(Квази)линейные методы классификации (LDA, QDA, дискриминант Фишера, логистическая регрессия)

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

Спецкурс «Теория статистических решений».
Лекция 2.

Список методов

- Линейный и квадратичный дискриминант.
- Дискриминант Фишера.
- Логистическая регрессия.
- Наивный байесовский классификатор.
- Машина опорных векторов.

Нормальное распределение

Одномерный случай, два класса, $y \in \{-1, 1\}$.

Пусть заданы $P(y) = P_y$ и условные плотности вероятности $\varphi_y(x)$.

$$\varphi_y(x) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(x-\mu_y)^2}{2(\sigma_y)^2}}.$$

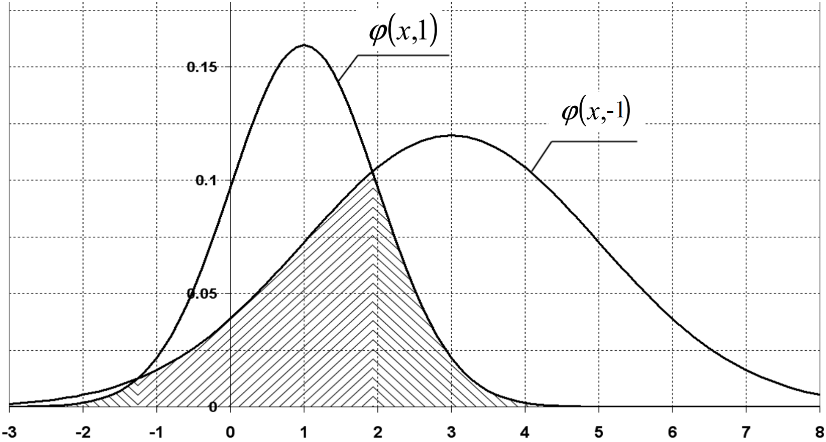
Байесовская решающая функция есть

$$y^*(x) = \arg\max_y \varphi(x, y),$$

где $\varphi(x, y) = \varphi_y(x)P(y)$ – совместная плотность вероятности.

Формула Байеса здесь не используется.

Пример совместной плотности



Функция условной вероятности

По формуле Байеса

$$g_1(x) = P(y = 1 \,|\, x) = \varphi_1(x) \frac{P_1}{\varphi(x)},$$

$$\varphi(x) = \varphi_1(x) P_1 + \varphi_{-1}(x) P_{-1}.$$

Функция условной вероятности часто выступает в качестве решения.

Разделяющая функция

Разделяющая функция

$$l(x) = \ln \varphi_1(x) - \ln \varphi_{-1}(x) + \ln \frac{P_1}{P_{-1}}.$$

Для нормальных распределений

$$l(x) = \frac{(x-\mu_{-1})^2}{2(\sigma_{-1})^2} - \frac{(x-\mu_1)^2}{2(\sigma_1)^2} + \ln \frac{\sigma_{-1}}{\sigma_1} + \ln \frac{P_1}{P_{-1}}.$$

Многомерное нормальное распределение

Плотности вероятности

$$\varphi_y(x) = \frac{1}{(2\pi)^{n/2}|\lambda_y|^{n/2}} e^{-\frac{1}{2}Q_y(x)},$$

где $Q_y(x) = (x - \mu_y)'(\lambda_y)^{-1}(x - \mu_y)$,
μ_y – вектор средних для класса y,
λ_y – ковариационная матрица.

Разделяющая функция

Разделяющая функция после подстановки нормальных плотностей

$$2l(x) = x'Ax + bx + c,$$

где

$$A = (\lambda_{-1})^{-1} - (\lambda_1)^{-1}$$

$$b = 2\mu_1(\lambda_{-1})^{-1} - 2\mu_{-1}(\lambda_1)^{-1}$$

$$c = \mu'_{-1}(\lambda_1)^{-1}\mu_{-1} - \mu'_1(\lambda_{-1})^{-1}\mu_1 + \ln\frac{|\lambda_1|}{|\lambda_{-1}|} + 2\ln\frac{P_1}{P_{-1}}.$$

Оценивание параметров

Безусловные (априорные) вероятности классов

$$\widetilde{P}_y = \frac{N_y}{N}.$$

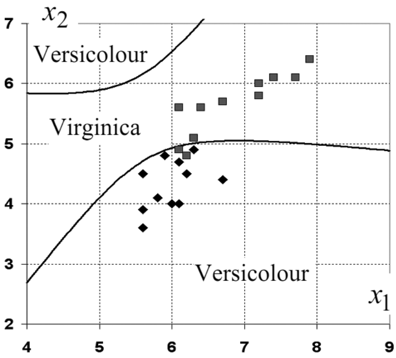
Компоненты вектора средних

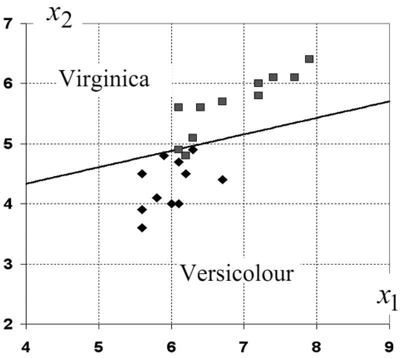
$$\widetilde{\mu}_j^y = \frac{1}{N_y} \sum_{i \in I_y} x_{ij},$$

где I_y – множество индексов объектов класса y из выборки.
Оценка ковариационной матрицы

$$\widetilde{\lambda}_{jl}^y = \frac{1}{N_y} \sum_{i \in I_y} (x_{ij} - \widetilde{\mu}_j^y)(x_{il} - \widetilde{\mu}_l^y).$$

Разделяющие кривые





Замечания

- В реальных задачах довольно редко можно обосновать нормальность распределений.
- Даже если известно, что распределения действительно нормальны, это не гарантирует оптимальность выборочной решающей функции.
- При большом числе переменных и малой выборке приходится строить решение как при равных матрицах ковариаций.

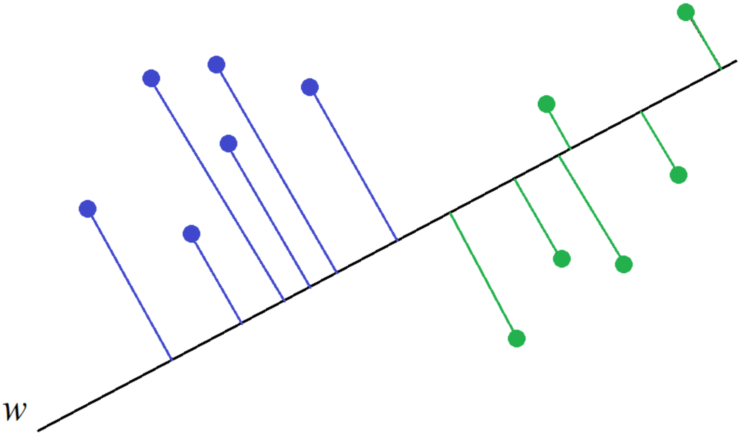
Дискриминант Фишера

Идея заключается в выборе направления, при проецировании выборки на которое образы классов оказываются наиболее удалёнными друг от друга. Это выражается в максимизации критерия

$$\Phi(w) = \frac{(\widetilde{\mu}_{w,1} - \widetilde{\mu}_{w,-1})^2}{\widetilde{S}_{w,1} + \widetilde{S}_{w,-1}},$$

где $\widetilde{\mu}_{w,y} = \frac{1}{N_y} \sum_{i \in I_y} w x_i$ – среднее,
а $\widetilde{S}_{w,y} = \frac{1}{N_y} \sum_{i \in I_y} (w x_i - \widetilde{\mu}_{w,y})^2$ – средний квадрат отклонений проекций.

Проецирование на направление



Оптимальное направление

Оказывается, что максимум $\Phi(w)$ достигается при

$$w_\Phi = \widetilde{S}^{-1}(\widetilde{\mu}_1 - \widetilde{\mu}_{-1}),$$

где $\widetilde{\mu}_y$ – среднее точек выборки y-го класса, а $\widetilde{S}_y$ – выборочная ковариационная матрица y-го класса, $\widetilde{S} = \widetilde{S}_1 + \widetilde{S}_{-1}$.

Замечания

- Метод не требует никаких вероятностных предположений.
- Выражение для $w_{Ф}$ очень похоже на выражение для нормали к разделяющей гиперплоскости для случая нормальных распределений с равными матрицами ковариаций.
- Метод предполагает оценивание по выборке только n параметров и не требователен к объёму выборки.
- Метод, выведенный при сильных предположениях, может оставаться пригодным и при нарушении предположений.

Основные понятия

$X = R^n$ – пространство значений прогнозирующих переменных,
$Y = \{-1, 1\}$ – прогнозируемая переменная,
$D = X \times Y$.

Решающая функция (алгоритм классификации)

$$f : X \to Y.$$

$V = \{(x^i, y^i) \in D \,|\, i = \overline{1, N}\}$ – случайная независимая выборка, $V \in D^N$.

$Q\colon D^N \to \Phi$ — метод построения решающих функций,
Φ — заданный класс решающих функций.

Постановка задачи

Эмпирический риск:

$$\widetilde{R}(V, f) = \frac{1}{N} \sum_{i=1}^{N} L(y^i, f(x^i)),$$

где $L(\cdot, \cdot)$ – эмпирическая функция потерь.
В качестве потерь может выступать $I(y^i \neq f(x^i))$, где $I(\cdot)$ – индикаторная функция.

Линейный пороговый классификатор

$$f(x) = \operatorname{sign}(wx - w_0).$$

Требуется найти вектор w и скаляр w_0, минимизирующие эмпирический риск при дополнительных требованиях.

Логистическая регрессия

Рассмотрим функцию условной вероятности для класса 1

$$g(x) = P(y = 1|x) = \frac{P(1)\varphi_1(x)}{P(1)\varphi_1(x) + P(-1)\varphi_{-1}(x)} = \frac{1}{1 + e^{-l(x)}},$$

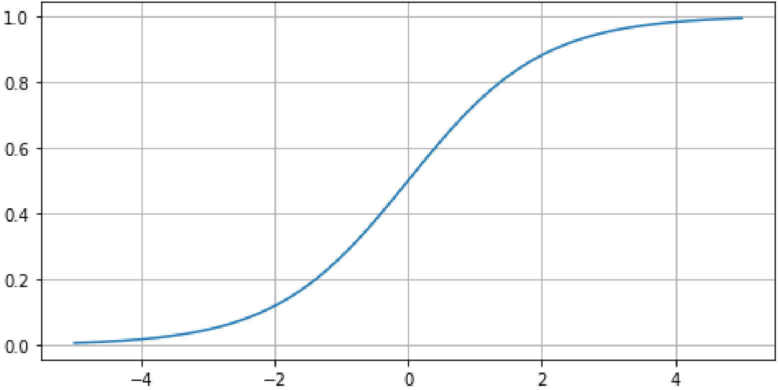
где $l(x)$ – разделяющая функция.
Подставив нормальные плотности при равных матрицах ковариаций, получим

$$g(x) = \frac{1}{1 + e^{-(wx + w_0)}} = \sigma(wx + w_0),$$

где w и w_0 есть b и c соответственно из линейного дискриминанта.

Сигмоид

Здесь $\frac{1}{1+e^{-z}}$ – так называемая логистическая функция (иногда также называемая сигмоидом или логит-функцией).



Оценивание параметров

Метод логистической регрессии основан на оценивании функции условной вероятности моделью $\widetilde{g}(x) = \sigma(\widetilde{w}x + \widetilde{w}_0)$, в которой $\widetilde{w}$ и $\widetilde{w}_0$ – настраиваемые параметры.
На практике параметры модели обычно оцениваются путём максимизации критерия правдоподобия (условной вероятности выборки).

$$-K_V(\widetilde{w}, \widetilde{w}_0) = \sum_{i \in I_1} \ln \widetilde{g}(x^i) + \sum_{i \in I_{-1}} \ln(1 - \widetilde{g}(x^i)).$$

Замечания

- Метод логистической регрессии похож на линейный дискриминант, но ослабляет вероятностные предположения.
- На практике (почти) никто вероятностные предположения не проверяет.
- По сравнению с линейным дискриминантом метод логистической регрессии более устойчив в «выбросам».
- Для повышения устойчивости требуется регуляризатор.

«Наивный» байесовский классификатор

Из формулы Байеса можем записать

$$g(x) = P(y=1|x) = \frac{P(dx, y=1)}{P(dx, y=1) + P(dx, y=-1)} = \frac{1}{1 + \frac{1-p}{p} \cdot \frac{P(dx|y=-1)}{P(dx|y=1)}},$$

где $p = P(1)$.

Независимость переменных

Пусть переменные независимы, т.е.

$$P(dx|y) = \prod_{j=1}^{n} P(dx_j|y).$$

После подстановки в предыдущее выражение и преобразований имеем

$$\frac{p}{1-p} \cdot \left(\frac{1}{g(x)} - 1\right) = \prod_{j=1}^{n} \frac{p}{1-p} \cdot \left(\frac{1}{g_j(x_j)} - 1\right),$$

где $g_j(x_j) = P(y = 1|x_j)$.

Решающая функция

Логарифмируем последнее выражение и получаем

$$\sigma^{-1}(g(x)) = (n-1)(\ln p - \ln(1-p)) + \sum_{j=1}^{n} \sigma^{-1}(g_j(x_j))$$

Получили решающую функцию в форме логистической регрессии

$$g(x) = \sigma\left(w_0 + \sum_{j=1}^{n} w_j z_j\right),$$

при $w_0 = (n-1)(\ln p - \ln(1-p))$, $w_j \equiv 1$, $z_j = \sigma^{-1}(g_j(x_j))$.

Замечания

- Предположение независимости переменных эквивалентно аддитивности в пространстве преобразованных переменных.
- Полученное преобразование является разумным вариантом target encoding, может использоваться и для вещественных переменных.
- Наивный байесовский классификатор является частным случаем логистической регрессии с использованием target encoding.
- Предположение о (строгой) независимости переменных обычно оказывается неоправданно сильным, однако в ослабленном виде гипотеза независимости очень полезна.