

Лекция 4. Линейные методы классификации

Основы интеллектуального анализа данных

Полузёров Т. Д.

БГУ ФПМИ

1 Общие идеи

- Постановка задачи
- Оценка сверху эмперического риска
- Линейная модель

2 Логистическая регрессия

- Обоснование модели
- Поиск решения

3 Метод опорных векторов

- Линейно разделимая выборка
- Линейно неразделимая выборка
- Трюк с ядром

Бинарная классификация

Рассмотрим задачу бинарной классификации,

$$X = (x_1, \dots, x_\ell) \in \mathbb{X}, Y \in \mathbb{Y} = \{-1, 1\}$$

Модель для классификации

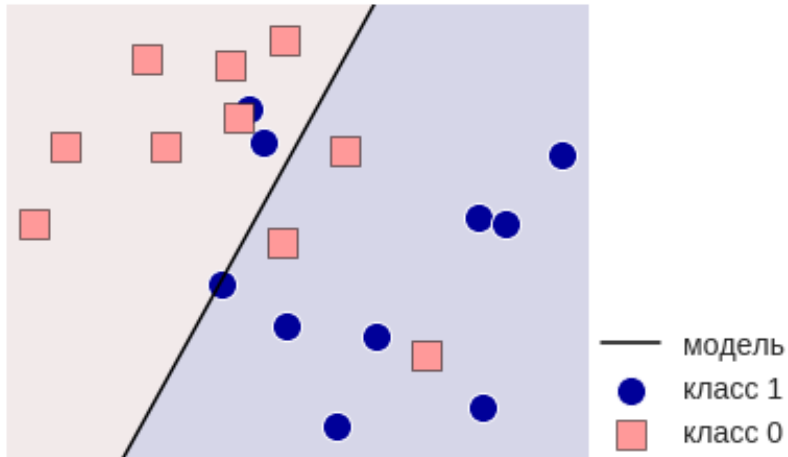
$$a(x, \theta) = \text{sign } f(x, \theta) = \begin{cases} 1, & f(x) > 0 \\ -1, & f(x) < 0 \end{cases}$$

- где $f(x, \theta)$ будем называть **дискриминантной функцией**.

Множество точек x где $f(x, \theta) = 0$ - **разделяющая поверхность**.

Задача состоит в настройке параметров θ в функции $f(x, \theta)$ по выборке (X, Y)

Пример



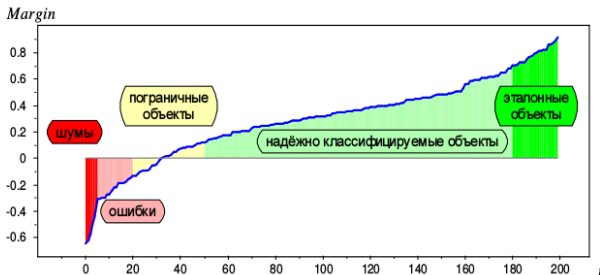
Отступ

Величина

$$M_i(\theta) = y_i f(x_i, \theta)$$

- **отступ** (margin) классификатора $a(x, \theta) = \text{sign } f(x, \theta)$ относительно объекта x_i .

Если $M_i(\theta) < 0$ то алгоритм a допускает ошибку на объекте x_i .



Функция потерь

Требуется подобрать параметры θ при которых классификатор a допускает как можно меньше ошибок:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [M_i(\theta) < 0] \rightarrow \min_{\theta}$$

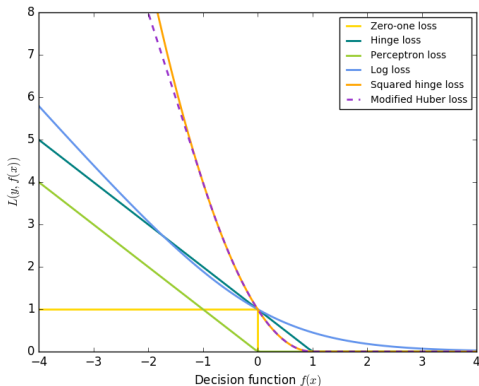
Однако в таком виде Q - кусочно постоянная функция

Идея - мажорирование (оценка сверху) индикатора ошибки $[M_i(\theta) < 0]$ с помощью "удобной" функцией потерь $\mathcal{L}(M_i)$:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [M_i(\theta) < 0] \leq \tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathcal{L}(M_i(\theta))$$

Популярные функции потерь

- 1 $[M < 0]$ - индикатор ошибки
- 2 $(1 - M)^2$ - квадратичная
- 3 $(1 - M)_+$ - кусочно линейная (Hinge loss)
- 4 $\frac{2}{1+e^M}$ - сигмоидная
- 5 $\log_2(1 + e^{-M})$ - логистическая
- 6 e^{-M} - экспоненциальная



Функция потерь и совместное распределение

Пусть множество $(\mathbb{X} \times \mathbb{Y})$ - вероятностное пространство. Имея выборку (X, Y) и предполагаемый вид совместной плотности $p(x, y; \theta)$, применим метод максимального правдоподобия

$$L(\theta) = \prod_{i=1}^{\ell} p(x_i, y_i; \theta) \rightarrow \max_{\theta}$$

$$\ln L = \sum_{i=1}^{\ell} \ln p(x_i, y_i; \theta) \rightarrow \max_{\theta}$$

$$-\ln p(x_i, y_i; \theta) = \mathcal{L}(y_i f(x_i, \theta))$$

По виду плотности $p(x, y; \theta)$ восстанавливается f и \mathcal{L} . И наоборот, используя некоторую разделяющую поверхность и функцию потерь - предполагаем определенное распределение в данных.

Линейная модель

Случай $f(x, \omega) = \langle x, \omega \rangle$ - класс линейных моделей классификации.

$$a(x, \omega) = \text{sign} \langle x, \omega \rangle$$

Разделяющая поверхность $\text{sign} \langle x, \omega \rangle = 0$ является гиперплоткостью в \mathbb{R}^n . Причем объекты по одну сторону от гиперплоткости относятся к одному классу, по другую - к другому.

Метод обучения

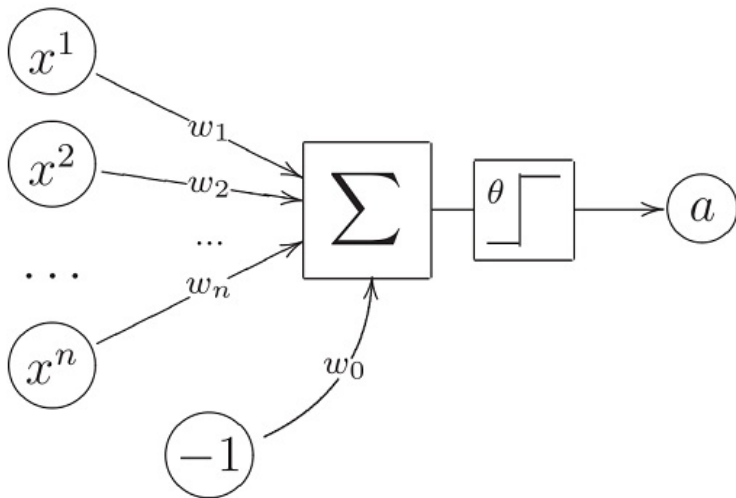
Метод минимизации мажорированного эмперического риска

$$\tilde{Q}(a, X) = \sum_{i=1}^{\ell} \mathcal{L}(\langle x_i, \omega \rangle y_i) \rightarrow \min_{\omega}$$

Необходимое условие минимума:

$$\frac{\partial Q}{\partial \omega} = \sum_{i=1}^{\ell} x_i y_i \mathcal{L}'(\langle x_i, \omega \rangle y_i) = 0$$

Модель нейрона МакКаллока-Питтса



Логистическая регрессия

Логистическая регрессия (Logistic regression) - линейный алгоритм бинарной классификации.

При выполнении достаточно сильных предположений обладает свойствами:

- оптимальный байесовский классификатор
- однозначно определена функция потерь
- возможность оценивать вероятности классов

Пусть $(\mathbb{X} \times \mathbb{Y}) = (\mathbb{R}^n \times \{-1, 1\})$ - вероятностное пространство с плотностью $p(x, y)$. Выборка (X, Y) получена из этого распределения.

Экспонентный класс распределений

Плотность $p(x)$, $x \in \mathbb{R}^n$ называется экспонентной, если

$$p(x) = \exp(c(\delta)\langle\theta, x\rangle + b(\delta, \theta) + d(x, \delta))$$

где θ - параметр сдвига, δ - масштаба, b, c, d - произвольные числовые функции.

Принадлежат к классу экспонентных:

- Равномерное, Нормальное, Гамма
- Гипергеометрическое, Пуассоновское, Биномиальное
- и другие

Обоснование линейной модели

Теорема

Если функции правдоподобия $p(x|y)$ принадлежат экспонентному классу, причем параметры d и δ одинаковы, а отличаются только параметры сдвига θ ,

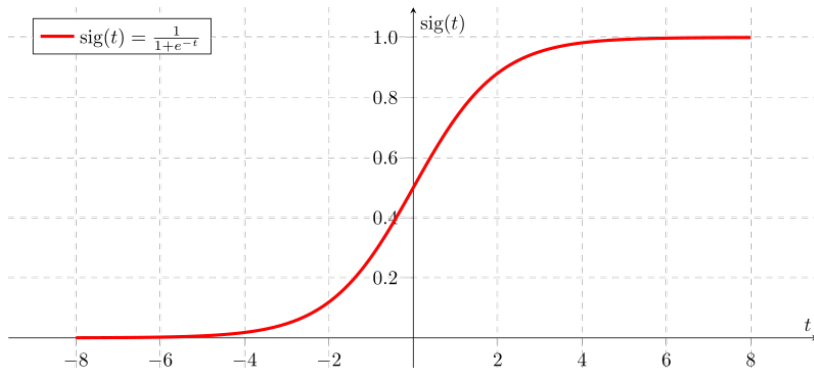
то:

① *оптимальный байесовский классификатор является линейным*

② *апостериорная вероятность $p(y|x) = \sigma(\langle \omega, x \rangle y)$*

где $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоидная функция, $\sigma(-z) = 1 - \sigma(z)$

Сигмоидная функция



Модель оценки вероятностей

Построим модель которая оценивает не сами метки классов, а **вероятности** принадлежности к ним.

$$b(x, \omega) = P(+1|x) = \sigma(\langle \omega, x \rangle)$$

Другими словами, в каждой точке x величина $y \sim \text{Bernoulli}(\sigma(\langle \omega, x \rangle))$

Задача классификации решается путем выбора порога $t \in [0, 1]$. И тогда итоговый классификатор имеет вид:

$$a(x, t) = \text{sign}(b(x, \omega) - t)$$

Метод максимального правдоподобия

С учетом вероятностной постановки задачи, воспользуемся методом максимального правдоподобия:

$$L(\omega) = \prod_{i=1}^{\ell} p(y_i|x_i) = \prod_{i=1}^{\ell} \sigma(\langle \omega, x_i \rangle y_i) \rightarrow \max_{\omega}$$

$$\begin{aligned} \ln L(\omega) &= \sum_{i=1}^{\ell} \ln \sigma(\langle \omega, x_i \rangle y_i) = \\ &= - \sum_{i=1}^{\ell} \ln(1 + e^{-y_i \langle \omega, x_i \rangle}) \rightarrow \max_{\omega} \end{aligned}$$

-совпадает с логистической функцией потерь

$$Q(\omega) = \sum_{i=1}^{\ell} \ln(1 + e^{-M_i}) \rightarrow \min_{\omega}$$

Решение оптимизационной задачи

Имеем

$$Q(\omega) = \sum_{i=1}^{\ell} \ln(1 + e^{-y_i \langle \omega, x_i \rangle}) \rightarrow \min_{\omega}$$

Аналитического решения нет, поэтому применяются градиентные методы

$$\nabla Q(\omega) = \sum_{i=1}^{\ell} x_i y_i \sigma(-\langle \omega, x_i \rangle)$$

Логистическая регрессия

Получена модель логистической регрессии $b : \mathbb{X} \rightarrow [0, 1]$

$$b(x) = \sigma(\langle x, \omega \rangle)$$

При обучении используется логистическая функция потерь

$$\mathcal{L}(y, a) = \ln(1 + e^{-y_i \langle \omega, x \rangle})$$

Выбирая порог $t \in [0, 1]$ получаем классификатор

$$a(x) = [b(x) > t]$$

Случай линейно разделимой выборки

По выборке (X, Y) будем строить классификатор вида

$$a(x, \omega) = \langle x, \omega \rangle - \omega_0$$

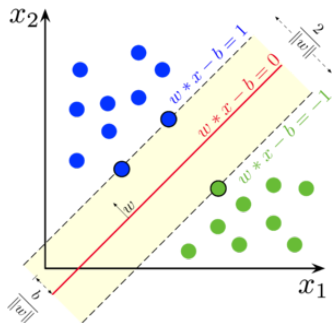
и предположим, что выборка (X, Y) линейна разделима. То есть найдутся такие веса ω^* , что

$$Q(\omega^*) = \frac{1}{\ell} [y_i \cdot a(x_i, \omega^*) < 0] = 0$$

Однако такое решение ω^* не единственное. Идея состоит в распоряжении этой свободой с умом.

Максимизация отступа

Среди всех подходящих разделяющих гиперплоскостей выберем ту, которая максимально удалена от "граничных" объектов.



Вследствие линейной разделимости
 $\exists \omega, \omega_0 : \{x : -1 < \langle \omega, x \rangle - \omega_0 < 1\}$
- полоса между классами.

Выберем наиболее широкую из всех доступных

$$\begin{cases} ||\omega||^2 \rightarrow \min \\ M_i(\omega) \geq 1, \quad i = 1, \dots, \ell \end{cases}$$

Линейно неразделимая выборка

В случае линейно неразделимой выборки смягчим требования

$$\begin{cases} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{\ell} \xi_i \rightarrow \min \\ M_i(\omega) \geq 1 - \xi_i, i = 1, \dots, \ell \\ \xi_i \geq 0, i = 1, \dots, \ell \end{cases}$$

Требование можно упростить

$$\begin{cases} \xi_i \geq 0 \\ \xi_i \geq 1 - M_i(\omega) \\ \sum_{i=1}^{\ell} \xi_i \rightarrow \min \end{cases} \Rightarrow \xi_i = (1 - M_i(\omega))_+$$

Итог - задача безусловной оптимизации

$$\frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{\ell} (1 - M_i(\omega))_+ \rightarrow \min_{\omega}$$

где C - гиперпараметр алгоритма

Метод опорных векторов

Получен **метод опорных векторов** (Support Vector Machine, SVM), где

$$a(x) = \text{sign}(\langle x, \omega \rangle)$$

$$\mathcal{L}(y, a) = (1 - M_i(\omega))_+$$

Т.е. линейный классификатор с квадратичной регуляризацией, обученный по функции потерь Hinge loss

Двойственная задача

По теореме Каруша-Куна-Такера, исходная задача эквивалентна двойственной задаче поиска седловой точки функции Лагранжа

$$\mathcal{L}(\omega, \lambda, \eta) = \frac{1}{2} \|\omega\|^2 - \sum_{i=1}^{\ell} \lambda_i (M_i(\omega) - 1) - \sum_{i=1}^{\ell} \xi_i (\lambda_i + \eta_i - C)$$

Приравнивая частные производные \mathcal{L} по ω, λ, η к нулю

$$\begin{cases} \omega = \sum_{i=1}^{\ell} \lambda_i y_i x_i \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0 \\ \eta_i + \lambda_i = C, i = 1, \dots, \ell \end{cases}$$

Типы объектов

При условии

$$\begin{cases} \eta_i + \lambda_i = C, i = 1, \dots, \ell \\ \eta_i \geq 0, \lambda_i \geq 0 \end{cases}$$

Существуют три ситуации

- 1 $\lambda_i = 0, \eta_i = C, \xi_i = 0, M_i \geq 1$ - **неинформативный** объект, классифицируется правильно и не влияет на ω
- 2 $0 < \lambda_i < C, 0 < \eta_i < C, \xi_i = 0, M_i = 1$ - **опорный** объект, классифицируется правильно и лежит на разделяющей полосе
- 3 $\lambda_i = C, \eta_i = 0, \xi_i > 0, M_i < 1$ - **опорный нарушитель**, либо неверно классифицируется, либо лежит внутри разделяющей полосы

Альтернативный вид классификатора

Подставляя ограничения в лагранжиан, переходим к двойственной задаче с двойственными переменными λ

$$\begin{cases} -\sum_{i=1}^{\ell} \lambda_i + \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle \rightarrow \min_{\lambda} \\ 0 \leq \lambda_i \leq C, i = 1, \dots, \ell \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0 \end{cases}$$

Получив решение λ , параметры классификатора равны

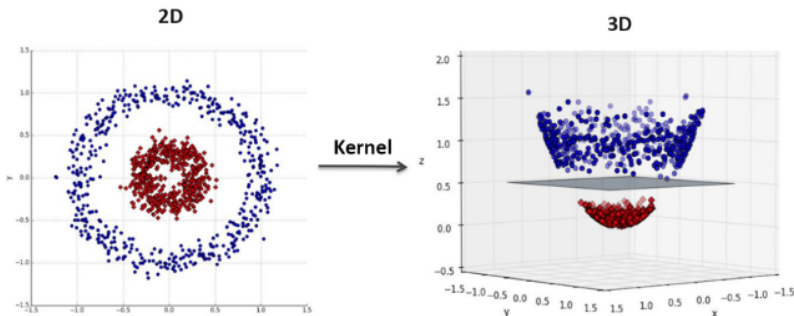
$$\begin{cases} \omega = \sum_{i=1}^{\ell} \lambda_i y_i x_i \\ \omega_0 = \langle x_i, \omega \rangle - y_i, \quad \forall i : \lambda_i > 0, M_i = 1 \end{cases}$$

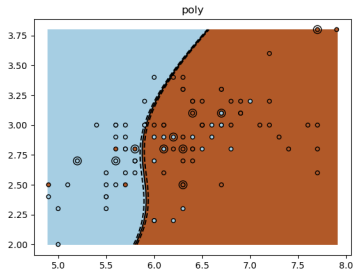
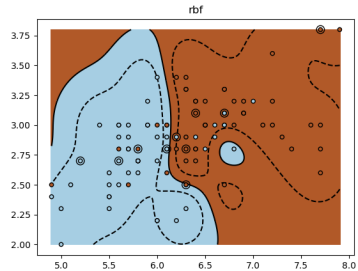
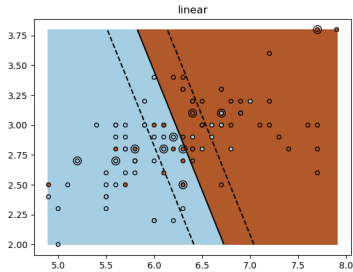
И итоговый классификатор имеет вид

$$a(x) = \text{sign} \left(\sum_{i=1}^{\ell} \lambda_i y_i \langle x, x_i \rangle - \omega_0 \right)$$

Трюк с ядром

Идея состоит в переходе от исходного признакового пространства \mathbb{X} в другое \mathbb{H} где, быть может, выборка является линейно разделимой, посредством $\psi : \mathbb{X} \rightarrow \mathbb{H}$.
Это выливается в использование вместо $\langle \psi(x_i), \psi(x_j) \rangle$ некоторого ядра $K(x_i, x_j)$





Заключение

- Линейный классификатор - модель нейрона
- Аппроксимация пороговой функции потерь
- Оценка вероятностей с помощью логистической регрессии
- SVM - очень сильный алгоритм классификации благодаря трюку с ядром