigma_n^2$, то оптимально взять в качестве оценки фактора среднее значение $n$ его зашумленных копий

**2.2**

Тройка мультиколлинеарных признаков, то есть

$l = 2, n = 3, \mathbf{G} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}^T, \mathbf{\Sigma}^{-1} = \mathrm{diag}(1/\sigma_1^2, 1/\sigma_2^2, 1/\sigma_3^2)$

Аналогично предыдущему, используя (2.2), получим

$$\mathbf{w} = \frac{1}{\frac{1}{\sigma_1^2\sigma_2^2} + \frac{1}{\sigma_1^2\sigma_3^2} + \frac{1}{\sigma_1^2\sigma_3^2}} \begin{pmatrix} \frac{v_1}{\sigma_1^2\sigma_2^2} + \frac{v_1 - v_2}{\sigma_1^2\sigma_3^2} \\ \frac{v_2}{\sigma_1^2\sigma_2^2} + \frac{v_2 - v_1}{\sigma_1^2\sigma_3^2} \\ \frac{v_1}{\sigma_2^2\sigma_3^2} + \frac{v_2}{\sigma_1^2\sigma_3^2} \end{pmatrix}$$

Так при $\sigma_1^2 \to \infty$, получим, что оптимально использовать только второй и третий признаки. Аналогично при $\sigma_2^2 \to \infty$ оптимально использовать только первый и третий признак, а при $\sigma_3^2 \to \infty$ только первый и второй. При этом даже, если, например, $\sigma_1^2, \sigma_2^2 \gg \sigma_3^2$ при $v_1 \neq v_2$ приходится использовать один из первых двух признаков, несмотря на их сильную зашумленность. Однако при $v_1 = v_2$ в той же ситуации оптимально будет использовать только третий признак.

### 2.3

Две пары мультиколлинеарных признаков, то есть
$l = 2, n = 4, \mathbf{G} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}^T, \mathbf{\Sigma}^{-1} = \mathrm{diag}(1/\sigma_1^2, 1/\sigma_2^2, 1/\sigma_3^2, 1/\sigma_4^2)$
Этот случай сводится к независимому применению результат случая 1 для первой и второй пары мультиколлинеарных признаков.

Таким образом, при наличии мультиколлинеарных признаков устранение мультиколлинеарности путем удаления избыточных признаков не является оптимальным, а информация из избыточных признаков может быть использована для улучшения качества прогноза. Поэтому наряду с оценкой важности отдельных признаков, важно оценивать и взаимосвязь между ними.