

Лекция III

Классификация

Т. Е. Абильдаев

Кафедра статистического моделирования
Матмех СПбГУ

2025-03-15

Задача классификации

Под классификацией обычно понимают задачу регрессии для переменной, принимающей дискретные значения. При этом эти значения могут не иметь естественного порядка, то есть переменная может быть категориальной.

Итак, мы имеем n наблюдений переменных $X_1, \dots, X_m, Y$ и хотим построить правило классификации h , такое, что $h(X_1, \dots, X_m) \approx Y$.

Но какую меру близости $h(X_1, \dots, X_m)$ и Y следует выбрать?

Задача классификации

Под классификацией обычно понимают задачу регрессии для переменной, принимающей дискретные значения. При этом эти значения могут не иметь естественного порядка, то есть переменная может быть категориальной.

Итак, мы имеем n наблюдений переменных $X_1, \dots, X_m, Y$ и хотим построить правило классификации h , такое, что $h(X_1, \dots, X_m) \approx Y$.

Но какую меру близости $h(X_1, \dots, X_m)$ и Y следует выбрать?

Так как данные имеют стохастическую природу, естественным желанием является минимизация ошибки классификации

$$L(h) = \mathbb{P}\{h(X_1, \dots, X_m) \neq Y\} \longrightarrow \inf_{h \in \mathcal{H}}, \quad (1)$$

что на эмпирическом языке соответствует

$$\hat{L}_n(h) = \frac{1}{n} \sum_{k=1}^n \mathbb{1}\{h(x_{1k}, \dots, x_{mk}) \neq y_k\} \longrightarrow \inf_{h \in \mathcal{H}}.$$

Какие проблемы возникают при решении задачи (1)?

Байесовский классификатор

Пусть сперва классификация бинарная, то есть Y принимает значения 0 и 1. Обозначим $\mathbb{P}\{Y = 1 \mid X = x\}$ через $r(x)$. Тогда по теореме Байеса

$$r(x) = \frac{f(x|Y=1)\mathbb{P}\{Y=1\}}{f(x)} = \frac{\pi f_1(x)}{\pi f_1(x) + (1-\pi)f_0(x)},$$

где $f_0(x) = f(x|Y=0)$, $f_1(x) = f(x|Y=1)$, $\pi = \mathbb{P}\{Y=1\}$.

Определение

Правило

$$h^*(x) = \begin{cases} 1, & r(x) > 0.5, \\ 0, & r(x) \leq 0.5 \end{cases}$$

называется *байесовским*. Множество $\{x \in \mathbb{R}^m \mid r(x) = 0.5\}$ называется *границей классификации* (decision boundary).

Байесовский классификатор (прод.)

Теорема 1

Байесовский классификатор h^* является оптимальным в том смысле, что для любого $h \in \mathcal{H}$

$$L(h^*) \leq L(h).$$

Как всё это обобщается на небинарный случай?

Пусть $\pi_r = \mathbb{P}\{Y = r\}$, $f_r(x) = f(x|Y = r)$.

Теорема 2

Правило

$$h^*(x) = \operatorname{argmax}_r \mathbb{P}\{Y = r|X = x\} = \operatorname{argmax}_r \pi_r f_r(x),$$

называемое байесовским, является оптимальным.

Но как найти или оценить π_r и f_r ?

Гауссовский дискриминантный анализ (GDA)

Предположим, что

$$f_r(x) = \frac{1}{(2\pi)^{d/2} |\Sigma_r|^{1/2}} \exp\left(-\langle \Sigma_r^{-1}(x - \mu_r), x - \mu_r \rangle / 2\right).$$

Тогда

$$\begin{aligned} \operatorname{argmax}_r \pi_r f_r(x) &= \operatorname{argmax}_r \frac{\pi_r}{(2\pi)^{d/2} |\Sigma_r|^{1/2}} \exp\left(-\langle \Sigma_r^{-1}(x - \mu_r), x - \mu_r \rangle / 2\right) \\ &= \operatorname{argmax}_r \left(-(\langle \Sigma_r^{-1}(x - \mu_r), x - \mu_r \rangle + \log |\Sigma_r|) / 2 + \log \pi_r \right). \end{aligned}$$

Что с границей классификации?

Гауссовский дискриминантный анализ (GDA)

Предположим, что

$$f_r(x) = \frac{1}{(2\pi)^{d/2} |\Sigma_r|^{1/2}} \exp\left(-\langle \Sigma_r^{-1}(x - \mu_r), x - \mu_r \rangle / 2\right).$$

Тогда

$$\begin{aligned} \operatorname{argmax}_r \pi_r f_r(x) &= \operatorname{argmax}_r \frac{\pi_r}{(2\pi)^{d/2} |\Sigma_r|^{1/2}} \exp\left(-\langle \Sigma_r^{-1}(x - \mu_r), x - \mu_r \rangle / 2\right) \\ &= \operatorname{argmax}_r \left(-(\langle \Sigma_r^{-1}(x - \mu_r), x - \mu_r \rangle + \log |\Sigma_r|) / 2 + \log \pi_r \right). \end{aligned}$$

Что с границей классификации?

В небинарном случае граница классификации это множество $x \in \mathbb{R}^m$, таких, что максимум по $\{\pi_r f_r(x)\}_r$ определён неоднозначно.

Гауссовский дискриминантный анализ (GDA): LDA и QDA

Возьмём, к примеру, все такие $x \in \mathbb{R}^m$, что максимум достигается при $r \in \{1, 2\}$. Предположим далее, что $\Sigma_1 = \dots = \Sigma_m = \Sigma$. Тогда границу классификации будут составлять те x , для которых

$$\begin{aligned} \langle \Sigma_1^{-1}(x - \mu_1), x - \mu_1 \rangle &= \langle \Sigma_2^{-1}(x - \mu_2), x - \mu_2 \rangle \\ \Leftrightarrow 2\langle x, \Sigma^{-1}\mu_1 \rangle - \langle \Sigma^{-1}\mu_1, \mu_1 \rangle &= 2\langle x, \Sigma^{-1}\mu_2 \rangle - \langle \Sigma^{-1}\mu_2, \mu_2 \rangle, \end{aligned}$$

то есть оно будет линейным многообразием. Поэтому этот метод называют *линейным дискриминантным анализом*, или *LDA*.

Предположим теперь, не умаляя общности, что $\Sigma_1 \neq \Sigma_2$. Тогда граница классификации будет задаваться квадратической формой, поэтому такой метод называют *квадратичным дискриминантным анализом*, или *QDA*.

Гауссовский дискриминантный анализ (GDA): оценка параметров

Как оценить π_r , μ_r , Σ_r ?

Разумеется, с помощью максимального правдоподобия. То есть

$$\hat{\pi}_r = \frac{1}{n} \sum_{k=1}^n \mathbb{1}\{y_k = r\}, \quad (\hat{\mu}_r)_j = \frac{1}{n\hat{\pi}_r} \sum_{s=1}^n x_{sj} \mathbb{1}\{y_k = r\},$$

$$(\hat{\Sigma}_r)_{st} = \frac{1}{n\hat{\pi}_r} \sum_{k=1}^n (x_{ks} - (\hat{\mu}_r)_s)(x_{kt} - (\hat{\mu}_r)_t) \mathbb{1}\{y_k = r\}.$$

В гомоскедастичном сеттинге

$$\hat{\Sigma}_r = \sum_r \hat{\pi}_r \hat{\Sigma}_r.$$

Линейный дискриминант Фишера

Рассмотрим линейную комбинацию признаков $Z = \sum_{j=1}^m w_j X_j$. Это некоторая случайная величина с каким-то средним и дисперсией

$$\mathbb{D}Z = \mathbb{D}\left[\sum_{j=1}^m w_j X_j\right] = \sum_{s=1}^m \sum_{t=1}^m w_s w_t \text{Cov}(X_s, X_t) = \langle \Sigma w, w \rangle.$$

Идея в том, чтобы подобрать $w_1, \dots, w_m$, «наилучшим образом разделяющие переменные». Мерой последнего возьмём

$$\begin{aligned} J(w) &= \frac{(\mathbb{E}(Z|Y=0) - \mathbb{E}(Z|Y=1))^2}{\langle \Sigma w, w \rangle} = \frac{(\langle w, \mu_0 \rangle - \langle w, \mu_1 \rangle)^2}{\langle \Sigma w, w \rangle} \\ &= \frac{(\langle w, \mu_0 - \mu_1 \rangle)^2}{\langle \Sigma w, w \rangle} = \frac{\langle (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T w, w \rangle}{\langle \Sigma w, w \rangle} \rightarrow \inf_{w \in \mathbb{R}^d}. \end{aligned}$$

Линейный дискриминант Фишера (прод.)

Функция $J(w)$ не может быть посчитана, так как мы не знаем μ_0, μ_1, Σ . Однако мы можем воспользоваться оценками последних величин и минимизировать

$$\hat{J}(w) = \frac{\langle S_B w, w \rangle}{\langle S_W w, w \rangle}, \quad (2)$$

где $S_B = (\hat{\mu}_0 - \hat{\mu}_1)(\hat{\mu}_0 - \hat{\mu}_1)^T$, $S_W = \hat{\Sigma}$.

Теорема 3

Вектор $w^* = S_W^{-1}(\hat{\mu}_0 - \hat{\mu}_1)$ является решением 2.

Определение

Правило

$$h(x) = \begin{cases} 0, & \langle w, x \rangle \geq m, \\ 1, & \langle w, x \rangle < m, \end{cases}$$

где $m = (\hat{\mu}_0 + \hat{\mu}_1)/2$, называется *линейный дискриминант Фишера*.

Оценка с помощью лог-регрессии

Рассмотрим бинарную классификацию. Вспомним, что оценка π , f_0 и f_1 нам была нужна потому, что

$$r(x) = \mathbb{P}\{Y = 1|X = x\} = \frac{\pi f_1(x)}{\pi f_1(x) + (1 - \pi)f_0(x)}.$$

А r нужна потому, что следующее правило является оптимальным. Полагаем $r(x)$

$$h^*(x) = \begin{cases} 1, & r(x) > 0.5, \\ 0, & r(x) \leq 0.5. \end{cases}$$

Можно оценить r напрямую, и естественным методом для этого является логистическая регрессия.

$$r(x) = \frac{\exp(\beta^T \mathbf{X})}{1 + \exp(\beta^T \mathbf{X})}, \quad \sum_{k=1}^n \left[y_k \sum_{j=0}^m x_{kj} \beta_j - \log \left(1 + \exp \left(\sum_{j=0}^m x_{kj} \beta_j \right) \right) \right] \longrightarrow \sup_{\beta_0, \dots, \beta_m}.$$

Гауссовский LDA vs Лог-регрессия

По определению логистической регрессии имеем

$$\log \left[\frac{\mathbb{P}\{Y = 1|X = x\}}{\mathbb{P}\{Y = 0|X = x\}} \right] = \beta^T \mathbf{x}.$$

Можно показать, что то же справедливо и для гауссовского LDA. То есть модели совпадают. В чём тогда разница?

В том, как мы оцениваем параметры. В LDA мы целиком оцениваем

$$p(X, Y) = p(Y|X)p(X),$$

а с помощью лог-регрессии мы оцениваем лишь $p(Y|X)$. При этом в последнем случае нет предположений о гауссовости.

Таким образом, если мы не уверены в гауссовости, использовать нужно лог-регрессию, если же уверены – гауссовский LDA.

Любое правило классификации это по сути дизъюнктивное разбиение $\mathbb{R}^m$. Поэтому правило классификации можно задать с помощью решающего дерева.

Пусть сперва у нас один признак. Выберем $a \in \mathbb{R}$ и положим $A_- = (-\infty, a]$, $A_+ = (a, \infty)$. Введём

$$\hat{p}_-(r) = \frac{\#\{k \in \overline{1, n} \mid y_k = r, x_k \in A_-\}}{\#\{k \in \overline{1, n} \mid y_k = r\}}, \quad \hat{p}_+(r) = \frac{\#\{k \in \overline{1, n} \mid y_k = r, x_k \in A_+\}}{\#\{k \in \overline{1, n} \mid y_k = r\}},$$

$$\gamma_- = 1 - \sum_r (\hat{p}_-(r))^2, \quad \gamma_+ = 1 - \sum_r (\hat{p}_+(r))^2.$$

Теперь, минимизируя γ_- и γ_+ по a получаем первое решающее правило. Процесс продолжается для поддеревьев до момента остановки.

Как выбрать классификатор?

Естественно смотреть на ошибку классификации, но нам нужно быть уверенными в способности модели к обобщению. Поэтому ошибку оценивают с помощью кросс-валидации. То есть делим наши данные на K частей и строим K классификаторов, каждый раз выбрасывая из данных одну из частей. В конце считаем ошибки классификации и берём среднее.